DISS. ETH N° 29913

Multimodal Representation Learning under Weak Supervision

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zurich)

presented by

IMANT DAUNHAWER

M.Sc. Social and Economic Data Analysis, University of Konstanz

born on 17.05.1991

Accepted on the recommendation of

Prof. Dr. Julia E. Vogt

Prof. Dr. Volker Roth

Prof. Dr. Karsten M. Borgwardt

2023

Abstract

Biological organisms experience a world of multiple modalities through a variety of sensory systems. For example, they may perceive physical or chemical stimuli through the senses of sight, smell, taste, touch, and hearing. Across species, the nervous system integrates heterogeneous sensory stimuli and forms *multimodal representations* that capture information shared between modalities. Analogously, machines can perceive their environment through different types of sensors, such as cameras and microphones. Yet, it is not sufficiently well understood how multimodal representations can be formed *in silico*, i.e., via computer simulation. In this thesis, we study how to leverage statistical dependencies between modalities to form multimodal representations computationally using machine learning.

We start from the premise that real-world data is generated from a few factors of variation. Given a set of observations, *representation learning* seeks to infer these latent variables, which is fundamentally impossible without further assumptions. However, when we have corresponding observations of different modalities, statistical dependencies between them can carry meaningful information about the latent structure of the underlying process. Motivated by this idea, we study multimodal learning under *weak supervision*, which means that we consider corresponding observations of multiple modalities without labels for what is shared between them. For this challenging setup, we design machine learning algorithms that transform observations into representations of shared and modality-specific information without explicit supervision by labels. Thus, we develop methods that infer latent structure from low-level observations using weak supervision in the form of multiple modalities.

We develop techniques for multimodal representation learning using two approaches—generative and discriminative learning. First, we focus on generative learning with variational autoencoders (VAEs) and propose a principled and scalable method for variational inference and density estimation on sets of modalities. Our method enhances the encoding and disentanglement of shared and modality-specific information and consequently improves

the generative performance compared to relevant baselines. Motivated by these results, we consider an explicit partitioning of the latent space into shared and modality-specific subspaces. We explore the benefits and pitfalls of partitioning and develop a model that promotes the desired disentanglement for the respective subspaces. Thereby, it further improves the generative performance compared to models with a joint latent space. On the other hand, we also establish fundamental limitations for generative learning with multimodal VAEs. We show that the sub-sampling of modalities enforces an undesirable bound on the approximation of the joint distribution. This limits the generative performance of mixture-based multimodal VAEs and constrains their application to settings where relevant information can be predicted in expectation across modalities on the level of observations.

To address these issues, we shift to discriminative approaches and focus on contrastive learning. We show that contrastive learning can be used to identify shared latent factors that are invariant across modalities up to a block-wise indeterminacy, even in the presence of non-trivial statistical and causal dependencies between latent variables. Finally, we demonstrate how the representations produced by contrastive learning can be used to transcend the limitations of multimodal VAEs, which yields a hybrid approach for multimodal generative learning and the disentanglement of shared and modality-specific information. Thus, we establish a theoretical basis for multimodal representation learning and explain in which settings generative and discriminative approaches can be effective in practice.

Zusammenfassung

Biologische Organismen erleben die Welt durch eine Vielzahl und Vielfalt an Sinnessystemen auf unterschiedliche Art und Weise, das heisst in Form von verschiedenen Modalitäten. Zum Beispiel können sie physikalische oder chemische Reize über mehrere verschiedene Sinnesmodalitäten—wie Sehen, Riechen, Schmecken, Tasten und Hören—wahrnehmen. Artenübergreifend verarbeitet das Nervensystem heterogene sensorische Reize und bildet *multimodale Repräsentationen*, welche zugrunde liegende Informationen erfassen, die zwischen den unterschiedlichen Modalitäten geteilt werden. Gleichermassen können auch Maschinen ihre Umgebung über verschiedene Arten von Sensoren, wie Kameras und Mikrofone, wahrnehmen. Es ist jedoch nicht ausreichend geklärt, wie multimodale Repräsentationen *in silico*, also durch computergestütze Rechenprozesse, erlernt werden können. In dieser Arbeit untersuchen wir, wie statistische Abhängigkeiten zwischen Modalitäten genutzt werden können, um multimodale Repräsentationen mit Hilfe von maschinellem Lernen zu erschliessen.

Wir nehmen an, dass die komplexen Daten, welche wir in vielen Anwendungen beobachten, prinzipiell auf wenige zugrunde liegende latente Faktoren zurückzuführen sind. Das Ziel beim *Repräsentationslernen* ist es, von einem Datensatz ausgehend, auf diese latenten Faktoren zu schliessen, was im Allgemeinen, ohne weitere Annahmen, unmöglich ist. Wenn wir jedoch mehrere Beobachtungen unterschiedlicher Modalitäten betrachten, können ihre statistischen Abhängigkeiten aufschlussreiche Informationen über die latenten Faktoren oder die Struktur des zugrunde liegenden generativen Prozesses enthalten. Auf Grundlage dieser Idee beschäftigen wir uns mit dem multimodalen maschinellen Lernen unter *schwacher Überwachung*. Das bedeutet, wir betrachten sinngemäss korrespondierende Beobachtungen mehrerer Modalitäten ohne zusätzliche Informationen darüber was zwischen ihnen geteilt wird. Für diese Problemstellung entwickeln wir Lernalgorithmen, welche die komplexen Beobachtungen in gemeinsame und modalitätsspezifische Faktoren zerlegen. Damit nutzen

wir die schwache Überwachung in Form von mehreren Modalitäten, um aus komplexen Daten die zugrunde liegende latente Struktur abzuleiten.

Wir entwickeln Methoden für das multimodale Repräsentationslernen basierend auf zwei Ansätzen—dem generativen und diskriminativen Lernen. Zunächst konzentrieren wir uns auf das generative Lernen mit dem Variational Autoencoder (VAE) und entwickeln eine skalierbare Methode für die approximative Inferenz und Dichteschätzung für Mengen von Modalitäten. Unsere Methode ermöglicht die Kodierung und Zerlegung gemeinsamer und modalitätsspezifischer Informationen und verbessert die generative Leistung im Vergleich zu den relevanten Grundmodellen. Darüber hinaus betrachten wir eine explizite Partitionierung des latenten Raums in gemeinsame und modalitätsspezifische Unterräume. Wir untersuchen die Vor- und Nachteile der Partitionierung und entwickeln ein Modell, das die gewünschte Zerlegung der Informationen für die jeweiligen Unterräume fördert. Dadurch erreicht das Modell eine weitere Verbesserung der generativen Leistung im Vergleich zu Modellen mit einem gemeinsamen latenten Raum. Andererseits stellen wir auch grundlegende Einschränkungen für das generative Lernen mit multimodalen VAEs fest. Wir zeigen, dass das Sub-Sampling der Modalitäten eine unerwünschte Schranke für die Approximation der gemeinsamen Verteilung impliziert. Dies beschränkt die generative Leistung von mischungsbasierten multimodalen VAEs. Zudem begrenzt es ihre Anwendbarkeit auf Probleme bei denen relevante Informationen im Erwartungswert über Modalitäten hinweg auf der Ebene von Beobachtungen vorhergesagt werden können.

Zusätzlich betrachten wir einen diskriminativen Ansätz, nämlich das kontrastive Lernen. Wir zeigen, dass damit geteilte latente Faktoren bis zu einer blockweisen Unbestimmtheit identifiziert werden können, wenn die geteilten Faktoren über Modalitäten hinweg invariant sind. Für das Resultat betrachten wir zusätzliche, nicht-triviale statistische und kausale Abhängigkeiten zwischen den latenten Variablen. Schliesslich zeigen wir, wie kontrastives Lernen genutzt werden kann, um die Einschränkungen multimodaler VAEs zu überwinden. Dafür entwickeln wir einen hybriden Ansatz für das multimodale generative Lernen und für die Zerlegung gemeinsamer und modalitätsspezifischer Informationen. Übergreifend schafft unsere Arbeit eine theoretische Grundlage für multimodales Repräsentationslernen und erklärt unter welchen Bedingungen generative und diskriminative Ansätze in der Praxis effektiv sein können.

Acknowledgments

I am deeply grateful to my advisor, Julia Vogt, for providing me with the opportunity, time, freedom, and resources required to do research. I benefitted immensely from Julia's mentorship and the excellent team she has created. Looking back at the evolution of the lab, I feel honored to have been part of it.

I thank Volker Roth and Karsten Borgwardt for kindly serving as members of my thesis committee. As my second advisor, Volker provided helpful comments and a uniquely sharp and sober perspective, which I greatly respect and appreciate.

As a member of the Medical Data Science group at ETH Zurich, I count myself lucky to have been part of such an intellectually stimulating yet warm and welcoming environment. To all members, past and present, of the MDS group: merci vielmal! I especially thank Thomas Sutter for being the most supportive and dependable companion since day one, Kieran Chin-Cheong for the valuable technical support and parenting advice, Ricards Marcinkevics for being the most reliable and authentic fellow, and Alexander Marx for memorable lessons in causality, information theory, and bouldering.

Considering my wonderful time in Basel, I am grateful to all members of the Biomedical Data Analysis group and the Computer Graphics and Vision group. They provided a stellar example of a cooperative environment that helped me ease into the doctorate. I especially thank Maxim Samarin and Mario Wieser for their help and enriching company.

My research was not done in isolation and would not have been possible without the help of countless others. I thank my direct collaborators Alice Bizeul, Kieran Chin-Cheong, Severin Kasser, Gilbert Koch, Mike Laszkiewicz, Ricards Marcinkevics, Alexander Marx, Juan Montoya, Ivan Ovinnikov, Emanuele Palumbo, Karthik Pattisapu, Marc Pfister, Amartya Sanyal, Alexander Sauter, Kai Schuhmacher, Yuge Shi, Bram Stieltjes, Martin Stocker, Thomas Sutter, Philip Torr, Thomas Weikert, Sven Wellmann, and Marco Wiering. I also

thank all collaborators behind the scenes whose contributions made our research possible in the first place.

Beyond my direct colleagues and collaborators, I thank Luigi Gresele, Adrián Javaloy, Julius von Kügelgen, Max Paulus, Petra Poklukar, Nicolò Ruggeri, and Miguel Vasco. Their unique contributions, even those brief and sporadic encounters, left a lasting impression and shaped my path in subtle yet meaningful ways.

Looking back further, I realize that this journey already started in Konstanz. I am deeply thankful for the advice and support from Christian Borgelt, Roxana Halbleib, Sven Kosub, Oliver Sampson, David Schoch, and Termeh Shafie, who nudged me in the direction of academia. Occasionally, all it takes is someone else's faith in your abilities.

I gratefully acknowledge the institutions and organizations that supported my research. In particular, the Department of Computer Science at ETH Zurich and the Department of Mathematics and Computer Science at the University of Basel. I thank the SNSF for funding a significant part of my doctorate and the IT service group and LeoMed cluster for the technical support and computational resources.

Finally, I would like to express my deepest gratitude to my family. I am grateful to my parents, Irina and Michael Daunhawer, for their boundless love and support. Likewise, I am indebted to Luisa Fleischhauer for her unwavering support throughout the journey. None of this would have been possible without my family.

Contents

Notation and Terminology

$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Random vectors.
$p(\mathbf{x})$	Probability distribution of $\mathbf{x}$; either probability mass function (p.m.f.) or probability density function (p.d.f.).
$p(\mathbf{y} \mid \mathbf{x})$	Conditional probability distribution (p.m.f. or p.d.f.) of $\mathbf{y}$ given $\mathbf{x}$.
$\{\mathbf{x}^{(n)}\}_{n=1}^{N} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$	An independent and identically distributed (i.i.d.) sample of size N.
$\mathbb{E}[\mathbf{x}]$	Expected value of $\mathbf{x}$.
$\mathbf{x} \perp\!\!\!\perp \mathbf{y}$	Independence between $\mathbf{x}$ and $\mathbf{y}$.
$\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$	Conditional independence between $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$.
$\mathbf{x}[d]$	The d-th component of the vector $\mathbf{x}$.
$\mathbf{x}[A]$	Vector of components $(\mathbf{x}[d])_{d \in A}$, where $A \subseteq \{1, \dots, D\}$ is an ordered set of integers.
$I(\mathbf{x}; \mathbf{y})$	Mutual information between $\mathbf{x}$ and $\mathbf{y}$.
$H(\mathbf{x})$	Entropy of $\mathbf{x}$.
$H(\mathbf{y} \mid \mathbf{x})$	Conditional entropy of $\mathbf{y}$ given $\mathbf{x}$.
$D_{\mathrm{KL}}(p(\mathbf{x}) \,\|\|\, p(\mathbf{y}))$	Kullback-Leibler divergence of a distribution $p(\mathbf{x})$ from another distribution $p(\mathbf{y})$.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2.
$\mathcal{P}(D)$	Powerset of the set of consecutive integers $\{1, \dots, D\}$; excluding the empty set.
$\mathbb{1}_{\{\cdot\}}$	Indicator function that maps to one if condition $\{\cdot\}$ is fulfilled and to zero otherwise.
$\exp(\cdot)$	Exponential function.
$\mathrm{sim}(\cdot, \cdot)$	Similarity function (e.g., negative Euclidean distance, cosine similarity, etc.).
ELBO	Evidence lower bound.
VAE	Variational autoencoder.
ICA	Independent component analysis.
NCE	Noise contrastive estimation.
PoE	Product of experts.
MoE	Mixture of experts.
MoPoE	Mixture of products of experts.

1

Introduction

By definition, a subject (e.g., an object or event) is *multimodal* if it exhibits different modes of occurrence or activity [c.f., OED]. Intuitively, it means that the subject appears *diversely*. However, the definition also entails a *permanence* of the subject despite its diverse appearance. This is not a contradiction.

To resolve the paradox and to illustrate the concept of multimodality, consider the following examples. The first describes a situation from everyday life; the second draws a connection to the related concept of multimodal distributions from statistics.

Example 1 (Multimodal perception)**.** *Imagine any tangible object that can be examined with multiple senses. First, using our eyes, we might observe it at some distance and from different angles. Then, with our hands, we could explore its shape and material. Finally, after sufficient exploration, we might notice that our perception of the object is multimodal in that the same object manifests in various forms (multiple perspectives, sensation of the left and right hand) and distinct formats (visual and tactile information).*

Example 2 (Multimodal probability distribution)**.** *Consider a multimodal distribution, i.e., any probability distribution with two or more (local) maxima. The distribution describes the probability of occurrence for different possible outcomes of an experiment. By definition, the outcomes can be diverse, yet they still relate to the same experiment.*

Equipped with an intuitive understanding of the central concept of this thesis, let us proceed with the research topic, which we denote as *multimodal representation learning*.

In representation learning, we start from the premise that observations are generated from a few latent factors of variation that can be inferred from data [BCV13; Sch+21]. The word "latent" means that these factors or variables are not directly observed; however, it is assumed that they are useful, and therefore we seek to infer them. For instance, the latent variables can be thought of as categories or abstract concepts that provide a meaningful description of the data.

The holy grail of representation learning is to infer the latent factors using only low-level observations (e.g., raw image data). Since this is known to be theoretically impossible without further assumptions [HP99], we instead consider the setup of *weak supervision* in the form of multiple modalities. Concretely, we consider corresponding observations of different modalities (e.g., image and text data) without labels for what is shared between them. Our goal is to design machine learning algorithms that leverage statistical dependencies between modalities to learn representations—i.e., data transformations—that encode the latent factors effectively and ideally recover them up to acceptable ambiguities.

1.1 Broad Context

Before we flesh out the concrete topic and research goal, let us illustrate the idea and significance of multimodal representation learning in the broader context of neuroscience, cognitive science, and computation.

Multisensory perception The formation of multimodal percepts or representations has a rich history in neuroscience and cognitive science [Ste12], where modalities stand for the stimuli perceived by distinct sensory systems. For example, consider the human visual and auditory systems, our eyes and ears, and how they respond to different forms of energy. The visual system detects light particles on the retina, perceiving visible light as electromagnetic waves within a certain frequency band. In contrast, the auditory system detects mechanical waves, i.e., the movement of molecules, as changes in air pressure. In this manner, information about a distal object is received by sensory receptors and transduced into electrical and chemical signals, which permeate throughout the nervous system and converge on individual neurons that produce multisensory responses [SM93].

Multisensory integration is observed at the level of individual, multisensory neurons [Mer02; SS08; SSR14] and even in low-level cortical areas [GS06; MS09a; MS09b], which suggests

that the integration occurs at multiple levels of the neuraxis. In addition, neuroimaging studies and cognitive experiments indicate that the human brain forms an abstract, modality-agnostic representation of the environment [EB04; Qui+05; Qui+09; Yil14], if not a unified perceptual experience, as suggested by some philosophers [Tye03; Bay12]. Despite the dawning interpretation that perception is *richly* multimodal [O12], there remains a gap between the understanding of the low-level neurophysiological mechanisms and the formation of higher-level multimodal percepts.

While we draw inspiration from neuroscience, in this thesis, we focus on *machine learning*. Analogous to animals, machines can perceive the world through different types of sensors of different modalities (e.g., cameras, microphones, tactile sensors [Eve95; SNS11]). From the perspective of an individual organism or agent, each sensory modality serves as a unique window of perception to the outside world [Uex92; Yon22] as it enables the sensation of novel, complementary, or redundant aspects of the environment. But besides the obvious advantage that is provided by experiencing the world through multiple sensory systems, there are theoretical aspects that justify *learning* from multimodal data.

Multimodal learning Multimodal learning describes the algorithmic process of learning from data to improve the performance on a set of tasks by using information from multiple modalities. From a computational perspective, crossmodal associations—such as statistical, structural, and semantic correspondences [Spe11]—can be used to integrate information across modalities to improve the performance on relevant tasks. For instance, the nervous system combines information from different sensory modalities to detect external objects faster and more accurately [Mil91; Hug+94], and to reduce uncertainty in conditions where an individual sense becomes unreliable [EB04; DC04; SMB05; NK10]. Moreover, statistical dependencies between modalities indicate the presence of common causes [Rei56; Kör+07], which can be thought of as latent factors of variation shared between modalities.[1] Thus, information that is shared between modalities provides a learning signal that can be used to enhance the perception of external stimuli and to guide adaptive behavior.

[1] The interpretation of perception as statistical inference of external causes from sensory cues can be traced back to Hermann von Helmholtz [von67, Part III]. More recently, models of the neural processing of sensory information have been developed under the umbrella of the *Bayesian brain* [KP04; Doy+07].

1.2 Research Goal

Based on the idea that statistical dependencies between modalities can carry meaningful information about the latent structure of the underlying process, we study multimodal learning under *weak supervision*, which means that we consider corresponding observations of multiple modalities without labels for what is shared between them.

Specifically, we consider the following setup and goal:

Given a dataset of corresponding observations of different modalities but no labels for what is shared between them, we investigate whether machine learning algorithms can draw inferences about the latent factors, particularly those shared between observations.

Therefore, we seek to address the following research questions:

Question 1. *How can we identify latent factors of variation shared between modalities?*

The first question describes our primary goal, which is to develop models that recover the latent factors shared between modalities. Therefore, we investigate whether the considered models encode mutually shared information effectively and whether the latent factors can be identified up to acceptable ambiguities.

Question 2. *How can we disentangle shared and modality-specific information?*

Based on the second question, our goal is to disentangle shared and modality-specific information, which poses an additional challenge because it requires that a model encodes modality-specific information in a form that separates it from shared information.

Question 3. *How can we learn generative models to draw inferences across modalities?*

For the third question, we investigate whether generative models can approximate the distribution of multimodal data in a way that allows models to draw inferences across modalities on the level of observations, specifically, to generate or impute missing modalities based on the learned representations.

In answering the above questions, we aim to establish a theoretical basis for multimodal representation learning and contribute to the overarching goal of discovering latent structure from low-level observations, which echoes through the field of machine learning.

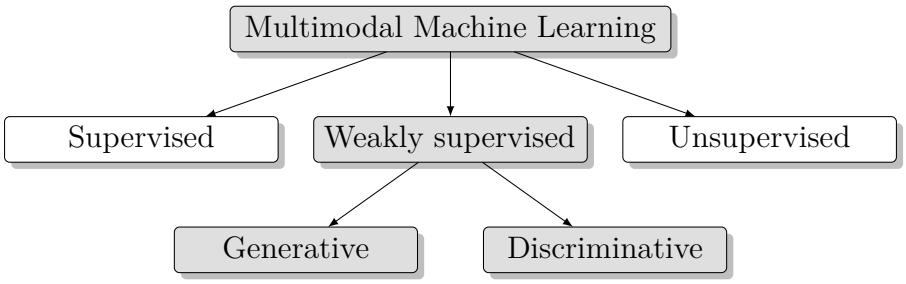


Figure 1.1: Overview of the subject area. Gray nodes denote the topics addressed in this thesis. The degree of supervision describes the *relation* between modalities (c.f., Example 3).

1.3 Topic and Scope

In Figure 1.1, we present a high-level overview of the subject area with the topics addressed in this thesis highlighted in gray. Broadly, we locate our research in the area of multimodal machine learning. More specifically, we focus on the setup of weak supervision and employ methods from generative and discriminative learning.

Weak supervision In multimodal learning, we can use the degree of supervision to describe the *relation* between modalities [WG18]. As in classic machine learning [Bis06], we distinguish between supervised and unsupervised learning, but in addition, we consider the setup of weak supervision, which lies in the middle of the two extremes.

The notion of weak supervision is formally introduced in Chapter 2, but intuitively, we describe it in terms of corresponding observations of different modalities without labels for what is shared between them. To clarify its relation to supervised and unsupervised learning in the context of multimodal data, consider the following example:

Example 3 (Degrees of supervision)**.** *Let $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ be a joint distribution for two modalities $\mathbf{x}_1, \mathbf{x}_2$ and labels $\mathbf{y}$ that describe what is shared between $\mathbf{x}_1$ and $\mathbf{x}_2$. We observe both modalities and labels in the supervised setup; i.e., $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y}) \sim p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ drawn from the joint distribution. Under weak supervision, we observe pairs $(\mathbf{x}_1, \mathbf{x}_2) \sim p(\mathbf{x}_1, \mathbf{x}_2)$ drawn from the joint distribution without labels. Finally, in the unsupervised setup, we observe unpaired examples $\mathbf{x}_1 \sim p(\mathbf{x}_1)$ and $\mathbf{x}_2 \sim p(\mathbf{x}_2)$ drawn from the marginal distributions.*

Note that in the given example, the distinction between the supervised and weakly supervised setup is expressed in terms of the presence or absence of labels, which is an

overly narrow characterization for the degree of supervision. In Chapter 2, we define the setup more generally in terms of latent variables and functional dependencies.

Considered methods Further, we make a classic distinction between discriminative and generative models [Jeb04]. In this thesis, we consider both types of methods because they provide complementary approaches for representation learning [Liu+23]. From the array of generative models, we employ the variational autoencoder [KW14; RMW14], and as a discriminative approach, we use contrastive learning [GH10; OLV18]. We introduce these methods in Chapter 3, but we already mention them here because the distinction provides a valuable perspective on our contributions, which we discuss next.

1.4 Contributions

As a last stop before we dive into the topic, let us summarize the contributions of our work and sketch out the structure of the thesis. First, we foreshadow the contributions of the individual chapters and relate them to the research questions. Then, we provide a list of our relevant publications and describe the structure of the manuscript.

Multimodal generative learning First, we focus on generative learning with variational autoencoders (VAEs). In Chapter 5, we introduce the mixture of products of experts multimodal variational autoencoder (MoPoE-VAE), a principled and scalable method for variational inference and density estimation on sets of modalities. The MoPoE-VAE learns a generative model for the joint distribution of multimodal data in a way that enables scalable inference across modalities. In its theoretical formulation, the model generalizes two existing, widely-used methods which only use specific subsets of modalities for the posterior approximation. Instead, our model considers all subsets of modalities in a scalable way using the mixture of products of experts. Empirically, the model enhances the encoding and disentanglement of shared and modality-specific information and the generation of missing modalities compared to relevant baselines from previous work.

In Chapter 6, we model shared and modality-specific information explicitly by partitioning the latent space of multimodal VAEs into respective subspaces. We explore the benefits and pitfalls of partitioning and develop the MMVAE+ as a method that promotes the desired disentanglement for the respective subspaces. Consequently, the model further improves

the generative performance compared to models with a joint latent space, especially for generating missing modalities.

On the contrary, in Chapter 7, we also establish fundamental limitations for generative learning with multimodal VAEs. We show that the sub-sampling of modalities enforces an undesirable bound for the approximation of the joint distribution, which limits the generative performance of mixture-based multimodal VAEs. In particular, we find that it restricts their utility to settings where shared information can be predicted in expectation across modalities on the level of observations.

Multimodal contrastive learning In the second part of the thesis, we transition to discriminative approaches and focus on contrastive learning as a particular method. In Chapter 8, we show that multimodal contrastive learning can be used to identify shared latent factors that are invariant across modalities up to block-wise ambiguities—even in the presence of non-trivial statistical and causal dependencies between latent variables. We empirically verify the theoretical result with numerical simulations and corroborate our findings on a complex multimodal dataset of image-text pairs.

Finally, in Chapter 9, we develop a hybrid approach that combines contrastive and generative learning to address the limitations of multimodal VAEs. We propose the DMVAE, which is designed to infer shared factors using contrastive learning and to disentangle modality-specific information using VAEs with a regularization technique that suppresses the encoding of shared information. We demonstrate promising results for multimodal generative learning and showcase the disentanglement capabilities of the model.

Research questions To address the research questions, we design several methods for multimodal representation learning and characterize their limitations and tradeoffs. We use multimodal VAEs (Chapters 5–7) to encode and disentangle shared and modality-specific information effectively (Questions 1 and 2) and to infer missing modalities (Question 3). However, we also show that mixture-based multimodal VAEs are restricted to settings where shared information can be predicted in expectation across modalities on the level of observations. To address the issue, we use contrastive learning (Chapters 8–9) to encode and, in some cases, provably recover shared latent factors even when shared information cannot be predicted across modalities on the level of observations. Consequently, compared to multimodal VAEs, we find that multimodal contrastive learning is more universally applicable for the inference of shared latent factors (Question 1) but not for the encoding

of modality-specific information. Finally, we demonstrate how contrastive learning can be combined with VAEs to disentangle shared and modality-specific information (Question 2) and to infer missing modalities (Question 3).

Thus, we develop techniques for multimodal representation learning based on generative and discriminative approaches and explain in which settings they can be effective in practice.

Relevant publications In Table 1.1, we list the publications on which this thesis is based. The first part of the table shows the publications that are directly integrated into individual chapters; the second part references publications and technical reports that, while not integrated directly, have influenced the work presented here.

Structure of the thesis Chapter 2 formalizes our problem setup and goal. Chapter 3 covers relevant background and introduces the methods we build upon. The rest of the thesis is structured according to the methods used. We start with multimodal generative models in Chapters 5–7. Then, we transition to contrastive learning as a discriminative approach in Chapter 8 and use it to develop a hybrid model in Chapter 9. Finally, in Chapter 10, we summarize our findings and discuss the limitations of our work as well as opportunities for future research.

Publication	Part of
Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., & Vogt, J. E. (2023). Identifiability Results for Multimodal Contrastive Learning. *ICLR 2023*.	Chapters 2, 8
Palumbo, E., **Daunhawer, I.**, & Vogt, J. E. (2023). MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. *ICLR 2023*.	Chapter 6
Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., & Vogt, J. E. (2022). On the Limitations of Multimodal VAEs. *ICLR 2022*.	Chapters 5, 7
Sutter*, T. M., **Daunhawer*, I.**, & Vogt, J. E. (2021). Generalized Multimodal ELBO. *ICLR 2021*.	Chapter 5
Daunhawer, I., Sutter, T. M., Marcinkevics, R., & Vogt, J. E. (2020). Self-supervised Disentanglement of Modality-specific and Shared Factors Improves Multimodal Generative Models. *GCPR 2020*.	Chapter 9
Daunhawer*, I., Schumacher*, K., Badura, A., Vogt, J. E., Michel, H., & Wellmann, S. (2023). Validating the Early Phototherapy Prediction Tool across Cohorts. *Frontiers in Pediatrics*.	–
Shi, Y., **Daunhawer, I.**, Vogt, J. E., Torr, P. H. S., & Sanyal, A. (2023). How Robust are Pre-trained Models to Distribution Shift? *ICLR 2023*.	–
Bizeul, A., **Daunhawer, I.**, Palumbo, E., Schölkopf, B., Marx, A., & Vogt, J. E. (2023). 3DIdentBox: A Toolbox for Identifiability Benchmarking. *CleaR 2023 (Dataset Track)*.	–
Stocker*, M., **Daunhawer*, I.**, Van Herk, W., El Helou, S., Dutta, S., Schuerman, F. A., ..., & Vogt, J. E. (2022). Machine Learning Used to Compare the Diagnostic Accuracy of Risk Factors, Clinical Signs and Biomarkers and to Develop a New Prediction Model for Neonatal Early-onset Sepsis. *The Pediatric Infectious Disease Journal*. LWW.	–
Montoya, J. M., **Daunhawer, I.**, Vogt, J. E., & Wiering, M. A. (2021). Decoupling State Representation Methods from Reinforcement Learning in Car Racing. *ICAART 2021*.	–
Sutter, T. M., **Daunhawer, I.**, & Vogt, J. E. (2020). Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence. *NeurIPS 2020*.	–
Pattisapu, V. K., **Daunhawer, I.**, Weikert, T., Sauter, A., Stieltjes, B., & Vogt, J. E. (2020). PET-guided Attention Network for Segmentation of Lung Tumors from PET/CT images. *GCPR 2020*.	–
Koch, G., Pfister, M., **Daunhawer, I.**, Wilbaux, M., Wellmann, S., & Vogt, J. E. (2020). Pharmacometrics and Machine Learning Partner to Advance Clinical Data Analysis. *Clinical Pharmacology and Therapeutics*. Wiley-Blackwell.	–
Daunhawer*, I., Kasser*, S., Koch, G., Sieber, L., Cakal, H., Tuetsch, J., Pfister, M., Wellmann, S., & Vogt, J. E. (2019). Enhanced Early Prediction of Clinically Relevant Neonatal Hyperbilirubinemia with Machine Learning. *Pediatric Research*. Nature.	–
Daunhawer, I., Sutter, T. M., & Vogt, J. E. (2019). Improving Multimodal Generative Models with Disentangled Latent Partitions. *NeurIPS 2019 Workshop on Bayesian Deep Learning*.	–

Table 1.1: List of publications on which this thesis is based. The first part of the table shows the publications directly integrated into the thesis; the second part lists those publications and technical reports that are relevant to the investigation but are not included in the thesis directly. Within each part, publications are sorted in reverse chronological order based on their date of publication. Equal contribution is denoted by an asterisk ($*$).

2

Problem Formulation

In the previous chapter, we introduced the setup of multimodal learning under weak supervision and motivated the goal of learning representations that recover latent factors of variation. In this chapter, we formalize the problem setup in terms of a generative process comprised of ground truth latent variables, which can be shared between observations of different modalities, and modality-specific mechanisms that produce the observations.

First, we provide a general definition of the generative process (Section 2.1) before we introduce additional assumptions on the relation between modalities (Section 2.2). Then, we consider a causal perspective and use the relational assumptions to connect our setup to the idea of common causes shared between modalities and to the so-called content-style separation (Section 2.3). We also relate our formulation to the concept of weak supervision and show that it encompasses our setup as a special case (Section 2.4). Finally, we introduce the evaluation metrics used in this thesis (Section 2.5).

2.1 Multimodal Generative Process

Let M be the number of observed modalities and let $i \in \{1, \dots, M\}$ denote the index of a modality. When it is clear from context, we simply use the index i to refer to a modality directly; for example, we typically write "modality i" instead of "the i-th modality".

We start from the assumption that the observations are generated by an unknown process that we want to learn about. Throughout this thesis, we assume the generative process given by Definition 1, for which the notion of independent mechanisms is specified in Section 2.3.

Definition 1 (Multimodal Generative Process)**.** *Let* $\mathbf{z}$ *be a random vector that takes values in the latent space* $\mathcal{Z}$*. Further, let* $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ *be a an ordered set of random vectors, each of which takes values in the ambient space* $\mathcal{X}_i$ *of modality* $i \in \{1, \ldots, M\}$ *respectively.*

The data generating process that produces the observations $(\mathbf{x}_1, \ldots, \mathbf{x}_M)$ *is defined as*

$$\mathbf{z} \sim p(\mathbf{z}) \,, \quad (\mathbf{x}_1, \ldots, \mathbf{x}_M) \sim p(\mathbf{x}_1, \ldots, \mathbf{x}_M) \,, \quad \mathbf{x}_i = \mathbf{f}_i(\mathbf{z}), \; i \in \{1, \ldots, M\} \,, \tag{2.1}$$

where each function $\mathbf{f}_i : \mathcal{Z} \to \mathcal{X}_i$ *represents an independent mechanism.*

In this context, we assume that the functions $\mathbf{f}_1, \ldots, \mathbf{f}_M$ as well as the latent variables (i.e., the components of $\mathbf{z}$) are unknown and that we only observe the functions' outputs. Specifically, we assume a dataset $\mathcal{D} = \{(\mathbf{x}_1^{(n)}, \ldots, \mathbf{x}_M^{(n)})\}_{n=1}^N$ of size N, where each element is an M-tuple of corresponding observations sampled from a joint distribution of multiple modalities, i.e., $(\mathbf{x}_1, \ldots, \mathbf{x}_M) \sim p(\mathbf{x}_1, \ldots, \mathbf{x}_M)$. Throughout this thesis, we usually drop the superscript "(n)" to simplify the notation when it is clear from context that we refer to a single observation.

Moreover, we typically assume that the space of observations is significantly more complex and of higher dimensionality than the latent space. For example, if we consider monochromatic images of size 64×64, where each pixel is an integer between 0 and 255, there are 256^{4096} possible combinations of values in "pixel space". In contrast, latent variables—i.e., the underlying factors of variation—can take values in a lower-dimensional space. For instance, they can represent class labels or attributes that provide a concise description of the contents of an image.

2.2 Relational Assumptions

While Definition 1 describes the algorithmic process that generates the data, it does not make further assumptions about the relation between modalities, i.e., their dependence structure. Throughout this thesis, we consider the following set of relational assumptions (Assumptions 1–3) and sometimes only a subset thereof.

First, we assume a *statistical dependence* between pairs of modalities (Assumption 1); otherwise, multimodal learning would have little benefit and (independent) modalities could be considered separately. Second, we assume *distinct mechanisms* (Assumption 2) to model the heterogeneity of multimodal data, which follows naturally when observations are produced by different types of sensors [BAM19]. Third, we assume *content invariance* (Assumption 3) to ensure that a subset of latent variables $\mathbf{c}_{ij}$, which take values on a subspace $\mathcal{C}_{ij} \subseteq \mathcal{Z}$, can be fully recovered from each modality $k \in \{i, j\}$.[2] Though, we do *not* assume conditional independence $(\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j) \,|\, \mathbf{c}_{ij}$ to allow for latent variables that are stochastically shared between modalities but are not strictly invariant.

Assumptions 1–3 (Relation between modalities)**.** *Let M be the number of modalities and let $i, j \in \{1, \dots, M\}$ index a pair of modalities. For each pair, $i \neq j$, we assume*

(1) $\mathbf{x}_i \not\perp\!\!\!\perp \mathbf{x}_j$ *(statistical dependence)*

(2) $\mathbf{f}_i \neq \mathbf{f}_j$ *(distinct mechanisms)*

(3) $\exists\, \mathbf{z}[d], d \in \{1, \dots, D\}$, *s.t.* $\mathbf{f}_i$ *and* $\mathbf{f}_j$ *are invertible in* $\mathbf{z}[d]$ *(content invariance)*

where $\mathbf{z}[d]$ denotes the d-th component of the latent vector $\mathbf{z}$ with D components. We denote the latent variables that satisfy Assumption (3) with the original order maintained as the vector $\mathbf{c}_{ij}$, which stands for "content" variables shared between the pair of modalities i, j.

For example, based on Definition 1 and Assumptions 1–3, we might observe a dataset of image-text pairs with the following properties. Based on Assumption 1, there would be some overlap between the contents of the images and corresponding texts—even though they were recorded with different devices (i.e., different generative mechanisms; Assumption 2). Following Assumption 3, each image-text pair would contain a certain feature that is shared between modalities. For instance, think of a class label that categorizes an object that is depicted in the image and described by the text.

Multimodal or multi-view? While these terms are sometimes used interchangeably in the literature, we distinguish between them and consider the multimodal setup to be more general. In the case of multiple modalities, we assume observations of different

[2] In other words, content invariance requires that there is no information loss with respect to $\mathbf{c}_{ij}$ when generating $\mathbf{x}_k = \mathbf{f}_k(\mathbf{c}_{ij}, \cdot)$ for each $k \in \{i, j\}$.

types or formats (e.g., image and audio data) produced by various types of devices (e.g., cameras and microphones), each of which represents a distinct generative mechanism (c.f., Assumption 2). In contrast, we regard the multi-view setup as a special case, where identical mechanisms (e.g., multiple cameras) produce observations of the same type and format (e.g., images from different perspectives). This distinction is essential, because in the multimodal setup the relation between observations can differ on the level of latent variables and generating functions, whereas in the multi-view setup it differs only with respect to the latent variables.

2.3 Causal Perspective

Next, we take a causal perspective on the considered setup to move beyond statistical dependencies and to provide more context with respect to ideas from existing literature.

In Section 2.3.1, we introduce two assumptions from the literature on causal inference—namely, the *common cause principle* and the *independent mechanisms principle*—and invoke them to derive the following implications for our setup. First, in Section 2.3.2, we show that given statistically dependent modalities (Assumption 1), the latent vector can be viewed as containing a set of *common causes* [Rei56; Kör+07]. Second, in Section 2.3.3, we show that content invariance (Assumption 3) implies a partition of the latent vector into subsets of invariant and changing variables, motivating the so-called problem of *content-style separation*, which stems from factor models [TF96; TF00; EL04; Vir+12; Kla+14] and appears in different flavors in the context of disentanglement, domain adaptation, and style transfer (e.g., [GEB16; ZZC18; Hua+18; BTN18; Pre+19]).

2.3.1 Common Causes and Independent Mechanisms

First, to establish a connection between statistical and causal dependencies, we consider Reichenbach's common cause principle [Rei56], which states that for any pair of statistically dependent variables there exists a third variable that causally influences both variables in the pair.

Principle 1 (Common cause principle, [Rei56; PJS17])**.** *If two random variables* $\mathbf{x}$ *and* $\mathbf{y}$ *are statistically dependent* ($\mathbf{x} \not\perp\!\!\!\perp \mathbf{y}$)*, then there exists a third variable* $\mathbf{z}$ *that causally influ-*

ences both (as a special case, $\mathbf{z}$ *may coincide with either* $\mathbf{x}$ *or* $\mathbf{y}$*). Furthermore,* $\mathbf{z}$ *screens* $\mathbf{x}$ *and* $\mathbf{y}$ *from each other in the sense that given* $\mathbf{z}$*, they become independent,* $\mathbf{x} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{z}$*.*

Second, we consider the principle of (physically) independent mechanisms [PJS17], which states that the individual functions or mechanisms in a generative process do not inform or influence each other. For our setup, the principle is helpful to reason about interventions, invariances, and conditional independence structures.

Principle 2 (Independent mechanisms, [PJS17])**.** *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other conditional distributions. In case we have only two variables, this reduces to an independence between the cause distribution and the mechanism producing the effect distribution.*

Generally, these principles are widely used assumptions in the literature on causal inference. Nevertheless, we would like to point out that there are also arguments why they should not be taken for granted (e.g., see Section 1.3 in [PJS17]).

2.3.2 From Statistical Dependencies to Common Causes

Based on the assumption of statistically dependent modalities (Assumption 1), it is easy to show that the common cause principle implies that the latent vector contains a set of common causes shared between modalities.

Proposition 1 (Common causes between modalities)**.** *Consider the multimodal generative process (Definition 1) with dependent modalities (Assumption 1) and assume that none of the components of the latent vector* $\mathbf{z}$ *is observed directly. Then, for each pair of modalities* $i, j \in \{1, \dots, M\}$*, the following causal structure holds:*

$$\mathbf{x}_i \leftarrow \mathbf{z} \rightarrow \mathbf{x}_j \ , \tag{2.2}$$

where $\mathbf{z}$ *contains a set of common causes shared between the pair of modalities* i, j*.*

Proof. Equation (2.2) follows from the common cause principle, since there is a statistical dependence between modalities i, j (Assumption 1) and both functions, $\mathbf{f}_i$ and $\mathbf{f}_j$, depend

on the latent vector $\mathbf{z}$ (Equation 2.1). Moreover, we can exclude direct causal effects $\mathbf{x}_i \to \mathbf{x}_j$ (or $\mathbf{x}_i \leftarrow \mathbf{x}_j$) because we assume that latent variables are not observed directly.

□

The assumption that latent variables are not observed directly corresponds to the perspective of indirect realism, i.e., the materialist view that distal objects are not perceived directly but only indirectly through some physical process or as an internal representation (e.g., [CF21]). A direct causal effect between modalities requires that at least one modality provides immediate access to a subset of latent variables—akin to the idea of direct perceptual experience of external objects, which would correspond to the position of direct realism in the philosophy of perception.

2.3.3 Content-style Separation of the Latent Space

Next, let us additionally consider the assumption of content invariance (Assumption 3). We find that the latent vector can be partitioned into a set of invariant components and a set changing components, which we refer to as *content* and *style* respectively.

Proposition 2 formalizes the content-style separation for the multimodal setup using an information-theoretic characterization of invariance. In the following, $H(\cdot)$ is the entropy of a random variable and $I(\cdot\,;\cdot)$ is the mutual information between two random variables. In this chapter, we assume discrete random variables and thus use the Shannon entropy.[3]

Proposition 2 (Content-style separation)**.** *Consider the multimodal generative process (Definition 1) with dependent modalities (Assumption 1) and assume that none of the components of the latent vector* $\mathbf{z}$ *is observed directly. Further, assume that there exists a set of invariant latent factors (Assumption 3). Then, for each pair of modalities* $i, j \in \{1, \dots, M\}$*, the following causal structure holds:*

$$\mathbf{x}_i \leftarrow (\mathbf{c}_{ij}, \mathbf{s}_{ij}) \to \mathbf{x}_j \tag{2.3}$$

where $(\mathbf{c}_{ij},\ \mathbf{s}_{ij})$ *is a unique partitioning of the latent vector* $\mathbf{z}$*, such that the following inequality relations hold:*

(i) $I(\mathbf{c}_{ij}; \mathbf{x}_k) = H(\mathbf{c}_{ij})$ *for each* $k \in \{i, j\}$*, and*

[3] Definitions of information-theoretic quantities are provided in Chapter 3, Section 3.3.4.1.

(ii) $I(\mathbf{s}_{ij}; \mathbf{x}_l) < H(\mathbf{s}_{ij})$ *for some* $l \in \{i, j\}$.

Consequently, we call the partitioning from Equation (2.3) *the content-style separation of* $\mathbf{z}$ *for the pair of modalities* i, j.

Proof. Assumption (3) states that the functions $\mathbf{f}_i$ and $\mathbf{f}_j$ are invertible with respect to $\mathbf{c}_{ij}$, which denotes a subset of components of the latent vector $\mathbf{z}$. Since $\mathbf{c}_{ij}$ takes values in $\mathcal{C}_{ij} \subseteq \mathcal{Z}$, i.e., a subspace of $\mathcal{Z}$, the invertibility of $\mathbf{f}_i$ and $\mathbf{f}_j$ with respect to $\mathbf{c}_{ij}$ implies that for each point $\mathbf{z} \in \mathcal{Z}$ the information content of its components $\mathbf{c}_{ij}$ is preserved throughout the functions $\mathbf{f}_i$ and $\mathbf{f}_j$. Hence, it follows that $I(\mathbf{x}_i; \mathbf{c}_{ij}) = H(\mathbf{c}_{ij})$ and $I(\mathbf{x}_j; \mathbf{c}_{ij}) = H(\mathbf{c}_{ij})$.

On the contrary, at least one function, $\mathbf{f}_i$ or $\mathbf{f}_j$, is not invertible with respect to the remaining components $\mathbf{s}_{ij}$, since otherwise these components would be considered part of the partition $\mathbf{c}_{ij}$ by definition. Hence, it follows that there are points $\mathbf{z} \in \mathcal{Z}$ for which some information content of its components $\mathbf{s}_{ij}$ is not preserved throughout at least one of the functions $\mathbf{f}_i$ or $\mathbf{f}_j$. Consequently, we have $I(\mathbf{x}_l; \mathbf{s}_{ij}) < H(\mathbf{s}_{ij})$ for some $l \in \{i, j\}$. □

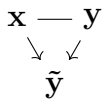
Thus, we made a connection to the existing problem of content-style separation, applied to the multimodal setup and based on the assumption of content invariance. We revisit the idea of content-style separation in Chapter 8, where we investigate how shared latent factors that are invariant across modalities can be identified up to acceptable ambiguities.

In summary, in this subsection we considered a causal perspective on the problem setup. We invoked the common cause principle to view the latent vector as containing a set of common causes shared between modalities. Further, we saw that if we additionally assume content invariance, the latent vector can be partitioned into a set of invariant and a set of changing components, akin to the idea of content-style separation.

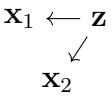
2.4 Weak Supervision

Next, we draw a connection to the concept of weak supervision. Specifically, we show that the considered generative process can be viewed as a special case of weak supervision.

In supervised learning, we consider pairs $(\mathbf{x}, \mathbf{y})$ of corresponding observations $\mathbf{x} \in \mathcal{X}$ and labels $\mathbf{y} \in \mathcal{Y}$ from which we want to learn a function $\mathbf{g} : \mathcal{X} \to \mathcal{Y}$ to model the conditional distribution $p(\mathbf{y} \mid \mathbf{x})$ and to make predictions on new, unlabeled data. In weakly supervised learning, we might be interested in learning the same relation, but instead of



(a) Weak supervision



(b) Multimodal setup

Figure 2.1: Comparison of the dependence structure underlying weak supervision (Figure 2.1a) and the multimodal setup with two modalities (Figure 2.1b). In Figure 2.1a, the edge from $\mathbf{x}$ to $\tilde{\mathbf{y}}$ is optional and the undirected edge between $\mathbf{x}$ and $\mathbf{y}$ represents a Markov equivalence class of causal DAGs

the ground truth labels, we are given corrupted or incomplete labels—hence the idea of *weak* supervision [HGIL16; Zho17], which, unfortunately, is an overloaded term [Sug+22].

To the best of our knowledge, there is no universal definition for weak supervision. In fact, the concept carries substantially different meaning across sub-disciplines. For example, Sugiyama et al. [Sug+22] point out that the term is used to denote incomplete supervision, inexact supervision, inaccurate supervision, and the use of heuristic labeling functions. For the purpose of our investigation, we see most overlap with the topic of learning from noisy or corrupted labels [Nat+13; RW17] and with weakly supervised disentangled representation learning [BTN18; Loc+20b; Shu+20; CB20; Bre+22].

We propose the following working definition, for which we take a causal perspective and assume that the weak labels are derived from the ground truth labels and that they can additionally depend on the observations.

Definition 2 (Weak supervision)**.** *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be random vectors that describe observations and labels respectively. We call* $\tilde{\mathbf{y}}$ *a weak label if it depends causally on* $\mathbf{y}$ *and, optionally, also on* $\mathbf{x}$*. Then, we say that* $\tilde{\mathbf{y}}$ *provides weak supervision w.r.t.* $p(\mathbf{y} \mid \mathbf{x})$*.*

The data generating process for weak labels is described by the Markov chain shown in Figure 2.1a, where the edge from $\mathbf{x}$ to $\tilde{\mathbf{y}}$ is optional and the undirected edge between $\mathbf{x}$ and $\mathbf{y}$ represents a Markov equivalence class of causal DAGs, meaning that it permits both causal directions between observations and labels, i.e., $\mathbf{x} \to \mathbf{y}$ and $\mathbf{x} \leftarrow \mathbf{y}$.

2.4.1 Weak Supervision Encompasses the Multimodal Setup

Based on Figure 2.1, it is easy to see that the multimodal setup represents a special case of weak supervision. First, consider the generative process from Definition 1 with two

statistically dependent modalities, $\mathbf{x}_1$ and $\mathbf{x}_2$, for which, by Proposition 1, we get the dependence structure shown in Figure 2.1b. To see the connection to weak supervision, compare the dependence structure to the one in Figure 2.1a. Since the Markov equivalence class of the relation $\mathbf{x} - \mathbf{y}$ includes the causal direction $\mathbf{x} \leftarrow \mathbf{y}$ and the dependence from $\mathbf{x}$ to $\tilde{\mathbf{y}}$ is optional, we can equate

$$\mathbf{y} = \mathbf{z} \,, \tag{2.4}$$

$$\mathbf{x} = \mathbf{x}_1 \,, \tag{2.5}$$

$$\tilde{\mathbf{y}} = \mathbf{x}_2 \,. \tag{2.6}$$

Hence, for the case of two modalities, we expressed the multimodal setup as a special case of weak supervision. Beyond two modalities, the same reasoning can be applied to any pair of subsets of modalities, assuming statistical dependence between at least two modalities (i.e., one from each subset) or between each pair of modalities (c.f., Assumption 1).

In summary, this section established a link between the considered setup and the topic of weak supervision by showing that weak supervision encompasses the multimodal setup as a special case.

2.5 Evaluation Metrics

In this section, we introduce the evaluation metrics that are used throughout this thesis. We describe the metrics used to assess the learned representations (Section 2.5.1) and to evaluate the generative performance of multimodal generative models (Section 2.5.2).

2.5.1 Evaluation of Learned Representations

To evaluate the quality of the learned representations, we use predictive models to probe the representations (i.e., the embeddings produced by the model) with respect to the ground truth information. Specifically, we use linear and nonlinear prediction models (i.e., linear and nonlinear probing [AB16]) to quantify how well the relevant information can be extracted from the embeddings and whether it is encoded at all. In all cases, we train predictive models on the embeddings of the training data and evaluate them on the embeddings of holdout data to control for overfitting.

Linear probing We use linear prediction models to evaluate whether the learned representations encode information about the ground truth factors in a linear separable format. Linear separability implies that the information can be read out using linear decoders [HRR22] and it is a common heuristic for disentanglement (e.g., [Hig+18]). When the target variable is continuous, we use linear regression and evaluate the R^2 coefficient of determination. When the target is discrete, we employ logistic regression and evaluate the classification accuracy. Strictly speaking, logistic regression is a generalized linear model, where the log-odds of the estimated probabilities are a linear function of the regression parameters. Nevertheless, we still use the term "linear probing", which became prevalent in machine learning (e.g., [AB16]).

Nonlinear probing We also consider nonlinear prediction functions (e.g., kernel machines or neural networks) to quantify whether the learned representations encode information about the ground truth factors in a more complex format than can be extracted by linear predictors. As for linear models, we assess the R^2 coefficient of determination if the target variable is continuous and the classification accuracy if the target is discrete. With nonlinear predictors, it is particularly important that we evaluate them on embeddings of holdout data that were not used to train the models.

2.5.2 Evaluation of Multimodal Generative Models

Next, we introduce the metrics used to evaluate the generative performance of multimodal generative models based on two criteria. On the one hand, we evaluate the *generative quality* of a model, i.e., the quality of its generated samples. On the other hand, we also consider the *semantic coherence* for the inference across modalities, through the generation of missing modalities based on the learned representations (c.f., Question 3).

Generative quality Since we use likelihood-based approaches, we evaluate the log-likelihood of the trained models on the test data. In addition, we also use the Fréchet inception distance (FID; [Heu+17]), a standard metric for evaluating the quality of generated images. The FID compares the distribution of generated images with the distribution of real images (e.g., the test dataset) in the feature space of a pre-trained network, typically using an Inception v3 network [Sze+15] trained on ImageNet [Den+09]. A lower FID

indicates that the generated samples are more similar to the real data and a value of zero means that the samples are indistinguishable in terms of their statistical features.

Hence, we employ two complementary metrics for generative quality. On the one hand, the log-likelihood provides a modality-agnostic metric, but it can be hard to interpret and does not always faithfully reflect the quality of samples [TOB16]. On the other hand, the FID is specific to images, but it correlates strongly with the perceived quality of generated samples as evaluated by humans [Heu+17]. Notably, both metrics are computed based on observations and generated samples, so they can be used to assess generative quality without ground truth labels.

Semantic coherence For models that can draw inferences across modalities, we estimate the semantic coherence of generated samples on the level of observations. Specifically, we evaluate the *generative coherence* [Shi+19], which measures whether generated samples agree in their semantic content across modalities. Notably, the computation of generative coherence requires ground truth labels that indicate what is shared between modalities as well as a pre-trained classifier that evaluates the semantic content of the generated samples.

Let $\mathbf{y}$ be a discrete target variable (e.g., a class label), let $\hat{\mathbf{x}}_i$ be a generated example of modality i, and let $\mathbf{h}_\varphi : \mathbf{x}_i \mapsto \mathbf{y}$ be a pre-trained classifier with parameters φ that predicts a class label for a given observation. Then, generative coherence can be computed as

$$\text{Coherence}(\mathbf{y}, \hat{\mathbf{x}}_i; \mathbf{h}_\varphi) = \mathbb{1}_{\{\mathbf{h}_\varphi(\hat{\mathbf{x}}_i) = \mathbf{y}\}} \,, \tag{2.7}$$

where the right-hand side is an indicator function that evaluates to one if the predicted class label for the generated example matches the ground truth value of the target variable.[4]

To assess the conditional generation, we compute the *conditional coherence accuracy*, for which we average the coherence values over a set of N conditionally generated examples, where N is typically the size of the test set. As a relevant special case, when $\hat{\mathbf{x}}_i$ is conditionally generated given a subset of modalities A, such that $A = \{1, \ldots, M\} \setminus i$, we call the metric *leave-one-out conditional coherence accuracy*, because the input consists of all modalities except the one being generated. Another special case is the *pairwise coherence accuracy*, for which we compute the average coherence across all pairs of distinct modalities.

[4]Here, we define coherence only for discrete factors, but the definition could easily be extended to continuous variables. For example, we could define the mean squared error between a continuous target and the output of a pre-trained regression model.

Analogously, we can also compute the coherence using unconditionally generated examples produced by a generative model (e.g., based on samples from a prior distribution). In this case, we do not need ground truth labels and instead compare the predicted class labels across all generated modalities. Only if all predicted class labels match, the *unconditional coherence accuracy* evaluates to one, which makes the task more difficult if the number of modalities is large.

When it is clear from context which type of coherence is used, we refer to the metric simply as "generative coherence". In each case, we describe the range as "coherence accuracy" because it always represents an accuracy measure with values between 0 and 1.

2.6 Summary

In this chapter, we formalized the setup in terms of a multimodal generative process with latent variables and modality-specific mechanisms and considered additional assumptions for the relation between modalities. Furthermore, we provided a causal perspective on the setup and drew a connection to the general concept of weak supervision. Finally, we introduced the evaluation metrics we use to assess the quality of the learned representations as well as the performance of multimodal generative models.

Next, we contextualize the problem formulation with respect to previous work and introduce the methods that we build upon in the following chapters.

3

Background

In the previous chapter, we presented our formulation for the considered setup of multimodal learning under weak supervision. In this chapter, we cover relevant background on the topic of representation learning and discuss related problem formulations from previous work. Further, we introduce the methods that we build upon in the following chapters.

We start by contextualizing our formulation with respect to previous work in representation learning (Section 3.1) and independent component analysis (Section 3.2). Then, we introduce relevant background with respect to the used methods. First, we give a brief introduction to the variational autoencoder (Section 3.3), which we use as a generative approach for multimodal representation learning. Second, we introduce contrastive learning (Section 3.4), which provides a discriminative approach. In both cases, we first introduce the general method before we discuss existing multimodal extensions thereof.

3.1 Representation Learning

In this section, we provide relevant background on the topic of representation learning. We first introduce some basic concepts (Section 3.1.1) before we discuss the historical context and locate our work in the research landscape of representation learning (Section 3.1.2).

3.1.1 Preliminaries

Broadly, the term "representation learning" is used to describe the process of learning encoders or feature extractors that map high-dimensional observations onto another, typically lower-dimensional space, which is referred to as the embedding or representation space.[5] The resulting embedding or representation can be thought of as a transformation of the data that preserves useful properties (e.g., information relevant for a downstream task) and filters out nuisance factors such as task-irrelevant information and random noise. Ideally, a representation should be useful for many different tasks [BCV13], but in general the usefulness of a representation depends on the application.

The encoder Based on the notation from Chapter 2, we denote the *encoder* for modality i as a function $\mathbf{g}_i : \mathcal{X}_i \to \hat{\mathcal{Z}}_i$ and the random vector that describes the embeddings as $\hat{\mathbf{z}}_i = \mathbf{g}_i(\mathbf{x}_i)$. Consequently, from the ground truth latents to the embeddings, we have the following Markov chain:

$$\mathbf{z} \xrightarrow{\mathbf{f}_i} \mathbf{x}_i \xrightarrow{\mathbf{g}_i} \hat{\mathbf{z}}_i \ , \tag{3.1}$$

such that $\mathbf{z} \perp\!\!\!\perp \hat{\mathbf{z}}_i \,|\, \mathbf{x}_i$ without further assumptions on the training of the encoder. We assume that the encoder is parameterized by a set of parameters ϕ_i specific to modality i and denote the parameterization explicitly as $\mathbf{g}_{\phi_i}$ when it is not clear from context.

The decoder Analogously, we denote the *decoder* from embeddings of modality i to observations of modality j as a function $\tilde{\mathbf{g}}_{i\to j} : \hat{\mathcal{Z}}_i \to \mathcal{X}_j$ and the random vector that describes the estimated observation as $\hat{\mathbf{x}}_{i\to j} = \tilde{\mathbf{g}}_{i\to j}(\hat{\mathbf{z}}_i)$. We assume that the decoder is parameterized by a set of modality-specific parameters $\theta_{i\to j}$ and denote the parameterization explicitly as $\tilde{\mathbf{g}}_{\theta_{i\to j}}$ when it is not clear from context. In the special case of $i = j$, the transformation $\tilde{\mathbf{g}}_{i\to j} \circ \mathbf{g}_i$ forms the basis of an *autoencoder* (see Section 3.3) and the estimated observation $\hat{\mathbf{x}}_{i\to j} = \tilde{\mathbf{g}}_{i\to j}(\mathbf{g}_i(\mathbf{x}_i))$ is called a *reconstruction* of the input $\mathbf{x}_i$.

3.1.2 Relevant Context

In the following, we provide some context for the topic of representation learning and the related idea of disentanglement. We touch upon the historical background and contextualize

[5] We often use the terms "representation", "embedding", and "encoding" synonymously.

our study with respect to previous work in the relevant subfields. Essentially, we situate our research in the neighborhood of multi-view and multimodal learning and in particular among works that seek to learn representations that recover underlying factors of variation.

Historical backdrop With the advent of deep neural networks, representation learning has become an increasingly popular research topic, as the field of machine learning shifted from manual feature engineering to automated approaches for feature extraction. The term "representation learning" became prevalent about a decade ago when a notable review by Bengio, Courville, and Vincent [BCV13] popularized the term and sparked the interest of the research community in the topic of disentangled representation learning.[6] A central theme of their review is the idea that certain representations of the data are better than others and that a good representation would ideally disentangle the underlying explanatory factors of variation. While the review provides a comprehensive, bird's eye perspective on the landscape and opportunities of the field at the time, it relies on intuitive but overly simple notions of disentanglement (e.g., statistically independent factors) and ultimately does not provide a formal definition of the concept. Following the idea, a myriad of disentanglement methods were proposed, many of which are heuristic and lack a proper theoretical foundation [c.f., HKM23]. In parallel, though at a slower pace, several theories were developed to build up a principled understanding of disentangled representation learning. Notably, there are approaches based on group theory [Hig+18; HRR22], causal inference [Sch+21], and independent component analysis or ICA [HKM23]. Our work also draws on concepts and ideas from causality and ICA, which we sketch out in the following and discuss more formally in Section 3.2.

Unsupervised disentanglement The topic of disentanglement (i.e., disentangled representation learning) became ubiquitous in the years following the review from Bengio, Courville, and Vincent [BCV13]. Across the different formulations of disentanglement, the holy grail represents what is sometimes called *unsupervised* disentanglement, which is analogous to the idea of recovering the factors of variation from observations alone. While the idea is compelling, the endeavor remains quixotic. In the general case, unsupervised disentanglement is proven to be theoretically impossible [HP99; Loc+19] and unsupervised model selection appears challenging in practice [Loc+19].

[6] In the same year, the International Conference on Learning Representations (ICLR) was initiated, which consistently ranks among the most impactful venues in machine learning and beyond [Goo].

Fortunately, there are more optimistic prospects for disentanglement, if we are willing to make additional assumptions. For example, Locatello et al. [Loc+20a] show that statistically independent factors can be disentangled successfully, when at least a few ground truth labels are observed and used to guide training or model selection. Moreover, a line of pioneering works establishes identifiability results—even in the challenging case of a nonlinear generative process—under additional assumptions, such as temporal structure [Har+03; SZW14; HM16; HM17], auxiliary variables [HST19; Khe+20], or weak supervision in the form of multiple views [Gre+19; Loc+20b; Zim+21]. Similar to these lines of research, our goal is to recover the latent factors up to acceptable ambiguities, but we focus on using weak supervision in the form of multiple modalities.

Causal Representation Learning Causal inference provides a formal language to ground the idea of discovering high-level causal variables given low-level observations—similar to the idea of identifying ground truth latent variables, which is central to ICA. However, causal representation learning goes beyond the classic setup of ICA and considers causal dependencies between latent variables (see Section 3.2). Moreover, as highlighted in an influential review from Schölkopf et al. [Sch+21], the causal perspective on disentanglement permits a formal characterization of the quality of a representation in terms of its robustness and out-of-distribution generalization performance. Therefore, causal inference offers the concept of *interventions* [Pea00; Spi+00], i.e., changes to a value or distribution of a variable in the underlying causal graph. In later chapters, we also draw on the tools of causal inference to reason about identifiability beyond ICA theory.

In summary, our work is situated in the neighborhood of multi-view and multimodal representation learning, specifically among studies that take a causal perspective on the problem and seek to learn representations that recover the underlying factors of variation up to acceptable ambiguities. Next, we formalize the relation to previous work through the framework of independent components analysis and extensions thereof.

3.2 Independent Component Analysis

Independent component analysis (ICA) is a statistical technique to decompose a signal into independent parts [HO00]. It was developed to address the problem of *blind source separation* (BSS), which seeks to separate a mixture of signals into its individual source

components, ideally using as few assumptions as possible [CJ10]. A classic example is the cocktail party problem, where a listener or listening device tries to tease apart the speech of one person from a cacophony of different sounds and speakers. Essentially, the problem of blind source separation is similar to the idea of identifying a set of ground truth latent variables, though BSS is more frequently encountered in signal processing, specifically in settings with separable temporal signals. In general, the BSS problem is highly underdetermined, but useful solutions exist under many realistic conditions (e.g., [CJ10]). ICA offers a theoretical framework that presents a solution to the BSS problem by finding transformations of the mixed signal that produce maximally uncorrelated or independent signals in a statistical or information-theoretic sense.

In ICA, we consider the following generative process [c.f., HKM23]:

$$\mathbf{x} = \mathbf{f}(\mathbf{z}), \quad \mathbf{z} \sim p(\mathbf{z}), \quad p(\mathbf{z}) = \prod_{d=1}^{D} p(z_d) \ , \tag{3.2}$$

where an observation $\mathbf{x}$ is generated from a mixing function $\mathbf{f}$ that is applied to a latent vector $\mathbf{z} = (z_1, \ldots, z_D)$ with independent components. The goal of ICA is to invert the mixing function in order to recover the latent variables (i.e., the components of $\mathbf{z}$). Specifically, it seeks to learn the inverse function given a dataset of observations. Our formulation from Chapter 2 resembles the setup of ICA, but in contrast to Equation (3.2), we consider multiple modalities and do not assume independent latent variables.

Challenges and compromises In many settings of ICA, it is impossible to fully recover the latent variables; specifically, when the observations are perturbed in such a way that the mixing function cannot be inverted. But even with an invertible mixing function, the latent variables might be *non-identifiable* without further assumptions.

While in many settings, full identifiability is impossible, certain ambiguities might be acceptable, which provides a potential compromise for representation learning. For example, identifiability might hold for a subset of components (i.e., partial identifiability), up to element-wise transformations and a permutation of the dimensions (i.e., component-wise indeterminacy), or up to groups of latent variables (i.e., block-wise indeterminacy).

Even in *linear* ICA, where $\mathbf{f}$ is a linear function, the latent vector can only be recovered up to a component-wise indeterminacy, which is usually considered acceptable. In other instances, such as independent subspace analysis [CW00; HH00; The06; Le+11], it can be sufficient to recover groups of variables up to a block-wise indeterminacy. In *nonlinear* ICA, where $\mathbf{f}$ is a nonlinear function, a landmark negative result establishes that the recovery of

the latent variables given i.i.d. observations is fundamentally impossible [HP99]. However, even in the challenging nonlinear case, identifiability is still feasible under additional assumptions (e.g., see the review from [HKM23]); specifically, also in the case of multiple views or modalities, as we discuss next.

3.2.1 Multi-view Nonlinear ICA

Among existing research in the field of ICA, our formulation (Chapter 2) is most similar to the setup of multi-view nonlinear ICA [Gre+19] and related formulations. Intuitively, a second view can resolve ambiguity introduced by the nonlinear mixing, if both views contain a shared signal but are otherwise sufficiently distinct [Gre+19]. For the case of two views, we can write the generative process as follows:

$$\mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}), \quad \mathbf{z} \sim p(\mathbf{z}), \quad p(\mathbf{z}) = \prod_{d=1}^{D} p(z_d) \; . \tag{3.3}$$

Hence, in the multi-view setup, the latent vector, or a subset of its components, is shared between *pairs* of observations $(\mathbf{x}_1, \mathbf{x}_2) \sim p(\mathbf{x}_1, \mathbf{x}_2)$, where the two views $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by two *nonlinear* mixing functions, $\mathbf{f}_1$ and $\mathbf{f}_2$ respectively.

The generative process in Equation (3.3) is similar to our formulation from Chapter 2, but we impose weaker assumptions on the mixing functions and dependencies between latent variables. The majority of previous works consider mutually independent latent variables [Son+14; Gre+19; Loc+20b] or independent groups of shared and view-specific variables [LF20; Lyu+22]. Moreover, some formulations do not consider view-specific mixing functions [e.g., Loc+20b; von+21]. Beyond the classic assumption of independent components, von Kügelgen et al. [von+21] consider additional causal and statistical dependencies between latent variables. Similarly, our formulation allows for more complex dependencies and additionally considers modality-specific mixing functions.

To summarize, in this section we touched upon the problem of BSS and introduced the framework of ICA, which offers a solution to the problem by decomposing a mixture of signals into independent parts. This can, under suitable conditions, lead to the identification of the latent variables up to acceptable ambiguities, which provides a theoretical basis for representation learning. Finally, we positioned our problem formulation relative to the setup studied in multi-view nonlinear ICA and related approaches. We return to the question of identifiability for the considered generative process in Chapter 8.

3.3 Variational Autoencoder

In this section, we introduce the variational autoencoder or VAE [KW14; RMW14], which provides a generative approach to representation learning based on reconstructions [Liu+23].

First, we describe the basic idea of an autoencoder, as we take a tour from the deterministic to the stochastic variant of the model (Section 3.3.1). Then, we turn to the VAE and focus on its objective, the evidence lower bound (Section 3.3.2), for which we consider a Bayesian and information-theoretic interpretation (Sections 3.3.3 and 3.3.4, respectively). Then, we touch upon the conditional VAE (Section 3.3.5) before we transition to the multimodal setup (Section 3.3.6), where the VAE is used to approximate a joint distribution of multiple modalities. In the context of multimodal VAEs, we discuss the problem of missing modalities and introduce two existing methods that are designed to tackle this problem and that we build upon in the subsequent chapters.

3.3.1 From Deterministic to Stochastic Autoencoders

The VAE is a deep latent variable model that can be viewed as a probabilistic version of the deterministic autoencoder (e.g., see [GBC16; Mur22]). In the following, we take a tour from the deterministic to the stochastic autoencoder before we return to the VAE.

An autoencoder is a type of artificial neural network that consists of two parts—the encoder and decoder networks. On a high level, an autoencoder is trained via reconstructions (c.f., Section 3.1.1). For a given observation, the encoder produces an embedding, which the decoder takes as input to reconstruct the observation. The reconstruction loss, which is typically averaged over a batch of examples, is then propagated back [Lin76; RHW86] through the autoencoder using stochastic gradient descent [RM51; KW52] to update the parameters of both networks in a single backwards pass.

Recall our notation from Section 3.1, where we denote the encoder for modality i as a function $\mathbf{g}_i : \mathcal{X}_i \to \hat{\mathcal{Z}}_i$ and the decoder as another function $\tilde{\mathbf{g}}_i : \hat{\mathcal{Z}}_i \to \mathcal{X}_i$.[7] Further, we denote the random vectors that describe the embeddings as $\hat{\mathbf{z}}_i = \mathbf{g}_i(\mathbf{x}_i)$ and the reconstructed observation as $\hat{\mathbf{x}}_i = \tilde{\mathbf{g}}(\hat{\mathbf{z}}_i)$ respectively. The inference process (i.e., the forward pass) of the

[7] To simplify the notation, we now denote the decoder with the subscript "i" (instead of "$i \to j$") because the decoder maps back to the same modality that the encoder takes as input.

autoencoder satisfies the following Markov chain:

$$\mathbf{x}_i \overset{\mathbf{g}_i}{\rightarrow} \hat{\mathbf{z}}_i \overset{\tilde{\mathbf{g}}_i}{\rightarrow} \hat{\mathbf{x}}_i \ . \tag{3.4}$$

We assume that the encoder and decoder are parameterized by sets of modality-specific parameters ϕ_i and θ_i respectively, and denote the parameterization explicitly as $\mathbf{g}_{\phi_i}$ and $\tilde{\mathbf{g}}_{\theta_i}$ when it is not clear from context.

Stochastic encoder In contrast to the deterministic autoencoder, the VAE uses a stochastic encoder, which produces a non-deterministic embedding of a given observation. A stochastic encoder provides more modeling flexibility, because it maps each observation to a distribution, whereas a deterministic encoder only yields point estimates.

More concretely, a stochastic encoder estimates the parameters of a specific distribution—the so-called variational posterior that is denoted by $q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i)$. Each sample $\mathbf{z} \sim q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i)$ drawn from the variational posterior can be thought of as a stochastic embedding of the given observation. For example, the encoder can be designed to output the mean and variance of a normal distribution $(\mu_{\mathbf{x}_i}, \sigma^2_{\mathbf{x}_i}) = \mathbf{g}_i(\mathbf{x}_i)$, where $\mu_{\mathbf{x}_i}$ and $\sigma^2_{\mathbf{x}_i}$ are the estimated parameters of the variational posterior for the given observation. This allows us to draw a set of K samples $\{\mathbf{z}^{(k)}\}_{k=1}^{K} \sim \mathcal{N}(\mu_{\mathbf{x}_i}, \sigma^2_{\mathbf{x}_i})$ for a given input. Moreover, the forward pass of the VAE satisfies the same Markov chain as the deterministic autoencoder (Equation 3.4), where the stochasticity induced by the sampling process can also be written explicitly by introducing a stochastic node $\boldsymbol{\varepsilon}$ that affects the embedding, i.e., $\mathbf{x}_i \rightarrow \hat{\mathbf{z}}_i \leftarrow \boldsymbol{\varepsilon}$.

Stochastic decoder Analogous to the encoder, the decoder of the VAE is typically stochastic as well. The stochastic decoder estimates the parameters of a distribution $p_{\theta_i}(\mathbf{x}_i \mid \mathbf{z})$ that approximates the generative model of the observations with a neural network parameterized by θ_i.[8] For example, the output is often assumed to be a normally distributed random variable with a constant variance parameter σ^2. In this case, the decoder takes as input a representation $\hat{\mathbf{z}}_i$ and outputs an estimated mean $\mu_{\hat{\mathbf{z}}_i} = \tilde{\mathbf{g}}_i(\hat{\mathbf{z}}_i)$, which is used to generate an observation $\hat{\mathbf{x}}_i \sim \mathcal{N}(\mu_{\hat{\mathbf{z}}_i}, \sigma^2 I)$. Notably, the decoder can be used for *conditional* generation, if the representation is produced by the encoder for a given observation or a corrupted version thereof. However, it can also be used for the *unconditional* generation of new observations, if the representation is sampled from a prior distribution (see Section 3.3.2).

[8] We further discuss the topic of generative modeling in Section 3.3.2.

Now that we introduced the general idea behind the autoencoder and its stochastic counterpart, we return to the VAE, which also uses a stochastic encoder and decoder. However, the VAE optimizes a slightly modified objective—the so-called evidence lower bound—that is anchored in a rich theoretical framework with connections to generative modeling, Bayesian inference, and information theory, as we discuss next.

3.3.2 Evidence Lower Bound

In this subsection, we introduce the evidence lower bound (ELBO), i.e., the objective optimized by the VAE. The ELBO represents an important theoretical concept in Bayesian inference and probabilistic modeling [Mur22]. First, we provide a definition of the ELBO before we formally derive it from a Bayesian and information-theoretic perspective in the following subsections.

To approximate a distribution $p(\mathbf{x})$, the VAE learns a generative model of the form $p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$, where $p_\theta(\mathbf{x} \mid \mathbf{z})$ is parameterized by a neural network and $p(\mathbf{z})$ is a prior distribution. For this purpose, it maximizes a lower bound on the log-evidence $\log p(\mathbf{x})$—hence, the term *evidence lower bound*.

Definition 3 (Evidence lower bound)**.** *Let* $\mathbf{x}$ *be random vector with distribution* $p(\mathbf{x})$*. Let* $q_\phi(\mathbf{z} \mid \mathbf{x})$ *be a stochastic encoder parameterized by* ϕ*, let* $p_\theta(\mathbf{x} \mid \mathbf{z})$ *be a stochastic decoder parameterized by* θ*, and let* $p(\mathbf{z})$ *be a prior distribution. The evidence lower bound (ELBO) on the log-evidence* $\log p(\mathbf{x})$ *is defined as*

$$\mathcal{L}_{ELBO}(\mathbf{x}; \phi, \theta) \coloneqq \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z})) \,, \tag{3.5}$$

where $D_{KL}(\cdot \,||\, \cdot)$ *denotes the Kullback-Leibler divergence (Definition 4).*

In practice, the VAE is trained by maximizing Equation (3.5) for a dataset of independent and identically distributed observations $\{\mathbf{x}^{(n)}\}_{n=1}^N \sim p(\mathbf{x})$.

The first term in Equation (3.5) represents the log-likelihood, which measures how well the model fits the observed data. Naturally, the log-likelihood is being maximized to improve the fit of the data. Since the log-likelihood term can be viewed as a (negative) reconstruction error [KW19], the VAE differs from the stochastic version of the autoencoder only by the second term (i.e., the KL-divergence), which is being minimized to approximate the prior. Thus, the second term can be viewed as a regularizer that penalizes the deviation of the variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ from the prior distribution $p(\mathbf{z})$.

Finally, to complete the definition of the ELBO, let us also define the KL-divergence.

Definition 4 (Kullback-Leibler Divergence)**.** *Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be two distributions defined on the same sample space $\mathcal{X}$. For discrete distributions, the Kullback-Leibler divergence from the distribution $p(\mathbf{x})$ to the reference distribution $q(\mathbf{x})$ is defined as a sum*

$$D_{KL}(p(\mathbf{x}) \,||\, q(\mathbf{x})) = \sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} \,. \tag{3.6}$$

For continuous distributions, it is defined as an integral

$$D_{KL}(p(\mathbf{x}) \,||\, q(\mathbf{x})) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \,. \tag{3.7}$$

Notably, The KL-divergence has a closed-form solution for Gaussian distributions, which are typically used in VAEs to model the variational posterior and prior distributions.

Thus, we specified the objective of the VAE—the so-called ELBO. Next, we provide a Bayesian perspective on the objective to explain why the VAE provides a principled approach for probabilistic modeling and Bayesian inference.

3.3.3 Bayesian Interpretation

From a probabilistic perspective, the VAE satisfies two complementary objectives. First, by maximizing the ELBO (Equation 3.5), the VAE approximates the marginal likelihood or evidence $p(\mathbf{x})$ with a generative model of the form $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ with parameters θ. A trained VAE yields a probabilistic generative model from which new observations can be generated by sampling $\mathbf{z} \sim p(\mathbf{z})$ and $\mathbf{x} \sim p_\theta(\mathbf{x} \mid \mathbf{z})$ from the prior and stochastic decoder respectively. Second, the VAE provides a principled, variational approach to Bayesian inference, because the encoder estimates the *posterior*

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})} \,, \tag{3.8}$$

which is typically intractable, because the evaluation of the denominator (i.e., the evidence) requires the computation of an integral

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) \, d\mathbf{z} \tag{3.9}$$

that is unavailable in closed form or requires exponential time to be computed exactly for many models including neural networks [BKM17]. The VAE approximates the (true)

posterior through amortized inference using a variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$, which can be grounded in the framework of variational Bayesian inference [KW14].

To clarify the twofold interpretation of the VAE in terms of variational inference and generative modeling, we restate a known derivation of the objective as a lower bound on the log-evidence (e.g., see [Mur23]).

Lemma 1. *Objective $\mathcal{L}_{ELBO}(\mathbf{x}; \phi, \theta)$ (Equation 3.5) forms a lower bound on $\log p(\mathbf{x})$, i.e.,*

$$\log p(\mathbf{x}) \geq \mathcal{L}_{ELBO}(\mathbf{x}; \phi, \theta) \; . \tag{3.10}$$

Proof. We start from the target quantity, the log-evidence, and introduce the variational posterior as follows:

$$\begin{aligned}
\log p(\mathbf{x}) &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\Big[\log p(\mathbf{x})\Big] && (3.11)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right] && (3.12)\\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z}) q_\phi(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x}) p(\mathbf{z} \mid \mathbf{x})}\right] && (3.13)\\
&= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]}_{\mathcal{L}_{\text{ELBO}}(\mathbf{x}; \phi, \theta)} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z} \mid \mathbf{x})}\right]}_{D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x}))} && (3.14)
\end{aligned}$$

The first term in Equation (3.14) is the ELBO and the second term the KL-divergence of the variational posterior from the true posterior. Since the KL-divergence is always non-negative, it follows that the ELBO forms a lower bound on the log-evidence. □

Notice that Equation (3.14) also contains the term $D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x}))$, which suggests that the training of the VAE also determines the gap between the variational posterior and the true posterior. This property forms the foundation for variational inference with respect to the true posterior $p(\mathbf{z} \mid \mathbf{x})$, as described by Lemma 2.

Lemma 2. *The maximization of the objective $\mathcal{L}_{ELBO}(\mathbf{x}; \phi, \theta)$ (Equation 3.5) with respect to the network parameters ϕ and θ is proportional to the minimization of the KL-divergence of the variational posterior from the true posterior, i.e.,*

$$\arg\max_{\phi, \theta} \mathcal{L}_{ELBO}(\mathbf{x}; \phi, \theta) \propto \arg\min_{\phi, \theta} D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x})) \; . \tag{3.15}$$

Proof. Re-write Equation (3.14) in terms of the ELBO

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}; \phi, \theta) = \log p(\mathbf{x}) - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x})) \tag{3.16}$$

and consider the optimization with respect to the network parameters ϕ and θ:

$$\begin{aligned}(\phi^*, \theta^*) &= \arg\max_{\phi, \theta} \mathcal{L}_{\text{ELBO}}(\mathbf{x}; \phi, \theta) && (3.17)\\ &= \arg\max_{\phi, \theta} \log p(\mathbf{x}) - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x})) && (3.18)\\ &\propto \arg\max_{\phi, \theta} -D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x}))\ . && (3.19)\end{aligned}$$

The last step (Equation 3.19) follows from the log-evidence being independent of the network parameters, which means that it represents a constant for the optimization. Hence, the maximization of the ELBO is proportional to the minimization of the KL-divergence of the variational posterior from the true posterior. □

In summary, we revisited the probabilistic perspective on the VAE, which justifies the application of the model for Bayesian inference and generative modeling. First, we showed that its objective forms a lower bound on the log-evidence (Lemma 1). Second, we demonstrated that the training of the VAE implicitly minimizes the KL-divergence of the variational from the true posterior distribution (Lemma 2).

3.3.4 Information-theoretic Interpretation

Next, we consider an information-theoretic perspective on the VAE. As before, we derive the ELBO as a variational lower bound on the log-evidence. However, in addition, we relate the ELBO to information-theoretic quantities, such as the entropy, conditional entropy, and mutual information. Let us define these terms before we proceed with the information-theoretic interpretation of the objective.

3.3.4.1 Information-theoretic Quantities

Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ denote the sample spaces of three discrete random vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ respectively, and let $p(\mathbf{x})$, $p(\mathbf{y})$, and $p(\mathbf{z})$ denote their marginal distributions.

The entropy of $\mathbf{x}$ is defined as

$$H(\mathbf{x}) = -\sum_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \log p(\mathbf{x})\ . \tag{3.20}$$

The conditional entropy of $\mathbf{x}$ given $\mathbf{y}$ is defined as

$$H(\mathbf{x} \mid \mathbf{y}) = -\sum_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} p(\mathbf{x},\mathbf{y}) \log p(\mathbf{x} \mid \mathbf{y}) \ . \tag{3.21}$$

The joint entropy of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$H(\mathbf{x},\mathbf{y}) = -\sum_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} p(\mathbf{x},\mathbf{y}) \log p(\mathbf{x},\mathbf{y}) \ . \tag{3.22}$$

The cross-entropy of the discrete probability distribution $q(\mathbf{x})$ from the discrete probability distribution $p(\mathbf{x})$ is defined as

$$CE(p(\mathbf{x}), q(\mathbf{x})) = -\sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) \log q(\mathbf{x}) \tag{3.23}$$

assuming that both distributions are defined on the same sample space $\mathcal{X}$.

The mutual information of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$I(\mathbf{x};\mathbf{y}) = D_{\mathrm{KL}}(p(\mathbf{x},\mathbf{y}) \,||\, p(\mathbf{x})p(\mathbf{y})) \ , \tag{3.24}$$

where $D_{\mathrm{KL}}(p(\mathbf{x},\mathbf{y}) \,||\, p(\mathbf{x})p(\mathbf{y}))$ denotes the Kullback-Leibler divergence (Definition 4) from the joint distribution to the product of marginals.

The conditional mutual information of $\mathbf{x}$ and $\mathbf{y}$ given $\mathbf{z}$ is defined as

$$I(\mathbf{x};\mathbf{y} \mid \mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{z}) D_{\mathrm{KL}}(p(\mathbf{x},\mathbf{y} \mid \mathbf{z}) \,||\, p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})) \ . \tag{3.25}$$

For continuous random vectors, the sums can be replaced with integrals. When we consider discrete random variables (e.g., pixel intensities, as in the case of RGB images), we can safely assume that the entropy, conditional entropy and conditional mutual information are non-negative. This is not generally true for continuous random variables [CT12].

3.3.4.2 Alternative Derivation of the ELBO

Next, we provide an alternative derivation of the ELBO based on an information-theoretic perspective. In the following, we denote the joint distribution between the random vectors $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$ and $\mathbf{x} \sim p(\mathbf{x})$ as

$$p(\mathbf{x},\mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z}) \tag{3.26}$$

$$= q_\phi(\mathbf{z} \mid \mathbf{x})p(\mathbf{x}) \ , \tag{3.27}$$

where $q_\phi(\mathbf{z} \mid \mathbf{x})$ is a stochastic encoder and $p(\mathbf{z})$ denotes its marginal distribution.

To begin with, notice that the *expected* log-evidence is equal to the negative entropy

$$-H(\mathbf{x}) = \mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] \ . \tag{3.28}$$

Given any random vector $\mathbf{z}$, the entropy can be decomposed as a sum of conditional entropy and mutual information terms:

$$H(\mathbf{x}) = H(\mathbf{x} \mid \mathbf{z}) + I(\mathbf{x}; \mathbf{z}) \ . \tag{3.29}$$

Using Equation (3.29), we can relate the expected log-evidence to the ELBO as follows:

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] &= -H(\mathbf{x} \mid \mathbf{z}) - I(\mathbf{x}; \mathbf{z}) && (3.30)\\
&\geq \mathbb{E}_{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \mathbb{E}_{p(\mathbf{x})}[D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, q(\mathbf{z}))] && (3.31)\\
&= \mathbb{E}_{p(\mathbf{x})}[\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}; \phi, \theta)] \ , && (3.32)
\end{aligned}$$

where the inequality follows from variational approximations of the respective terms. Following Alemi et al. [Ale+17], we can use the following variational bounds:

First, for the conditional entropy, we have

$$\begin{aligned}
-H(\mathbf{x} \mid \mathbf{z}) &= \mathbb{E}_{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] && (3.33)\\
&= \mathbb{E}_{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] + \mathbb{E}_{p(\mathbf{z})}\left[D_{\mathrm{KL}}(p(\mathbf{x} \mid \mathbf{z}) \,||\, p_\theta(\mathbf{x} \mid \mathbf{z}))\right] && (3.34)\\
&\geq \mathbb{E}_{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] \ , && (3.35)
\end{aligned}$$

where $p_\theta(\mathbf{x} \mid \mathbf{z})$ is a *variational decoder* parameterized by θ.

Second, for the mutual information, we have

$$\begin{aligned}
-I(\mathbf{x}; \mathbf{z}) &= -\mathbb{E}_{p(\mathbf{x})}\left[D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z}))\right] && (3.36)\\
&= -\mathbb{E}_{p(\mathbf{x})}\left[D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, q(\mathbf{z}))\right] + D_{\mathrm{KL}}(p(\mathbf{z}) \,||\, q(\mathbf{z})) && (3.37)\\
&\geq -\mathbb{E}_{p(\mathbf{x})}\left[D_{\mathrm{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, q(\mathbf{z}))\right] \ , && (3.38)
\end{aligned}$$

where $q(\mathbf{z})$ is a prior that does not necessarily equal the marginal distribution $p(\mathbf{z})$.

Hence, the ELBO forms a variational lower bound on the expected log-evidence:

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x})}[\log p(\mathbf{x})] &= \mathbb{E}_{p(\mathbf{x})}[\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}; \phi, \theta) + \Delta_{\mathrm{VA}}(\mathbf{x}; \theta)] && (3.39)\\
&\geq \mathbb{E}_{p(\mathbf{x})}[\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}; \phi, \theta)] \ , && (3.40)
\end{aligned}$$

where

$$\Delta_{\mathrm{VA}}(\mathbf{x};\theta) = \mathbb{E}_{p(\mathbf{z})}\left[D_{\mathrm{KL}}(p(\mathbf{x}\mid\mathbf{z}) \,||\, p_\theta(\mathbf{x}\mid\mathbf{z}))\right] + D_{\mathrm{KL}}(p(\mathbf{z}) \,||\, q(\mathbf{z})) \tag{3.41}$$

denotes the variational approximation gap, which is non-negative.

In summary, the information-theoretic perspective reveals an alternative derivation of the ELBO, which allows us to interpret the individual terms of the objective as variational estimators of the conditional entropy and mutual information respectively.

3.3.5 Conditional VAE

A widely-used extension of the variational autoencoder is the conditional VAE [SLY15], which incorporates conditioning information into the generative process.

For each observation, we consider an additional variable $\mathbf{c}$, i.e., we assume pairs of observations $(\mathbf{x}, \mathbf{c}) \sim p(\mathbf{x}, \mathbf{c})$. For example, $\mathbf{c}$ can be a ground truth label associated with $\mathbf{x}$.

The conditional VAE approximates $\log p(\mathbf{x}\mid\mathbf{c})$, which is the log-evidence conditioned on $\mathbf{c}$, by maximizing the objective

$$\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}|\mathbf{c};\phi,\theta) := \mathbb{E}_{q_\phi(\mathbf{z}\mid\mathbf{x})}\left[\log p_\theta(\mathbf{x}\mid\mathbf{z},\mathbf{c})\right] - D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \,||\, p(\mathbf{z}\mid\mathbf{c})\right) \ . \tag{3.42}$$

In contrast to the (unconditional) VAE, the decoder and prior in Equation (3.42) are conditioned on the variable $\mathbf{c}$, which can therefore be seen as a *context vector* that is provided as an additional input to the decoder.

Formally, Equation (3.42) forms a variational lower bound on the conditional log-evidence $\log p(\mathbf{x}\mid\mathbf{c})$ as described by Lemma 3.

Lemma 3 (Conditional ELBO)**.** *Objective $\mathcal{L}_{ELBO}(\mathbf{x}\mid\mathbf{c};\phi,\theta)$ (Equation 3.42) forms a lower bound on $\log p(\mathbf{x}\mid\mathbf{c})$, i.e.,*

$$\log p(\mathbf{x}\mid\mathbf{c}) \geq \mathcal{L}_{ELBO}(\mathbf{x}\mid\mathbf{c};\phi,\theta) \ . \tag{3.43}$$

Proof. We proceed analogously to the derivation of the ELBO in Lemma 1. Starting from

the target quantity $\log p(\mathbf{x} \mid \mathbf{c})$, we introduce the variational posterior as follows:

$$\log p(\mathbf{x} \mid \mathbf{c}) = \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\Big[\log p(\mathbf{x} \mid \mathbf{c})\Big] \tag{3.44}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{c})}{p(\mathbf{z} \mid \mathbf{x}, \mathbf{c})}\right] \tag{3.45}$$

$$= \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x})}{q_\phi(\mathbf{z} \mid \mathbf{x}) p(\mathbf{z} \mid \mathbf{x}, \mathbf{c})}\right] \tag{3.46}$$

$$= \underbrace{\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{c})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right]}_{\mathcal{L}_{\text{ELBO}}(\mathbf{x} \mid \mathbf{c}; \phi, \theta)} + \underbrace{\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z} \mid \mathbf{x})}{p(\mathbf{z} \mid \mathbf{x}, \mathbf{c})}\right]}_{D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{x}, \mathbf{c}))} \tag{3.47}$$

The first term in Equation (3.47) forms a lower bound on $\log p(\mathbf{x} \mid \mathbf{c})$, because the second term, the KL-divergence, is always non-negative.

It is easy to see that the first term corresponds to the objective from Equation (3.42), if we apply the chain rule $p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{c}) = p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c}) p(\mathbf{z} \mid \mathbf{c})$ to decompose the first term from Equation (3.47) as follows:

$$\mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}, \mathbf{z} \mid \mathbf{c})}{q_\phi(\mathbf{z} \mid \mathbf{x})}\right] = \mathbb{E}_{q_\phi(\mathbf{z} \mid \mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{c})] - D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,||\, p(\mathbf{z} \mid \mathbf{c}))\,. \tag{3.48}$$

The right-hand side of Equation (3.48) is equal to the objective in Equation (3.42). Thus, the objective $\mathcal{L}_{\text{ELBO}}(\mathbf{x} \mid \mathbf{c}; \phi, \theta)$ forms a lower bound on the conditional log-evidence $\log p(\mathbf{x} \mid \mathbf{c})$.

□

In some formulations of the conditional VAE, the variational posterior (i.e., the encoder) is also conditioned on $\mathbf{c}$, which can be justified if we additionally assume that $\mathbf{z} \perp\!\!\!\perp \mathbf{c} \mid \mathbf{x}$.

In the context of this thesis, we revisit the conditional VAE in Chapter 9.

3.3.6 Multimodal VAEs

Next, we introduce multimodal variants of the variational autoencoder. Multimodal VAEs are designed to approximate a joint distribution of multiple modalities in a way that enables inference across modalities based on the learned representations, which motivates their use in the context of this thesis (c.f., Question 3).

First, we start with the approximation of a joint distribution of a set of modalities, for which we define a multimodal ELBO. Then, we consider the problem of missing modalities and

discuss a naive solution that does not scale well as a function of the number of modalities. Finally, we introduce two approaches that handle missing modalities efficiently by using certain types of variational posteriors.

Multimodal ELBO In its basic formulation, the multimodal ELBO is a straightforward extension of the unimodal ELBO (Definition 3). In fact, we can directly apply Definition 3, if we define the observation as a set of multiple modalities. Specifically, let $\bar{\mathbf{x}} \coloneqq \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ and let $p(\bar{\mathbf{x}}) \coloneqq p(\mathbf{x}_1, \dots, \mathbf{x}_M)$ denote the joint distribution, which we want to approximate with a generative model by maximizing the multimodal ELBO.

Definition 5 (Multimodal ELBO)**.** *Let $\bar{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ be a set of random vectors that describe M modalities and let $p(\bar{\mathbf{x}})$ be the joint distribution. Further, let $q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})$ be a stochastic encoder parameterized by ϕ, let $p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})$ be a stochastic decoder parameterized by θ, and let $p(\mathbf{z})$ be a prior distribution. The evidence lower bound on the log-evidence $\log p(\bar{\mathbf{x}})$ is defined as*

$$\mathcal{L}_{ELBO}(\bar{\mathbf{x}}; \phi, \theta) \coloneqq \mathbb{E}_{q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{KL}(q_\phi(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \,, \tag{3.49}$$

where $D_{KL}(\cdot \,||\, \cdot)$ denotes the Kullback-Leibler divergence (Definition 4).

In this formulation, the multimodal ELBO is equivalent to its unimodal counterpart. We simply redefined the observation $\bar{\mathbf{x}}$ as a set of multiple modalities and adjusted the encoder and decoder accordingly.

Missing modalities Suppose we observe only one modality, say $\mathbf{x}_1$. To compute its embedding, it requires a modality-specific encoder for the (unimodal) variational posterior $q_\phi(\mathbf{z} \mid \mathbf{x}_1)$. However, in Definition 5 we only defined the variational joint posterior $q_\phi(\mathbf{z} \mid \mathbf{x}_1, \dots, \mathbf{x}_M)$, i.e., an encoder that takes *all* modalities as input.

To handle missing modalities, we could define a distinct inference model for each possible subset of modalities (e.g., [SNM16; Ved+18; HG18]). However, such an approach would quickly become intractable as the number of inference networks required grows exponentially with the number of modalities (c.f., [WG18]). Hence, it requires a more elaborate approach to handle missing modalities efficiently.

Scalable multimodal VAEs Scalable multimodal VAEs are designed to approximate the multimodal ELBO while handling missing modalities efficiently. Using only M inference

networks (i.e., one for each modality), they enable efficient posterior inference for different subsets of modalities. Consequently, they facilitate the inference across modalities based on the learned representations.

The core idea behind scalable multimodal VAEs is to design a variational joint posterior that can be decomposed in terms of the unimodal variational posteriors. Starting with the seminal work from Wu and Goodman [WG18], the variational joint posterior was defined as a product of unimodal posteriors, a so-called product of experts (PoE [Hin02])

$$q_\phi^{\text{PoE}}(\mathbf{z} \mid \mathbf{x}_1, \ldots, \mathbf{x}_M) = \prod_{i=1}^{M} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i) \ , \tag{3.50}$$

which has a simple, closed-form solution in the case of Gaussian variational posteriors [CF14; WG18]. In the follow-up work from Shi et al. [Shi+19], the variational joint posterior was defined as a mixture of unimodal posteriors, i.e., a mixture of experts (MoE)

$$q_\phi^{\text{MoE}}(\mathbf{z} \mid \mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{1}{M} \sum_{i=1}^{M} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i) \ . \tag{3.51}$$

Since both decompositions describe valid probability distributions, they can be used to define the variational joint posterior in the multimodal ELBO. Consequently, the resulting generative models approximate the joint distribution of multiple modalities with a variational joint posterior that can be decomposed in terms of the unimodal variational posterior distributions. In Chapter 5, we generalize the existing approaches by formulating the variational joint posterior as a mixture of products of experts.

With this, we conclude our recap of VAEs. To summarize, in this section we introduced autoencoders and VAEs and provided a probabilistic and information-theoretic perspective on the ELBO. Then, we returned to the topic of multimodal learning and discussed the problem of missing modalities, which is addressed by (scalable) multimodal VAEs. In the following chapters, we build upon multimodal VAEs by generalizing existing formulations (Chapter 5), modeling modality-specific information explicitly (Chapter 6), and formalizing their limitations (Chapter 7).

3.4 Contrastive Learning

In this section, we introduce contrastive learning and explain why it provides a particularly suitable method for multimodal learning under weak supervision. In contrast to the

VAE and other types of generative models, which learn representations based on the reconstruction or masked prediction error computed on the level of observations, contrastive learning takes a more direct approach to representation learning. Specifically, contrastive learning provides a discriminative approach that maximizes the similarity between different embeddings and thus casts representation learning as a prediction problem in latent space [Liu+23].

First, we review some basic ideas and developments underlying contrastive learning (Section 3.4.1). Then, we define the widely used InfoNCE objective (Section 3.4.2), which is the objective we adopt in this thesis. Next, we address two theoretical interpretations of the InfoNCE objective (Section 3.4.3). Finally, we return to the setup of multimodal learning and explain how contrastive learning can be used in this context (Section 3.4.4).

3.4.1 Preliminaries

Contrastive learning emerges from an extensive line of research, which dates back to the early 1990s [LHS20], i.e., long before the term "representation learning" became prevalent. The review from Le-Khac, Healy, and Smeaton [LHS20] examines the history of the method, discussing different formulations of the objective function and various interpretations thereof, including mutual information maximization [BH92; OLV18; Hje+19], distance metric learning [Bro+93; CHL05; HCL06; CW08; Che+10; Wan+14], and the estimation of unnormalized statistical models [GH10; GH12].

Self-supervised learning More recently, contrastive learning became widely adopted in applications of self-supervised learning [Wu+18; Che+20; He+20; Hén+20], where models are trained to solve pretext tasks (e.g., denoising, masked prediction or transformation prediction) for which labels or corresponding observations can be produced cheaply, e.g., by using data augmentations [SK19]. However, the effectiveness of contrastive learning is determined by the design of suitable pretext tasks, which in turn depend on the downstream applications [Tia+20; Mit+21]. Thus, it typically requires human expertise to design suitable pretext tasks and significant fine-tuning to choose the relevant hyperparameters, such as the composition and strength of augmentations used [Che+20].

Conveniently, in the setup of multimodal learning under weak supervision, a suitable pretext task is *inherent* to the data: given corresponding observations of different modalities, the pretext task can be defined as the prediction of the similarities between observations of

different modalities. From the perspective of representation learning, the prediction of similarities would ideally correspond to recovering the latent factors of variation shared between modalities (c.f., Question 1).

Basic concepts At its core, contrastive learning uses weak supervision in the form of corresponding observations to learn an encoder that maps similar inputs (i.e., positive pairs) closer to each other and dissimilar inputs (negative pairs) further apart (see Section 3.4.2). There exist many related algorithms in the context of multi-view representation learning [Mur23], but even non-contrastive methods (e.g., [Gri+20; CH21; Zbo+21; Car+21; BPL22]) seek to maximize the similarity between positive pairs and merely differ with respect to the strategy used to prevent a "collapsed" representation—i.e., a degenerate solution where the encoder maps to a constant [Jin+22]. In contrastive learning, this is prevented by using negative pairs in the objective.

In the context of this thesis, we consider contrastive learning based on the method of noise-contrastive estimation [GH10] that inspired the widely-used InfoNCE objective [OLV18], which we use throughout our experiments. Next, we introduce the InfoNCE objective and then discuss its theoretical underpinnings as well as extensions to multimodal learning.

3.4.2 The InfoNCE Objective

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be random vectors that take values in the same space $\mathcal{X}$ and let $p(\mathbf{x}_1, \mathbf{x}_2)$ denote their joint distribution. As such, $\mathbf{x}_1$ and $\mathbf{x}_2$ represent different views of the same modality; for example, think of two images or different augmentations of the same image. Further, let $\mathbf{g}_\phi : \mathcal{X} \to \hat{\mathcal{Z}}$ be an encoder that is parameterized by ϕ.

For a given sample $\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K \overset{\text{i.i.d.}}{\sim} p(\mathbf{x}_1, \mathbf{x}_2)$, the InfoNCE objective [OLV18] is defined as

$$\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi) := -\sum_{k=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{g}_\phi(\mathbf{x}_1^k), \mathbf{g}_\phi(\mathbf{x}_2^k))/\tau\}}{\sum_{l=1}^{K} \exp\{\text{sim}(\mathbf{g}_\phi(\mathbf{x}_1^k), \mathbf{g}_\phi(\mathbf{x}_2^l))/\tau\}} , \tag{3.52}$$

where τ is a temperature hyperparameter and $\text{sim}(\cdot, \cdot)$ is a similarity metric (e.g., the cosine similarity). Each term in the outer sum can be viewed as (the logarithm of) a softmax function that takes as input a vector of length K, comprised of the similarity between a positive pair (i.e., the numerator) and the similarity between $K-1$ negative pairs, all of which are in the denominator of the softmax function.[9]

[9] To be precise, the denominator includes one positive pair and $K-1$ negative pairs.

In terms of the distribution $p(\mathbf{x}_1, \mathbf{x}_2)$, we specify the objective as

$$\mathbb{E}_{\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K \sim p(\mathbf{x}_1, \mathbf{x}_2)} \Big[\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi) \Big] \,, \tag{3.53}$$

where the integer K becomes an additional hyperparameter that controls the number of negative pairs used for contrasting. In practice, we sample data pairs from a finite dataset $\mathcal{D} = \{(\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)})\}_{n=1}^N$ of size $N \gg K$.

Intuitively, the minimization of Equation (3.53) with respect to the parameters ϕ trains the encoder to learn a transformation that accentuates similarities between data points, i.e., a space where positive pairs are closer to each other, while negative pairs are further apart [WI20]. To formalize these properties, let us consider two theoretical interpretations of the InfoNCE objective.

3.4.3 Theoretical Interpretations

We discuss two theoretical interpretations of contrastive learning with a special focus on the InfoNCE objective. First, we consider an information-theoretic perspective; second, an interpretation of the objective in terms of alignment and entropy regularization.

Mutual information estimation From an information-theoretic perspective, Equation (3.53) can be interpreted as a variational lower bound on the mutual information $I(\mathbf{x}_1; \mathbf{x}_2)$. Specifically, for any encoder $\mathbf{g}_\phi$, it holds that

$$\begin{aligned} I(\mathbf{x}_1; \mathbf{x}_2) &\geq I(\mathbf{g}_\phi(\mathbf{x}_1); \mathbf{g}_\phi(\mathbf{x}_2)) && (3.54) \\ &\geq \mathbb{E}_{\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K \sim p(\mathbf{x}_1, \mathbf{x}_2)} \Big[- \mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi) \Big] + \log K \,, && (3.55) \end{aligned}$$

where the first bound follows from the data processing inequality and the second inequality is based on the derivation of the InfoNCE objective as a variational lower bound of the mutual information between two random variables [OLV18; Poo+19]. Thus, by minimizing the InfoNCE objective (Equation 3.53), we maximize a lower bound on the mutual information $I(\mathbf{g}_\phi(\mathbf{x}_1); \mathbf{g}_\phi(\mathbf{x}_2))$. Consequently, the encoder is trained to maximize the similarity between the embeddings $\mathbf{g}_\phi(\mathbf{x}_1)$ and $\mathbf{g}_\phi(\mathbf{x}_2)$ in the sense of maximizing their mutual dependence.

It is worth noting that a fundamental limitation of the InfoNCE objective for estimating the mutual information is that its estimate is upper-bound by $\log K$, as can be seen in Equation (3.55). In fact, as shown by McAllester and Stratos [MS20], any distribution-free

high-confidence lower bound on mutual information cannot be larger than $\mathcal{O}(K)$, where K is the size of the sample. Though, in practice, one can use relatively large K (e.g., a large batch size) or compute the moving average over multiple batches of samples [He+20].

In the context of representation learning, Tschannen et al. [Tsc+20] demonstrate that estimating tighter bounds on the mutual information can result in *worse* representations. Consequently, they argue that the success of representation learning with InfoNCE cannot be fully attributed to the property of it being a mutual information estimator and that the quality of the learned representation with respect to several downstream tasks is influenced by other factors such as the choice of the encoder architecture. Along the same lines, Tian et al. [Tia+20] show that *minimizing* the mutual information between views, while keeping task-relevant information intact, can improve the quality of the representation for the respective tasks. Similar findings regarding the removal of task-irrelevant information were reported in parallel studies [Fed+20; Tsa+21], which also consider an information-theoretic perspective on the objective. Overall, these results are not entirely unexpected, because the "quality" of a representation depends on the downstream task.

Alignment and uniformity Asymptotically, the InfoNCE objective (Equation 3.53) can be interpreted as the alignment of positive pairs (numerator) with approximate entropy regularization (denominator), which produces embeddings that are uniformly distributed on a hypersphere [WI20]. Intuitively, maximizing uniformity prevents a degenerate solution in which the encoder maps to a constant.

Formally, when instantiating the objective in Equation (3.53) with $\tau = 1$ and $\text{sim}(a, b) = -(a - b)^2$, it asymptotically behaves like the objective

$$\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim p(\mathbf{x}_1, \mathbf{x}_2)} \left[\|\mathbf{g}_\phi(\mathbf{x}_1) - \mathbf{g}_\phi(\mathbf{x}_2)\|_2 \right] - H(\mathbf{g}_\phi(\mathbf{x}_1)) \tag{3.56}$$

when $K \to \infty$ [von+21]. Thus, by minimizing Equation (3.56) with respect to the parameters ϕ, the encoder is trained to align positive pairs while maximizing the entropy (i.e., the uniformity) of the learned representation.

3.4.4 Multimodal Contrastive Learning

Thus far, we have described the InfoNCE objective as it is typically used for self-supervised learning with multi-view data. However, when the data are comprised of heterogeneous observations of different modalities (e.g., image and text data), the objective needs to

be adapted accordingly. In the following, we describe how the InfoNCE objective can be adapted for *multimodal* contrastive learning.

InfoNCE for multimodal data Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be random vectors that take values in *distinct* spaces, $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively, and let $p(\mathbf{x}_1, \mathbf{x}_2)$ be the joint distribution. In this setup, it is natural to employ separate encoders $\mathbf{g}_{\phi_1} : \mathcal{X}_1 \to \hat{\mathcal{Z}}_1$ and $\mathbf{g}_{\phi_2} : \mathcal{X}_2 \to \hat{\mathcal{Z}}_2$, which makes it easier to process inputs with different dimensionalities and to implement suitable architectures (e.g., convolutional networks for images; recurrent networks for text).

Thus, for multimodal data, we express the InfoNCE objective from Equation (3.52) using two encoders. For a given sample $\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K \overset{\text{i.i.d.}}{\sim} p(\mathbf{x}_1, \mathbf{x}_2)$, we define

$$\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi_1, \phi_2) \coloneqq -\sum_{k=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{g}_{\phi_1}(\mathbf{x}_1^k), \mathbf{g}_{\phi_2}(\mathbf{x}_2^k))/\tau\}}{\sum_{l=1}^{K} \exp\{\text{sim}(\mathbf{g}_{\phi_1}(\mathbf{x}_1^k), \mathbf{g}_{\phi_2}(\mathbf{x}_2^l))/\tau\}} , \tag{3.57}$$

where we specify a different encoder for each of the two modalities.

Further, it is intuitive to use a symmetrized version of the objective, which can be obtained by computing the loss in both directions, i.e.,

$$\tfrac{1}{2}\,\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi_1, \phi_2) + \tfrac{1}{2}\,\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_2^k, \mathbf{x}_1^k\}_{k=1}^K; \phi_2, \phi_1) , \tag{3.58}$$

to balance the contributions of the individual modalities with respect to the overall loss value. For instance, the symmetrized objective was used for representation learning with image-text pairs in medical imaging [Zha+22] and contrastive pre-training on a large-scale image-captions dataset [Rad+21]. We revisit this objective in Chapter 8, where we discuss the application of multimodal contrastive learning for the identification of latent factors shared between observations of different modalities.

3.5 Summary

In this chapter, we covered relevant background to contextualize our problem formulation and methods. We introduced the topic of representation learning and positioned our research in the neighborhood of multi-view and multimodal representation learning and in particular among works that seek to learn "disentangled" representations that recover the underlying factors of variation. We also discussed the related topic of ICA, which provides a theoretical framework for representation learning in terms of the recovery of

latent variables. In this context, we also touched upon the topic of causal representation learning, which goes beyond the classic assumptions of ICA theory. Finally, we introduced two methods—the variational autoencoder and contrastive learning—as two complementary approaches for representation learning. In this thesis, we draw on both approaches to design suitable methods for *multimodal* representation learning.

4

Datasets

In this chapter, we describe the datasets used in this thesis. We include synthetic datasets, designed to provide simple benchmarks and to address conceptual problems without unnecessary complexity, as well as multiple real-world datasets for the evaluation of more realistic use cases.

First, we introduce the PolyMNIST dataset (Section 4.1) as a simple benchmark for multimodal generative models. It is complemented by the Translated-PolyMNIST dataset, which provides a salient demonstration of the limitations of existing approaches. Then, we describe two real-world datasets: the Bimodal-CelebA dataset (Section 4.2), which comprises facial images and associated attributes in the form of textual descriptions, and the CUB Image-Captions dataset (Section 4.3) that consists of natural images of birds with corresponding captions. Finally, we introduce the Multimodal3DIdent dataset (Section 4.4), which enables a principled assessment of the recovery of ground truth latent factors with a controllable data generating process that produces image-text pairs.

4.1 PolyMNIST

The PolyMNIST dataset is based on the MNIST database [LCB98; LeC+98] that is comprised of black and white images of handwritten digits. PolyMNIST combines the

MNIST digits with different background images to generate five synthetic image modalities. It is designed to be a conceptually simple yet exemplary dataset for multimodal learning.

Each example in the PolyMNIST dataset is a set of five images with the same digit label but different backgrounds and handwriting styles. For instance, see Figure 4.1, where each column shows one example and each row represents a different modality, expressed by a characteristic type of background. In total, there are 60 000 training and 10 000 test examples. Notably, we used distinct sets of MNIST images to construct the training and test sets respectively.

Figure 4.1: Ten examples from the PolyMNIST dataset. Each column depicts one training example that consists of five image modalities that share the same digit label.

Applications The PolyMNIST dataset is designed for our work presented in Chapter 5, where we benchmark multimodal generative models in a controlled setup with more than two modalities. In this context, the dataset allows to measure the aggregation of shared information across multiple modalities and to assess the generative performance with missing modalities. Further, it facilitates the comparison of different methods, as it removes the need for modality-specific architectures and hyperparameters because the observations of different modalities have similar characteristics (e.g., dimensionality and complexity) and discernable features of shared and modality-specific information. Consequently, PolyMNIST provides a testbed that enables a fair comparison across methods. For this purpose, the dataset is used in Chapters 5 to 7. Additionally, in Chapter 6, we use it to assess how much shared information (i.e., information about the digit label) is encoded in the shared and modality-specific subspaces of models with a partitioned latent space.

4.1.1 Translated-PolyMNIST

As a simple extension of PolyMNIST, we introduce the Translated-PolyMNIST dataset, which adds a simple transformation to each image—namely, a fixed downscaling and random translation (i.e., spatial positioning) of digits—as illustrated in Figure 4.2. The resulting dataset is conceptually similar to PolyMNIST in that a digit label is shared between five synthetic modalities.

Technically, in the generation of the dataset, we merely change the size and position of each digit. So, instead of overlaying a full-sized 28×28 MNIST digit on the respective background image, we scale down the digit by a factor of 0.7 and place it at a random (x, y)-coordinate before we combine it with the background image. Conceptually, these transformations leave shared information (i.e., the digit label) unaffected but make it more difficult to predict shared information in expectation across modalities on the level of observations. Specifically, for unseen examples, shared information *cannot* be predicted across modalities if the predictions are evaluated using a pixel-wise reconstruction loss (e.g., mean squared error) between the reconstructed and actual observations.

Figure 4.2: Ten examples from the Translated-PolyMNIST dataset. Each column depicts one example that consists of five different "modalities" that share the same digit label.

Applications In Chapter 7, we use the Translated-PolyMNIST to showcase the limitations of multimodal VAEs. In Chapter 9, we revisit the dataset and demonstrate how the limitations can be addressed using contrastive learning to estimate the shared information.

4.2 Bimodal-CelebA

The Bimodal-CelebA dataset [SDV20] provides a more realistic use case with image-text pairs. It extends the CelebA dataset [Liu+15], which is comprised of images depicting faces of celebrities and labeled attributes that describe the properties of a person's face. In the Bimodal-CelebA dataset, textual descriptions are constructed from the labeled attributes.

Figure 4.3: Three examples from the Bimodal-CelebA dataset, which is comprised of images and corresponding textual descriptions.

Each image is labeled according to 40 attributes (e.g., male, blond, sunglasses, beard). Based on these attributes, textual descriptions are constructed, as illustrated in the examples shown in Figure 4.3. The dataset has a highly imbalanced distribution of attributes,

which poses additional challenges for machine learning models. Moreover, missing attributes (e.g., no sunglasses) are not annotated and hence there is no fixed position for a specific attribute in the text. This introduces additional variability in the data and makes the text modality even more challenging. In total, the dataset is comprised of 202 600 examples, split into 162 771 training, 19 961 validation, 19 868 test examples.

Applications We use the Bimodal-CelebA dataset in Chapter 5 to provide an experiment with a realistic multimodal dataset.

4.3 CUB Image-Captions

The CUB Image-Captions dataset was introduced in [Shi+19] as an extension of the Caltech-Birds dataset (CUB-200-2011 [Wah+11]). It provides a real-world dataset with two modalities—images of birds of 200 different species with corresponding captions describing the birds' appearance, as illustrated in Figure 4.4.

Each image from the CUB-200-2011 dataset is coupled with 10 crowdsourced descriptions of the respective bird. In total, there are 11 788 images that are combined with the text into distinct sets of 88 550 training and 29 330 test examples. It is important to note that we use the CUB Image-Captions dataset with *real images* instead of the simplified version of the dataset used in previous works [Shi+19; Shi+21], where images were replaced by ResNet-features [He+16].

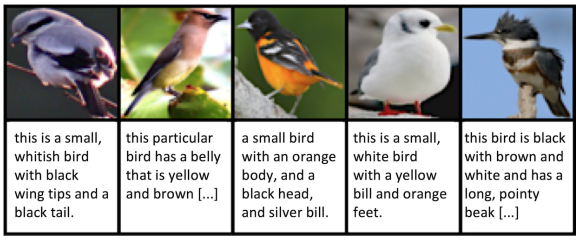


Figure 4.4: Five examples from the CUB Image-Captions dataset, which is comprised of natural images of birds with corresponding captions.

Applications The CUB Image-Captions dataset is used in Chapter 6 to provide a realistic experiment. Additionally, in Chapter 7, we use it to showcase the limitations of multimodal VAEs on a real-world dataset, for which shared information cannot be predicted in expectation across modalities on the level of observations.

4.4 Multimodal3DIdent

The Multimodal3DIdent dataset provides an identifiability benchmark with image-text pairs. Both modalities are generated from controllable ground truth factors, some of which are shared between image and text examples, as illustrated in Figure 4.5. Crucially, the dataset provides full control over the data generating process, because we sample the ground truth factors and render the observations of both modalities as functions thereof.

Multimodal3DIdent provides an extension of the Causal3DIdent dataset [von+21; Zim+21], which was designed to render high-dimensional images based on ground truth factors, which represent the parameters of a complex generative process. We adapt the data generation to a multimodal setup with image-text pairs and therefore synthesize complex observations of two modalities via distinct generating mechanisms. The training, validation, and test sets contain 125 000, 10 000, and 10 000 image-text pairs and ground truth factors, respectively. In the following, we describe the latent factors, the generative mechanisms used to render image and text observations, and the dependencies between modalities.

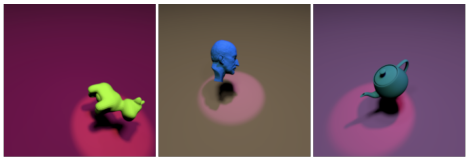


Figure 4.5: Three examples from the Multimodal3DIdent dataset. Emphasized words indicate factors of variation shared between image-text pairs.

Latent factors We define the *ground truth* latent variables used to generate image and text observations according to Table 4.1. Each factor is sampled from a uniform distribution defined on the specified set of values for the respective factor.

Image rendering We use the Blender rendering engine [Ble18] to create visually complex images that depict a three-dimensional scene. Each image in the dataset shows a colored object of a certain shape or class (i.e., teapot, hare, cow, armadillo, dragon, horse, or head). Specifically, the object is positioned in front of a colored background and illuminated by a differently colored spotlight that is focused on the object and located on a semicircle above the scene. The resulting RGB images are of size $224 \times 224 \times 3$.

Text generation We generate a sentence describing the respective scene. Each sentence describes the object's shape or class (e.g., teapot), position (e.g., bottom-left), and color.

The color is represented in a human-readable form (e.g., “lawngreen”, “xkcd:bright aqua”) based on the name of the color (from a given palette[9]) that is closest to a sampled color value in RGB space. The sentence is constructed from one of five pre-configured phrases with placeholders for the respective ground truth factors.

Relation between modalities Three latent factors are shared between image-text pairs, namely the shape of the object and its position in the image (x- and y-coordinates). Additionally, the object color also exhibits a dependence between modalities; however, it is not a 1-to-1 correspondence because the color palette is sampled randomly from a set of multiple palettes.[9] Moreover, the color of the object is causally influenced by position of the object. We split the range of hue values $[0, 1]$ into three equally sized intervals, each of which is associated with a fixed x-position of the object. For instance, if x-position is “left”, we sample the hue value from the interval $[0, 1/3]$. Consequently, the color of the object can be predicted to some degree from the position of the object.

Applications In Chapter 8, we study whether latent factors of variation can be identified using contrastive learning. In that context use the Multimodal3DIdent dataset to provide an experiment on a complex multimodal dataset of image-text pairs.

4.5 Summary

In this chapter, we introduced and described the datasets used in this thesis. We presented a collection of synthetic and real-world multimodal datasets that are used for the experiments in the following chapters.

For the novel methods we develop, we demonstrate their utility on both synthetic and real-world datasets. For some of the existing methods we study, their practical utility has already been established, so we evaluate them on synthetic datasets to investigate their effectiveness in a controlled setup and to identify possible failure cases.

[9] We use the following three color palettes from the `matplotlib.colors` API [Mat]: Tableau colors (10 values), CSS4 colors (148 values), and XKCD colors (949 values).

Modality	Latent Factor	Values	Details
Image	Object shape	{0, 1, . . . , 6}	Mapped to Blender shapes like “Teapot”, “Hare”, etc.
Image	Object x-position	{0, 1, 2}	Mapped to {-3, 0, 3} for Blender
Image	Object y-position	{0, 1, 2}	Mapped to {-3, 0, 3} for Blender
Image	Object z-position	{0}	Constant
Image	Object alpha-rotation	[0, 1]-interval	Linearly transformed to [-pi/2, pi/2] for Blender
Image	Object beta-rotation	[0, 1]-interval	Linearly transformed to [-pi/2, pi/2] for Blender
Image	Object gamma-rotation	[0, 1]-interval	Linearly transformed to [-pi/2, pi/2] for Blender
Image	Object color	[0, 1]-interval	Hue value in HSV transformed to RGB for Blender
Image	Spotlight position	[0, 1]-interval	Transformed to a unique position on a semicircle
Image	Spotlight color	[0, 1]-interval	Hue value in HSV transformed to RGB for Blender
Image	Background color	[0, 1]-interval	Hue value in HSV transformed to RGB for Blender
Text	Object shape	{0, 1, . . . , 6}	Mapped to strings like “teapot”, “hare”, etc.
Text	Object x-position	{0, 1, 2}	Mapped to strings “left”, “center”, “right”
Text	Object y-position	{0, 1, 2}	Mapped to strings “top”, “mid”, “bottom”
Text	Object color	string values	Color names from 3 different color palettes[9]
Text	Text phrasing	{0, 1, . . . , 4}	Mapped to 5 different English sentences

Table 4.1: Description of the latent factors used to generate the Multimodal3DIdent dataset. The first 11 factors represent the parameters of the generative process for the image rendering and the remaining 5 factors represent the parameters used to generate text observations. Under “details”, we describe the transformations used to adapt the value range of a given factor for the generation of the respective modality.

5

Generalized Multimodal ELBO

In this chapter, we propose a new multimodal generative model within the framework of variational inference. We introduce the *mixture of products of experts multimodal variational autoencoder* (MoPoE-VAE)—a principled and scalable method for approximate inference and density estimation on sets of modalities. Our formulation of the model generalizes two existing, widely-used approaches, namely the MVAE [WG18] and MMVAE [Shi+19], which model the variational joint posterior as a product of experts (PoE) and mixture of experts (MoE) respectively. Compared to these baselines, we demonstrate that the MoPoE-VAE enhances the encoding and disentanglement of shared and modality-specific information and consequently improves the generation of missing modalities.

First, we introduce relevant background (Section 5.1). Second, we describe the proposed method and explain how it generalized existing approaches (Section 5.2). Finally, we evaluate our method with experiments on synthetic and real-world data (Section 5.3).

5.1 Motivation and Background

Among the class of scalable multimodal VAEs, which was introduced in Section 3.3.6, the two main approaches are the multimodal variational autoencoder [MVAE, WG18] and the mixture of experts multimodal variational autoencoder [MMVAE, Shi+19]. We show that these approaches differ merely in the form of the variational joint posterior and draw a

theoretical connection between these models, showing that they can be subsumed under a unified formulation as a mixture of products of experts (MoPoE).

This insight has practical implications, because the form of the variational posterior influences the properties of a model. The MVAE is based on the product of experts (PoE), which can aggregate information across multiple modalities. For the MVAE, we observe a good approximation of the joint distribution but a lack of semantic coherence for the generation of missing modalities. On the other hand, the MMVAE is based on the mixture of experts (MoE), which can approximate a multimodal posterior (i.e., a distribution with multiple local maxima) even if the individual experts are unimodal distributions. We find that the MMVAE produces a good approximation of the unimodal and pairwise conditional distributions but a worse approximation of the joint distribution compared to the MVAE.

We generalize these formulations and introduce the mixture of products of experts (MoPoE) variational joint posterior and thus the MoPoE-VAE, which combines the benefits of the MVAE and MMVAE without significant tradeoffs. In contrast to the existing baselines, our method approximates the joint posterior for *all* subsets of modalities—a property that improves the learned representations and the generation of missing modalities, as we empirically demonstrate in Section 5.3.

Table 5.1 summarizes the properties of existing multimodal VAEs and highlights the benefits of the proposed MoPoE-VAE, namely (*i*) the ability to aggregate information across multiple modalities through the PoE; (*ii*) to approximate a multimodal posterior, similar to the MMVAE but for all subsets of modalities; and (*iii*) to efficiently handle missing modalities at test time, like the MVAE but without a heuristic extension of the training procedure.[10]

5.2 Method: MoPoE-VAE

First, we introduce the mixture of products of experts (MoPoE) variational joint posterior, designed to use all subsets of modalities for the posterior approximation (Section 5.2.1). Second, we propose the MoPoE-VAE and show that it is a principled approach for approximate inference and density estimation on sets of modalities (Section 5.2.2). Finally, we

[10]To handle missing modalities, a version of the MVAE uses ELBO sub-sampling during training, which is empirically motivated [WG18] but can produce an invalid bound on the joint log-evidence [WG19]. Throughout this thesis, we use the MVAE *without* ELBO sub-sampling, unless explicitly stated otherwise.

Model	Posterior form	Aggregate modalities	Multimodal posterior	Missing modalities
MVAE	PoE	✓	✗	(✓)
MMVAE	MoE	✗	✓	✓
MoPoE-VAE	MoPoE	✓	✓	✓

Table 5.1: Properties of existing multimodal VAEs and our proposed model, the MoPoE-VAE. For the MVAE, we denote its ability to handle missing modalities in parentheses, because in practice it requires an empirically-motivated extension of the objective.[10]

define the family of mixture-based multimodal VAEs and show that it provides a general formulation that encompasses the MVAE, MMVAE, and MoPoE-VAE (Section 5.2.3).

5.2.1 Mixture of Products of Experts (MoPoE)

Let M be the number of modalities, let $\mathcal{P}(M)$ be the powerset of the set of consecutive integers $\{1, \dots, M\}$ excluding the empty set and let $|\mathcal{P}(M)|$ denote its cardinality. Based on the set $\mathcal{P}(M)$, which enumerates all subsets of modalities, we define the variational joint posterior as a mixture of products of unimodal variational posteriors:

$$q_\phi^{\text{MoPoE}}(\mathbf{z} \mid \mathbf{x}_1, \dots, \mathbf{x}_M) = \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \prod_{i \in A} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i) \tag{5.1}$$

$$= \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} q_{\phi_A}^{\text{PoE}}(\mathbf{z} \mid \mathbf{x}_A) \ . \tag{5.2}$$

Thus, we define a mixture over *all* subsets of modalities and, for each subset $A \in \mathcal{P}(M)$, we take a product of the unimodal posteriors of the modalities indexed by A. In Equation (5.2), the set of parameters ϕ_A is fully determined by the parameters of the unimodal encoders, i.e., $\phi_A = \{\phi_i \mid i \in A\}$.

Generalized formulation It is easy to see how the MoPoE posterior in Equation (5.1) corresponds to either the PoE or MoE posterior when we consider only a specific subset of $\mathcal{P}(M)$. If we consider only the complete set of modalities and thus replace $\mathcal{P}(M)$ with $\{\{1, \dots, M\}\}$, which is a set with cardinality one, then the sum in Equation (5.1) would be reduced to a single term—a product of *all* unimodal posteriors—which corresponds exactly to the PoE posterior (Equation 3.50).

Conversely, if we consider only the unimodal subsets and therefore replace $\mathcal{P}(M)$ with $\{\{1\}, \dots \{M\}\}$, which is a set with cardinality M, then each product in the mixture would be reduced to a single factor—a unimodal posterior. Hence, Equation (5.1) would be a mixture over all unimodal posteriors, which corresponds to the MoE posterior (Equation 3.51).

Implications The MoPoE posterior (Equation 5.1) defines a mixture distribution over *all* subsets of modalities based on the powerset $\mathcal{P}(M)$, whereas existing formulations only consider certain subsets of modalities (see Section 5.2.3). This can have implications for multimodal generative learning, since the form of the variational joint posterior can influence the properties of the resulting model. For example, aggregation through the PoE results in a sharp joint posterior [Hin02], but it can hinder the optimization of the unimodal posteriors [WG18; KGS19]. In contrast, the formulation of the MoE defines the joint posterior as a mixture of unimodal posteriors, which can jointly approximate a multimodal distribution but cannot aggregate information across different modalities.

5.2.2 MoPoE-VAE: Scalable Inference with Missing Modalities

To develop a generative model that is scalable in the number of modalities—i.e., a model that efficiently handles missing modalities—previous works decompose the variational joint posterior $q_\phi(\mathbf{z} \mid \mathbf{x}_1, \dots, \mathbf{x}_M)$ as product or mixture of unimodal posteriors (c.f., Section 3.3.6). In the following, we define the objective of the MoPoE-VAE based on the MoPoE decomposition from Equation (5.1), which considers *all* subsets of modalities and can still be computed efficiently when modalities are missing.

To define the objective of the MoPoE-VAE, we start from the definition of the multimodal ELBO (Equation 3.49) and simply replace the variational joint posterior $q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})$, where $\bar{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, with the MoPoE posterior from Equation (5.1):

$$\mathcal{L}_{\text{ELBO}}(\bar{\mathbf{x}}; \phi, \theta) = \mathbb{E}_{q_\phi^{\text{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}(q_\phi^{\text{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \tag{5.3}$$

$$= \mathbb{E}_{\frac{1}{|\mathcal{P}(M)|}\sum_{A \in \mathcal{P}(M)} q_{\phi_A}^{\text{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}(q_\phi^{\text{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})). \tag{5.4}$$

Then, using the linearity of expectation (c.f., [Shi+19, p. 5]), we rewrite Equation (5.4) by taking the weighted sum outside the expectation to obtain

$$\frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \mathbb{E}_{q_{\phi_A}^{\text{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}(q_\phi^{\text{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \,. \tag{5.5}$$

Since the above KL-divergence cannot be computed in closed form, we resort to a lower bound of Equation (5.5). For this, we upper-bound the KL-divergence as follows:

$$D_{\mathrm{KL}}(q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) = \mathbb{E}_{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}\left[\log \frac{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}{p(\mathbf{z})}\right] \tag{5.6}$$

$$= \mathbb{E}_{\frac{1}{|\mathcal{P}(M)|}\sum_{A \in \mathcal{P}(M)} q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}{p(\mathbf{z})}\right] \tag{5.7}$$

$$= \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}{p(\mathbf{z})}\right] \tag{5.8}$$

$$\leq \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}{p(\mathbf{z})}\right] \tag{5.9}$$

$$= \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} D_{\mathrm{KL}}(q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A) \,||\, p(\mathbf{z})) \,. \tag{5.10}$$

In Equation (5.9), we construct a variational upper bound for each term in the sum. In particular, for each term, we approximate $q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})$ with $q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)$ as follows:

$$\mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})}{p(\mathbf{z})}\right] \tag{5.11}$$

$$= \mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})\, q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}{p(\mathbf{z})\, q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\right] \tag{5.12}$$

$$= \mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}{p(\mathbf{z})}\right] - D_{\mathrm{KL}}(q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A) \,||\, q_\phi^{\mathrm{MoPoE}}(\mathbf{z} \mid \bar{\mathbf{x}})) \tag{5.13}$$

$$\leq \mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}\left[\log \frac{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}{p(\mathbf{z})}\right], \tag{5.14}$$

where the inequality follows from the property that the KL-divergence is non-negative.

Hence, we can use the inequality from Equation (5.9) to lower-bound the multimodal ELBO (i.e., Equation 5.5).

Objective of the MoPoE-VAE Consequently, following the steps described, we obtain the objective of the MoPoE-VAE:

$$\mathcal{L}_{\mathrm{ELBO}}^{\mathrm{MoPoE}}(\bar{\mathbf{x}}; \phi, \theta) \coloneqq \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \left\{\mathbb{E}_{q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\mathrm{KL}}(q_{\phi_A}^{\mathrm{PoE}}(\mathbf{z} \mid \mathbf{x}_A) \,||\, p(\mathbf{z}))\right\} . \tag{5.15}$$

Since Equation (5.15) lower-bounds the multimodal ELBO (Equation 5.5), it follows that Equation (5.15) forms a lower bound on the log-evidence $\log p(\mathbf{x}_1, \ldots, \mathbf{x}_M)$, as desired for a multimodal generative model. Moreover, if we use Gaussian unimodal variational posteriors, the objective of the MoPoE-VAE can be computed efficiently for any subset of modalities, as the PoE has a closed-form solution in that case [CF14; WG18]. Hence, the MoPoE-VAE provides a principled and scalable method for approximate inference and density estimation on sets of modalities.

5.2.3 The Family of Mixture-based Multimodal VAEs

In this subsection, we provide a generalized formulation of existing multimodal VAEs. Specifically, we introduce the family of *mixture-based* multimodal VAEs and show that it subsumes the MVAE, MMVAE, and the proposed MoPoE-VAE.

First, we define a mixture-based decomposition of the variational joint posterior that generalizes the decompositions used by existing models. In the following, $\mathcal{S}$ denotes a set of pairs (A, ω_A), where $A \subseteq \{1, \ldots, M\}$ is a subset of modalities and $\omega_A \in [0, 1]$ is the corresponding mixture coefficient for the subset A. To simplify the notation, we write $A \in \mathcal{S}$ to abbreviate $(A, \omega_A) \in \mathcal{S}$ in Definition 6 and throughout this thesis.

Definition 6 (Mixture posterior over $\mathcal{S}$)**.** *Let $\mathcal{S}$ be a given set of non-empty subsets of modalities and corresponding mixture coefficients that satisfies*

$$\mathcal{S} \subseteq \Big\{(A, \omega_A) \,|\, A \subseteq \{1, \ldots, M\},\ A \neq \emptyset,\ \omega_A \in [0, 1]\Big\} \text{ and } \sum\nolimits_{A \in \mathcal{S}} \omega_A = 1. \tag{5.16}$$

Then, we define the variational joint posterior as a mixture distribution over $\mathcal{S}$, i.e.,

$$q_\phi^{\mathcal{S}}(\mathbf{z} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_M) \coloneqq \sum_{A \in \mathcal{S}} \omega_A \, q_{\phi_A}(\mathbf{z} \,|\, \mathbf{x}_A)\ , \tag{5.17}$$

where $\phi_A = \{\phi_i \,|\, i \in A\}$ for each $A \in \mathcal{S}$.

It is easy to see how the above definition encompasses all formulations of the variational joint posterior for the existing models. Concretely, we have

$$q_\phi^{\mathcal{S}}(\mathbf{z} \,|\, \bar{\mathbf{x}}) = \begin{cases} q_\phi(\mathbf{z} \,|\, \bar{\mathbf{x}}), & \text{if } \mathcal{S} = \Big\{(\{1, \ldots, M\}, 1)\Big\} \quad \text{(MVAE)} \\ \frac{1}{M} \sum_{i=1}^{M} q_{\phi_i}(\mathbf{z} \,|\, \mathbf{x}_i), & \text{if } \mathcal{S} = \Big\{(\{i\}, \frac{1}{M})\Big\}_{i=1}^{M} \quad \text{(MMVAE)} \\ \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} q_{\phi_A}(\mathbf{z} \,|\, \mathbf{x}_A), & \text{if } \mathcal{S} = \Big\{(A, \frac{1}{|\mathcal{P}(M)|})\Big\}_{A \in \mathcal{P}(M)} \quad \text{(MoPoE-VAE)} \end{cases}$$

where the MVAE and MoPoE-VAE further assume that $q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})$ and $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ factorize as a product of experts (c.f., Equation 3.50). Hence, Definition 6 generalizes the formulation of the variational joint posteriors used by the MVAE, MMVAE, and MoPoE-VAE.

Next, we define the family of mixture-based multimodal VAEs, for which we restrict the class of models optimizing the multimodal ELBO to the subclass of models that use a mixture variational joint posterior.

Definition 7 (Mixture-based multimodal VAEs)**.** *The family of mixture-based multimodal VAEs comprises all models that maximize the multimodal ELBO using a mixture variational joint posterior* $q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}})$ *consistent with Definition 6. That is, all models that maximize*

$$\mathcal{L}^{\mathcal{S}}_{ELBO}(\bar{\mathbf{x}}; \phi, \theta) = \sum_{A \in \mathcal{S}} \omega_A \, \mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{KL}(q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \,, \tag{5.18}$$

or a lower bound thereof.

It is easy to see how the family of mixture-based multimodal VAEs subsumes the MVAE, MMVAE, and MoPoE-VAE. For each model, we can plug in the model-specific set $\mathcal{S}$ into Equation (5.18) and use uniform mixture coefficients $\omega_A = \frac{1}{|\mathcal{S}|}$ for all $A \in \mathcal{S}$. For example, for the MoPoE-VAE, we can plug in $\mathcal{S} = \{(A, \frac{1}{|\mathcal{P}(M)|})\}_{A \in \mathcal{P}(M)}$ into Equation (5.18) to recover the objective from Equation (5.5). Alternatively, we can use the lower bound

$$\mathcal{L}^{\mathcal{S}}_{\text{ELBO}}(\bar{\mathbf{x}}; \phi, \theta) \geq \sum_{A \in \mathcal{S}} \omega_A \left\{ \mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}\left(q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A) \,||\, p(\mathbf{z})\right) \right\} \tag{5.19}$$

to recover the objective from Equation (5.15). In both cases, it is further assumed that $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ factorizes as a product of experts.

Summary and implications To summarize, in this subsection we have introduced the family of mixture-based multimodal VAEs as a generalization of existing approaches. Compared to previous formulations of multimodal VAEs based on the mixture of experts [Shi+19; SDV20], our formulation is more general in that it allows for arbitrary subsets of modalities with non-uniform mixture coefficients. Specifically, our formulation subsumes multiple existing models—namely, the MVAE, MMVAE, and our proposed MoPoE-VAE—as well as possible extensions thereof.

From a computational perspective, an interesting characteristic of mixture-based multimodal VAEs is the *sub-sampling* of modalities during training, which is a direct consequence of defining the variational joint posterior as a mixture distribution over subsets of modalities

(Equation 5.17). The only member of the family of mixture-based multimodal VAEs that forgoes sub-sampling, defines a trivial mixture over a single subset—namely, the complete set of modalities. In Chapter 7, we return to the family of mixture-based multimodal VAEs and show that the sub-sampling of modalities enforces an undesirable bound for the approximation of the joint distribution.

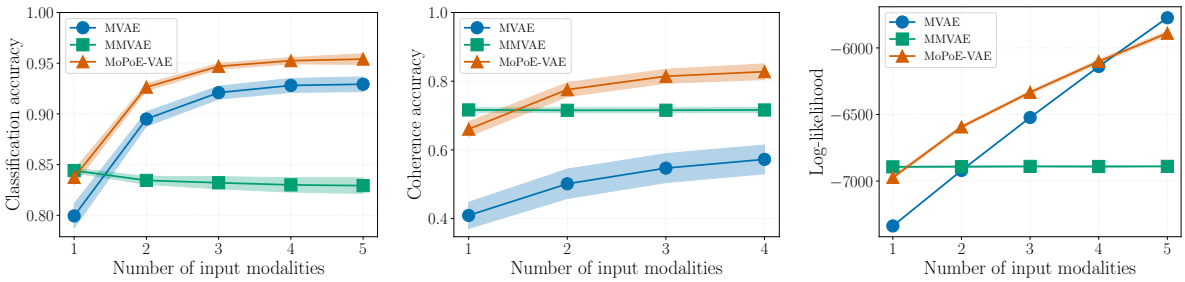
5.3 Experiments

Datasets Primarily, we evaluate the models in a controlled setup using PolyMNIST—a dataset that features up to five semi-synthetic modalities. Additionally, we investigate a more realistic setup using the Bimodal-CelebA dataset, which is a more complex multimodal dataset comprised of image-text pairs. All datasets are described in Chapter 4.

Evaluation metrics Following previous work [WG18; Shi+19], we evaluate each model in terms of the following three metrics. First, we assess the information content and disentanglement of the embeddings using linear probing (i.e., logistic regression) with respect to the shared latent factors. Second, we evaluate the generative coherence, for which we assess whether the generated samples agree in their semantic content across modalities using pre-trained classifiers. Third, we evaluate the approximation of the data distribution using log-likelihoods computed on the test set. Further information about the individual metrics is provided in Section 2.5.

Hypotheses We expect that both the MVAE and the proposed MoPoE-VAE learn a joint representation that aggregates shared information from multiple modalities effectively through the PoE, and thus benefit from additional modalities. For the MMVAE, we anticipate that it approximates the unimodal and pairwise conditional distributions well, but that it does not benefit from additional modalities. Further, we expect that the MVAE achieves the best generative performance when all modalities are present, but that its performance deteriorates with an increasing number of missing modalities. In contrast, the MoPoE-VAE should perform well given any subset of modalities.

Implementation details For a fair comparison, we use the same architecture for all models. For the MVAE, we use the objective with sub-sampling of unimodal ELBO terms, which is a heuristic extension of the model that achieves better empirical results [WG18].



(a) Linear probing (b) Generative coherence (c) Joint log-likelihood

Figure 5.1: Performance on PolyMNIST as a function of the number of input modalities. For each point, we average the results over all subsets of modalities of the respective size. The performance is measured in terms of three different metrics (in each case, larger is better) shown in separate subplots. Markers denote the means and error bands the standard deviations over five runs with different seeds. Figure 5.1a shows the classification accuracy for logistic regression models trained to predict the digit label given the embeddings inferred from the respective subset of modalities. In Figure 5.1b, evaluate the semantic coherence for the generation of missing modalities. Figure 5.1c shows the joint log-likelihood values.

5.3.1 PolyMNIST

First, we use the PolyMNIST dataset to compare the performance across methods in a controlled setup with up to five modalities. We simulate missing modalities at test time and consider all subsets of modalities—from one modality (i.e., four inputs missing) up to five modalities (none missing).

Performance with missing modalities Figure 5.1 presents our results on the PolyMNIST dataset, for which we simulate randomly missing modalities at test time. In each subplot, we show a different metric and plot the performance as a function of the number of input modalities. As expected, both the MVAE and MoPoE-VAE benefit from additional inputs through the PoE aggregation, whereas the performance of the MMVAE remains constant across all metrics. When all five input modalities are present, the log-likelihood of the MoPoE-VAE is on par with the MVAE, but the former is clearly superior in the case of missing modalities and also in terms of linear probing and generative coherence. Conversely, with a single input modality, the MoPoE-VAE matches the performance of the MMVAE, but it performs significantly better with two or more modalities present.

Overall, the results on the PolyMNIST dataset illustrate that the proposed MoPoE-VAE does not only theoretically encompass the MVAE and MMVAE, but that it also exhibits superior results in the case of missing modalities and even matches the performance in special cases that favor the baselines.

Linear probing To assess how well shared information is encoded in the learned representations, we use linear probing (i.e., logistic regression) to predict the digit label from the embeddings. We train the classifiers separately for each model and evaluate the classification accuracy on the embeddings of the test data. Figure 5.1a shows that the shared information can be extracted relatively well from the embeddings using linear probing. We find that the MVAE and MoPoE-VAE aggregate information across multiple modalities effectively, as the performance improves with additional modalities.

Overall, the MoPoE-VAE exhibits a significantly better encoding of shared information compared to the MVAE and MMVAE. Consequently, the results suggest that the information is encoded in a linearly separable format, which indicates a better disentanglement of shared and modality-specific information.

Generative performance To assess the generative performance, we follow previous work [Shi+19; SDV20] and evaluate the generative coherence (Figure 5.1b) and joint log-likelihood (Figure 5.1c) on the test set. We observe that the MoPoE-VAE exhibits a significantly better tradeoff in terms of both metrics compared to the baselines. Hence, the results suggest that the model approximates the joint distribution for all subsets of modalities effectively and that it enhances the generation of missing modalities based on the learned representations.

5.3.2 Bimodal-CelebA

Next, we consider the Bimodal-CelebA dataset to evaluate the models in a more realistic multimodal setup with image-text pairs. The dataset is comprised of images depicting faces of celebrities and textual descriptions of the properties of a person's face. Similar to previous work [SDV20], we use modality-specific latent spaces, which were found to improve the generative quality of a model [HG18; SDV20; Dau+20].[11]

[11] We revisit the idea of using modality-specific latent spaces in Chapter 6, where we discuss the benefits and disadvantages of partitioning the latent space into shared and modality-specific subspaces.

	Linear probing			Generative coherence	
	Image	Text	Joint	Image → Text	Text → Image
MVAE	0.30	0.31	0.32	**0.26**	0.33
MMVAE	0.35	0.38	0.35	0.14	0.41
MoPoE-VAE	**0.40**	**0.39**	**0.39**	0.15	**0.43**

Table 5.2: Quantitative results on the Bimodal-CelebA dataset. We evaluate linear probing and generative coherence in terms of the average precision to account for rare attributes. Specifically, we compute the mean average precision over all attributes.

Quantitative evaluation Table 5.2 shows the results for linear probing and conditional coherence in terms of the mean average precision over all attributes. We observe that all models perform comparably well, but the MoPoE-VAE shows a slightly better tradeoff in performance even though the data is comprised of two modalities, which is favorable for the baselines. Additionally, in Figure 5.3, we present the results of an attribute-specific evaluation for the MoPoE-VAE. Using linear probing, we find that salient attributes like "gender" or "smiling" are encoded relatively well whereas subtle and infrequent attributes are more difficult to recover from the embeddings.

Qualitative results Figure 5.2 presents the qualitative results for the MoPoE-VAE, showing the conditional generation of images given textual descriptions. Overall, the generated images are relatively coherent with respect to the given text. Again, we find that salient attributes like "gender" and "smiling" are learned well as they manifest clearly and consistently in the generated samples, whereas subtle and infrequent attributes appear to be more difficult to generate consistently.

5.4 Summary

In this chapter, we introduced the MoPoE-VAE as a principled and scalable method for multimodal generative learning within the framework of variational inference. We showed that the MoPoE-VAE generalizes the MVAE and MMVAE and combines their benefits without significant tradeoffs, as it considers all subsets of modalities for the

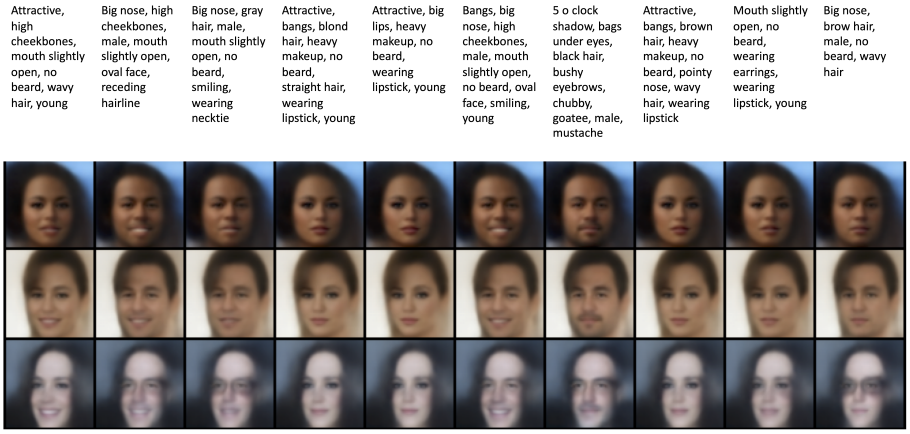


Figure 5.2: Qualitative results for the MoPoE-VAE on the Bimodal-CelebA dataset. The images are conditionally generated given the text shown in respective column.

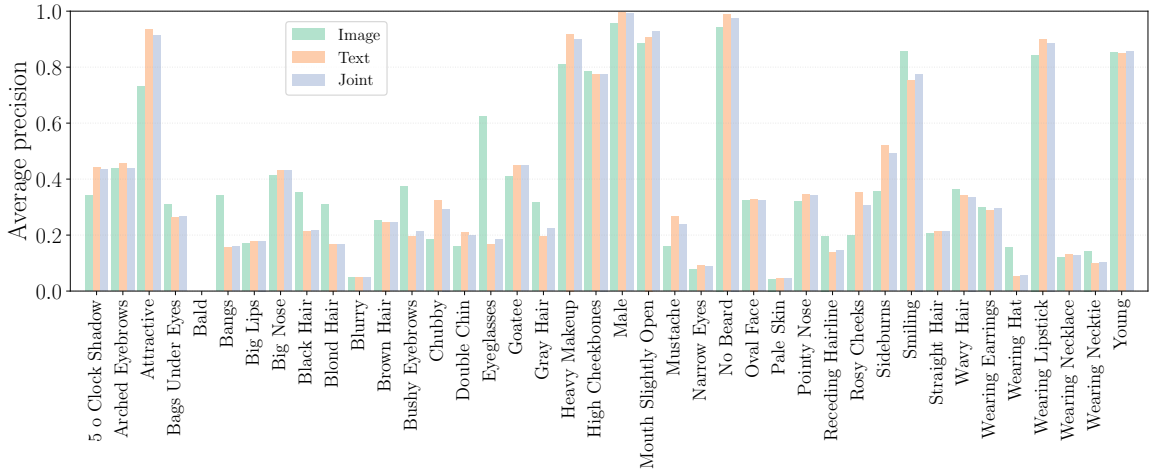


Figure 5.3: Detailed evaluation of the learned representation of the MoPoE-VAE on the Bimodal-CelebA dataset. We use linear probing with respect to each attribute and evaluate the average precision to account for the skewed distribution of attributes in the data.

posterior approximation. Empirically, we demonstrated the advantages of the MoPoE-VAE compared to the MVAE and MMVAE. Specifically, we showed that it improves the encoding and disentanglement of shared and modality-specific information as well as the generative performance, as validated on synthetic and real-world multimodal datasets.

With respect to the research questions, our results demonstrate how the MoPoE-VAE can be used to encode and disentangle shared and modality-specific information effectively (c.f., Questions 1 and 2). Additionally, we presented promising results for the inference across

modalities (Question 3), as showcased by the generation of missing modalities.

In the scope of this thesis, the MoPoE-VAE is the first method we propose. In the following chapters, we include it as a baseline when we evaluate the performance of multimodal VAEs. In the next chapter, we further investigate the perks and pitfalls of partitioning the latent space into shared and modality-specific subspaces. Further, in Chapter 7, we analyze the tradeoffs between the different models presented in this chapter and discuss their limitations.

6

Partitioned Latent Spaces

The previous chapter focused on generative learning with multimodal VAEs that are characterized by a *joint* latent space. While we demonstrated promising preliminary results for the disentanglement of shared and modality-specific information, we noticed that a joint latent space necessarily encodes both sources of information, since it is learned by reconstructing the observations of all modalities. In this chapter, we consider an *explicit partitioning* of the latent space into shared and modality-specific subspaces. Thus, we present a targeted approach to improve the disentanglement of shared and modality-specific information through partitioning.

First, in Section 6.1, we conceptualize the idea of a multimodal generative model with a partitioned latent space. We introduce the notation for this chapter and illustrate how partitioned models differ from models with a joint latent space.

Second, we present a motivating example to illustrate the idea as well as the challenges of partitioning the latent space into shared and modality-specific subspaces. In Section 6.2, we introduce a naive approach for partitioning the latent space that works in principle, but in practice it exhibits a failure mode, which we describe in Section 6.2.3 as the *shortcut problem*—a degenerate solution, where all information (i.e., both shared and modality-specific) is encoded in the modality-specific subspaces and is *not* disentangled. As a proof of concept, we demonstrate how a small number of labeled examples can be used for selecting the hyperparameter values of the naive partitioning approach to disentangle shared and modality-specific information effectively.

Finally, in Section 6.3, we develop a technique for disentangling shared and modality-specific information without additional labels. We propose the MMVAE+, which extends the MMVAE with a partitioned latent space and disincentivizes shared information from being encoded in the modality-specific subspaces with a suitable regularizer. Our method mitigates the shortcut problem through regularization of the crossmodal-reconstruction terms, for which we replace the modality-specific embeddings with samples from a prior distribution with a learned variance parameter. We demonstrate that the MMVAE+ improves the disentanglement of shared and modality-specific information and enhances the conditional generation of missing modalities, as validated through experiments on synthetic and real-world datasets.

6.1 Preliminaries

In the previous chapter, we considered multimodal generative models with a joint latent space that was denoted by $\mathbf{z} \in \mathcal{Z}$. In this chapter, we assume that the latent space can be partitioned into $M+1$ subspaces $(\mathbf{c}, \mathbf{m}_1, \ldots, \mathbf{m}_M)$, where $\mathbf{c} \in \mathcal{C}$ denotes a vector of shared components and $\mathbf{m}_i \in \mathcal{M}_i$ represents a vector of modality-specific components for each modality $i \in \{1, \ldots, M\}$. We define the *partitioning* of $\mathbf{z}$ as a dimension-wise permutation

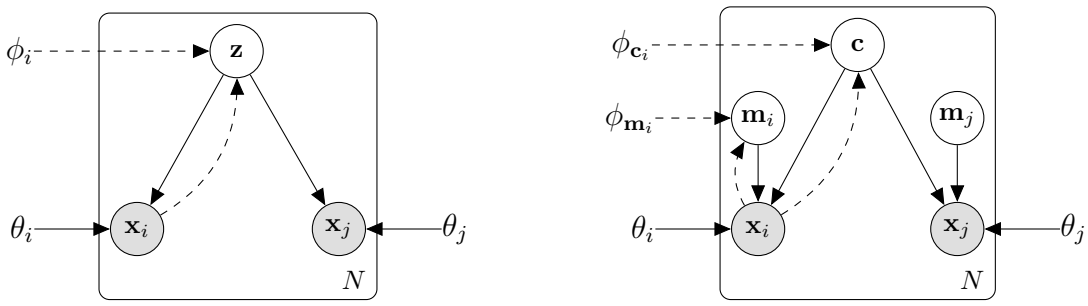
$$\mathbf{z} = \rho(\mathbf{c}, \mathbf{m}_1, \ldots, \mathbf{m}_M) \,, \tag{6.1}$$

where ρ denotes a dimension-wise rearrangement of the concatenated input vectors and thus a bijection of $\mathcal{C} \times \mathcal{M}_1 \times \cdots \times \mathcal{M}_M$ onto itself. In the following, we assume, w.l.o.g., that ρ is the identity mapping and thus write $\mathbf{z} = (\mathbf{c}, \mathbf{m}_1, \ldots, \mathbf{m}_M)$ when we define the partitioning of the inferred latent space for a given model.

Generative model In Chapter 5, we assumed the conditional independence $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{z}$ for all pairs of modalities $i, j \in \{1, \ldots, M\}, i \neq j$. In this chapter, we consider the conditional independence $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{c}$ for $i \neq j$ and design the generative model accordingly.

In Figure 6.1, we show the graphical models that describe multimodal VAEs with a joint latent space (Figure 6.1a) and partitioned latent space (Figure 6.1b) respectively. With the partitioning, we assume a generative model with the following factorization:

$$p_\theta(\bar{\mathbf{x}}, \mathbf{c}, \mathbf{m}_1, \ldots, \mathbf{m}_M) = p(\mathbf{c}) \prod_{i=1}^{M} p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) p(\mathbf{m}_i) \,, \tag{6.2}$$



(a) Joint latent space (b) Partitioned latent space

Figure 6.1: Graphical models for multimodal VAEs with a joint and partitioned latent space. Observations are denoted by filled nodes and latent variables by clear nodes. The plate notation indicates amortized inference, i.e., the optimization of a shared set of parameters given a dataset of size N. To simplify the illustration, we show only two modalities, since we assume the same graphical model for each pair of modalities $i, j \in \{1, \ldots, M\}, i \neq j$, and we only depict the inference network for modality i.

where θ_i are the parameters of a modality-specific decoder and $\theta = \{\theta_1, \ldots, \theta_M\}$ denotes the set of all decoder parameters.

Inference networks As illustrated by dashed lines in Figure 6.1, the inference networks are designed to mirror the generative process. We denote the variational joint posterior as $q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}})$ and the variational posterior specific to modality i as $q_{\phi_i}(\mathbf{m}_i \mid \mathbf{x}_i)$ for each $i \in \{1, \ldots, M\}$. By $\phi \coloneqq \{\phi_\mathbf{c}, \phi_{\mathbf{m}_1}, \ldots, \phi_{\mathbf{m}_M}\}$, we denote the set of all encoder parameters.

In summary, in this section we conceptualized the idea of a multimodal generative model with a partitioned latent space and clarified the distinction between models with a joint and partitioned space. The described model already provides the blueprint for multimodal VAEs with a partitioned latent space; in the following, we merely specify the objective function. First, we describe a "naive" approach without further adjustments of the model (Section 6.2) and, in the same breath, discuss its limitations (Section 6.2.3). Finally, we propose a more sophisticated method (Section 6.3), for which we extend the model with a regularization technique to incentivize the disentanglement of shared and modality-specific information.

6.2 A Naive Approach

In this section, we introduce a naive approach for partitioning the latent space of a multimodal VAE into shared and modality-specific subspaces. We describe it as "naive", because later, in Section 6.2.3, we discuss a failure mode that thwarts the disentanglement of shared and modality-specific information for the given model. However, at the end of this section, we also present a proof of concept that demonstrates the feasibility of disentanglement *in principle* by using small number of labeled examples to guide model selection.[12] Ultimately, the supervised model selection only serves as a motivating example; in Section 6.3 we introduce a more sophisticated method that does *not* depend on additional labels to learn a disentangled representation.

To implement a partitioned multimodal VAE, the straightforward approach is to split the latent space into shared and modality-specific subspaces without changing the objective function, which corresponds to what we call the naive partitioning. In this section, we describe specify the network architecture and objective function for the naive approach.

6.2.1 Model Architecture

As illustrated in Figure 6.1b, we partition the inferred latent space into a vector $\mathbf{c}$ that is shared between the decoders of different modalities and M modality-specific vectors $\mathbf{m}_1, \ldots, \mathbf{m}_M$. We specify individual networks to model the variational joint posterior, modality-specific posteriors, and modality-specific decoders, respectively. Concretely, we use separate networks to model the following distributions:

$$q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \mathbf{x}_1, \ldots, \mathbf{x}_M); \tag{6.3}$$

$$q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i), \qquad \text{for } i \in \{1, \ldots, M\}; \tag{6.4}$$

$$p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i), \qquad \text{for } i \in \{1, \ldots, M\}. \tag{6.5}$$

To model the variational joint posterior (Equation 6.3), we use the decomposition of the respective type of multimodal VAE for which we partition the latent space; for the MVAE, we take the product of experts, whereas for the MMVAE, we use the mixture of experts.[13] Each modality-specific variational posterior (Equation 6.4) is modeled by a separate

[12] Notably, the labels are only used for model selection but not during training.

[13] The decompositions are defined in Chapter 3, Section 3.3.6.

inference network, i.e., a modality-specific encoder for each modality $i \in \{1, \dots, M\}$. For the generative part of the model (Equation 6.5), we use M decoder networks, each of which is conditioned on both the shared and modality-specific embeddings; for example, the decoder of modality i takes $(\mathbf{c}, \mathbf{m}_i)$ as input to generate $\mathbf{x}_i \sim p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)$.

6.2.2 Objective Function

Despite the partitioning, the objective function does not change compared to previous approaches that use a joint latent space. We can simply rewrite the multimodal ELBO (Definition 5) by replacing the joint latent space with a partitioned space, i.e., $\mathbf{z} = (\mathbf{c}, \bar{\mathbf{m}})$, where $\bar{\mathbf{m}} \coloneqq (\mathbf{m}_1, \dots, \mathbf{m}_M)$.

We start from the definition of the multimodal ELBO (Definition 5) and adjust the notation by replacing the variational posterior $q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})$ with the partitioned variant $q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}})$, i.e.,

$$\mathbb{E}_{q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})}\Big[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})\Big) \tag{6.6}$$

$$= \mathbb{E}_{q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}})}\Big[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{c}, \bar{\mathbf{m}})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c}, \bar{\mathbf{m}})\Big)\,. \tag{6.7}$$

Then, we use the (conditional) independence assumptions from Equation (6.2) to rewrite Equation (6.7) as follows:

$$\mathbb{E}_{q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}})}\Big[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{c}, \bar{\mathbf{m}})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c}, \bar{\mathbf{m}})\Big) \tag{6.8}$$

$$= \sum_{i=1}^{M} \mathbb{E}_{q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}})}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c}, \bar{\mathbf{m}})\Big) \tag{6.9}$$

$$= \sum_{i=1}^{M} \mathbb{E}_{q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}})}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c}) \textstyle\prod_{j=1}^{M} p(\mathbf{m}_j)\Big)\,. \tag{6.10}$$

More precisely, in the first step we use $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{c}$ and $\mathbf{x}_i \perp\!\!\!\perp \mathbf{m}_j$ for $i \neq j$, while the second step follows from the independence of prior distributions, $p(\mathbf{c}, \bar{\mathbf{m}}) = p(\mathbf{c}) \prod_{j=1}^{M} p(\mathbf{m}_j)$.

For the inference network, we assume a mean-field factorization

$$q_\phi(\mathbf{c}, \bar{\mathbf{m}} \mid \bar{\mathbf{x}}) = q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)\,, \tag{6.11}$$

hence, we sample $\mathbf{c} \sim q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}})$ and $\mathbf{m}_j \sim q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)$ independently but conditioned on the same (subset of) observations.[14]

[14] To be more precise, we sample each dimension of these variational posteriors independently conditioned on a given observation, similar to the sampling procedure of the (unimodal) VAE.

We plug the right-hand side of Equation (6.11) into Equation (6.10) to obtain

$$\begin{aligned}&\sum_{i=1}^{M} \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ &- D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \prod_{j=1}^{M} p(\mathbf{m}_j)\Big) \,,\end{aligned} \tag{6.12}$$

where we can simplify

$$\sum_{i=1}^{M} \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \tag{6.13}$$

$$= \sum_{i=1}^{M} \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \tag{6.14}$$

and decompose

$$D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \prod_{j=1}^{M} p(\mathbf{m}_j)\Big) \tag{6.15}$$

$$= D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c})\Big) + \sum_{j=1}^{M} D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \,||\, p(\mathbf{m}_j)\Big) \,, \tag{6.16}$$

as described in more detail in Appendix A.1.

Hence, we reformulated the partitioned objective (Equation 6.12) as a sum of M likelihood terms (Equation 6.14) and $M + 1$ KL-divergence terms (Equation 6.16), one for each variational posterior. Finally, we combine the two sums and define the objective of the *partitioned* multimodal VAE as follows:

$$\begin{aligned}\mathcal{L}_{\text{ELBO-Part.}}(\bar{\mathbf{x}}; \phi, \theta) \coloneqq \sum_{i=1}^{M} \Big\{ &\mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ &- D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i)\Big) \Big\} - D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c})\Big) \,.\end{aligned} \tag{6.17}$$

Crucially, the objective remains a valid multimodal ELBO, i.e., a lower bound on the log-evidence $\log p(\bar{\mathbf{x}})$, which justifies its use for training a generative model to approximate the joint distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_M)$.

Nevertheless, we refer to the objective in Equation (6.17) as a naive approach. This is because, in principle, the described model could disentangle shared and modality-specific information in the respective subspaces, but in practice it fails to achieve the desired disentanglement because the optimization of Equation (6.17) permits a degenerate solution—the so-called shortcut problem, which we discuss next.

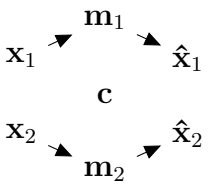
6.2.3 The Shortcut Problem

An Achilles' heel of the naive approach is that the objective in Equation (6.17) can be minimized without the desired disentanglement of shared and modality-specific information. For example, suppose that *all* information (i.e., shared and modality-specific) is encoded in the modality-specific embeddings while the shared latent space is ignored by the model. In this case, the model learns to approximate the data distribution even if the encoder fails to disentangle shared and modality-specific information in the respective subspaces. Hence, the optimization of the objective in Equation (6.17) with respect to θ and ϕ has no unique solution that coincides with the desired disentanglement. We call such an undesired behavior a *shortcut problem*, because it thwarts the disentanglement of shared and modality-specific information despite the suitable inductive bias of the partitioning.

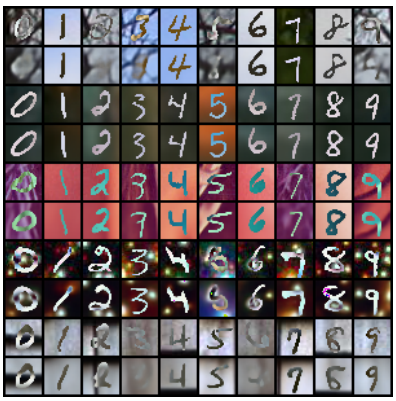
Figure 6.2 illustrates the shortcut problem and how it manifests for a model that exhibits the described shortcomings. In Figure 6.2a, we provide a graphical representation of the figurative "shortcut", which illustrates how the shared subspace is ignored by the model as most information flows through the modality-specific subspaces. Additionally, we present qualitative results showing how the problem typically manifests for a partitioned model trained on the PolyMNIST dataset.[15] While the model achieves high likelihood values and therefore exhibits decent reconstructions (see Figure 6.2b), it encodes shared information in the modality-specific embeddings (for the given example, we measure 65% classification accuracy using linear probing) and consequently the model exhibits a lack of semantic coherence for the conditional generation across modalities (see Figure 6.2c).

Supervised model selection On the bright side, we can demonstrate how the naive partitioning can, in principle, disentangle shared and modality-specific information despite the shortcut problem. In our experiments, we found that the dimensionality of the modality-specific latent space is an important hyperparameter that can be used to control the performance in a desirable manner. Based on this insight, we demonstrate how to conduct model selection when we have access to a handful of ground truth labels for the value of the shared latent variable. Notably, we do not use labels for training but only for model selection.

[15] Concretely, we use a partitioned MMVAE with an equal split of the latent space, which means that we allocate the same number of dimensions to each subspace. The dataset is described in Chapter 4.



(a) Graphical illustration

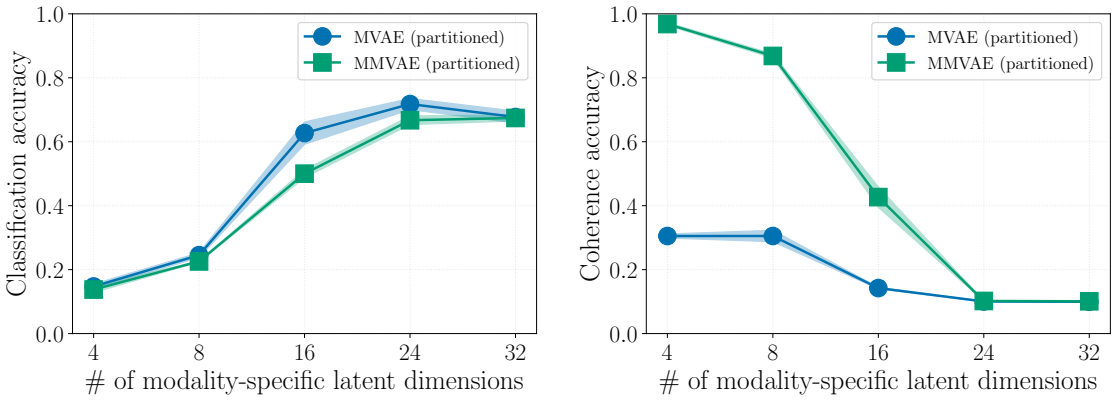


(b) Reconstruction

(c) Conditional generation

Figure 6.2: Illustration of the shortcut problem. Figure 6.2a presents a graphical illustration, which depicts the flow of information from inputs $\mathbf{x}_i$ of modality $i \in \{1, 2\}$ to the reconstructions $\hat{\mathbf{x}}_i$ through the partitioned latent space $(\mathbf{c}, \mathbf{m}_i)$. Despite the partitioning, the model only passes information through the modality-specific subspaces $\mathbf{m}_1$ and $\mathbf{m}_2$ but not through the shared subspace $\mathbf{c}$. If a model exhibits such an undesired behavior, it typically shows decent reconstructions (Figure 6.2b) but a lack of semantic coherence for the conditional generation across modalities (Figure 6.2c), as illustrated here using a partitioned MMVAE trained on the PolyMNIST dataset.[15] Figure 6.2b shows original samples and reconstructions for each modality. Figure 6.2c presents qualitative results for the generation across modalities; each column depicts ten conditionally generated images of one modality given the respective image of another modality shown in the first row.

In Figure 6.3, we plot the results for the MVAE and MMVAE (both using a naive partitioning) as a function of the dimensionality of the modality-specific latent space. Figure 6.3a shows how much shared information can be extracted from the modality-specific embeddings using linear probing. As we increase the dimensionality, we observe that significantly more shared information can be extracted, which indicates a *worse* disentanglement for the modality-specific embeddings. In terms of the conditional coherence (Figure 6.3b), we also find that the performance decreases as a function of the modality-specific dimensionality. Overall, these results suggest that we can constrain the capacity of the modality-specific latent space to incentivize the disentanglement and consequently improve the semantic coherence for the generation across modalities.



(a) Linear probing

(b) Generative coherence

Figure 6.3: Evaluation of models with a naive partitioning using the PolyMNIST dataset. We plot the classification accuracy and conditional coherence as a function of the number of modality-specific latent dimensions. Figure 6.3a shows the classification accuracy of a logistic regression model trained to predict the shared information (i.e., the digit label) from the modality-specific embeddings. Thereby, we estimate how well shared information can be extracted from the modality-specific embeddings, and thus a lower value indicates a better disentanglement. Figure 6.3b shows the conditional coherence accuracy averaged over all pairs of distinct modalities. Higher values indicate a better semantic coherence for the conditional generation across modalities.

6.3 Method: MMVAE+

In this section, we introduce the MMVAE+ as a partitioned variant of the mixture of experts multimodal variational autoencoder or MMVAE. The model combines the partitioned architecture with the mixure of experts variational joint posterior and a new regularization technique to alleviate the shortcut problem resulting from the naive partitioning.

First, we define the partitioned MMVAE (Section 6.3.1), for which we take the naive approach and specify the variational joint posterior as a mixture of experts. Then, we propose a regularization technique (Section 6.3.2), for which we decompose the objective of the partitioned MMVAE into two types of reconstructions (self and crossmodal) and regularize the crossmodal-reconstruction terms by sampling the modality-specific embeddings from a prior distribution with a learned variance parameter.

6.3.1 Partitioned MMVAE

The MMVAE defines the variational joint posterior as a mixture of experts (MoE), i.e., a mixture distribution over unimodal variational posteriors (see Equation 3.51). Analogously, for the partitioned MMVAE, we specify the variational joint posterior (i.e., Equation 6.3) using the MoE-decomposition

$$q_{\phi_{\mathbf{c}}}^{\text{MoE}}(\mathbf{c} \mid \mathbf{x}_1, \ldots, \mathbf{x}_M) = \frac{1}{M} \sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,, \tag{6.18}$$

where the subscript j indicates the modality and the parameters of the respective encoder.

To derive the objective of the partitioned MMVAE, we apply the MoE-decomposition from Equation 6.18 to the variational joint posterior in the objective of the naive approach (i.e., Equation 6.17). Thus, we have

$$\begin{aligned} &\sum_{i=1}^{M} \Big\{ \mathbb{E}_{\frac{1}{M}\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ &- D_{\text{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i) \Big) \Big\} - D_{\text{KL}}\Big(\tfrac{1}{M} \textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \Big) \,, \end{aligned} \tag{6.19}$$

for which, using the linearity of expectation, we take the weighted sum from the variational joint posterior out of the expectation (c.f., [Shi+19, p. 5]) to define the objective of the partitioned MMVAE as follows:

$$\begin{aligned} \mathcal{L}_{\text{ELBO-Part.}}^{\text{MMVAE}}(\bar{\mathbf{x}}; \phi, \theta) \coloneqq \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} &\Big\{ \mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ &- D_{\text{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i) \Big) \Big\} \\ &- D_{\text{KL}}\Big(\tfrac{1}{M} \textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \Big) \,. \end{aligned} \tag{6.20}$$

Notably, Equation (6.20) corresponds to the objective used in Section 6.2.3 to showcase the shortcut problem resulting from the naive partitioning of the MMVAE.[16]

Next, we extend the above objective with a regularization technique to incentivize the disentanglement of shared and modality-specific information.

[16] To be precise, the last term from Equation (6.20) is generally not available in closed form, but it can be approximated by sampling uniformly at random from the set of unimodal posteriors [Shi+19].

6.3.2 Regularization of Crossmodal-reconstructions

To derive the objective of the MMVAE+, we take the objective of the partitioned MMVAE (Equation 6.20) and add a simple condition that changes some of the log-likelihood terms. Specifically, we categorize the log-likelihood terms into two types—self- and crossmodal-reconstructions—and regularize the latter by sampling the modality-specific embeddings from an auxiliary prior instead of the modality-specific variational posterior.

Two types of reconstructions Each of the log-likelihood terms in Equation (6.20) can represent either one of two types of reconstructions, as defined by the following cases:

$$\mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)\Big] = \begin{cases} \text{self-reconstruction}, & \text{if } i = j; \\ \text{crossmodal-reconstruction}, & \text{if } i \neq j. \end{cases}$$

In both cases, we evaluate $p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)$, i.e., the log-likelihood of a (target) observation $\mathbf{x}_i$ under our model, given the pair of embeddings $(\mathbf{c}, \mathbf{m}_i) \sim q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)$. In the former case (i.e., $i = j$), we describe the log-likelihood as a "self-reconstruction", because both $q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)$ and $q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)$ are conditioned on the (target) observation, since $\mathbf{x}_i = \mathbf{x}_j$. Conversely, when $i \neq j$, we describe it as a "crossmodal-reconstruction", because $q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)$ is conditioned on an observation of a different modality, i.e., $\mathbf{x}_i \neq \mathbf{x}_j$.

Intuitively, for the partitioned MMVAE, the shortcut problem described in Section 6.2.3 is reinforced through the optimization of crossmodal-reconstructions as part of the objective. For self-reconstructions, both embeddings, $\mathbf{c}$ and $\mathbf{m}_i$, are conditioned on the (target) observation $\mathbf{x}_i$ and thus they can, in principle, both encode the complete information about the given input. However, for crossmodal-reconstructions, only the modality-specific embedding $\mathbf{m}_i \sim q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)$ is conditioned on $\mathbf{x}_i$ and thus it is more informative than $\mathbf{c} \sim q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j)$ with respect to $\mathbf{x}_i$. Thus, through the optimization of crossmodal-reconstructions, the model is incentivized to encode *all* information (i.e., both shared and modality-specific) in the modality-specific embeddings. This corresponds to the failure mode of the partitioned MMVAE that we illustrated in Figures 6.2 and 6.3.

Regularization of crossmodal-reconstructions To block the figurative "shortcut" for crossmodal-reconstructions through the modality-specific encoder, our idea is to draw the modality-specific embeddings from an auxiliary prior instead of the modality-specific variational posterior. The resulting model is called the MMVAE+.

For each crossmodal-reconstruction term in Equation (6.20), we replace the modality-specific variational posterior $q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)$ with an auxiliary prior $q_{\psi_i}(\mathbf{m}_i)$ parameterized by ψ_i. Thus, during training, the MMVAE+ samples

$$\mathbf{m}_i \sim \begin{cases} q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i), & \text{if } i = j; \\ q_{\psi_i}(\mathbf{m}_i), & \text{if } i \neq j. \end{cases} \tag{6.21}$$

In this way, the MMVAE+ can be derived from the objective of the MMVAE with a naive partitioning (Equation 6.20) if we apply the condition defined in Equation (6.21). In terms of indicator functions, we define the objective of the MMVAE+ as follows:

$$\begin{aligned} \mathcal{L}^{\text{MMVAE+}}_{\text{ELBO-Part.}}(\bar{\mathbf{x}}; \phi, \theta, \psi) := \frac{1}{M} \sum_{i=1}^{M} \Big\{ & \mathbb{E}_{q_{\phi_{\mathbf{c}_i}}(\mathbf{c} \mid \mathbf{x}_i) q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ & + \sum_{j=1}^{M} \mathbb{1}_{\{i \neq j\}} \mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) q_{\psi_i}(\mathbf{m}_i)} \Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i) \Big] \\ & - D_{\text{KL}} \Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i) \Big) \Big\} \\ & - D_{\text{KL}} \Big(\tfrac{1}{M} \textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \Big) \,, \end{aligned} \tag{6.22}$$

where $\mathbb{1}_{\{i \neq j\}}$ is an indicator function, which evaluates to one if $i \neq j$ and zero otherwise, and $\psi = \{\psi_1, \ldots, \psi_M\}$ are the parameters of the auxiliary priors.

Thus, we gave an intuitive explanation of how the MMVAE+ objective relates to the objective of the MMVAE with a naive partitioning. More formally, to justify the application of the model for multimodal generative learning, the objective can be derived as a lower bound on $\log p(\bar{\mathbf{x}})$, as described in Lemma 4.

Lemma 4. *Objective $\mathcal{L}^{MMVAE+}_{ELBO\text{-}Part.}(\bar{\mathbf{x}}; \phi, \theta, \psi)$ (Equation 6.22) forms a lower bound on $\log p(\bar{\mathbf{x}})$, i.e.,*

$$\log p(\bar{\mathbf{x}}) \geq \mathcal{L}^{MMVAE+}_{ELBO\text{-}Part.}(\bar{\mathbf{x}}; \phi, \theta, \psi) \,. \tag{6.23}$$

A proof of Lemma 4 is provided in Appendix A.2.

Design of the auxiliary prior The sampling procedure in Equation (6.21) already yields a generic form of the MMVAE+ objective without further specification of the auxiliary prior. However, it is vital to select a suitable distribution for the prior, because it can affect the approximation of the log-evidence. A natural choice is to select an auxiliary

prior that is compatible with the modality-specific prior $p(\mathbf{m}_i)$, for which we use a standard normal distribution. Analogously, we define the auxiliary prior as a normal distribution

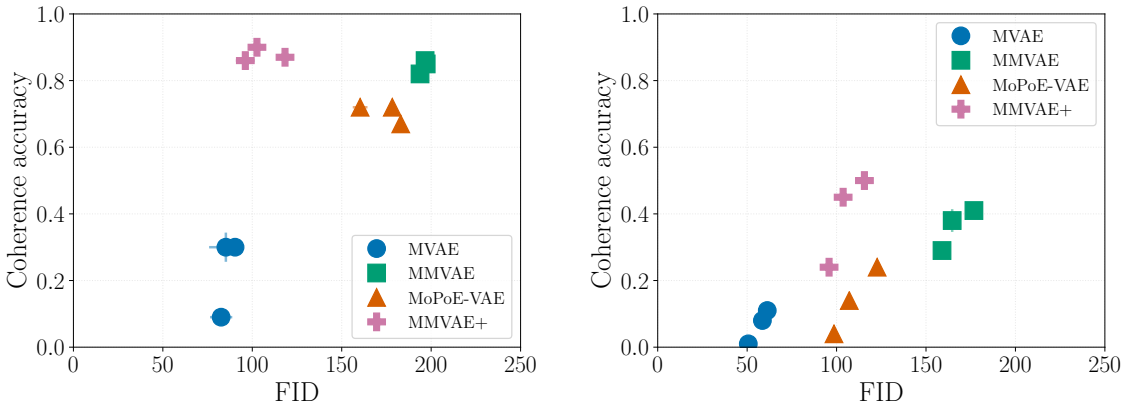
$$q_{\psi_i}(\mathbf{m}_i) \coloneqq \mathcal{N}(0, \sigma_i^2) \tag{6.24}$$

that is centered at zero and has a learned variance parameter (i.e., $\psi_i = \sigma_i^2$). Hence, by design, the auxiliary prior has the same location as the modality-specific prior, but its scale remains flexible for the modeling of crossmodal-reconstructions. At test time, the MMVAE+ samples $\mathbf{m}_i \sim p(\mathbf{m}_i)$, but in the experiments, we also investigate the effects of sampling from the auxiliary prior and of interpolating between the two distributions.

6.4 Experiments

Datasets For the experiments, we make use of two datasets. First, we use the PolyMNIST dataset to evaluate the MMVAE+ in a controlled setup with five synthetic modalities and compare its performance to previous multimodal VAEs. Second, we demonstrate the utility of the proposed model on CUB Image-Captions, which is a challenging multimodal dataset comprised of images of birds paired with matching textual descriptions of the birds in natural language. While previous work [Shi+19; Shi+21; Joy+22] tackled a simplified version of this dataset by using pre-trained ResNet-features, we train the models on actual images. For more information on the datasets, see Chapter 4.

Evaluation metrics For the quantitative evaluation of the generated samples, we mainly consider two metrics that capture different aspects of generative performance in a multimodal setup, namely *generative coherence* and *generative quality* (c.f., Section 2.5). Generative coherence measures if the generated samples agree in their semantic content across modalities, which we evaluate using the conditional coherence accuracy (defined in Section 2.5) averaged over all pairs of distinct modalities. Conversely, generative quality measures how well the model approximates the data distribution. We use the Fréchet inception distance [FID; Heu+17], a standard metric used to evaluate sample quality for generative models in image domains. All quantitative results in this section are averaged over three seeds and include standard deviations.



(a) Conditional generation

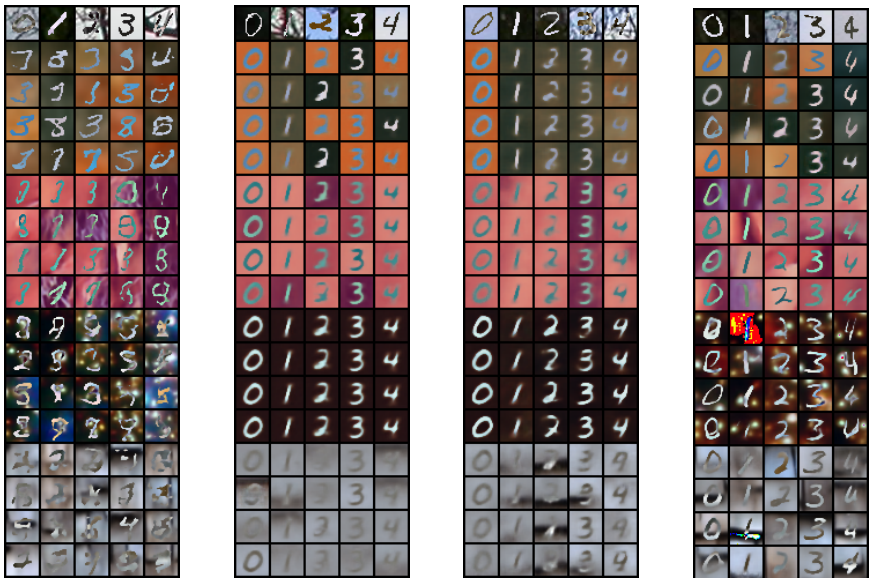
(b) Unconditional generation

Figure 6.4: Evaluation of generative quality and coherence on PolyMNIST. In Figure 6.4a, we evaluate the *conditional* generation, for which we compute the average FID (horizontal axis) or average coherence accuracy (vertical axis) over all pairs of distinct modalities. In Figure 6.4a, we assess the *unconditional* generation performance, for which we plot the FID (averaged over modalities) on the horizontal and the unconditional coherence on the vertical axis. For each model, we include multiple β values (shown as individual points).[17] Standard deviations are shown as error bars around points, if the values are sufficiently large. A capable model should exhibit high coherence and low FID; an optimal performance would correspond to a point in the top-left corner of each scatterplot.

6.4.1 PolyMNIST

Quantitative results In Figure 6.4, we show the results of a quantitative evaluation of the proposed MMVAE+ compared to the baseline approaches, namely the MMVAE [Shi+19], MVAE [WG18], and MoPoE-VAE [SDV21]. In each scatterplot, we visualize the generative coherence on the horizontal and generative quality on the vertical axis. We observe that the MMVAE+ achieves a decent generative quality (i.e., low FID) and high coherence for both conditional and unconditional generation (Figures 6.4a and 6.4b respectively), while the baselines underperform in at least one of the two performance criteria. Overall, the MMVAE+ exhibits a better and more consistent tradeoff, as the described trend holds for a range of β values (shown as individual points).[17]

[17]The hyperparameter β weights the KL-divergence (or the sum of KL-divergence terms); thus, it can be used to control the tradeoff between reconstruction and matching of the prior distribution [Hig+17].

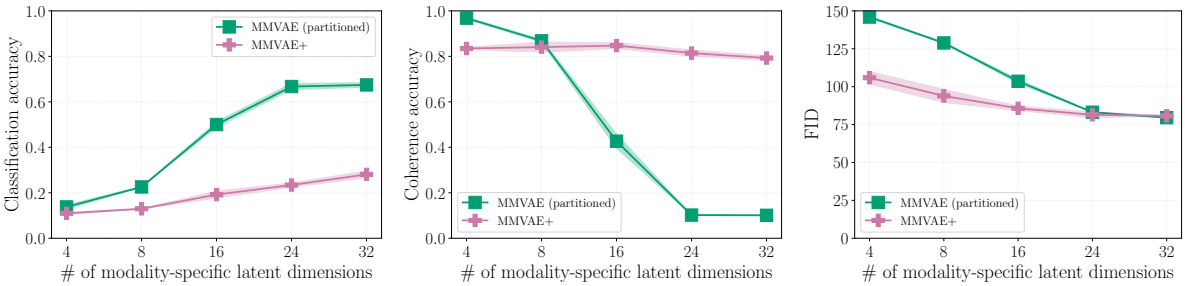


(a) MVAE (b) MMVAE (c) MoPoE-VAE (d) MMVAE+

Figure 6.5: Qualitative results for the conditional generation across modalities. Along each column of a subfigure, we visualize the input image (first row) and the generated images for each of the missing modalities (remaining rows). Specifically, we show four generated images per modality, all of which were conditionally generated given the respective input image shown in the first row. A performant model should produce semantically coherent examples (i.e., consistent digits along each column) of sufficiently high diversity.

Qualitative results In Figure 6.5, we showcase examples for the conditional generation across modalities. Consistent with the quantitative results, we observe that the MVAE exhibits a lack of semantic coherence as the conditionally generated samples show incoherent digits, whereas the MMVAE and MoPoE-VAE generate average-looking samples that indicate a lack of sample diversity. In contrast, the MMVAE+ shows a significantly better generative quality with an improved sample diversity and high semantic coherence. In direct comparison to the MMVAE, the results validate that our approach yields an improved generative model with a superior generative quality and semantic coherence.

Effectiveness of the regularization In Figure 6.6, we compare the MMVAE+ to the MMVAE with a naive partitioning of the latent space (i.e., the model based on Equation 6.20). The only difference between the two approaches is the regularization of crossmodal-reconstructions using the auxiliary priors. In each subplot, we show the



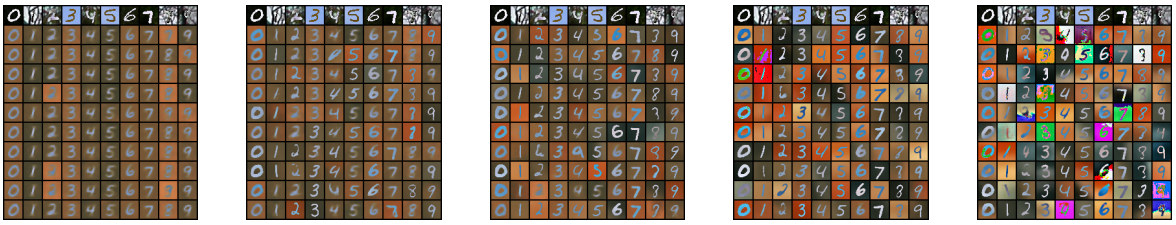
(a) Linear probing (b) Generative coherence (c) Generative quality

Figure 6.6: Comparison of the MMVAE+ to the MMVAE with a naive partitioning. We evaluate the performance in terms of three metrics, for each of which we plot the results as a function of the modality-specific dimensionality.[18] Figure 6.6a shows the classification accuracy of a logistic regression model trained to predict the shared information (i.e., the digit label) from the modality-specific embeddings; thus, a lower value indicates a better disentanglement. Figure 6.6b evaluates the generation across modalities, for which we compute the average coherence over all pairs of distinct modalities. Figure 6.6c assesses the unconditional generation performance, for which we compute the average FID across all modalities generated from prior samples. Ideally, a multimodal generative model should exhibit high semantic coherence and generative quality, i.e., high coherence and low FID.

performance as a function of the number of modality-specific dimensions.[18] We find that the MMVAE+ exhibits a significantly better performance over most of the value range. Only in terms of generative coherence (Figure 6.6b), the partitioned MMVAE performs slightly better, but only when the modality-specific dimensionality is small and at the expense of generative quality (see Figure 6.6c). Additionally, we use linear probing to assess how much shared information is encoded in the modality-specific embeddings (Figure 6.6a). We observe that significantly less shared information can be extracted from the modality-specific embeddings produced by the MMVAE+ compared to those of the MMVAE with a naive partitioning, which indicates that the regularization incentivizes the disentanglement of shared and modality-specific information effectively.[19]

[18]For each approach, we use a shared latent space with 32 dimensions and only vary the number of modality-specific latent dimensions. Specifically, we train separate models, each of which uses a different dimensionality for the modality-specific subspaces.

[19]Using linear probing, we verified that both the MMVAE+ and the partitioned MMVAE encode shared information well in the shared subspace. Though, for the partitioned MMVAE, we found that the classification accuracy drops significantly if the number of modality-specific dimensions is larger than 16.



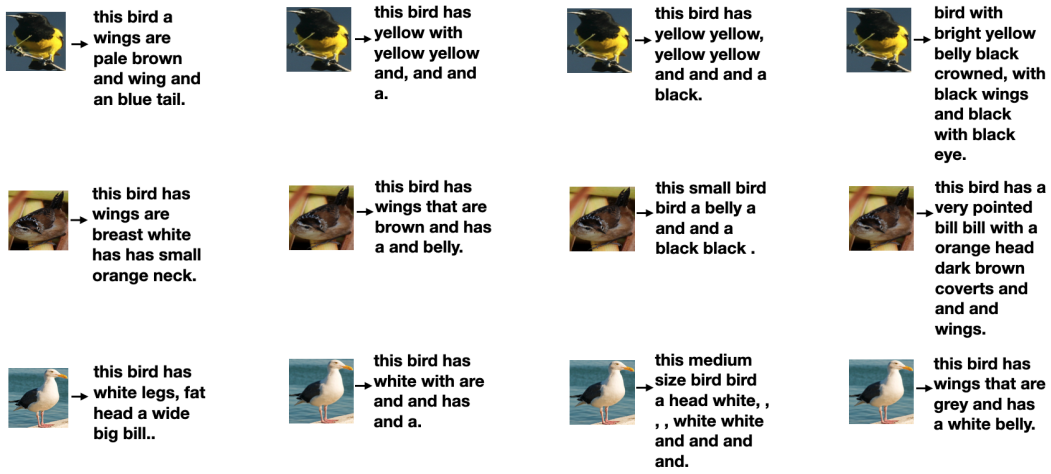
(a) $\alpha = 0.001$ (b) $\alpha = 0.025$ (c) $\alpha = 0.5$ (d) $\alpha = 1$ (e) $\alpha = 2$

Figure 6.7: Varying the scale of the modality-specific prior at test time. In each subfigure, we show the qualitative results for the conditional generation from one modality to another for a specific value of α, which is used to draw the modality-specific embeddings $\mathbf{m}_i \sim \mathcal{N}(0, \alpha^2)$. Along each column, we visualize the input image (first row) and ten generated images (remaining rows) using the MMVAE+ trained on PolyMNIST. For each generated image, we draw a different sample from the modality-specific prior scaled by α.

Overall, the results support that the proposed regularization of crossmodal-reconstructions addresses the shortcut problem, that it improves the disentanglement of shared and modality-specific information, and consequently enhances the generative quality and coherence compared to the partitioned MMVAE and the baselines with a joint latent space.

Varying the scale of the modality-specific prior We further investigate the effects of the auxiliary prior with an analysis for which we vary the scale parameter of the prior distribution at test time. For a trained model, we draw $\mathbf{m}_i \sim \mathcal{N}(0, \alpha^2)$, where we set a specific value for α to scale the prior distribution. A value of $\alpha = 0.01$ approximates the learned scale parameter of the auxiliary prior $q_{\psi_i}(\mathbf{m}_i)$, which is concentrated around zero. A value of $\alpha = 1.0$ corresponds to the modality-specific prior $p(\mathbf{m}_i)$.

In Figure 6.7, we show the qualitative results for the conditional generation with a fixed model that was trained on PolyMNIST. Starting from $\alpha = 0.01$, we increase the value to interpolate between the two distributions and further increase the value up to $\alpha = 2$. For $\alpha = 0.01$, we observe coherent samples but with a pronounced lack of diversity, like previously observed for the MMVAE. As we increase α, we observe more diversity and less coherence. Thus, the results suggest that the scale of the modality-specific prior can be adapted at test time to fine-tune the generative performance of the MMVAE+.



(a) MVAE (b) MoPoE-VAE (c) MMVAE (d) MMVAE+

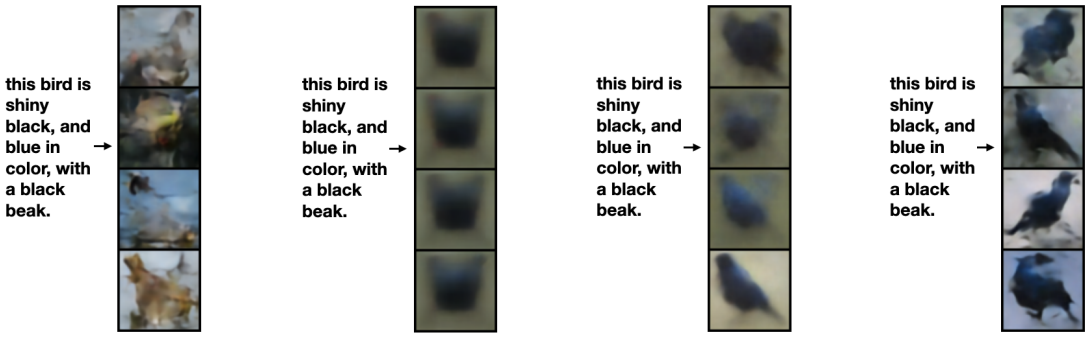
Figure 6.8: Conditional image-to-caption generation on the CUB Image-Captions dataset.

6.4.2 CUB Image-Captions

Next, we evaluate the MMVAE+ on the CUB Image-Captions dataset [Shi+19], a multimodal dataset comprised of natural images of birds with corresponding captions, i.e, textual descriptions of the birds' appearance.

Qualitative results For the conditional caption-to-image generation (Figure 6.8), we again observe that the MMVAE and MoPoE-VAE tend to generate average-looking samples with a lack of sample diversity, whereas the MVAE exhibits a decent generative quality but a severe lack of semantic coherence with respect to the given caption. Similar results can be observed for the conditional image-to-caption generation (Figure 6.9). In comparison, the MMVAE+ shows significantly better results for both caption-to-image and image-to-caption generation, as the model produces coherent samples with high sample diversity for both modalities. Hence, the qualitative results demonstrate the effectiveness of our approach on a complex real-world dataset with heterogeneous modalities.

Quantitative results Table 6.1 presents a quantitative evaluation of the caption-to-image generation in terms of generative coherence and generative quality. To evaluate the generative coherence, we use a proxy to determine the color of the bird (see Appendix A.4).



(a) MVAE (b) MoPoE-VAE (c) MMVAE (d) MMVAE+

Figure 6.9: Conditional caption-to-image generation on the CUB Image-Captions dataset. For each model, the shown caption is used as input to conditionally generate four images.

We find that the MMVAE+ performs favorably in terms of both generative coherence and quality compared to the baselines, which corroborates our previous results.

	Generative coherence	**Generative quality**
MVAE	0.271 (±0.007)	172.21 (±39.6)
MMVAE	0.713 (±0.057)	232.20 (±2.14)
MoPoE-VAE	0.579 (±0.158)	265.55 (±4.01)
MMVAE+	**0.721** (±0.090)	**164.94** (±1.50)

Table 6.1: Evaluation of generative coherence (higher is better) and generative quality (in terms of FID; lower is better) for caption-to-image generation using CUB Image-Captions.

6.5 Summary

In this chapter, we designed multimodal VAEs with an explicit partitioning of the latent space into shared and modality-specific subspaces. First, we introduced a naive approach for partitioning the latent space and discussed its shortcomings. Specifically, we highlighted the shortcut problem and illustrated the issue based on a concrete example. Yet, we also presented empirical results suggesting that the naive partitioning can, in principle, disentangle shared and modality-specific information in the respective subspaces, if we

have access to a handful of labels to guide model selection. Finally, we proposed the MMVAE+ and demonstrated that it effectively averts the shortcut problem and promotes the disentanglement without additional labels. We presented experiments on synthetic and real-world datasets, showing that the MMVAE+ achieves a superior tradeoff in terms of generative coherence and generative quality compared to the baselines.

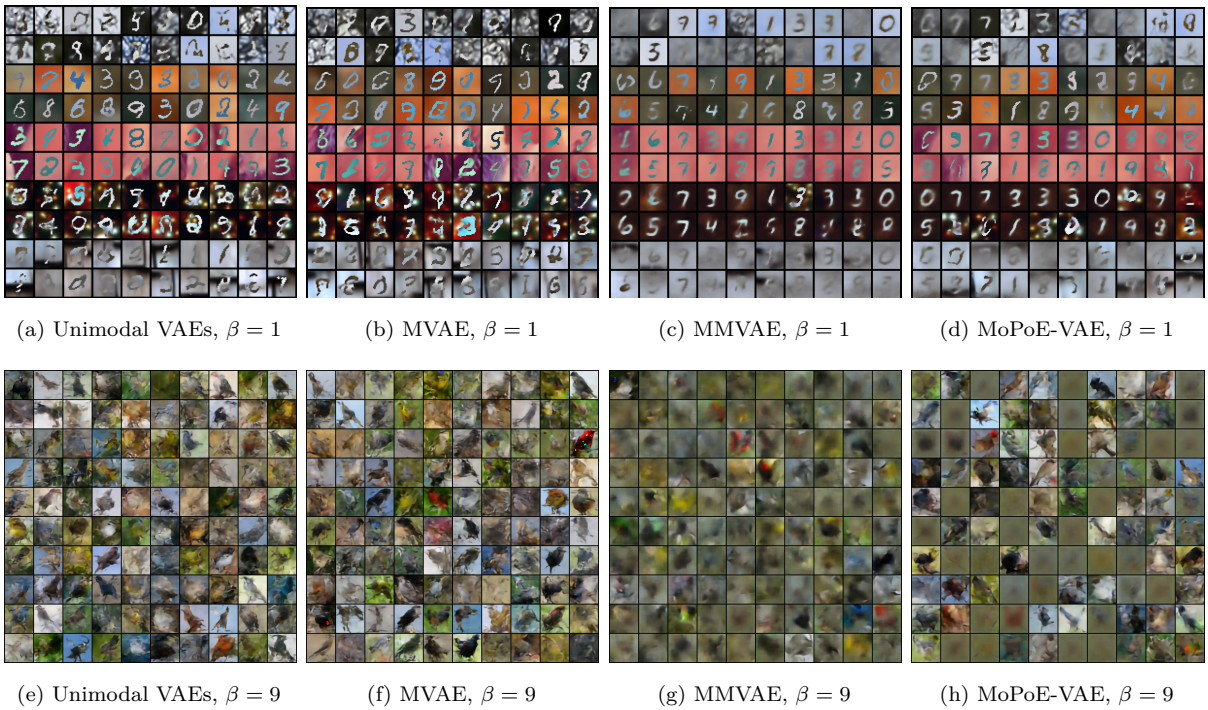
In the scope of the thesis, this chapter presented a method to promote the disentanglement of shared and modality-specific information (c.f., Question 2) using multimodal VAEs. Additionally, it also provided a foretaste of the challenges for representation learning with multimodal VAEs. In the next chapter, we further investigate these limitations, before we turn to viable alternatives with a discriminative approach (Chapter 8) and a hybrid method (Chapter 9).

7

Limitations of Multimodal VAEs

In the previous chapters, we developed novel models for multimodal generative learning within the framework of variational inference. In this chapter, we pause to reflect on the previous results and assess the limitations of multimodal VAEs, including some of the models we proposed. Thereby, we seek to obtain a broader and more nuanced perspective to gauge our progress with respect to the considered research questions.

First, in Section 7.1, we highlight a counterintuitive result we observed in previous experiments, showing that some multimodal VAEs exhibit a *worse* generative quality compared to unimodal VAEs despite the advantage of weak supervision in the multimodal setup. In our attempt to explain this gap, we uncover a fundamental limitation that applies to a large family of multimodal VAEs. We prove that the sub-sampling of modalities used by mixture-based multimodal VAEs enforces an undesirable upper bound on the multimodal ELBO and thereby limits the generative quality of the respective models (Section 7.2). Empirically, we showcase the gap in generative quality on both synthetic and real data and discuss the tradeoffs between different variants of multimodal VAEs (Section 7.3). We find that none of the existing approaches fulfills all desired criteria of an effective multimodal generative model when applied on more complex datasets than those used in previous benchmarks—specifically, when shared information cannot be predicted in expectation across modalities on the level of observations. More broadly, we identify, formalize, and validate fundamental limitations of VAE-based approaches for modeling weakly supervised data and discuss implications for real-world applications.



(a) Unimodal VAEs, $\beta = 1$ (b) MVAE, $\beta = 1$ (c) MMVAE, $\beta = 1$ (d) MoPoE-VAE, $\beta = 1$

(e) Unimodal VAEs, $\beta = 9$ (f) MVAE, $\beta = 9$ (g) MMVAE, $\beta = 9$ (h) MoPoE-VAE, $\beta = 9$

Figure 7.1: Qualitative results for the unconditional generation using prior samples. For models trained on the PolyMNIST dataset (Figures (a) to (d)), we show 20 samples for each of the five modalities. For models trained on the CUB Image-Captions dataset, we show 100 generated images respectively (Figures (e) to (h)). The results indicate that the MVAE performs on par with unimodal VAEs in terms of generative quality, whereas the MMVAE and MoPoE-VAE perform significantly worse in comparison.

7.1 Example and Overview

Multimodal VAEs have shown great potential as efficient generative models for weakly supervised data, e.g., for pairs of images or paired images and captions. Previous studies [WG18; Shi+19; SDV20; SDV21], including our own work from Chapter 5, demonstrate that multimodal VAEs leverage weak supervision to learn generalizable representations useful for downstream tasks, for example in biomedical applications [Dor+19; Min+21; LS21]. However, in our experiments with multimodal VAEs, we also noticed several shortcomings, which we seek to formalize in this chapter.

One counterintuitive example of the observed limitations is shown in Figure 7.1, which compares the generated samples from different types of multimodal VAEs to those generated

by unimodal VAEs.[20] These results illustrate that some multimodal VAEs underperform in terms of generative quality compared to unimodal VAEs, despite the advantage of weak supervision in the multimodal setup. A notable exception is the MVAE, which performs on par with unimodal VAEs; however, it exhibits a lack of generative coherence, as discussed in the previous chapters. This apparent tradeoff serves as a starting point for our analysis, which aims to explain the observed lack of generative quality in terms of a fundamental limitation that underlies the family of mixture-based multimodal VAEs.

Next, we recall the definition of the family of mixture-based multimodal VAEs and briefly discuss the tradeoffs between different types of models.

7.1.1 Mixture-based Multimodal VAEs

As previously noted in Chapter 5, different types of multimodal VAEs use different decompositions of the variational joint posterior in terms of the unimodal variational posteriors. Specifically, they decompose it as a product, mixture, or mixture of products of experts respectively. In Section 5.2.3, we showed that existing approaches can be generalized and therefore defined the *family of mixture-based multimodal VAEs*, which subsumes the MVAE [WG18], MMVAE [Shi+19], and the proposed MoPoE-VAE.

Recall our definition of the family of mixture-based multimodal VAEs (Definition 7), where

$$q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}}) = \sum_{A \in \mathcal{S}} \omega_A \, q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A) \tag{7.1}$$

defines the variational joint posterior as a mixture distribution over a given set $\mathcal{S}$, which is comprised of subsets of modalities $A \subseteq \{1, \ldots, M\}$ and corresponding mixture coefficients ω_A, such that

$$\mathcal{S} \subseteq \{(A, \omega_A) \mid A \subseteq \{1, \ldots, M\},\ A \neq \emptyset,\ \omega_A \in [0, 1]\}, \text{ and } \textstyle\sum_{A \in \mathcal{S}} \omega_A = 1. \tag{7.2}$$

Based on the above formulation of the variational joint posterior, we defined the ELBO for mixture-based multimodal VAEs in Definition 7 as the objective

$$\mathcal{L}_{\text{ELBO}}^{\mathcal{S}}(\bar{\mathbf{x}}; \phi, \theta) = \sum_{A \in \mathcal{S}} \omega_A \, \mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}(q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \,, \tag{7.3}$$

[20] By "unimodal VAEs" we simply mean separate VAEs, each of which is trained independently of the others on the observations of a single modality.

	Model properties		Generative performance	
Model	**Decomposition of $q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}})$**	**Modality sub-sampling**	**Quality**	**Coherence**
MVAE [WG18]	$\prod_{i=1}^{M} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i)$	No sub-sampling	Good	Poor
MMVAE [Shi+19]	$\frac{1}{M} \sum_{i=1}^{M} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i)$	Individual modalities	Poor	Good ↯
MoPoE-VAE [SDV21]	$\frac{1}{\lvert\mathcal{P}(M)\rvert} \sum_{A \in \mathcal{P}(M)} \prod_{i \in A} q_{\phi_i}(\mathbf{z} \mid \mathbf{x}_i)$	All subsets of modalities	Poor	Good ↯

Table 7.1: Overview of multimodal VAEs. Entries for generative quality and generative coherence are based on the observed empirical results. The lightning symbol (↯) denotes previous findings for which this chapter presents contrary evidence. For the MVAE, we consider the model without ELBO sub-sampling, as discussed in Chapter 5.

or any lower bound thereof. Given a specific choice of $q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}})$, we can plug it into Equation (7.3) to recover the objective of the MVAE, MMVAE, and MoPoE-VAE respectively, as shown in Section 5.2.3 and summarized in Table 7.1.

Modality sub-sampling From a computational perspective, a characteristic of mixture-based multimodal VAEs is the sub-sampling of modalities during training, which is a direct consequence of defining the variational joint posterior as a mixture distribution over subsets of modalities (Equation 7.1). In particular, the MMVAE sub-samples individual modalities $i \in \{1, \ldots, M\}$, whereas the MoPoE-VAE considers all subsets of modalities, i.e., each subset $A \in \mathcal{P}(M)$, where $\mathcal{P}(M)$ is the powerset of $\{1, \ldots, M\}$. The only member of the family of mixture-based multimodal VAEs that forgoes sub-sampling is the MVAE, which defines a trivial mixture over a single subset—namely, the complete set of modalities.

Tradeoffs between models Table 7.1 provides an overview of the different variants of mixture-based multimodal VAEs and the empirical results for the respective models. Notably, there appears to be a tradeoff between the *generative quality* (i.e., the fidelity of generated samples) and *generative coherence* (i.e., the ability to generate semantically related samples across modalities). While a lack of generative coherence for the MVAE was already observed in previous works [e.g., Shi+19], the tradeoff between generative quality and coherence was not yet established. Visually, a lack of generative quality or coherence can be discerned from the qualitative results we presented previously (e.g., Figures 6.5 and 7.1) and from our previous quantitative evaluations (e.g., Figure 6.4).

In this chapter, we explain *why* the generative quality is worse for models that sub-sample

modalities (Theorem 1) and show that a tighter approximation of the joint distribution can be achieved without sub-sampling (Corollary 1). Through systematic ablations, we validate the proposed theoretical limitations and showcase the tradeoff between generative quality and generative coherence (Section 7.3.1). Our experiments also reveal that generative coherence cannot be guaranteed in applications on more complex datasets than those used in previous benchmarks (Section 7.3.2).

7.2 Theoretical Results

In this section, we present our theoretical results on the limitations of multimodal VAEs. First, we introduce some preliminary notions (Section 7.2.1) and give an intuition about the problem (Section 7.2.2). Then, we formalize the limitations with a theoretical result (Section 7.2.3) and discuss its implications for individual models (Section 7.2.4).

7.2.1 Preliminaries

Let M be the number of modalities, let $\bar{\mathbf{x}} \coloneqq \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$ be a set of discrete random vectors that describe different modalities, and let $p(\bar{\mathbf{x}})$ denote their joint distribution. As in the previous chapters, we denote subsets of modalities using subscripts; in particular, we write $\mathbf{x}_A$ to index a subset of modalities $A \subseteq \{1, \dots, M\}$.

Throughout this chapter, we assume that the observations are described by *discrete* random vectors (e.g., pixel intensities, as in the case of RGB images), so that we can safely assume non-negative entropy and conditional entropy terms. Definitions for all required information-theoretic quantities are provided in Section 3.3.4.1.

7.2.2 Intuition about the Problem

Before we delve into the details, let us illustrate how modality sub-sampling can affect the likelihood estimation that is part of the multimodal ELBO and therefore influence the approximation of the log-evidence.

Recall the definition of the ELBO for mixture-based multimodal VAEs (Equation 7.3) and pay attention to the likelihood estimation, i.e., the term

$$\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] \,, \tag{7.4}$$

where $A \subseteq \{1, \ldots, M\}$ indexes a subset of modalities.

Crucially, in Equation (7.4) the variational posterior $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ is conditioned on a *subset* of modalities. This can have a profound impact on the likelihood estimation, because the precise estimation of $\log p(\bar{\mathbf{x}} \mid \mathbf{z})$ might depend on information from *all* modalities. However, the embedding $\mathbf{z}$ is sampled from $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ and thus it only contains information from a subset of modalities.

Consequently, for the reconstruction of missing modalities, a model might be able to predict information that is shared between observations of different modalities. However, it cannot reliably predict modality-specific information, such as the background details in an image given a concise verbal description of its content. Hence, the sub-sampling of modalities can affect the likelihood estimation and thus the approximation of the log-evidence through the maximization of the multimodal ELBO.

In the following, we formalize the above intuition by showing that, in the presence of modality-specific variability, modality sub-sampling enforces an undesirable upper bound on the multimodal ELBO and therefore prevents a tight approximation of the log-evidence.

7.2.3 Formalization of the Problem

Theorem 1 states our main theoretical result for this chapter. It describes a non-trivial limitation of mixture-based multimodal VAEs because it shows that the sub-sampling of modalities enforces an undesirable upper bound on the approximation of the joint distribution (i.e., the expected log-evidence) when there is modality-specific variability in the data. This limitation conflicts with the goal of modeling real-world multimodal data, which typically exhibits a considerable degree of modality-specific variability.

Theorem 1. *Any mixture-based multimodal VAE (Definition 7) approximates the expected log-evidence up to an irreducible discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ that depends on the model-specific mixture distribution $\mathcal{S}$ and on the amount of modality-specific information in the data. For the maximization of $\mathcal{L}^{\mathcal{S}}_{ELBO}(\bar{\mathbf{x}}; \phi, \theta)$ and any value of ϕ and θ, the following inequality holds:*

$$\mathbb{E}_{p(\bar{\mathbf{x}})}[\log p(\bar{\mathbf{x}})] \geq \mathbb{E}_{p(\bar{\mathbf{x}})}[\mathcal{L}^{\mathcal{S}}_{ELBO}(\bar{\mathbf{x}}; \phi, \theta)] + \Delta(\bar{\mathbf{x}}, \mathcal{S}) \tag{7.5}$$

where

$$\Delta(\bar{\mathbf{x}}, \mathcal{S}) = \sum_{A \in \mathcal{S}} \omega_A \, H(\mathbf{x}_{\{1,\ldots,M\} \setminus A} \mid \mathbf{x}_A) \; . \tag{7.6}$$

In particular, the discrepancy is always greater than or equal to zero and it is independent of ϕ and θ and thus remains constant during the optimization.

Before we begin with the proof, recall that the multimodal ELBO forms a variational lower bound on the log-evidence $\log p(\bar{\mathbf{x}})$, and if we take the expectation over $p(\bar{\mathbf{x}})$, we get a variational lower bound on the expected log-evidence, which is the target quantity we want to approximate with a multimodal generative model.

For the family of mixture-based multimodal VAEs, we consider the objective from Equation (7.3), which also forms a lower bound on the expected log-evidence, namely

$$\mathbb{E}_{p(\bar{\mathbf{x}})}[\log p(\bar{\mathbf{x}})] \geq \mathbb{E}_{p(\bar{\mathbf{x}})}[\mathcal{L}^{\mathcal{S}}_{\text{ELBO}}(\bar{\mathbf{x}}; \phi, \theta)] \tag{7.7}$$

$$= \mathbb{E}_{p(\bar{\mathbf{x}})}\left[\sum_{A \in \mathcal{S}} \omega_A \, \mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] - D_{\text{KL}}(q^{\mathcal{S}}_\phi(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \right] . \tag{7.8}$$

We can reformulate Equation (7.8) using the linearity of expectation to obtain

$$\sum_{A \in \mathcal{S}} \omega_A \, \mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] \right] - \mathbb{E}_{p(\bar{\mathbf{x}})}\left[D_{\text{KL}}(q^{\mathcal{S}}_\phi(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})) \right] , \tag{7.9}$$

which is comprised of a sum of likelihood terms, one for each $A \in \mathcal{S}$, and one KL-divergence.

Proof sketch In the following, we show that for any $A \in \mathcal{S}$ the respective likelihood term from Equation (7.9) is upper-bound by an irreducible error that corresponds to the uncertainty for the prediction of the complete set of observations given a subset thereof. First, we notice that the embedding is a function of $\mathbf{x}_A$ since it is sampled from the variational posterior $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ that is conditioned on $\mathbf{x}_A$, which implies the conditional independence $\mathbf{z}_A \perp\!\!\!\perp \mathbf{x}_{\{1,\dots,M\} \setminus A} \mid \mathbf{x}_A$, where $\mathbf{z}_A \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$. Then, we draw a link to the conditional entropy $H(\bar{\mathbf{x}} \mid \mathbf{z}_A)$ and use the conditional independence to decompose it into a reducible and irreducible error. The latter corresponds to $H(\mathbf{x}_{\{1,\dots,M\} \setminus A} \mid \mathbf{x}_A)$, which is non-negative and independent of the optimization with respect to ϕ and θ. Specifically, we notice that the irreducible error remains even if we assume an optimal variational approximation, i.e., $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A) = p(\mathbf{z} \mid \mathbf{x}_A)$ and $p_\theta(\bar{\mathbf{x}} \mid \mathbf{z}) = p(\bar{\mathbf{x}} \mid \mathbf{z})$ for $\mathbf{z} \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$. Finally, we extend the result to the weighted sum of likelihood terms in Equation (7.9).

Proof. Consider the objective in Equation (7.9). Take any subset $A \in \mathcal{S}$ and focus on the likelihood estimation, i.e., the term

$$\mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})] \right] . \tag{7.10}$$

Notice that the variational posterior in Equation (7.10) is conditioned on the observations of a *subset* of modalities $A \subseteq \{1, \dots, M\}$. In particular, we sample $\mathbf{z} \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ and thus the embedding is a function of $\mathbf{x}_A$, i.e.,

$$\mathbf{z}_A = \mathbf{g}_{\phi_A}(\mathbf{x}_A, \boldsymbol{\varepsilon}_{\mathbf{x}_A}) \,, \tag{7.11}$$

where $\mathbf{z}_A$ is the embedding produced by an encoder $\mathbf{g}_{\phi_A} : (\mathbf{x}_A, \boldsymbol{\varepsilon}_{\mathbf{x}_A}) \mapsto \mathbf{z}_A$ for which $\boldsymbol{\varepsilon}_{\mathbf{x}_A}$ is an optional noise vector that can also be a function of the input $\mathbf{x}_A$.[21] In Equation (7.11), we denote the embedding $\mathbf{z}_A$ with a subscript to indicate the functional dependence on $\mathbf{x}_A$ explicitly.

Let $\{1, \dots, M\} \setminus A$ be the complement of A and let $\mathbf{x}_{\{1,\dots,M\}\setminus A}$ be the set of *unobserved* modalities as a complement of $\mathbf{x}_A$. The encoding process satisfies the Markov chain

$$\mathbf{z}_A \leftarrow \mathbf{x}_A \multimap \mathbf{x}_{\{1,\dots,M\}\setminus A} \tag{7.12}$$

because the embedding $\mathbf{z}_A$ is a function of $\mathbf{x}_A$ (i.e., the observed modalities) and it depends on the remaining (i.e., unobserved) modalities only through $\mathbf{x}_A$. Therefore, we have the conditional independence

$$\mathbf{z}_A \perp\!\!\!\perp \mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A \tag{7.13}$$

and consequently it holds that $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A) = q_{\phi_A}(\mathbf{z} \mid \bar{\mathbf{x}})$, where $\bar{\mathbf{x}} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$.

Next, we draw a connection to the conditional entropy. As in the information-theoretic derivation of the ELBO, the log-likelihood term can be viewed as a variational lower bound on the conditional entropy (c.f., Section 3.3.4). Specifically, we have

$$\mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})]\right] \leq -H(\bar{\mathbf{x}} \mid \mathbf{z}_A) \,, \tag{7.14}$$

where $H(\bar{\mathbf{x}} \mid \mathbf{z}_A)$ is the conditional entropy of $\bar{\mathbf{x}}$ given $\mathbf{z}_A \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ with respect to the joint distribution of the data and encoding, i.e., $p(\bar{\mathbf{x}}, \mathbf{z}_A) = p(\bar{\mathbf{x}}) q_{\phi_A}(\mathbf{z}_A \mid \mathbf{x}_A)$, for which $q_{\phi_A}(\mathbf{z}_A \mid \mathbf{x}_A) = q_{\phi_A}(\mathbf{z}_A \mid \bar{\mathbf{x}})$ due to the conditional independence in Equation (7.13).

Next, we decompose the conditional entropy as follows:

$$\begin{aligned}
H(\bar{\mathbf{x}} \mid \mathbf{z}_A) &= H(\mathbf{x}_A, \mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{z}_A) && (7.15)\\
&= H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{z}_A, \mathbf{x}_A) + H(\mathbf{x}_A \mid \mathbf{z}_A) && (7.16)\\
&= H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) - I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{z}_A \mid \mathbf{x}_A) + H(\mathbf{x}_A \mid \mathbf{z}_A) && (7.17)\\
&= H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) + H(\mathbf{x}_A \mid \mathbf{z}_A) \,. && (7.18)
\end{aligned}$$

[21] For example, for the VAE with a Gaussian variational posterior, we have $\mathbf{z}_A = \mu_{\mathbf{x}_A} + \boldsymbol{\varepsilon}_{\mathbf{x}_A}$ with $\boldsymbol{\varepsilon}_{\mathbf{x}_A} \sim \mathcal{N}(0, \sigma^2_{\mathbf{x}_A})$, where $\mu_{\mathbf{x}_A}$ and $\sigma^2_{\mathbf{x}_A}$ are the estimated parameters of the variational posterior for the given observation (e.g., see [Doe16]).

In Equation (7.16), we apply a standard decomposition of the conditional entropy of two random variables given a third (e.g., [CT12, p. 18]). Equation (7.17) can be derived from the definition of the conditional mutual information.[22] The last step follows from the conditional independence in Equation (7.13), which implies $I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{z}_A \mid \mathbf{x}_A) = 0$.

Thus, the likelihood term from Equation (7.10) forms the following lower bound:

$$\mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})]\right] \overset{(7.14)}{\leq} -H(\bar{\mathbf{x}} \mid \mathbf{z}_A) \tag{7.19}$$

$$\overset{(7.18)}{=} -H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) - H(\mathbf{x}_A \mid \mathbf{z}_A)\ . \tag{7.20}$$

While the second term in Equation (7.20) can be minimized as a function of the data, the first term remains independent of the optimization with respect to ϕ and θ as it represents the aleatoric uncertainty with respect to the missing modalities. Specifically, for the maximization of the log-likelihood, we have

$$\max_{\phi,\theta} \mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})]\right] \leq -H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A)\ , \tag{7.21}$$

even if we assume infinite training data and an optimal variational approximation, i.e., $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A) = p(\mathbf{z} \mid \mathbf{x}_A)$ and $p_\theta(\bar{\mathbf{x}} \mid \mathbf{z}) = p(\bar{\mathbf{x}} \mid \mathbf{z})$ for $\mathbf{z} \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$.

We can repeat the same argument for any $A \in \mathcal{S}$. Consequently, it holds that

$$\max_{\phi,\theta} \sum_{A \in \mathcal{S}} \omega_A\, \mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})]\right] \leq \underbrace{-\sum_{A \in \mathcal{S}} \omega_A\, H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A)}_{\Delta(\bar{\mathbf{x}}, \mathcal{S})}\ . \tag{7.22}$$

Therefore, for the objective in Equation (7.9), we have

$$\max_{\phi,\theta} \sum_{A \in \mathcal{S}} \omega_A\, \mathbb{E}_{p(\bar{\mathbf{x}})}\left[\mathbb{E}_{q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)}[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})]\right] - \mathbb{E}_{p(\bar{\mathbf{x}})}\left[D_{\mathrm{KL}}(q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z}))\right] \tag{7.23}$$

$$\leq -\Delta(\bar{\mathbf{x}}, \mathcal{S}) - \mathbb{E}_{p(\bar{\mathbf{x}})}\left[D_{\mathrm{KL}}(q_\phi^{\mathcal{S}}(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z}))\right] \tag{7.24}$$

$$\leq -\Delta(\bar{\mathbf{x}}, \mathcal{S})\ , \tag{7.25}$$

because the KL-divergence is always non-negative.

Thus, for the approximation of the expected log-evidence (Equation 7.7) through the objective in Equation (7.9), it holds that

$$\mathbb{E}_{p(\bar{\mathbf{x}})}[\log p(\bar{\mathbf{x}})] \geq \mathbb{E}_{p(\bar{\mathbf{x}})}[\mathcal{L}_{\mathrm{ELBO}}^{\mathcal{S}}(\bar{\mathbf{x}}; \phi, \theta)] + \Delta(\bar{\mathbf{x}}, \mathcal{S}) \tag{7.26}$$

[22] In general, the conditional mutual information can be defined in terms of conditional entropies (e.g., see [CT12, p. 23]); specifically, $I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{z}_A \mid \mathbf{x}_A) = H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) - H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{z}_A, \mathbf{x}_A)$.

for any value of ϕ and θ.

The exact value of $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ depends on the subsets and weights in $\mathcal{S}$ as well as on the amount of modality-specific variability in the data. In particular, $\Delta(\bar{\mathbf{x}}, \mathcal{S}) > 0$, if there is any subset $A \in \mathcal{S}$ with $\omega_A > 0$ for which $H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) > 0$.

□

Implications Theorem 1 formalizes the rationale that the prediction across modalities cannot recover information that is specific to the target modalities that are unobserved due to modality sub-sampling. The conditional entropy $H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A)$ measures the amount of information in one subset of random vectors $\mathbf{x}_{\{1,\dots,M\}\setminus A}$ that is not shared with another subset $\mathbf{x}_A$. For the approximation of the joint distribution (i.e., the expected log-evidence), the sub-sampling of modalities produces an irreducible error $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ that is a weighted average of conditional entropies $H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A)$ of the unobserved modalities $\mathbf{x}_{\{1,\dots,M\}\setminus A}$ given the observed modalities $\mathbf{x}_A$. Hence, $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ describes the modality-specific information that cannot be recovered by crossmodal prediction, averaged over all subsets of modalities.

Theorem 1 applies to the MVAE, MMVAE, and MoPoE-VAE, since each of these models belongs to the class of mixture-based multimodal VAEs (see Section 5.2.3). However, $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ can vary significantly between different models, depending on the mixture distribution defined by the respective model and on the amount of modality-specific variability in the data. Next, we discuss the implications for specific models before we corroborate our results with experiments in Section 7.3.

7.2.4 Implications for Specific Models

First, we consider the case of no modality sub-sampling, for which it is easy to show that the discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ vanishes.

Corollary 1. *Without modality sub-sampling,* $\Delta(\bar{\mathbf{x}}, \mathcal{S}) = 0$ *.*

Proof. Without modality sub-sampling, $\mathcal{S}$ is comprised of only one subset, the complete set of modalities $\{1, \dots, M\}$, and therefore $\mathbf{x}_A = \bar{\mathbf{x}}$ and $\mathbf{x}_{\{1,\dots,M\}\setminus A} = \emptyset$. It follows that $\Delta(\bar{\mathbf{x}}, \mathcal{S}) = H(\bar{\mathbf{x}}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) = H(\emptyset \mid \bar{\mathbf{x}}) = 0$, since the (conditional) entropy of the empty set is zero. □

The result from Corollary 1 applies to the MVAE without ELBO sub-sampling and suggests that this model achieves a tighter approximation of the log-evidence and hence a better generative quality compared to mixture-based multimodal VAEs that sub-sample modalities. Note that this does not imply that a model without modality sub-sampling is superior to one that uses sub-sampling; for instance, there can be an inductive bias that favors sub-sampling despite the approximation error it incurs. Specifically, Corollary 1 does not imply that the variational approximation is tight for the MVAE; for instance, the model can be underparameterized or simply misspecified due to simplifying assumptions, such as the PoE-factorization [c.f., KGS19].

In Corollary 2, we analyze how an additional modality would affect the discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$. It shows that for the MMVAE and MoPoE-VAE the discrepancy increases if the new modality is sufficiently diverse in the sense of Equation (7.28).

Corollary 2. *Let $\mathbf{x}_{M+1}$ be a random vector that describes an additional modality and let $\bar{\mathbf{x}}^+ := \bar{\mathbf{x}} \cup \{\mathbf{x}_{M+1}\}$ denote the extended set of modalities. Further, let $\mathcal{S}^+$ denote the model-specific set of subsets of modalities and corresponding mixture coefficients for $\bar{\mathbf{x}}^+$.*

For the MMVAE and MoPoE-VAE, the discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ increases, i.e.,

$$\Delta(\bar{\mathbf{x}}^+, \mathcal{S}^+) > \Delta(\bar{\mathbf{x}}, \mathcal{S}) \,, \tag{7.27}$$

if $\mathbf{x}_{M+1}$ is sufficiently diverse in the following sense:

$$\begin{aligned}\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{x}_{M+1} \mid \mathbf{x}_A) \; < \; & \frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} H(\mathbf{x}_A \mid \mathbf{x}_{M+1}) \\ & + \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \,.\end{aligned} \tag{7.28}$$

A proof is provided in Appendix A.3.

The result from Corollary 2 suggests an increased discrepancy (and hence, a decline of generative quality) when we increase the number of modalities for the MMVAE and MoPoE-VAE. Intuitively, a new modality is sufficiently diverse, if it does *not* add too much redundant information with respect to the existing modalities. When there is a lot of redundant information, $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ can decrease given an additional modality, but it never vanishes. Only if there is zero modality-specific information for each modality, we have $\Delta(\bar{\mathbf{x}}, \mathcal{S}) = 0$ for the MMVAE and MoPoE-VAE. This would require each pair of modalities to be fully redundant, which is violated in most multimodal datasets as $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ typically represents a significant part of the total variability.

(a) PolyMNIST (5 modalities)

(b) Translated-PolyMNIST (5 modalities)

(c) CUB Image-Captions (2 modalities)

Figure 7.2: The three datasets considered in this chapter. The two PolyMNIST datasets are conceptually similar in that a digit label is shared between the observations of different modalities. The CUB Image-Captions dataset provides a more realistic application with image-text pairs, i.e., natural images of birds and corresponding captions.

In summary, Theorem 1 formalizes how the family of mixture-based multimodal VAEs is fundamentally limited for the task of approximating the joint distribution, and Corollaries 1 and 2 connect this result to specific models—namely the MVAE, MMVAE, and the proposed MoPoE-VAE. We now turn to the experiments, where we present empirical support for our theoretical results.

7.3 Experiments

Figure 7.2 presents the three considered datasets. PolyMNIST is comprised of sets of images where a digit label is shared between five synthetic modalities. It is a conceptually simple dataset that we use to conduct experiments in a controlled setup. Additionally, we introduce the Translated-PolyMNIST dataset that adds a simple transformation—namely, the downscaling and random translation (i.e., spatial positioning) of each digit—to demonstrate the limitations of existing methods when shared information cannot be predicted in expectation across modalities on the level of observations.[23] Finally, we use CUB Image-Captions to validate the limitations on a real-world dataset comprised of image-caption pairs. Notably, we use CUB with *real images* and not the simplified version

[23]To be precise, we use Translated-PolyMNIST with a downscaling factor of 0.7. While we used a downscaling factor of 0.5 for the results in our original publication [Dau+22], in this chapter, we reproduce the results with a less severe downscaling to showcase the robustness of our findings.

of the dataset based on precomputed ResNet-features that was used in previous works [e.g., Shi+19; Shi+21]. For a more detailed description of the considered datasets, see Chapter 4.

In Section 7.3.1, we showcase the gap in generative quality for mixture-based multimodal VAEs in a large-scale empirical study. In Section 7.3.2, we demonstrate that generative coherence cannot be guaranteed for more complex datasets where the information shared between modalities cannot be predicted in expectation across modalities on the level of observations. Finally, we evaluate the quality of the learned representation in Section 7.3.3.

Implementation details For all experiments in this chapter, we average the results over three seeds and include standard deviations. We use the same implementation as for our experiments in Chapter 5 with one notable difference, which is the network architecture. In this chapter, we employ the ResNet architecture [He+16], because we found that the previously used convolutional networks underperformed on the more complex datasets considered in this chapter.[24] In total, we trained more than 400 models using approximately 1.5 GPU years of compute on a single NVIDIA GeForce RTX 2080 Ti GPU.

All models were trained using the Adam optimizer [KB15] with learning rate 5e-4, batch size 256, and using 500, 1000, and 150 epochs on PolyMNIST, Translated-PolyMNIST, and CUB Image-Captions respectively. Similar to previous work, we use Gaussian priors and a latent space with 512 dimensions for PolyMNIST and 64 dimensions for CUB Image-Captions. For a fair comparison, we reduce the latent dimensionality of unimodal VAEs proportionally (w.r.t. the number of modalities) to control for capacity. For image modalities, we use Laplace distributions to estimate the log-likelihood, and for captions, we employ one-hot categorical distributions. For the β-ablations[25], we use $\beta \in \{3\text{e-}4, 3\text{e-}3, 3\text{e-}1, 1, 3, 9\}$ and, in addition, 32 for CUB Image-Captions. Note that the MVAE could not be trained with values of $\beta > 3$ on the PolyMNIST dataset due to numerical instabilities.

7.3.1 The Generative Quality Gap

First, we compare the generative quality of multimodal VAEs across different methods and to unimodal VAEs. We measure the generative quality in terms of the Fréchet inception

[24] To ensure consistency, in this chapter we use ResNets for PolyMNIST as well. We verified that there is no significant difference compared to our results from Chapter 5 when we use ResNets instead.

[25] The regularization coefficient β weights the KL-divergence term of the multimodal ELBO and is an important hyperparameter for VAEs [Hig+17].

distance (FID, [Heu+17], a standard metric used for evaluating the quality of generated images (see Section 2.5.2). In addition, in Appendix A.5 we provide log-likelihood values, as well as qualitative results for all modalities.

In the context of our theoretical results, we hypothesize that a higher discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ is associated with a lower generative quality (i.e., higher FID) because it implies a larger gap for the approximation of the expected log-evidence. For the interpretation, we assume that the variational approximation[26] is sufficiently good, such that the observed difference in generative quality between models can be attributed to the discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$.

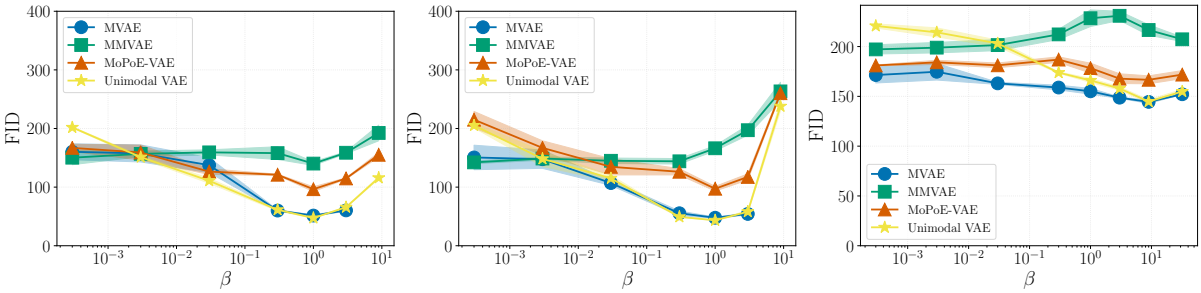
In Figure 7.3, we evaluate the generative quality for each dataset and all considered methods across a wide range of β values. Across different methods, we can compare models with the lowest FID value respectively. We find that mixture-based approaches that sub-sample modalities (MMVAE and MoPoE-VAE) exhibit a pronounced gap in generative quality compared to unimodal VAEs. When we compare the best models, we observe a gap of more than 60 points on both PolyMNIST and Translated-PolyMNIST, and about 30 points on CUB.[27] In contrast, the MVAE achieves a similar generative quality as unimodal VAEs. Additionally, Figure 7.4 examines how the generative quality is affected when we vary the number of modalities used for training. Notably, for the MMVAE and MoPoE-VAE, the generative quality deteriorates continuously with the number of modalities, whereas the performance of the MVAE remains stable.

7.3.2 Lack of Generative Coherence

Besides generative quality, another desired criterion for multimodal generative models is *generative coherence*, i.e., the ability to generate semantically related examples across modalities. Specifically, we measure the leave-one-out generative coherence (see Section 2.5.2), which means that the input to each model consists of all modalities except the one that is being conditionally generated. On CUB Image-Captions, we resort to a qualitative evaluation of coherence because there are no ground truth labels that indicate what is shared between modalities.

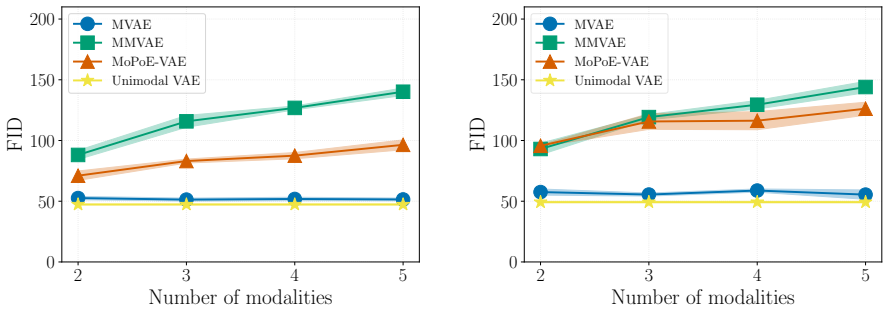
[26] In particular, we assume that $p(\mathbf{z} \mid \mathbf{x}_A)$ is approximated well by the variational posterior $q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$ and likewise for the approximation of $p(\bar{\mathbf{x}} \mid \mathbf{z})$ by $p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})$ for $\mathbf{z} \sim q_{\phi_A}(\mathbf{z} \mid \mathbf{x}_A)$.

[27] In Appendix A.5, we include qualitative results that verify that the gap in generative quality is clearly visible in the generated samples and that it applies not only to images but also to generated captions.



(a) PolyMNIST (b) Translated-PolyMNIST (c) CUB Image-Captions

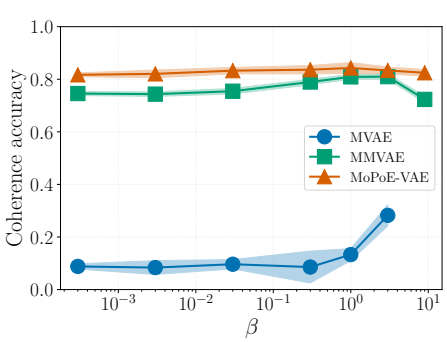
Figure 7.3: Generative quality for one modality over a range of β values. Lower FID values indicate a better generative quality. Results for other modalities are shown in Figure A.1.



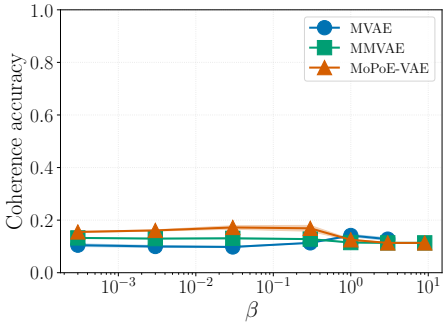
(a) PolyMNIST (b) Translated-PolyMNIST

Figure 7.4: Generative quality as a function of the number of modalities. For each model, we show the results for the same modality and therefore all values are on the same scale. For the unimodal VAE, which uses only a single modality, the average and standard deviation are plotted as a constant. All models were trained using $\beta = 1$.

In terms of generative coherence, Figure 7.5 reveals that the promising results from previous work do not necessarily apply to more complex datasets. As a baseline, for PolyMNIST (Figure 7.5a) we replicate the coherence results from Chapter 5 for a range of β values. Consistent with previous work [Shi+19; Shi+21] and our results from Chapter 5, we find that the MMVAE and MoPoE-VAE exhibit a superior coherence compared to the MVAE. Though, it was not apparent from previous work that MVAE's coherence can improve significantly with increasing β values, which can be of independent interest for future work. On Translated-PolyMNIST (Figure 7.5b), the stark decline of the performance across all models makes it evident that coherence cannot be guaranteed when shared information



(a) PolyMNIST



(b) Translated-PolyMNIST

MVAE, $\beta = 9$

MMVAE, $\beta = 9$

MoPoE-VAE, $\beta = 9$

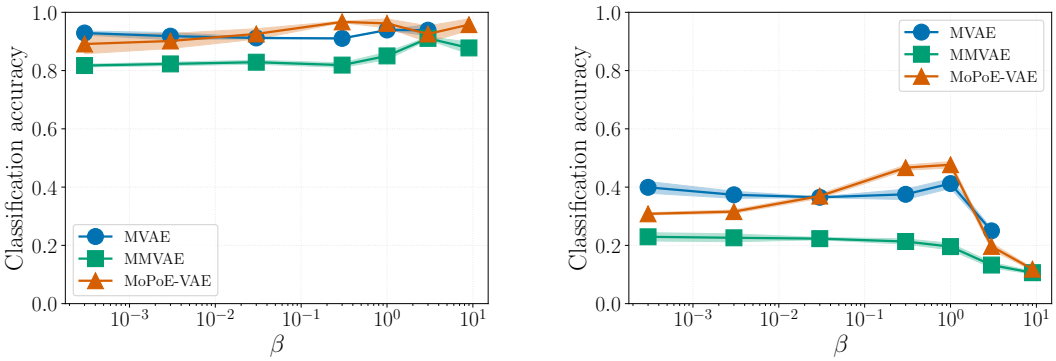
(c) CUB Image-Captions

Figure 7.5: Evaluation of generative coherence. For PolyMNIST (Figures 7.5a and 7.5b), we plot the average leave-one-out coherence. For CUB Image-Captions (Figure 7.5c), we show qualitative results for the conditional generation of images given captions.

cannot be predicted in expectation across modalities on the level of observations. In Appendix A.5, we include qualitative results showing that not a single multimodal VAE generates coherent examples across modalities. We observe similar results on CUB Image-Captions (Figure 7.5c), where the qualitative results for conditional generation verify that none of the existing approaches generates images that are both of sufficiently high quality and coherent with respect to the given caption. Overall, the negative results on Translated-PolyMNIST and CUB Image-Captions showcase the limitations of existing approaches when applied to more complex datasets than those used in previous benchmarks.

7.3.3 Quality of the Learned Representations

Like in the previous chapters, we use linear probing (c.f., Section 2.5) to assess the quality of the learned representations and to estimate how well the encoded information decomposes



(a) PolyMNIST (b) Translated-PolyMNIST

Figure 7.6: Linear probing of the learned representations. For each model, logistic regression models were trained to predict the shared information (i.e., the digit label) based on the embeddings from 500 randomly sampled training examples.

into shared and modality-specific features.

In Figure 7.6, we show the results of linear probing with respect to the shared information (i.e., the digit label). Similar to the decline in generative coherence (Section 7.3.2), we find that also the quality of the learned representations deteriorates when we transition from PolyMNIST to Translated-PolyMNIST. While a low classification accuracy does not imply that there is no digit information encoded in the latent representation (after all, digits show up in self-reconstructions), the results indicate that a *linear* classifier cannot extract the digit information particularly well. Thus, shared and modality-specific information is not encoded in a linearly separable format.

7.4 Discussion

First, we assess our findings from the experiments in relation to the theoretical results from Section 7.2. The experiments reveal a gap in generative quality for mixture-based multimodal VAEs on both synthetic and real-world datasets (Section 7.3.1). Specifically, the results from Figure 7.3 and Figure 7.4 showcase the gap in generative quality for the MMVAE and MoPoE-VAE compared to unimodal VAEs. For the MVAE, we observe no gap compared to unimodal VAEs, which is in line with Corollary 1. Moreover, the gap increases disproportionally with each additional modality for both the MMVAE and

MoPoE-VAE (c.f., Corollary 2). Thus, the experiments lend support to the theoretical results presented in Section 7.2. Though, it should be noted that our experiments focused on image data. For future work, it would be interesting to generate simulated data for which the discrepancy from Theorem 1 can be measured exactly and where the variational approximation is controlled for.

Beyond the theoretical results, our experiments also reveal that none of the existing method (including those without modality sub-sampling) fulfills all desired criteria of an effective multimodal generative model. This was demonstrated by the lack of generative coherence (Section 7.3.2) and insufficient disentanglement of shared and modality-specific information (Section 7.3.3). More broadly, our results showcase the limitations of VAE-based approaches for modeling weakly supervised data in the presence of modality-specific information, and in particular when shared information cannot be predicted in expectation across modalities on the level of observations. The Translated-PolyMNIST dataset demonstrates this problem in a controlled setup, while the results on CUB Image-Captions verify that similar issues can be expected in real-world applications.

Model selection and generalization Our results raise fundamental questions regarding model selection and generalization for multimodal VAEs. Since the results for generative quality and generative coherence can differ, this raises the question which metric to use for model selection. On the one hand, our experiments reveal that FIDs and log-likelihoods do not reflect the problem of lacking coherence (e.g., consider the performance of the MVAE). On the other hand, without access to ground truth labels for what is shared between modalities, coherence cannot be evaluated. Consequently, it can be particularly difficult to use coherence for model selection in applications with less interpretable types of modalities, such as DNA sequences. Hence, for future work it could be interesting to design alternative metrics for generative coherence that can be applied when shared information is not annotated. Moreover, for future work it might be worthwhile to consider model selection metrics for out-of-distribution generalization (e.g., [Mon+21]) in addition to generative quality and coherence.

Limitations The limitations and tradeoffs presented in this chapter apply to a large family of VAEs, but not necessarily to other types of generative models. For example, implicit likelihood models (e.g., generative adversarial networks [Goo+14]) might not be subject to the same limitations. Similar to previous work [WG18; Shi+19], we only

considered models with simple priors, such as Gauss and Laplace distributions with independent dimensions. Further, we have not included models with modality-specific latent spaces that we discussed in Chapter 6. While modality-specific latent spaces can significantly enhance the generative quality for each of the considered models, more work is required to overcome the lack of coherence on datasets where shared information cannot be predicted in expectation across modalities on the level of observations.

7.5 Summary

In this chapter, we have identified, formalized, and demonstrated several limitations of multimodal VAEs. Our analysis revealed an irreducible gap in the approximation of the joint distribution and empirically investigated the gap in generative quality between unimodal and mixture-based multimodal VAEs across multiple datasets. We offered the explanation that the sub-sampling of modalities enforces an undesirable upper bound on the multimodal ELBO and therefore limits the generative quality of the respective models. While the sub-sampling of modalities allows multimodal VAEs to learn the inference networks for different subsets of modalities efficiently, it leads to a tradeoff in terms of generative quality. Finally, we demonstrated an additional failure case, showing that all models exhibit a lack of generative coherence when shared information cannot be predicted in expectation across modalities on the level of observations.

In the scope of the thesis, this chapter discussed the limitations of multimodal VAEs, including our own work presented in previous chapters. Specifically, it shows that the utility of mixture-based multimodal VAEs is restricted to settings where shared information can be predicted in expectation across modalities on the level of observations. With respect to the research questions, the results of this and the previous two chapters suggest that multimodal VAEs can be effective for the encoding of shared information (Question 1). However, this chapter demonstrates that in applications on more complex multimodal datasets, shared information might be encoded in a nonlinear format, which would imply a lack of disentanglement of shared and modality-specific information (Question 2). Moreover, in applications where shared information cannot be predicted in expectation across modalities, we would expect a lack of coherence for the generation of missing modalities (Question 3).

To address these limitations, in the following two chapters we instead consider contrastive learning as a discriminative approach for multimodal representation learning. In Chapter 8,

we demonstrate that contrastive learning provides a viable alternative for the estimation of information that is shared between modalities—even if the information cannot be predicted in expectation across modalities on the level of observations. In Chapter 9, we develop a hybrid method for the disentanglement of shared and modality-specific information and present promising results on the Translated-PolyMNIST dataset.

8

Multimodal Contrastive Learning

Thus far, we addressed the problem of multimodal representation learning with generative models, specifically using multimodal VAEs. In the preceding chapter, we analyzed the limitations of the generative capabilities of multimodal VAEs and discussed the ramifications for the learned representations. In this chapter, we adopt a different strategy and focus on *contrastive learning* as a discriminative approach to multimodal representation learning. Specifically, we examine how contrastive learning can be used to identify latent factors of variation shared between modalities up to acceptable ambiguities. In the following chapter, we build on these results to demonstrate how contrastive learning can be used to address some of the limitations of multimodal VAEs.

In Section 8.1, we motivate the use of contrastive learning for multimodal representation learning and cover relevant background. In Section 8.2, we revisit our problem formulation from Chapter 2 and consider additional assumptions to describe a realistic yet tractable setup that can be analyzed theoretically. For this setup, we derive an identifiability result, which shows that, asymptotically, contrastive learning can recover the shared latent factors that are invariant across modalities up to a block-wise indeterminacy (Section 8.3). Then, we verify the identifiability result with numerical simulations and corroborate our findings on a complex dataset of image-text pairs (Section 8.4). Finally, we discuss potential limitations and opportunities for future work (Section 8.5).

8.1 Motivation and Background

Essentially, contrastive learning uses weak supervision in the form of corresponding observations of different views to learn an encoder that maps similar inputs closer to each other and dissimilar inputs further apart (c.f., Section 3.4) Thus, it matches the setup of multimodal learning under weak supervision, where we consider corresponding observations of different modalities. In practice, contrastive learning has become a cornerstone in multimodal learning, as for example witnessed by the contribution of CLIP [Rad+21] to recent advancements in text-to-image generation [Ram+21; Ram+22; Rom+22; Sah+22]. However, despite its empirical success, the effectiveness of contrastive learning is not sufficiently well understood in theory.

Following a line of recent work, we examine contrastive learning through the lens of independent component analysis (ICA), which offers an explanation of the effectiveness of contrastive learning in terms of the identification of latent variables shared between views [Gre+19; Zim+21; von+21]. To provide a better understanding of the effectiveness of contrastive learning for *multimodal* representation learning, we investigate under which assumptions contrastive learning can be used to identify shared latent factors up to acceptable ambiguities in the context of our problem formulation from Chapter 2.

Based on the formulation from Chapter 2, we consider a generative process with independent mechanisms that produce heterogeneous observations of different modalities. However, in this chapter, we consider additional assumptions to provide a formulation that can be analyzed theoretically with respect to the *identifiability* of latent variables. Specifically, we draw inspiration from previous work on multi-view nonlinear ICA [Gre+19] and consider additional dependencies between latent variables [von+21] to describe a more realistic setup beyond the classic assumptions of ICA.

Contributions Primarily, we show that contrastive learning can be used to identify shared latent factors that are invariant across modalities up to a block-wise indeterminacy. In the context of previous work, we generalize existing identifiability results [von+21; Lyu+22] to model real-world multimodal data. Based on our results, we provide a better understanding of the assumptions required to recover latent factors of variation in the context of multimodal learning under weak supervision. Moreover, within the scope of this thesis, we demonstrate that contrastive learning can transcend some of the limitations of multimodal VAEs described in Chapter 7. Specifically, we present promising results in

applications with a high degree of modality-specific variation where shared information cannot be predicted in expectations across modalities on the level of observations.

In summary, in this chapter, we study identifiability in the context of multimodal representation learning and focus on contrastive learning as a particular algorithm for which we derive an identifiability result. Next, we review relevant background on identifiability and contrastive learning and contextualize related work. Then, we present our problem formulation (Section 8.2) and identifiability result (Section 8.3). To verify the result, we conduct numerical simulations and experiments on a dataset of image-text pairs (Section 8.4). Finally, we discuss potential limitations and opportunities for future work (Section 8.5).

8.1.1 Identifiability of Latent Variables

Generally, identifiability poses the question whether specific parts of a generative process can be estimated from observations. In statistics, a model is identifiable if given an infinite number of observations it is theoretically possible to recover the underlying parameters of the model or values of the latent variables [LC06]. Thus, the question of identifiability lies at the heart of statistical inference and it applies to problems such as ICA, causal discovery, and inverse problems, among others.

In the framework of ICA, we consider the relation $\mathbf{x} = \mathbf{f}(\mathbf{z})$, where an observation $\mathbf{x}$ is generated from a mixing function $\mathbf{f}$ that is applied to a latent vector $\mathbf{z}$. The goal of ICA is to invert the mixing function in order to recover the latent variables, i.e., the individual components of the latent vector. In particular, ICA algorithms attempt to learn the inverse function given a dataset of observations, ideally using as few assumptions as possible. In the following, we focus on related work on the topic of multi-view nonlinear ICA, which is most relevant to the problem we address in this chapter. For more information on the general topic of ICA, see Section 3.2.

Multi-view nonlinear ICA Our problem formulation builds on the setup of multi-view nonlinear ICA and related approaches that we discuss in Section 3.2.1. Crucially, a setup with multiple views or modalities offers a solution to the challenging problem of *nonlinear* ICA, where it is theoretically impossible to recover the latent variables without further assumptions [HP99]. Intuitively, a second view can resolve ambiguity introduced by a nonlinear mixing function if both views contain a shared signal but are otherwise

sufficiently distinct [Gre+19]. This property can be leveraged to identify latent variables shared between observations of different views or modalities.[28]

In the context of multi-view nonlinear ICA, previous work shows that the latent variables can be identified up to a component-wise indeterminacy in a setting with mutually independent variables [Gre+19; Loc+20b] or up to block-wise indeterminacies in the case of independent groups of shared and view-specific variables [LF20; Lyu+22]. Beyond the strict assumption of independent (groups of) variables, a related line of work considers shared variables that are *invariant* across views or domains, showing that these variables can be identified up to a block-wise indeterminacy even when there are additional dependencies between latent variables [von+21; Kon+22].

We advance in the same direction but with a focus on multimodal learning. To describe real-world multimodal data, we consider a generative process with heterogeneous modalities and optional statistical and causal dependencies between latent variables. Therefore, we combine formulations from previous works (specifically, from von Kügelgen et al. [von+21] and Lyu et al. [Lyu+22]) and relax their assumptions to model the properties of real-world multimodal data. Consequently, we generalize existing identifiability results and show that shared latent factors that are invariant across modalities can be identified up to a block-wise indeterminacy in a novel setting with modality-specific mixing functions, modality-specific latent variables, and additional dependencies between latent variables.

8.1.2 Multimodal Contrastive Learning

We introduced contrastive learning in Section 3.4 and described the widely-used InfoNCE objective [GH10; OLV18] in that context. In the following, we revisit the symmetrized version of the objective that is used for all experiments in this chapter. Additionally, we consider an asymptotic form of the objective [WI20; von+21], which we use to derive an identifiability result in Section 8.3.

[28] While the terms "view" and "modality" have been used interchangeably in previous works, we distinguish them to differentiate between the multi-view setting with one generative mechanism (e.g., multiple cameras of the same type) and the multimodal setting that is characterized by *distinct* mechanisms (e.g., cameras and microphones). We address the distinction in Section 2.2, in the context of Assumption 2.

Objective function Recall from Equation (3.57) that the InfoNCE objective is given by

$$\mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi_1, \phi_2) = -\sum_{k=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{g}_{\phi_1}(\mathbf{x}_1^k), \mathbf{g}_{\phi_2}(\mathbf{x}_2^k))/\tau\}}{\sum_{l=1}^{K} \exp\{\text{sim}(\mathbf{g}_{\phi_1}(\mathbf{x}_1^k), \mathbf{g}_{\phi_2}(\mathbf{x}_2^l))/\tau\}}\,, \tag{8.1}$$

where the set of observations $\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K$ consists of data pairs sampled from the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2)$, the functions $\mathbf{g}_{\phi_1}$ and $\mathbf{g}_{\phi_2}$ are encoders parameterized by ϕ_1 and ϕ_2, and the hyperparameters are the temperature τ and similarity metric $\text{sim}(\cdot, \cdot)$.

For contrastive learning with multimodal data, a rational choice is to use distinct encoders $\mathbf{g}_{\phi_1} \neq \mathbf{g}_{\phi_2}$ and a symmetrized version of the objective (c.f., Section 3.4.4). Following the same principle, in this chapter, we consider the symmetrized InfoNCE objective (Equation 3.58), which we specify in terms of the joint distribution $p(\mathbf{x}_1, \mathbf{x}_2)$ as

$$\mathbb{E}_{\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K \sim p(\mathbf{x}_1, \mathbf{x}_2)} \Big[\tfrac{1}{2}\, \mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_1^k, \mathbf{x}_2^k\}_{k=1}^K; \phi_1, \phi_2) + \tfrac{1}{2}\, \mathcal{L}_{\text{InfoNCE}}(\{\mathbf{x}_2^k, \mathbf{x}_1^k\}_{k=1}^K; \phi_2, \phi_1) \Big]\,, \tag{8.2}$$

where the integer K becomes an additional hyperparameter that controls the number of negative pairs used for contrasting. The encoders are trained by minimizing Equation (8.2) with respect to the parameters ϕ_1 and ϕ_2. In practice, we sample data pairs from a finite dataset $\mathcal{D} = \{(\mathbf{x}_1^{(n)}, \mathbf{x}_2^{(n)})\}_{n=1}^N$ of size $N \gg K$, but in numerical simulations we can draw samples from $p(\mathbf{x}_1, \mathbf{x}_2)$ directly.

Asymptotic behavior Asymptotically, the minimization of the InfoNCE objective (Equation 8.1) with respect to ϕ_1 and ϕ_2 can be interpreted as the alignment of positive pairs (numerator) with approximate entropy regularization (denominator), which produces encoders that are aligned and map to a uniform distribution (c.f., Section 3.4.3). Formally, when instantiating the symmetrized InfoNCE objective (Equation 8.2) with $\tau = 1$ and $\text{sim}(a, b) = -(a - b)^2$, it asymptotically behaves like

$$\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_2) \sim p(\mathbf{x}_1, \mathbf{x}_2)} \Big[\|\mathbf{g}_{\phi_1}(\mathbf{x}_1) - \mathbf{g}_{\phi_2}(\mathbf{x}_2)\|_2 \Big] - \tfrac{1}{2} \Big(H(\mathbf{g}_{\phi_1}(\mathbf{x}_1)) + H(\mathbf{g}_{\phi_2}(\mathbf{x}_2)) \Big) \tag{8.3}$$

for $K \to \infty$ [von+21; WI20]. In this sense, the objective can be interpreted as the alignment of positive pairs subject to the maximization of the entropy (i.e., the uniformity) of the learned representation, if we minimize Equation (8.3) with respect to ϕ_1 and ϕ_2.

For the symmetrized InfoNCE objective, the approximation of the alignment term is identical for both loss terms in Equation (8.2), since the similarity measure is symmetric. Further, each entropy term is approximated via the denominator of the respective loss term in Equation (8.2), which can be viewed as a nonparametric entropy estimator [WI20].

For the experiments, we use the objective in Equation (8.2) with large K to approximately match the form of Equation (8.3). For the theoretical analysis, we use the asymptotic form (Equation 8.3) to derive an identifiability result.

8.2 Problem Formulation

In the following, we define the generative process (Section 8.2.1) and then specify our technical assumptions on the relation between modalities (Section 8.2.2). Compared to our problem formulation from Chapter 2, in this chapter, we focus on *pairs* of modalities. Further, we introduce an alternative notion of content invariance in terms of conditional distributions and consider additional dependencies between subsets of latent variables. Given these additional assumptions, we derive an identifiability result in Section 8.3.

8.2.1 Multimodal Generative Process

Let $\mathbf{z}$ be a continuous random vector that takes values in $\mathcal{Z} \subseteq \mathbb{R}^n$ with density $p(\mathbf{z})$. In this chapter, we assume that $\mathbf{z}$ can be uniquely partitioned as

$$\mathbf{z} = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1, \mathbf{m}_2) \ , \tag{8.4}$$

which is comprised of

(i) an invariant part $\mathbf{c}$, which is shared across modalities, and which we refer to as *content*;

(ii) a variable part $\mathbf{s}$, which may change across modalities, and which we refer to as *style*;

(iii) two modality-specific parts, $\mathbf{m}_1$ and $\mathbf{m}_2$, each unique to the respective modality.

Let $\mathbf{z}_1$ and $\mathbf{z}_2$ be two random vectors that take values in $\mathcal{Z}_1 \subset \mathcal{Z}$ and $\mathcal{Z}_2 \subset \mathcal{Z}$ respectively and let $p(\mathbf{z}_1, \mathbf{z}_2)$ be their joint density. Further, let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two random vectors that represent different modalities that take values in $\mathcal{X}_1 \subseteq \mathbb{R}^{d_1}$ and $\mathcal{X}_2 \subseteq \mathbb{R}^{d_2}$ respectively. We define the generative process in terms of the latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ as follows:

$$\mathbf{z}_1 \sim p(\mathbf{z}_1), \quad \mathbf{z}_2 \sim p(\mathbf{z}_2 \mid \mathbf{z}_1), \quad \mathbf{x}_1 = \mathbf{f}_1(\mathbf{z}_1), \quad \mathbf{x}_2 = \mathbf{f}_2(\mathbf{z}_2), \tag{8.5}$$

where $\mathbf{f}_1 : \mathcal{Z}_1 \to \mathcal{X}_1$ and $\mathbf{f}_2 : \mathcal{Z}_2 \to \mathcal{X}_2$ are two smooth and invertible mixing functions with smooth inverse (i.e., diffeomorphisms). Generally, we assume that observations from different modalities are generated by distinct mechanisms $\mathbf{f}_1 \neq \mathbf{f}_2$ (c.f., Assumption 2).

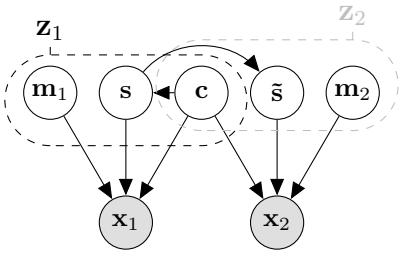


Figure 8.1: Illustration of the generative process. Latent variables are denoted by clear nodes and observations by shaded nodes. We partition the latent space into $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 = (\tilde{\mathbf{c}}, \tilde{\mathbf{s}}, \mathbf{m}_2)$, where $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 4) and hence we only visualize $\mathbf{c}$. Further, $\tilde{\mathbf{s}}$ is a perturbed version of $\mathbf{s}$ (Assumption 5) and $\mathbf{m}_1$, $\mathbf{m}_2$ are modality-specific variables. The observations $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated by two distinct mixing functions $\mathbf{f}_1 \neq \mathbf{f}_2$, which are applied to the subsets of latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$ respectively.

8.2.2 Relation between Modalities

Based on the partitioning in Equation (8.4), let $\mathbf{z}_1 = (\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{z}_2 = (\tilde{\mathbf{c}}, \tilde{\mathbf{s}}, \mathbf{m}_2)$, where $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere and $\tilde{\mathbf{s}}$ is generated by perturbations that we specify in Section 8.2.2. We assume that $p(\mathbf{z}_1, \mathbf{z}_2)$ is a smooth density that factorizes as

$$p(\mathbf{z}_1, \mathbf{z}_2) = p(\mathbf{c})p(\mathbf{s} \mid \mathbf{c})p(\tilde{\mathbf{s}} \mid \mathbf{s})p(\mathbf{m}_1)p(\mathbf{m}_2) \ , \tag{8.6}$$

as illustrated in Figure 8.1.

Unlike in the classic setting of ICA, we do not assume independent components. Instead, we draw inspiration from previous work [von+21] and consider statistical dependencies within blocks of variables (e.g., between components of $\mathbf{c}$) as well as causal dependencies from $\mathbf{c}$ to $\mathbf{s}$. Following previous work, we assume that content is invariant, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere (Assumption 4), and that $\tilde{\mathbf{s}}$ is a perturbed version of $\mathbf{s}$ (Assumption 5). However, our formulation considers modality-specific mixing functions $\mathbf{f}_1 \neq \mathbf{f}_2$ and modality-specific latent variables $\mathbf{m}_1$ and $\mathbf{m}_2$ in order to model real-world multimodal data (see Section 8.2.3).

Conditional distributions Next, we define the conditional distributions $p(\mathbf{z}_2 \mid \mathbf{z}_1)$ and $p(\tilde{\mathbf{s}} \mid \mathbf{s})$, which describe the dependence between modalities. First, we formalize the concept of content invariance[29] with respect to the conditional distribution $p(\mathbf{z}_2 \mid \mathbf{z}_1)$:

[29] In Chapter 2, we introduced a notion of content invariance in terms of functions that are invertible with respect to a subset of their arguments (Assumption 3). Instead, in this chapter we define content invariance with respect to the conditional distribution $p(\mathbf{z}_2 \mid \mathbf{z}_1)$.

Assumption 4 (Content invariance w.r.t. $p(\mathbf{z}_2 \mid \mathbf{z}_1)$)**.** *The conditional density $p(\mathbf{z}_2 \mid \mathbf{z}_1)$ over $\mathcal{Z}_2 \times \mathcal{Z}_1$ takes the form*

$$p(\mathbf{z}_2 \mid \mathbf{z}_1) = \delta(\tilde{\mathbf{c}} - \mathbf{c}) p(\tilde{\mathbf{s}} \mid \mathbf{s}) p(\mathbf{m}_2) \tag{8.7}$$

for some continuous density $p(\tilde{\mathbf{s}} \mid \mathbf{s})$ on $\mathcal{S} \times \mathcal{S}$, where $\delta(\cdot)$ denotes the Dirac delta function used to express the assumption that $\tilde{\mathbf{c}} = \mathbf{c}$ almost everywhere.

To fully specify $p(\mathbf{z}_2 \mid \mathbf{z}_1)$, it remains to define the style changes, which are described by the conditional distribution $p(\tilde{\mathbf{s}} \mid \mathbf{s})$:

Assumption 5 (Style changes)**.** *Let $\mathcal{A}$ be the powerset of style components $\{1, \dots, n_s\}$ and let p_A be a distribution on $\mathcal{A}$. Then, the conditional density $p(\tilde{\mathbf{s}} \mid \mathbf{s})$ that specifies style changes is obtained by conditioning on a set A:*

$$p(\tilde{\mathbf{s}} \mid \mathbf{s}) = \sum_{A \in \mathcal{A}} p_A(A) \Big(\delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) p(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A) \Big) \tag{8.8}$$

where $p(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A)$ is a continuous density on $\mathcal{S}_A \times \mathcal{S}_A$, $\mathcal{S}_A \subseteq \mathcal{S}$ denotes the subspace of changing style variables specified by A, and $A^c = \{1, \dots, n_s\} \backslash A$ denotes the complement of A. Further, for any style variable $l \in \{1, \dots, n_s\}$, there exists a set $A \subseteq \{1, \dots, n_s\}$ with $l \in A$, s.t.

(i) $p_A(A) > 0$,

(ii) $p(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A)$ *is smooth w.r.t. both* $\mathbf{s}_A$ *and* $\tilde{\mathbf{s}}_A$*, and*

(iii) for any $\mathbf{s}_A$, $p(\tilde{\mathbf{s}}_A \mid \mathbf{s}_A) > 0$, *in some open non-empty subset containing* $\mathbf{s}_A$.

The rationale behind Assumption 5 is to describe a stochastic relation between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ in general terms. For example, one could view $\tilde{\mathbf{s}}$ to be a noisy version of $\mathbf{s}$ or as the result of an intervention on $\mathbf{s}$ (e.g., data augmentation). Further, note that the asymmetry between $\mathbf{s}$ and $\tilde{\mathbf{s}}$ (or between $\mathbf{z}_1$ and $\mathbf{z}_2$ for that matter) is not strictly required. It merely simplifies the notation and ensures consistency with previous work [von+21]. Instead, one could model *both* $\mathbf{z}_1$ and $\mathbf{z}_2$ via perturbations of $\mathbf{z}$, as described in Appendix A.7.

Intuitively, to generate $\tilde{\mathbf{s}} \sim p(\tilde{\mathbf{s}} \mid \mathbf{s})$, we independently flip a biased coin for each component in $\mathbf{s}$ to select a subset of style features, which are *jointly* perturbed to obtain $\tilde{\mathbf{s}}$. Condition *(i)* ensures that every style component has a positive probability to be perturbed,[30] while *(ii)* and *(iii)* are technical smoothness conditions that will be used for the proof of Theorem 2.

[30] If a style variable would be perturbed with zero probability, it would be a content variable.

8.2.3 Interpretation of the Model

Our formulation is designed to capture the complexities of real-world multimodal data with a generative process that describes not only invariances between modalities, but also stochastic effects and modality-specific variability.

The content invariance (Assumption 4) describes a shared phenomenon that is not directly observed but manifests in the observations of different modalities. Style changes (Assumption 5) represent shared influences that are not robust across modalities. For example, style variables can describe spurious correlations between modalities induced by the (non-deterministic) effects of an unobserved confounder between $\mathbf{x}_1$ and $\mathbf{x}_2$. Moreover, the conditional distribution $p(\tilde{\mathbf{s}} \mid \mathbf{s})$ can represent non-invertible transformations, such as data augmentation between different views (c.f., [von+21]). Modality-specific factors can be viewed as variables that describe the heterogeneity or noise inherent to a specific modality, e.g., the unique aspects of visual data compared to other types of signals.

With this, we conclude the problem formulation. In summary, we have specified the generative process (Section 8.2.1) and formalized our assumptions on the relation between modalities (Section 8.2.2). Next, we return to the topic of representation learning and show that, for the specified generative process, contrastive learning can identify the content factors up to a block-wise indeterminacy.

8.3 Identifiability Result

First, we need to define block-identifiability [von+21] for the multimodal setup, in which we consider modality-specific mixing functions and encoders. In the following, n_c denotes the number of content variables (i.e., the components of $\mathbf{c}$) and the subscript $1{:}n_c$ indicates the subset of content dimensions (indexed from 1 to n_c w.l.o.g.).

Definition 8 (Block-identifiability of content)**.** *The true content partition* $\mathbf{c} = \mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c}$ *is block-identified by a function* $\mathbf{g}_i : \mathcal{X}_i \to \mathcal{Z}_i$*, with* $i \in \{1, 2\}$*, if there exists an invertible function* $\mathbf{h}_i : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$*, such that for the inferred content partition* $\hat{\mathbf{c}}_i = \mathbf{g}_i(\mathbf{x}_i)_{1:n_c}$ *it holds that* $\hat{\mathbf{c}}_i = \mathbf{h}_i(\mathbf{c})$*.*

Notably, block-identifiability does not require the identification of *individual* factors, as required for some definitions of disentanglement [BCV13; Hig+18; Shu+20]. Instead, it

is sufficient to isolate the group of invariant factors (i.e., the content partition) from the remaining factors of variation in the data.

Next, we show that contrastive learning can block-identify the content variables for the multimodal generative process described in Section 8.2. We formalize this in Theorem 2 based on the asymptotic form of the InfoNCE objective (Equation 8.3).

Theorem 2. *Consider the data generating process described in Section 8.2, i.e., data pairs* $(\mathbf{x}_1, \mathbf{x}_2)$ *generated according to (8.5) with* $p(\mathbf{z}_2 \mid \mathbf{z}_1)$ *as defined in Assumptions 4 and 5. Further, assume that* $p(\mathbf{z}_1, \mathbf{z}_2)$ *is a smooth and continuous density on* $\mathcal{Z}_1 \times \mathcal{Z}_2$ *with* $p(\mathbf{z}_1, \mathbf{z}_2) > 0$ *almost everywhere. Let* $\mathbf{g}_1 : \mathcal{X}_1 \to (0,1)^{n_c}$ *and* $\mathbf{g}_2 : \mathcal{X}_2 \to (0,1)^{n_c}$ *be smooth functions that minimize the functional*

$$\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p(\mathbf{x}_1,\mathbf{x}_2)}\Big[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2\Big] - \tfrac{1}{2}\Big(H\left(\mathbf{g}_1\left(\mathbf{x}_1\right)\right) + H\left(\mathbf{g}_2\left(\mathbf{x}_2\right)\right)\Big) \,. \tag{8.9}$$

Then, $\mathbf{g}_1$ *and* $\mathbf{g}_2$ *block-identify the true content variables in the sense of Definition 8.*

Proof. To prove Theorem 2, we follow the proof structure from von Kügelgen et al. [von+21, Theorem 4.4] and divide the proof into three steps. First, we show that there exists a pair of smooth functions $\mathbf{g}_1^*, \mathbf{g}_2^*$ that attain the global minimum of Equation (8.9). Further, in Equations (8.16–8.18), we derive invariance conditions that must hold almost surely for any pair of smooth functions $\mathbf{g}_1, \mathbf{g}_2$ attaining the global minimum of Equation (8.9). In Step 2, we use the invariance conditions derived in Step 1 to show by contradiction that any pair of smooth functions $\mathbf{g}_1, \mathbf{g}_2$ that attain the global minimum in Equation (8.9) can only depend on content and not on style or modality-specific information. In the third and final step, for $\mathbf{h}_1 := \mathbf{g}_1 \circ \mathbf{f}_1$ and $\mathbf{h}_2 := \mathbf{g}_2 \circ \mathbf{f}_2$, we show that both functions must be bijections and hence that $\mathbf{c}$ is block-identified by $\mathbf{g}_1$ and $\mathbf{g}_2$ respectively.

Step 1. The global minimum of Equation (8.9) is reached when the first term is minimized and the second term is maximized. The first term is minimized when the encoders $\mathbf{g}_1$ and $\mathbf{g}_2$ are perfectly aligned, i.e., when $\mathbf{g}_1(\mathbf{x}_1) = \mathbf{g}_2(\mathbf{x}_2)$ holds for all pairs $(\mathbf{x}_1, \mathbf{x}_2) \sim p(\mathbf{x}_1, \mathbf{x}_2)$. The second term attains its maximum when $\mathbf{g}_1$ and $\mathbf{g}_2$ map to a uniformly distributed random variable on $(0,1)^{n_c}$ respectively.[31]

[31] We define the range of $\mathbf{g}_1$ and $\mathbf{g}_2$ on $(0,1)^{n_c}$ merely to simplify the notation. Generally, the uniform distribution $\mathcal{U}(a,b)$ is the maximum entropy distribution on the interval $[a,b]$.

To show that there *exists* a pair of functions that minimize Equation (8.9), let

$$\mathbf{g}_1^* \coloneqq \boldsymbol{d}_1 \circ \mathbf{f}_{1,1:n_c}^{-1} \,, \tag{8.10}$$

$$\mathbf{g}_2^* \coloneqq \boldsymbol{d}_2 \circ \mathbf{f}_{2,1:n_c}^{-1} \,, \tag{8.11}$$

where the subscript $1{:}n_c$ indicates the subset of content components w.l.o.g. and where $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ will be defined using the Darmois construction [Dar51; HP99].

First, recall that $\mathbf{f}_1^{-1}(\mathbf{x}_1)_{1:n_c} = \mathbf{c}$ and that $\mathbf{f}_2^{-1}(\mathbf{x}_2)_{1:n_c} = \tilde{\mathbf{c}}$ by definition. Second, for $i \in \{1, 2\}$, let us define $\boldsymbol{d}_i : \mathcal{C} \mapsto (0,1)^{n_c}$ using the Darmois construction, such that $\boldsymbol{d}_i$ maps $\mathbf{c}$ and $\tilde{\mathbf{c}}$ to a uniform random variable respectively. It follows that $\mathbf{g}_1^*, \mathbf{g}_2^*$ are smooth functions, because any function $\boldsymbol{d}_i$ obtained via the Darmois construction is smooth and $\mathbf{f}_1^{-1}, \mathbf{f}_2^{-1}$ are smooth as well (each being the inverse of a smooth function).

Next, we show that the pair of functions $\mathbf{g}_1^*, \mathbf{g}_2^*$, as defined in Equations (8.10) and (8.11), attains the global minimum of the functional in Equation (8.9), i.e.,

$$\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p(\mathbf{x}_1,\mathbf{x}_2)} \left[\|\mathbf{g}_1^*(\mathbf{x}_1) - \mathbf{g}_2^*(\mathbf{x}_2)\|_2\right] - \tfrac{1}{2}\left(H(\mathbf{g}_1^*(\mathbf{x}_1)) + H(\mathbf{g}_2^*(\mathbf{x}_2))\right) \tag{8.12}$$

$$= \mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p(\mathbf{x}_1,\mathbf{x}_2)} \left[\|\boldsymbol{d}_1(\mathbf{c}) - \boldsymbol{d}_2(\tilde{\mathbf{c}})\|_2\right] - \tfrac{1}{2}\left(H(\boldsymbol{d}_1(\mathbf{c})) + H(\boldsymbol{d}_2(\tilde{\mathbf{c}}))\right) \tag{8.13}$$

$$= 0 \,. \tag{8.14}$$

Concretely, by Assumption 4, $\mathbf{c} = \tilde{\mathbf{c}}$ almost surely, which implies that the first term in Equation (8.13) is zero almost surely. Further, $\boldsymbol{d}_i$ maps $\mathbf{c}, \tilde{\mathbf{c}}$ to uniformly distributed random variables on $(0,1)^{n_c}$, which implies that the differential entropy of $\boldsymbol{d}_1(\mathbf{c})$ and $\boldsymbol{d}_2(\tilde{\mathbf{c}})$ is zero as well. Consequently, there exists a pair of functions $\mathbf{g}_1^*, \mathbf{g}_2^*$ that minimizes Equation (8.9).

Next, let $\mathbf{g}_1 : \mathcal{X}_1 \mapsto (0,1)^{n_c}$ and $\mathbf{g}_2 : \mathcal{X}_2 \mapsto (0,1)^{n_c}$ be *any* pair of smooth functions that attains the global minimum of Equation (8.9), i.e.,

$$\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p(\mathbf{x}_1,\mathbf{x}_2)} \left[\|\mathbf{g}_1(\mathbf{x}_1) - \mathbf{g}_2(\mathbf{x}_2)\|_2\right] - \tfrac{1}{2}\left(H(\mathbf{g}_1(\mathbf{x}_1)) + H(\mathbf{g}_2(\mathbf{x}_2))\right) = 0 \,. \tag{8.15}$$

Let $\mathbf{h}_1 \coloneqq \mathbf{g}_1 \circ \mathbf{f}_1$ and $\mathbf{h}_2 \coloneqq \mathbf{g}_2 \circ \mathbf{f}_2$. Notice that both are smooth functions since all involved functions are smooth by definition. Since Equation (8.15) is a global minimum, it implies the following invariance conditions for the individual terms:

$$\mathbb{E}_{(\mathbf{x}_1,\mathbf{x}_2)\sim p(\mathbf{x}_1,\mathbf{x}_2)} \left[\|\mathbf{h}_1(\mathbf{z}_1) - \mathbf{h}_2(\mathbf{z}_2)\|_2\right] = 0 \,, \tag{8.16}$$

$$H(\mathbf{h}_1(\mathbf{z}_1)) = 0 \,, \tag{8.17}$$

$$H(\mathbf{h}_2(\mathbf{z}_2)) = 0 \,. \tag{8.18}$$

Therefore, $\mathbf{h}_1(\mathbf{z}_1) = \mathbf{h}_2(\mathbf{z}_2)$ must hold almost surely w.r.t. $p(\mathbf{x}_1, \mathbf{x}_2)$. Additionally, Equation (8.17) implies that $\hat{\mathbf{c}}_1 = \mathbf{h}_1(\mathbf{z}_1)$ must be uniform on $(0,1)^{n_c}$. Likewise, for Equation (8.18) and $\hat{\mathbf{c}}_2 = \mathbf{h}_2(\mathbf{z}_2)$.

Step 2. Next, we show that any pair of functions that minimize Equation (8.9) depends only on content information. Since style is independent of $\mathbf{m}_1$ and $\mathbf{m}_2$, we first show that $\mathbf{h}_1(\mathbf{z}_1)$ does not depend on $\mathbf{m}_1$ and, likewise, that $\mathbf{h}_2(\mathbf{z}_2)$ does not depend on $\mathbf{m}_2$. We then show that $\mathbf{h}_1$ and $\mathbf{h}_2$ also cannot depend on style, based on a result from previous work.

First note that we can exclude all degenerate solutions where $\mathbf{g}_1$ maps a component of $\mathbf{m}_1$ to a constant, since $\mathbf{g}_1$ would not be invertible anymore and such a solution would violate the invariance in Equation (8.17).

To prove a contradiction, suppose that, w.l.o.g., $\mathbf{h}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)_{1:n_c} := \mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ depends on some component in $\mathbf{m}_1$ in the sense that the partial derivative of $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ w.r.t. some modality-specific variable $m_{1,l}$ is non-zero for some point $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*) \in \mathcal{Z}_1$. Specifically, it implies that the partial derivative is positive, i.e.,

$$\frac{\partial \mathbf{h}_1(\mathbf{z}_1)_{1:n_c}}{\partial m_{1,l}} > 0 \tag{8.19}$$

in a neighborhood around $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$, which is a non-empty open set, since $\mathbf{h}_1$ is smooth. On the other hand, due to the independence of $\mathbf{z}_2$ and $\mathbf{m}_1$, the fact that $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ cannot depend on $\mathbf{m}_1$, and that $p(\mathbf{z}_1, \mathbf{z}_2) > 0$ almost everywhere, we come to a contradiction. That is, there exists an open set of points with positive measure, namely the neighborhood around $(\mathbf{c}^*, \mathbf{s}^*, \mathbf{m}_1^*)$, on which

$$|\, (\mathbf{h}_1(\mathbf{z}_1)_{1:n_c} - \mathbf{h}_2(\mathbf{z}_2)_{1:n_c}) \,| \; > 0 \tag{8.20}$$

almost surely, which contradicts the invariance in Equation (8.16). The statement does not change, if we add further dependencies for $\mathbf{h}_1$ on components of $\mathbf{m}_1$, or for $\mathbf{h}_2$ on components of $\mathbf{m}_2$, because $\mathbf{m}_1$ and $\mathbf{z}_2$ are independent, and $\mathbf{m}_2$ and $\mathbf{z}_1$ are independent as well. Hence, we show that *any* encoder that minimizes the objective in Equation (8.9) cannot depend on modality-specific information.

Having established that neither $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$, nor $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ can depend on modality-specific information, it remains to show that style information is also not encoded. Leveraging Assumption 5, we can show that the strict inequality in Equation (8.20) holds with probability greater than zero if $\mathbf{h}_1(\mathbf{z}_1)_{1:n_c}$ or $\mathbf{h}_2(\mathbf{z}_2)_{1:n_c}$ were dependent on a component in $\mathbf{s}$ or $\tilde{\mathbf{s}}$ respectively. This would again lead to a violation of the invariance derived in Equation (8.16), as shown in von Kügelgen et al. [von+21, Proof of Theorem 4.2].

Step 3. It remains to show that $\mathbf{h}_1, \mathbf{h}_2$ are bijections. We know that $\mathcal{C}$ and $(0,1)^{n_c}$ are simply connected and oriented C^1 manifolds, and we have established in Step 1 that $\mathbf{h}_1$

and $\mathbf{h}_2$ are smooth and hence differentiable functions. Since $p(\mathbf{c})$ is a regular density, the uniform distributions w.r.t. the pushthrough functions $\mathbf{h}_1$ and $\mathbf{h}_2$ are regular densities. Thus, $\mathbf{h}_1$ and $\mathbf{h}_2$ are bijections [Zim+21, Proposition 5].

Step 3 concludes the proof. We have shown that for any pair of smooth functions $\mathbf{g}_1, \mathbf{g}_2$ that attain the global minimum of Equation (8.9), we have that $\mathbf{c}$ is *block-identified* (Definition 8) by $\mathbf{g}_1$ and $\mathbf{g}_2$. □

Limitations While Theorem 2 suggests that contrastive learning can identify the content variables up to a block-wise indeterminacy, the result rests on two key assumptions. First, it is based on the asymptotic form of the InfoNCE objective (Equation 8.3). Second, it assumes that the number of content variables is known or that it can be estimated.

To address these limitations, we verify the result with suitable experiments. Using the symmetrized InfoNCE objective (Equation 8.2), we conduct numerical simulations (Section 8.4.1) as well as experiments on a dataset of image-text pairs (Section 8.4.2). As part of these experiments, we vary the encoding size and explore whether we can estimate the number of content variables. In Section 8.5, we continue the discussion of the assumptions in the context of the experimental results.

8.4 Experiments

The goal of our experiments is to test whether contrastive learning can block-identify content in the multimodal setting, as described by Theorem 2. First, we verify identifiability in a fully controlled experiment with numerical simulations (Section 8.4.1). Second, we corroborate our findings on a complex multimodal dataset of image-text pairs (Section 8.4.2).

8.4.1 Numerical Simulations

The numerical simulation is designed to assess identifiability with full control over the generative process. We extend the numerical simulation from von Kügelgen et al. [von+21] and implement the multimodal generative process described in Section 8.2 using modality-specific mixing functions ($\mathbf{f}_1 \neq \mathbf{f}_2$) and modality-specific latent variables $\mathbf{m}_1$ and $\mathbf{m}_2$.

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content c	Style s
1.0	✗	✗	**1.00** (±0.00)	0.00 (±0.00)
0.75	✗	✗	**0.99** (±0.01)	0.00 (±0.00)
0.75	✓	✗	**0.99** (±0.00)	0.52 (±0.09)
0.75	✗	✓	**1.00** (±0.00)	**0.79** (±0.04)
0.75	✓	✓	**0.99** (±0.01)	**0.81** (±0.04)

(a) Multi-view setting

Generative process			R^2 (nonlinear)		
p(chg.)	Stat.	Cau.	Content c	Style s	Modality $\mathbf{m}_i$
1.0	✗	✗	**0.99** (±0.00)	0.00 (±0.00)	0.00 (±0.00)
0.75	✗	✗	**1.00** (±0.00)	0.00 (±0.00)	0.00 (±0.00)
0.75	✓	✗	**0.95** (±0.01)	0.56 (±0.23)	0.00 (±0.00)
0.75	✗	✓	**0.98** (±0.00)	**0.87** (±0.04)	0.00 (±0.00)
0.75	✓	✓	**0.95** (±0.03)	**0.89** (±0.07)	0.00 (±0.00)

(b) Multimodal setting

Table 8.1: Results of the numerical simulations. We compare the multi-view setting ($\mathbf{f}_1 = \mathbf{f}_2$, left table) with the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, right table). Only the multimodal setting includes modality-specific latent variables. Each row presents the results of a different setup with varying style-change probability p(chg.) and possible statistical (Stat.) and/or causal (Caus.) dependencies. Each value represents the R^2 coefficient of determination (averaged across 3 seeds) for a nonlinear regression model that predicts the respective ground truth factor ($\mathbf{c}, \mathbf{s}$, or $\mathbf{m}_i$) from the embeddings produced by the encoder.

Implementation For the data generation, we sample $\mathbf{c} \sim \mathcal{N}(0, \Sigma_\mathbf{c})$, $\mathbf{m}_i \sim \mathcal{N}(0, \Sigma_{\mathbf{m}_i})$, and $\mathbf{s} \sim \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_\mathbf{s})$. Statistical dependencies within blocks (e.g., among components of $\mathbf{c}$) are induced by non-zero off-diagonal entries in the corresponding covariance matrix (e.g., in $\Sigma_\mathbf{c}$). To induce a causal dependence from content to style, we set $a_i, B_{ij} \sim \mathcal{N}(0, 1)$; otherwise, we set $a_i, B_{ij} = 0$. For style changes, Gaussian noise is added with probability π independently for each style component, i.e., $\tilde{\mathbf{s}}_i = \mathbf{s}_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \Sigma_\epsilon)$ with probability π. We generate the observations $\mathbf{x}_1 = \mathbf{f}_1(\mathbf{c}, \mathbf{s}, \mathbf{m}_1)$ and $\mathbf{x}_2 = \mathbf{f}_2(\mathbf{c}, \tilde{\mathbf{s}}, \mathbf{m}_2)$ using two *distinct* nonlinear mixing functions; specifically, for each $i \in \{1, 2\}$, $\mathbf{f}_i : \mathbb{R}^d \to \mathbb{R}^d$ is a separate, invertible 3-layer MLP with LeakyReLU activations. The invertible MLP is constructed similar to previous work [HM16; HM17; Zim+21; von+21] by resampling square weight matrices until their condition number surpasses a threshold value.

We train the encoders for 300 000 iterations using the symmetrized InfoNCE objective (Equation 8.2) and the hyperparameters listed in Table A.2a. We evaluate block-identifiability using nonlinear probing (c.f. Section 2.5.1), i.e., by predicting the ground truth factors from the embeddings produced by the model. Specifically, we use kernel ridge regression and report the R^2 coefficient of determination on holdout data. In Table A.2a, we specify the main hyperparameters for the numerical simulation.

Results We compare the multi-view setting ($\mathbf{f}_1 = \mathbf{f}_2$, Table 8.1a) with the multimodal setting ($\mathbf{f}_1 \neq \mathbf{f}_2$, Table 8.1b) and find that content can be block-identified in *both* settings, as the R^2 score is close to one for the prediction of content and close to chance-level for the prediction of style and modality-specific information. Consistent with previous work, we observe that some style information can be predicted when there are statistical and/or causal dependencies; this is expected because statistical dependencies decrease the effective dimensionality of content, while the causal dependence from $\mathbf{c}$ to $\mathbf{s}$ makes style partially predictable from the encoded content information. In Appendix A.6, we include a more rigorous empirical evaluation using interventions, which provides further support for the block-identifiability of content even in the case of causal dependencies.

Overall, the results of the numerical simulations are consistent with our theoretical result from Theorem 2, showing that contrastive learning based on the InfoNCE objective can block-identify content, if the generative process satisfies the assumptions from Section 8.2.

8.4.2 Multimodal3DIdent

Next, we test whether block-identifiability holds in a more realistic setting, for which we use a dataset of image-text pairs. Specifically, we use the Multimodal3DIdent dataset, which provides an identifiability benchmark with image-text pairs generated from controllable ground truth factors, some of which are shared between image and text modalities, as illustrated in Figure 8.2. For each pair, there are three content factors that are invariant across modalities: object position (x- and y-coordinates) and object shape. As a style factor, we have the object color, which is shared between modalities but not invariant. Furthermore, the object color depends causally on the position of the object. The remaining factors are modality-specific. For more information about the dataset, see Section 4.4.

Implementation We train the encoders for 100 000 iterations using the symmetrized InfoNCE objective (Equation 8.2) and the hyperparameters listed in Table A.2b. For the image encoder we use a ResNet-18 architecture [He+16] and for the text we use a convolutional network. As for the numerical simulation, we evaluate block-identifiability using nonlinear

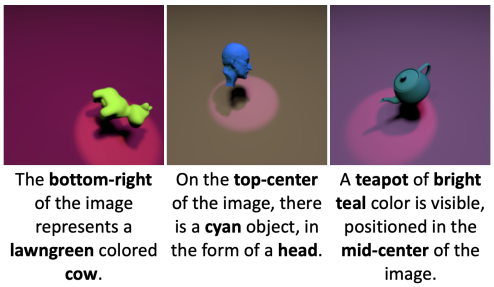
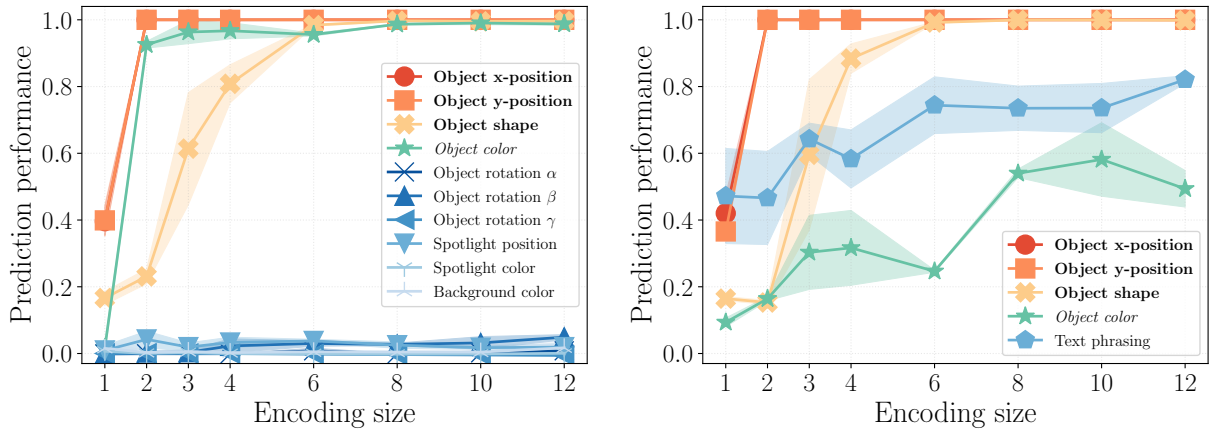


Figure 8.2: Examples of image-text pairs.



(a) Prediction of image factors

(b) Prediction of text factors

Figure 8.3: Results on the Multimodal3DIdent dataset as a function of the encoding size of the model. Using nonlinear probing, we assess the prediction of ground truth image factors (Figure 8.3a) and text factors (Figure 8.3b) to quantify how well the embeddings encode the respective information. Along the x-axis, we vary the encoding size, i.e., the output dimensionality of the model. We measure the prediction performance in terms of the R^2 coefficient of determination for continuous factors and classification accuracy for discrete factors respectively. Each point denotes the average across three seeds and bands show one standard deviation. In each figure legend, content factors are denoted in bold and style factors in italic.

probing, i.e., by predicting the ground truth factors from the embeddings produced by the encoders. For continuous factors, we use kernel ridge regression and report the R^2 coefficient of determination; for discrete factors, we report the classification accuracy of a shallow MLP. We train the predictors on training embeddings and evaluate them on embeddings of holdout data.

Results Figure 8.3 presents the results on the Multimodal3DIdent dataset with a dimensionality ablation, where we vary the size of the encoding. We observe that content factors (object position and shape) are always encoded well, unless the encoding size is too small (i.e., smaller than 3-4 dimensions). When there is sufficient capacity, style information (object color) is also encoded, partly because there is a causal dependence from

content to style and partly because of the excess capacity.[32] Image-specific information (object rotation, spotlight position, background color) is mostly discarded, independent of the encoding size. Text-specific information (phrasing) is encoded to a moderate degree (48–80% accuracy), which we attribute to the fact that phrasing is a discrete factor that violates the assumption of continuous latent variables. This hints at possible limitations in the presence of discrete latent factors, which we further discuss in Section 8.5.

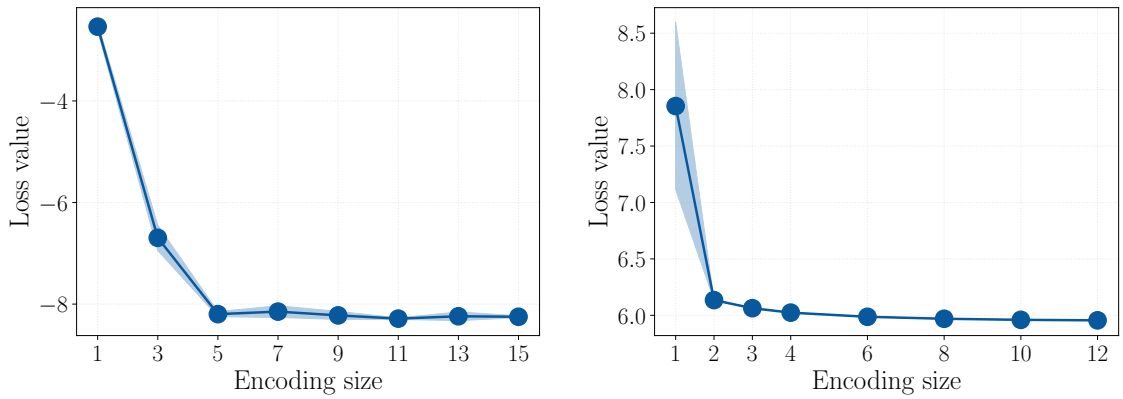
Overall, the results indicate that contrastive learning can be used to block-identify content factors in a complex multimodal setting with image-text pairs.

8.4.3 Estimating the Number of Content Factors

The estimation of the number of content factors is an important aspect, because Theorem 2 assumes that the number of content factors is known or that it can be estimated. In practice, the number of content factors can be viewed as a single hyperparameter that can be tuned with respect to a suitable model selection metric (e.g. [Loc+20b]). For example, the validation loss would be a convenient metric, because it only requires a holdout dataset and no additional labels. In the following, we explore the idea of using the validation loss to select the number of content factors.

In Figure 8.4, we plot the validation loss (averaged over 2000 validation samples) as a function of the encoding size for both experiments used in this chapter. Results for the numerical simulation are shown in Figure 8.4a and for the image-text experiment in Figure 8.4b. For both datasets, we observe that the validation loss increases most significantly in the neighborhood of the true number of content factors. For the numerical simulation, the results show a clear "elbow" [Jam+13] at the correct value of 5, which corresponds to the true number of content factors. The results are less clear for the image-text experiment, where the elbow method might suggest the range of 2-4 content factors, whereas the true value is 3.

[32] In Appendix A.6, we provide additional results on a version of the dataset with mutually independent factors, for which the encoding of content factors over style factors can be observed more clearly.



(a) Numerical simulation

(b) Multimodal3DIdent

Figure 8.4: Estimation of the number of content factors using the validation loss. The validation loss corresponds to the value of the objective in Equation (8.2) computed on a holdout dataset. Since we are interested in estimating the true number of content factors to select the encoding size appropriately, we plot the validation loss as a function of the encoding size. We show the validation loss for the numerical simulation with independent factors (Figure 8.4a) and for the Multimodal3DIdent dataset (Figure 8.4b) respectively.

8.5 Discussion

Implications and scope We investigated whether contrastive learning can block-identify content factors—i.e., shared latent variables that are invariant across modalities—under the assumptions described in Section 8.2. Therefore, in Theorem 2, we derived an identifiability result based on the asymptotic form of the InfoNCE objective. We used numerical simulations (Section 8.4.1) to verify the theoretical result in a controlled setup, demonstrating that contrastive learning can identify content factors up to a block-wise indeterminacy (Definition 8) when the size of the encoding matches the number of content factors. With experiments on a dataset of image-text pairs, we corroborated the results in a more realistic setting and even under some violations of the assumptions underlying Theorem 2. Specifically, we investigated the mismatch between the encoding size and the number of content factors and included experiments with discrete factors. Overall, we observed that contrastive learning encodes the content factors effectively across all considered settings. When there is sufficient capacity, stochastically shared information (i.e., style) is also encoded to a moderate degree, but seemingly without affecting the encoding of content factors, which could still be recovered using nonlinear probing in all considered settings.

For practical applications, our results suggest that contrastive learning without capacity constraints can encode *any* shared factor, regardless of whether the factor is truly invariant across modalities or whether its effect on the observations is stochastic. This is in line with the theoretical interpretation that the optimization of the InfoNCE objective maximizes the mutual information between representations (see Section 3.4.3). As a measure of mutual dependence, the mutual information quantifies *any* amount of information that is shared between two random variables, regardless of the invariance. Yet, our results also demonstrate that the size of the encoding can be reduced to learn a representation that encodes only the invariant factors. In practice, this can be leveraged for representation learning in settings of content-preserving distribution shifts, where information relevant for a downstream task remains unchanged across domains [Roj+18; Mit+21; FTF21].

Limitations and outlook A potential limitation of our results is the assumption of content invariance, which might not be satisfied in real-world settings. For example, there could be pairs of observations for which the invariance is inadvertently violated, e.g., due to measurement errors, occlusions, or other inaccuracies in the data. On the one hand, such a violation can be viewed as a mere artifact of the data collection, which could be addressed through interventions on the data generating process. On the other hand, violations of content invariance blur the line between content and style factors, which suggests a possible generalization of the problem formulation in terms of only stochastically shared factors as an opportunity for future work.

Another possible limitation is that Theorem 2 assumes the number of content factors is known or that it can be estimated. In Section 8.4.3, we explored this idea based on the validation loss, but more work is required to verify these results on real-world data. We believe that the estimation of the number of latent factors and the design of suitable heuristics are interesting directions for future research.

Moreover, Theorem 2 assumes that all latent factors are continuous. While this assumption prevails in related work [HP99; HM16; HST19; Gre+19; Loc+19; Loc+20b; Zim+21; von+21; Kli+21], our results in Figure 8.3b indicate that in the presence of discrete content factors, some style or modality-specific information can be encoded. In Appendix A.6, we include additional results based on numerical simulations that support these findings.

Finally, our problem formulation could be extended to more than two modalities—a setting for which there are intriguing identifiability results [Gre+19; Sch+16] as well as suitable learning objectives [TKI20; Lyu+22].

8.6 Summary

In this chapter, we focused on contrastive learning as a discriminative approach for multimodal representation learning. Specifically, we investigated whether contrastive learning can be used to identify latent factors of variation shared between modalities up to acceptable ambiguities.

First, we theoretically showed that, asymptotically, contrastive learning can identify shared latent factors that are invariant across modalities up to a block-wise indeterminacy. While the theoretical result is based on an asymptotic form of the InfoNCE objective and requires that the number of content factors is known or that it can be estimated, our empirical findings suggest that the desired results can be achieved in practice even when some of the assumptions are violated. Specifically, we used numerical simulations to verify the identifiability result and corroborated our findings on a dataset of image-text pairs. Thus, we provided a better understanding of the assumptions required to recover latent factors of variation in the context of multimodal learning under weak supervision.

In the scope of the thesis, this chapter analyzed contrastive learning as a discriminative approach to address Question 1, i.e., to identify latent factors of variation shared between modalities up to acceptable ambiguities. As part of the experiments, we demonstrated promising results even in settings where shared information cannot be predicted in expectation across modalities on the level of observations. In the following chapter, we build on insights from this and the previous chapters to develop a hybrid approach for multimodal generative learning that combines contrastive learning and VAEs.

9

A Hybrid Approach

In the previous chapter, we showed that contrastive learning can be used to identify latent factors shared between modalities. In this chapter, we leverage this property and introduce a hybrid approach that combines contrastive learning with VAEs. We combine insights and techniques from the previous chapters to develop a generative model that approximates the joint distribution of multiple modalities and disentangles shared and modality-specific information effectively. As an outlook, we demonstrate how the model can address some of the limitations of multimodal VAEs that we covered in Chapter 7.

In Section 9.2, we propose the disentangling multimodal variational autoencoder (DMVAE), which learns a multimodal generative model with an inference network that disentangles shared and modality-specific information. The model consists of two parts that are trained in separate stages. First, to infer shared information, we use contrastive learning and extend the InfoNCE objective with a mixture of product of experts to model the variational joint posterior in a way that enables efficient inference given any subset of modalities. Second, to infer modality-specific information, we use VAEs conditioned on the embeddings produced by the variational joint posterior and a regularization technique to suppress the encoding of shared information for the VAEs. In Section 9.3, we evaluate the DMVAE on the Translated-PolyMNIST dataset, for which we present promising results beyond the capabilities of existing multimodal VAEs. Specifically, we find that the DMVAE reduces the gap in generative quality between unimodal and multimodal VAEs and achieves substantial improvements for the generation of missing modalities.

9.1 Motivation and Background

Multimodal VAEs are designed to approximate a joint distribution of multiple modalities in a way that enables inference across modalities based on the learned representations (c.f., Sections 3.3.6 and 5.1). While multimodal VAEs have demonstrated promising results in real-world applications [Dor+19; LS21; Min+21; GZP21], we showed that they exhibit severe limitations in settings where shared information cannot be predicted in expectation across modalities on the level of observations (Chapter 7).

In this chapter, we seek to address this limitation and design a method that decouples the inference of shared information from generative modeling. Therefore, we build on ideas from the previous chapters; specifically, the design of a variational joint posterior using a mixture of products of experts (Chapter 5), the partitioning of the latent space into shared and modality-specific subspaces (Chapter 6), and the inference of shared latent factors using contrastive learning (Chapter 8). With the DMVAE, we propose a method that combines these ideas and yields a hybrid approach for multimodal generative learning with an inference network that disentangles shared and modality-specific information.

9.2 Method: DMVAE

The DMVAE is comprised of two parts that are trained in separate stages. First, to infer information that is shared between modalities, we model the variational joint posterior with a shared inference network trained using contrastive learning (Section 9.2.2). Second, to infer modality-specific information, for each modality we train a conditional VAE (Section 9.2.3) with a regularizer that suppresses the encoding of shared information (Section 9.2.4). Jointly, these components yield a multimodal generative model that disentangles shared and modality-specific information and that can be used to generate missing modalities. After we introduce the individual components of the model, we describe the training procedure for the DMVAE (Section 9.2.5).

9.2.1 Preliminaries

Let M be the number of modalities, let $\bar{\mathbf{x}} \coloneqq \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$ be a set of random vectors that describe different modalities, and let $p(\bar{\mathbf{x}})$ denote their joint distribution. As in the

previous chapters, we denote subsets of modalities using subscripts; specifically, we write $\mathbf{x}_A$ to index a subset of modalities $A \subseteq \{1, \dots, M\}$.

9.2.2 Part 1: Multimodal Contrastive Learning

The first part of the DMVAE is an inference network trained using contrastive learning to infer the information shared between observations of different modalities. In the following, we build on our idea from Chapter 5 and define the variational joint posterior as a mixture of product of experts (MoPoE) to aggregate information across different subsets of modalities efficiently.

Variational joint posterior Let $q_{\psi_i}(\mathbf{c}|\mathbf{x}_i)$ denote the variational posterior for modality i with parameters ψ_i and let the complete set of parameters be denoted by $\psi \coloneqq \{\psi_1, \dots, \psi_M\}$. Let $\mathcal{P}(M)$ be the powerset of the set of consecutive integers $\{1, \dots, M\}$ excluding the empty set and let $|\mathcal{P}(M)|$ denote its cardinality.

To aggregate information across all subset of modalities efficiently, we define the variational joint posterior as a mixture of products of unimodal posterior—a so-called mixture of product of experts (MoPoE; see Chapter 5):

$$q_{\psi}^{\text{MoPoE}}(\mathbf{c} \mid \bar{\mathbf{x}}) = \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} \prod_{i \in A} q_{\psi_i}(\mathbf{c} \mid \mathbf{x}_i) \tag{9.1}$$

$$= \frac{1}{|\mathcal{P}(M)|} \sum_{A \in \mathcal{P}(M)} q_{\psi_A}^{\text{PoE}}(\mathbf{c} \mid \mathbf{x}_A) \ , \tag{9.2}$$

where the set of parameters ψ_A is fully determined by the parameters of the unimodal encoders, i.e., $\psi_A = \{\psi_i \,|\, i \in A\}$ for each $A \in \mathcal{P}(M)$, and where "PoE" stands for product of experts [Hin02].

To sample from $q_{\psi}^{\text{MoPoE}}(\mathbf{c} \mid \bar{\mathbf{x}})$, we draw a set A uniformly at random from $\mathcal{P}(M)$ and then sample $\mathbf{c} \sim q_{\psi_A}^{\text{PoE}}(\mathbf{c} \mid \mathbf{x}_A)$. Following previous work [WG18], we use Gaussian unimodal posteriors, so that $q_{\psi_A}^{\text{PoE}}(\mathbf{c} \mid \mathbf{x}_A)$ is also Gaussian and can be computed efficiently in closed form for any subset of modalities.

InfoNCE with the MoPoE posterior To train the shared inference network using contrastive learning, we need to extend the InfoNCE objective from Equation (3.53), which is only defined for two views or modalities.

Based on the MoPoE variational joint posterior (Equation 9.1), we define the objective as

$$\mathcal{L}^{\text{MoPoE}}_{\text{InfoNCE}}(\{\bar{\mathbf{x}}^k\}_{k=1}^K;\psi) = -\sum_{k=1}^{K} \log \frac{\exp\{\text{sim}(\mathbf{c}_A^k, \mathbf{c}^k)/\tau\}}{\sum_{l=1}^{K} \exp\{\text{sim}(\mathbf{c}_A^k, \mathbf{c}^l)/\tau\}} \,, \tag{9.3}$$

where $\mathbf{c}_A^k \sim q^{\text{PoE}}_{\psi_A}(\mathbf{c} \mid \mathbf{x}_A^k)$ and $\mathbf{c}^k \sim q^{\text{PoE}}_{\psi}(\mathbf{c} \mid \bar{\mathbf{x}}^k)$. For each training example, we draw A uniformly at random from $\mathcal{P}(M)$. Thus, the embedding $\mathbf{c}_A^k$ is computed based on a random subset of modalities, whereas $\mathbf{c}^k$ is based on the complete set. Since the objective combines the sub-sampling of modalities with the PoE-aggregation, it can be viewed as a mixture of products of experts (c.f., Chapter 5).

Intuitively, the objective in Equation (9.3) contrasts between a positive pair of embeddings $(\mathbf{c}_A^k, \mathbf{c}^k)$ drawn from corresponding examples and $K-1$ negative pairs $\{(\mathbf{c}_A^k, \mathbf{c}^l)\}$ drawn from different examples $k \neq l$. Through the minimization of the objective with respect to ψ, the shared inference network is trained to learn a representation where positive pairs are closer to each other while negative pairs are further apart (c.f., Section 3.4).

9.2.3 Part 2: Conditional VAEs with Coupled Decoders

The second part of the DMVAE is a multimodal generative model comprised of a set of conditional VAEs. For each modality, we train a separate VAE, but the decoders are *coupled* across modalities because we condition each decoder on a context vector that is shared between modalities. In this subsection, we define the objective of the conditional VAE before we extend the objective with an additional regularizer to incentivize the desired disentanglement of shared and modality-specific information in Section 9.2.4.

Broadly, our idea is to train a conditional VAE for each modality $i \in \{1, \ldots, M\}$ to learn a generative model $p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)$, where $\mathbf{c}$ is a context vector that is shared between modalities, $\mathbf{m}_i$ is a modality-specific latent vector, and θ_i denotes the parameters of the decoder for modality i. Since $\mathbf{c}$ is shared between modalities, the decoders are coupled and thus we implicitly train a multimodal generative model $p_\theta(\bar{\mathbf{x}} \mid \mathbf{c})$, where $\theta = \{\theta_1, \ldots, \theta_M\}$. In general, we assume that $\mathbf{c}$ is not known and instead use the embeddings produced by the shared inference network we introduced in Section 9.2.2

Conditional VAE With a conditional VAE for modality i, we approximate $\log p(\mathbf{x}_i \mid \mathbf{c})$, i.e., the log-evidence conditioned on a context vector $\mathbf{c}$. For each modality $i \in \{1, \ldots, M\}$,

we train a conditional VAE by maximizing the objective

$$\mathcal{L}_{\text{ELBO}}(\mathbf{x}_i|\mathbf{c};\phi_i,\theta_i) := \mathbb{E}_{q_{\phi_i}(\mathbf{m}_i \mid \mathbf{x}_i)}\left[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{m}_i, \mathbf{c})\right] - D_{\text{KL}}\left(q_{\phi_i}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i)\right) \ , \quad (9.4)$$

where ϕ_i denotes the parameters of the encoder for modality i. Compared to the standard objective of the conditional VAE (Equation 3.42), we merely replace the notation for the observation (i.e., $\mathbf{x}_i$ instead of $\mathbf{x}$) as well as for the latent vector ($\mathbf{m}_i$ instead of $\mathbf{z}$), and additionally assume that $p(\mathbf{m}_i \mid \mathbf{c}) = p(\mathbf{m}_i)$ to remove the conditioning for the prior distribution.[33] Consequently, Equation (9.4) forms a variational lower bound on the conditional log-evidence $\log p(\mathbf{x}_i \mid \mathbf{c})$.

To summarize, we train a conditional VAE independently for each modality $i \in \{1, \dots, M\}$ using the objective in Equation (9.4) with an additional regularizer that we introduce in Section 9.2.4. While each VAE is trained independently, all decoders are conditioned on the same context vector $\mathbf{c}$, which provides a coupling between the decoders of different modalities. Thus, we can independently train M conditional VAEs—each with a separate, modality-specific latent space—and still associate the generative models of different modalities through the conditioning on the shared context vector. Consequently, the DMVAE can generate observations of different modalities conditioned on the same context vector (e.g., to generate missing modalities).

9.2.4 Disentanglement by Mutual Information Minimization

In Chapter 6, we discussed the benefits and pitfalls of partitioning the latent space of multimodal VAEs into shared and modality-specific subspaces. Specifically, we showed that partitioning without additional constraints does not suffice to achieve the desired disentanglement of shared and modality-specific information. Similarly, for the DMVAE, we found that the model tends to encode not only modality-specific but also shared information in the modality-specific embeddings $\mathbf{m}_i \sim q_{\phi_i}(\mathbf{m}_i \mid \mathbf{x}_i)$ if we train the VAEs using the objective in Equation (9.4) without additional regularization.

To incentivize the disentanglement of shared and modality-specific information, we add a regularization term to the objective in Equation (9.4). Specifically, we introduce a regularization technique that suppresses the encoding of shared information for the VAEs through the minimization of the mutual information between the context vector and the

[33] We introduce the latter assumption because we use the model to infer modality-specific information which, by definition, is independent of information shared between modalities.

modality-specific embeddings, but only with respect to the latter. In this subsection, we describe the mutual information estimator, which we use to define the regularized training objective of the conditional VAE in Section 9.2.5.

Mutual information estimation Let $\mathbf{c}$ and $\mathbf{m}_i$ be two random vectors for which we want to estimate the mutual information. The mutual information is given by the KL-divergence between their joint distribution and the product of marginals; thus, it can be expressed as a ratio of two densities

$$I(\mathbf{c}; \mathbf{m}_i) = D_{\mathrm{KL}}(q(\mathbf{c}, \mathbf{m}_i) \,||\, q(\mathbf{c})q(\mathbf{m}_i)) \tag{9.5}$$

$$= \mathbb{E}_{q(\mathbf{c},\mathbf{m}_i)} \left[\log \frac{q(\mathbf{c}, \mathbf{m}_i)}{q(\mathbf{c})q(\mathbf{m}_i)} \right] . \tag{9.6}$$

For estimating the density ratio in Equation (9.6), we use a common technique for which a discriminator or classifier is trained to distinguish between samples drawn from these distributions [NWJ10; SSK12; KM18].

For each modality $i \in \{1, \dots, M\}$, we define a discriminator $D_i : (\mathbf{c}, \mathbf{m}_i) \mapsto [0, 1]$, which estimates the probability that a given pair of embeddings $(\mathbf{c}, \mathbf{m}_i)$ is sampled from the joint distribution $q(\mathbf{c}, \mathbf{m}_i)$ as opposed to the product of marginals $q(\mathbf{c})q(\mathbf{m}_i)$. The discriminator is trained to minimize the binary cross-entropy loss

$$-\mathbb{E}_{q(\mathbf{c},\mathbf{m}_i)} \left[\log D_i(\mathbf{c}, \mathbf{m}_i)\right] - \mathbb{E}_{q(\mathbf{c})q(\mathbf{m}_i)} \left[\log(1 - D_i(\mathbf{c}, \mathbf{m}_i))\right] \tag{9.7}$$

based on a finite sample from the two distributions.

Following Kim and Mnih [KM18], we define the mutual information estimator as

$$\hat{I}(\mathbf{c}; \mathbf{m}_i) \coloneqq \mathbb{E}_{q(\mathbf{c},\mathbf{m}_i)} \left[\log \frac{D_i(\mathbf{c}, \mathbf{m}_i)}{1 - D_i(\mathbf{c}, \mathbf{m}_i)} \right] , \tag{9.8}$$

with the difference that the estimator in Equation (9.8) does not estimate the total correlation[34] between all components of a latent vector, but the mutual information between two vectors $\mathbf{c}$ and $\mathbf{m}_i$ that represent a specific partitioning of the latent space.

Mutual information minimization Concurrent to the training of the DMVAE, we train the discriminator for each modality i based on samples $(\mathbf{c}, \mathbf{m}_i) \sim q_\psi^{\mathrm{MoPoE}}(\mathbf{c}|\bar{\mathbf{x}}) q_{\phi_i}(\mathbf{m}_i|\mathbf{x}_i)$

[34] In the case of two random variables, total correlation is equivalent to mutual information [e.g., CT12].

and use batch-wise permutations to create samples from the marginal distributions. In the optimization of the DMVAE, we use the estimated mutual information to minimize the statistical dependence between the shared and modality-specific embeddings, but only with respect to the parameters ϕ_i of the modality-specific encoder. Thus, while the discriminator is trained to estimate the mutual information, its estimates affect only the parameters ϕ_i of the modality-specific encoder in order to suppress the encoding of shared information for the respective VAE. Next, we specify the complete training procedure of the DMVAE and describe the optimization in more detail.

9.2.5 Training Procedure

To recapitulate, the DMVAE is comprised of two parts: it has a shared inference network that infers the information shared between modalities using contrastive learning and, for each modality, it uses a conditional VAE to infer modality-specific information.

The DMVAE is trained in two stages, using the following objectives[35]:

$$\min_{\psi} \; \mathbb{E}_{\{\bar{\mathbf{x}}^k\}_{k=1}^K \sim p(\bar{\mathbf{x}})} \left[\mathcal{L}^{\text{MoPoE}}_{\text{InfoNCE}}(\{\bar{\mathbf{x}}^k\}_{k=1}^K; \psi) \right], \tag{Stage 1}$$

$$\max_{\phi_i, \theta_i} \; \mathbb{E}_{p(\bar{\mathbf{x}}) q_\psi^{\text{MoPoE}}(\mathbf{c} \mid \bar{\mathbf{x}})} \left[\mathcal{L}_{\text{ELBO}}(\mathbf{x}_i | \mathbf{c}; \phi_i, \theta_i) \right] - \delta_i \hat{I}(\mathbf{c}; \mathbf{m}_i), \quad \text{for } i \in \{1, \ldots, M\}. \tag{Stage 2}$$

In the first stage, we only train the shared inference network using the objective in Equation (9.3). In the second stage, we train a conditional VAE independently for each modality, optimizing a Lagrangian function where the ELBO of the conditional VAE (Equation 9.4) is maximized subject to the disentanglement regularization (Equation 9.8), which is weighted by a Lagrange multiplier $\delta_i \geq 0$.[36] We update the discriminator parameters once after every training step of the VAE, but in principle other update schedules would be possible. Notably, the parameters ψ of the shared inference network remain fixed during the second stage of training; thus, the regularization only affects the parameters $\{\phi_1, \ldots, \phi_M\}$ of the modality-specific encoders to suppress the encoding of shared information.

[35] In theory, we define the objectives in terms of expectations over the joint distribution $p(\bar{\mathbf{x}})$. In practice, we sample observations from a finite dataset $\mathcal{D} = \{(\bar{\mathbf{x}}^{(n)})\}_{n=1}^N$ of size N.

[36] Notably, $\delta_i = 0$ corresponds to training a conditional VAE without additional regularization. We investigate the effect of δ_i in the experiments (Section 9.3.2).

9.3 Experiments

Next, we evaluate the DMVAE as a multimodal generative model in comparison to multimodal VAEs. Specifically, we include the MVAE [WG18], MMVAE [Shi+19], as well as the MoPoE-VAE and MMVAE+ that we introduced in the previous chapters (Chapters 5 and 6, respectively). We conduct experiments on the Translated-PolyMNIST dataset (described in Section 4.1.1), for which shared information cannot be predicted in expectation across modalities on the level of observations. Thus, with the DMVAE, we seek to address limitations of existing multimodal VAEs that we discussed in Chapter 7.

To analyze the robustness of the model, we evaluate the DMVAE in multiple configurations. In Section 9.3.1, we vary the number of modality-specific dimensions, which we previously found to be an important hyperparameter for multimodal VAEs with partitioned latent spaces (Chapter 6). Moreover, in Section 9.3.2, we include an ablation study for the hyperparameter δ_i that controls the disentanglement of shared and modality-specific information.

Evaluation metrics As in the previous chapters, we evaluate the encoding of shared information, the disentanglement of shared and modality-specific information, and the generative performance of the model. To evaluate the encoding of shared information, we use *linear probing* with logistic regression to measure how well the shared information (i.e., the digit label) is encoded in the learned representation. For models with a partitioned latent space (i.e., the MMVAE+ and DMVAE), we evaluate the embeddings of the variational joint posterior that describes the shared latent space. As before, we interpret the classification accuracy of linear probing as a proxy for disentanglement in the sense of linear separability. Additionally, to assess the disentanglement, we use *nonlinear probing* with MLPs to measure how much shared information is encoded in the modality-specific embeddings. To assess the generative performance, we consider both *generative quality* (measured in terms of FID) and *semantic coherence* for the conditional generation (measured in terms of generative coherence) to discern potential tradeoffs between these performance criteria (c.f., Section 7.1.1). For more information on the evaluation metrics, see Section 2.5.

Implementation details For the baselines, we employ the same configurations as in the previous chapters. For the MVAE, MMVAE, and MoPoE-VAE, we use the configurations described in Chapter 7, and for the MMVAE+, we use the same configuration as in Chapter 6. We employ ResNets [He+16] for the encoders and decoders of all models. The

DMVAE is trained for 1000 epochs in each training stage. In the first stage, we use a batch size of $K = 1024$ for contrastive learning; in the second stage, we set $\delta_i = 400$ for each modality $i \in \{1, \ldots, M\}$, based on the ablation study in Section 9.3.2.

For unconditional generation with the DMVAE, one needs to define a prior distribution for the variational joint posterior. In principle, we could include a prior distribution in the formulation. However, we found it to be more convenient to fit a Gaussian mixture model (GMM) on the embeddings of the training data and use the resulting model as a prior distribution (c.f., [Gho+20]). Specifically, we used a GMM with 100 mixture components, though we found that even a single component would suffice to achieve similar results.

9.3.1 Translated-PolyMNIST

In the following, we present the results on the Translated-PolyMNIST dataset. First, we describe the quantitative results for all models and present the qualitative results for the DMVAE. For the baselines, we reference the qualitative results from the previous chapters.

Quantitative results Table 9.1 presents the quantitative evaluation for all models. In terms of linear probing, the DMVAE shows a significantly better performance compared to the baselines. In fact, we find that a logistic regression can predict the shared information (i.e., the digit label) *perfectly* from the embeddings produced by the model. Even more strikingly, we observe a marked improvement in generative coherence, as the DMVAE exhibits a coherence accuracy between 63–69% for the conditional generation of missing modalities. In contrast, the other models perform close to the chance-level of 10% on the Translated-PolyMNIST dataset, as previously highlighted in Chapter 7. Finally, we observe that the DMVAE's generative quality for unconditional generation is on par with the performance of the MVAE and thus it approaches the feasible limit of what can be achieved by unimodal VAEs. To recognize this, we can directly compare the FID values (for which lower is better) to the results in Figure A.1.

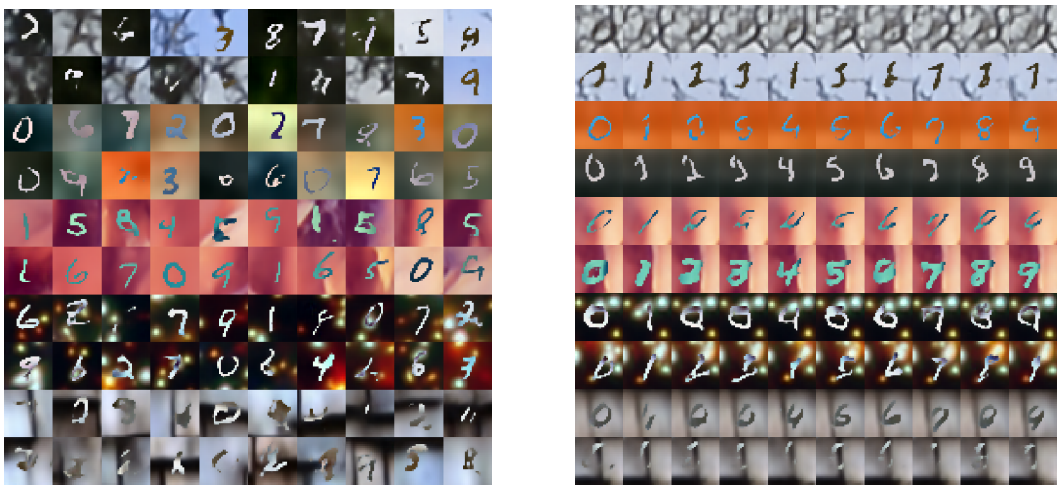
Overall, in both configurations of the DMVAE, the model exhibits a favorable tradeoff in terms of generative quality and coherence compared to the baselines. Nevertheless, we also notice a tradeoff for the DMVAE when we increase the dimensionality of the modality-specific latent space from 32 to 64 dimensions. We find that the generative quality slightly improves while the coherence decreases by about 6–9 percentage points.

	Linear probing	Generative coherence	Generative quality
MVAE	0.828 (±0.007)	0.146 (±0.007)	**54.600** (±0.764)
MMVAE	0.388 (±0.010)	0.114 (±0.001)	168.887 (±2.099)
MoPoE-VAE	0.837 (±0.009)	0.123 (±0.001)	107.320 (±0.871)
MMVAE+	0.171 (±0.010)	0.111 (±0.001)	91.497 (±3.563)
DMVAE (dim=32)	**1.000** (±0.000)	**0.692** (±0.007)	57.721 (±1.447)
DMVAE (dim=64)	**1.000** (±0.000)	0.630 (±0.020)	57.220 (±1.273)

Table 9.1: Evaluations on the Translated-PolyMNIST dataset. For linear probing, we evaluate the classification accuracy of a logistic regression model trained to predict the shared information (i.e., the digit label) from the learned embeddings sampled from the variational joint posterior conditioned on the full set of modalities. For the generative coherence, we compute the leave-one-out conditional coherence accuracy averaged across all modalities. Only for the MMVAE and MMVAE+, we report the pairwise conditional coherence accuracy, because they cannot aggregate information across multiple modalities. To assess generative quality, we compute the average FID (lower is better) across all modalities generated from prior samples. For the DMVAE, we include two configurations, where "dim" denotes the dimensionality of the modality-specific space. For each model, the results are averaged over three seeds and in parentheses we show the standard deviation.

Qualitative results To supplement the quantitative results, in Figure 9.1 we examine the generated samples produced by the DMVAE. Figure 9.1a shows the unconditionally generated samples, for which we observe a decent quality and diversity, significantly better than for the MMVAE and MoPoE-VAE and on par with the MVAE and unimodal VAEs (e.g., compare to the results in Figure A.3, subplots (e) to (h)).

Figure 9.1b shows the conditionally generated samples, specifically the leave-one-out conditional generation of missing modalities. Notably, the DMVAE generates coherent samples across modalities, while the baselines fail to accomplish this task on the Translated-PolyMNIST dataset (e.g., see Figure A.4, subplots (d) to (f)). Specifically, the DMVAE generates coherent samples with respect to both the inferred shared information (i.e., consistent digits along each column) and the inferred modality-specific information. The latter manifests in terms of a coherent "style" in the generated images, as witnessed by the consistent background and digit appearance (e.g., the location, slope and thickness of the digits) along each row of Figure 9.1b.



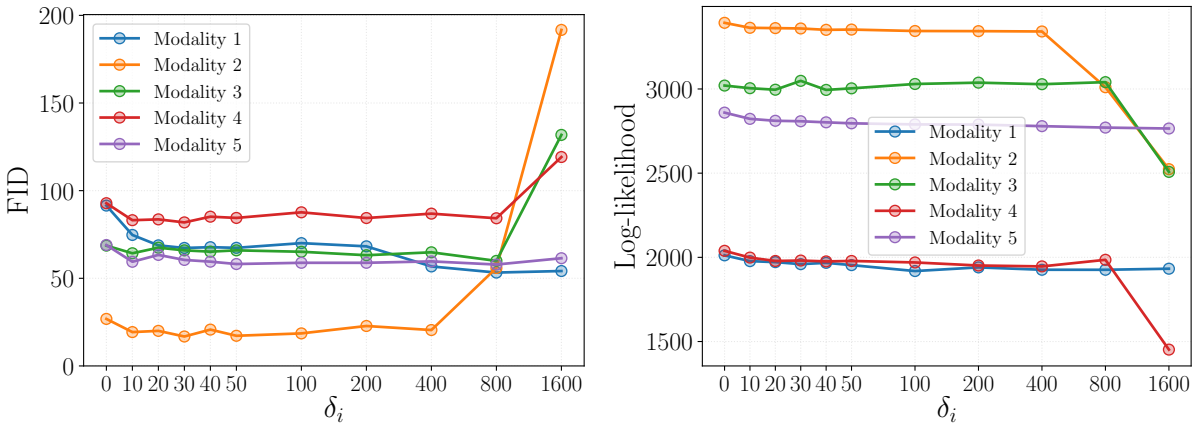
(a) Unconditional generation (b) Conditional generation

Figure 9.1: Qualitative results for the DMVAE trained on the Translated-PolyMNIST dataset. In Figure 9.1a, we showcase the unconditional generation with 20 samples for each of the five modalities, generated using modality-specific embeddings drawn from the prior distribution and shared embeddings drawn from a GMM trained on the embeddings of the training dataset. In Figure 9.1b, we present qualitative results for the (leave-one-out) conditional generation, showing 20 samples for each modality. Along each column, we keep the shared embedding fixed to demonstrate the coherence with respect to the inferred shared information; likewise, for each row, we keep the modality-specific embedding fixed to demonstrate the coherence with respect to the inferred modality-specific information.

9.3.2 Ablation Study

In this subsection, we investigate the effects of the hyperparameter δ_i that controls the regularization of the DMVAE (see Stage 2). Recall that the parameter weights the estimate of the mutual information between the inferred shared and modality-specific information independently for each modality $i \in \{1, \dots, M\}$. Thus, its purpose is to incentivize the disentanglement of shared and modality-specific information by suppressing the encoding of shared information for the VAEs.

For the ablation study, we explore a large range of values, $\delta_i \in [0, 1600]$, where $\delta_i = 0$ corresponds to training a conditional VAE without additional regularization. Ideally, the value of δ_i should be selected based on a metric that does not require labels that indicate what is shared between modalities, since we generally assume that this information



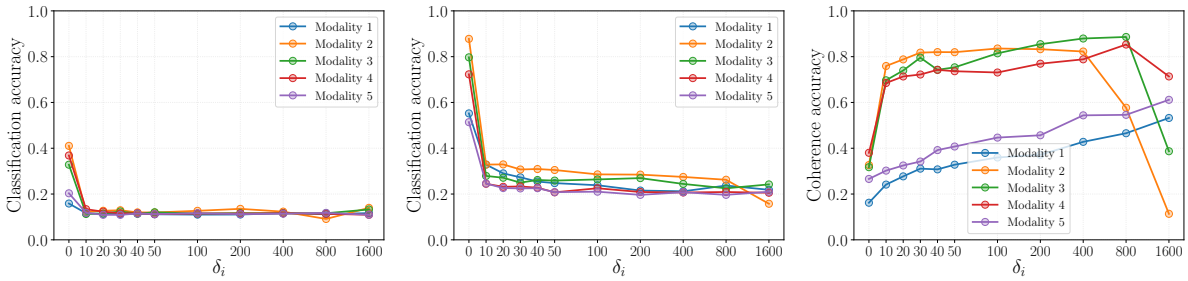
(a) Unconditional generation

(b) Log-likelihood

Figure 9.2: Ablation study for the DMVAE with respect to the hyperparameter δ_i that weights the regularization term. Figure 9.2a shows the generative quality for the unconditional generation in terms of FID (lower is better). Figure 9.2b shows the log-likelihood values of the models on holdout data (higher is better). Each point represents a different model trained with the respective value of δ_i. On the horizontal axis, we use a linear scale for the range $[0, 50]$ and a logarithmic scale for values larger than 50.

is not available in the considered setup. Therefore, we first assess two "unsupervised" metrics (FID and log-likelihood) before we evaluate metrics that require ground truth labels (latent probing and semantic coherence).

FID and log-likelihood In Figure 9.2, we present the results of the ablation study in terms of "unsupervised" metrics, which estimate how well the generative model approximates the data distribution. First, we assess the generative quality of the unconditional generation in terms of FID (Figure 9.2a). We find that the performance of the model in terms of FID remains relatively stable up to a value of $\delta_i = 400$, beyond which it starts to diverge for some modalities. In Figure 9.2b, we observe a similar trend in terms of the log-likelihood, which remains relatively stable up to the same value of $\delta_i = 400$, beyond which it deteriorates for individual modalities. It is a promising sign to observe a common trend across the two metrics, because the FID correlates well to the perceived quality of generated images [Heu+17], whereas the log-likelihood is modality-agnostic but does not always faithfully reflect the quality of samples [TOB16].



(a) Linear probing (b) Nonlinear probing (c) Generative coherence

Figure 9.3: Ablation study for the DMVAE with respect to the hyperparameter δ_i that weights the regularization term. In contrast to Figure 9.2, the evaluation metrics shown here require ground truth labels. In Figures 9.3a and 9.3b, we assess how much shared information is encoded in the modality-specific embeddings using linear and nonlinear probing and evaluate the classification accuracy respectively. Notably, lower values indicate a better disentanglement. We show the results for linear probing using logistic regression (Figure 9.3a) and nonlinear probing using MLP classifiers (Figure 9.3b). In Figure 9.3c, we evaluate the semantic coherence for the generation of missing modalities in terms of the leave-one-out conditional coherence accuracy (higher is better). Each point represents a different model trained with the respective value of δ_i. On the horizontal axis, we use a linear scale for the range $[0, 50]$ and a logarithmic scale for values larger than 50.

Latent probing and semantic coherence In Figure 9.3, we show the results of the ablation study in terms of metrics that, to be computed, require ground truth labels that indicate what is shared between modalities. First, we assess the disentanglement using linear and nonlinear probing to predict the shared information (i.e., the digit label) from the modality-specific embeddings. Thereby, we estimate how well shared information can be extracted from the modality-specific embeddings and thus a lower value indicates a better disentanglement. Second, we evaluate the semantic coherence for the conditional generation of missing modalities in terms of the leave-one-out generative coherence.

We show the results for linear probing using logistic regression (Figure 9.3a) as well as nonlinear probing using MLP classifiers (Figure 9.3b). Generally, we find that with values of $\delta_i > 0$, the encoding of shared information is suppressed effectively. Even with $\delta = 10$, significantly less shared information can be recovered from the modality-specific embeddings compared to the model without regularization (i.e., $\delta = 0$). Nevertheless, when we use nonlinear probing, we notice that the modality-specific embeddings still encode shared

information to a small but measurable extent ($\leq 30\%$ classification accuracy). In terms of the generative coherence (Figure 9.3c), which estimates the semantic coherence for the conditional generation across modalities, we observe that the performance improves up to a value of $\delta_i = 400$, beyond which it decreases for individual modalities—consistent with the results of the FID and log-likelihood evaluation.

Overall, for the ablation study, the consistency of the results across all considered metrics is a promising sign, suggesting that model selection with respect to δ_i might be feasible without ground truth labels that indicate what is shared between modalities.

9.4 Summary

In this chapter, we presented a hybrid approach for multimodal learning that combines generative and discriminative approaches. We proposed the DMVAE as a multimodal generative model that approximates the joint distribution of multiple modalities and disentangles shared and modality-specific information effectively.

The proposed model integrates the ideas and insights from the previous chapters. First, the model leverages the effectiveness of contrastive learning to infer information that is shared between modalities, which builds on our results from Chapter 8. Second, it overcomes some of the limitations of multimodal VAEs that we discussed in Chapter 7. Specifically, it reduces the gap in generative quality between unimodal and multimodal VAEs, while improving the conditional generation of missing modalities—even in the challenging setup where shared information cannot be predicted in expectation across modalities on the level of observations, as showcased by our results on the Translated-PolyMNIST dataset. Nevertheless, more work is required to assess the effectiveness in real-world applications.

In the scope of the thesis, this chapter touches on each of the considered research questions. It provides a model that leverages weak supervision to infer latent factors of variation shared between modalities (Question 1), to disentangle shared and modality-specific information (Question 2), and to generate missing modalities based on the learned representations (Question 3). Thereby, this chapter provides a demonstration and outlook for how multimodal representation learning can benefit from a hybrid approach. In the following chapter, we conclude this thesis with a discussion of the results and limitations of our work.

10

Discussion and Conclusion

In this final chapter, we take a bird's eye perspective as we revisit the core aspects of the thesis, reflect on the significance of our results, and highlight potential research directions. First, we summarize our contributions and key insights in the context of our research goals (Section 10.1). Second, we do not only consolidate the contributions made but also discuss the limitations of our work and identify opportunities for future research (Section 10.2). Finally, we close the discussion and conclude the thesis (Section 10.3).

10.1 Summary of Contributions

We considered the problem of multimodal learning under weak supervision—a challenging setup, where we are given a dataset comprised of corresponding observations of different modalities without labels for what is shared between them. For this setup, we designed machine learning methods to improve the performance on a set of tasks by leveraging information from multiple modalities. Specifically, we developed techniques to encode and identify latent factors shared between modalities (Question 1), to disentangle shared and modality-specific information (Question 2), and to learn generative models of multimodal data in a way that enables inference across modalities (Question 3). By formalizing the problem setup, exploring and designing suitable methods, and conducting extensive

empirical evaluations, we contributed meaningful insights to representation learning with multiple modalities. In the following, we summarize the main contributions.

10.1.1 Encoding and Identification of Shared Factors

First, we made several contributions to the encoding and identification of latent factors shared between modalities. This corresponds to our goal specified in Question 1.

Throughout this thesis, we evaluated the learned representations of different models by using auxiliary classification or regression tasks with respect to the ground truth factors. Specifically, we used the prediction performance (i.e., classification accuracy for discrete and the coefficient of determination for continuous factors) to measure how well the learned representation encodes shared information (c.f., Section 2.5).

First, we tackled the problem using generative models. In Chapter 5, we proposed a multimodal VAE for variational inference and density estimation on sets of modalities. We found that by using all subsets of modalities for the posterior approximation our model produces a better encoding of shared information compared to state-of-the-art approaches and consequently improves the generation of missing modalities based on the learned representations. On the contrary, in Chapter 7, we observed that multimodal VAEs fail to encode shared information in a linearly separable format if we consider a more challenging setup where shared information cannot be predicted in expectation across modalities on the level of observations.

To address the issue, in Chapters 8 and 9 we demonstrated that discriminative approaches provide a viable alternative to infer shared information even in the challenging setup that restricts multimodal VAEs. We formally showed that contrastive learning can recover shared factors that are invariant across modalities up to a block-wise indeterminacy and corroborated our findings with numerical simulations and complex multimodal datasets.

Thus, we have addressed Question 1 theoretically and empirically, demonstrating how different approaches can be used to encode and, in some cases, provably recover latent factors of variation shared between modalities. Thereby, our work provides a stepping stone for multimodal integration and techniques for learning an abstract, modality-agnostic representation of the environment using machine learning.

10.1.2 Disentanglement of Shared and Modality-specific Factors

We also made several contributions with respect to the second goal (Question 2), i.e., the disentanglement of shared and modality-specific information. The disentanglement poses an additional challenge because it requires that a model infers modality-specific information in a form that separates it from shared information.

As for the encoding of shared information, we evaluated the learned representations by using auxiliary classification or regression tasks with respect to the ground truth factors to measure disentanglement. Primarily, we used *linear probing* (i.e., linear regression or logistic classification), for which a high prediction performance implies that the respective information is encoded in a linearly separable format (c.f., Section 2.5). Additionally, for models with a partitioned latent space, we used separate predictors for the shared and modality-specific subspaces to assess whether each subspace encodes the relevant information. For some experiments, modality-specific information was not annotated. In these cases, we used qualitative and quantitative evaluations to assess the disentanglement, specifically through the generation of missing modalities based on the learned representations.

For generative learning with multimodal VAEs, we developed models that outperform existing baselines in terms of linear probing and the generation of missing modalities. In Chapter 5, we found that using all subsets of modalities for the posterior approximation can improve these performance criteria compared to baseline approaches that use only certain subsets of modalities. In Chapter 6, we devised a model with a partitioned latent space, for which linear probing showed that shared information was encoded primarily in the shared and not in the modality-specific subspaces. Consequently, the disentanglement through partitioning also improved the generative performance compared to models with a joint latent space. Finally, in Chapter 9, we developed a hybrid model that infers shared information using contrastive learning and modality-specific information using VAEs and thereby demonstrated further improvements in terms of linear probing and the generation of missing modalities. Notably, even with *nonlinear* probing, scarcely any shared information could be extracted from the modality-specific embeddings produced by the model.

Hence, to address Question 2, we presented generative and discriminative models that leverage weak supervision to encode shared and modality-specific information in a disentangled format. Thereby, we made progress towards the inference of modality-specific factors of variation, i.e., not only factors that are shared between modalities.

10.1.3 Contributions to Multimodal Generative Models

The last key aspect was the development of generative models that can draw inferences across modalities (Question 3). In this respect, we made the following contributions.

To assess the generative performance of multimodal generative models, we used two types of metrics. To assess the *generative quality*, we evaluated the log-likelihood of the models and additionally computed the Fréchet inception distance (FID) for the conditional and unconditional generation. To measure the *semantic coherence* across modalities, we evaluated the generative coherence (c.f., Section 2.5) with classifiers that were trained using ground truth labels.

Based on the framework of the variational autoencoder, we introduced several new types of multimodal VAEs. In Chapter 5, we developed the MoPoE-VAE, which generalizes two existing, widely-used approaches and found that it improved the conditional generation of missing modalities compared to these baselines. Second, in Chapter 6, we designed multimodal VAEs with a partitioned latent space and discussed the benefits and pitfalls of partitioning. Based on these insights, we developed the MMVAE+, which promotes the disentanglement for the respective subspaces and consequently improves the generative performance compared to models with a joint latent space.

On the flip side, in Chapter 7, we found that the sub-sampling of modalities enforces an undesirable bound for the approximation of the joint distribution, which limits the generative performance of mixture-based multimodal VAEs and constrains their application to settings where shared information can be predicted in expectation across modalities. In this context, we highlighted the tradeoffs between existing models in terms of generative quality and generative coherence with an empirical study comprising multiple datasets.

Finally, in Chapter 7, we developed the DMVAE as a hybrid approach that combines contrastive learning with conditional VAEs. In a proof of concept, we demonstrated promising results in comparison to multimodal VAEs in a setting where shared information *cannot* be predicted across modalities on the level of observations. Specifically, we found that the model reduces the gap in generative quality compared to unimodal VAEs and shows substantial improvements for the generation of missing modalities.

Thus, to address Question 3, we developed several techniques for multimodal generative learning that can draw inferences across modalities based on the learned representations. Yet, we also established fundamental limitations and tradeoffs for the proposed models.

10.2 Limitations and Future Directions

In this section, we discuss the limitations of our results and developed methods. By thoroughly examining the limitations, we aim to promote a comprehensive understanding of the scope of our research and to identify fruitful directions for future work.

Real-world applications While we used various datasets to evaluate the developed models and theoretical results, there is more work required to identify practical use cases and to develop impactful real-world applications.

Throughout this thesis, we mainly used synthetic datasets and numerical simulations, though we also included common machine learning datasets, such as CelebA and CUB, to showcase more realistic use cases. However, in many real-world applications with multimodal data, we encounter significantly more complex datasets. For example, in robotics applications, we might face continuous streams of data from multiple modalities (e.g., cameras, microphones, and tactile sensors). Thus, to draw definitive conclusions about the utility of our results and methods for real-world applications, more dedicated effort and research in the particular application domain is required.

Nevertheless, we can offer a perspective on the potential impact of our results. For example, multimodal VAEs have been successfully used in biomedical applications [Dor+19; LS21; Min+21; GZP21]. In these or similar applications, we would expect to see further improvements based on our methods and insights. Additionally, our contributions with respect to the limitations of multimodal VAEs can help practitioners identify suitable applications for the existing approaches. Finally, we also presented new results to better understand the effectiveness of existing methods, such as contrastive learning, which are already widely adopted in many applications (e.g., [Rad+21]).

As a concrete example for future work, we envision applications in medicine and healthcare. Medical examinations often produce multiple types of measurements of different modalities (e.g., vital signs, blood tests, medical images). Notably, these measurements offer a natural source of weak supervision when they belong to the same patient. Our methods and insights could be used to develop tools that use information from multiple modalities to learn meaningful representations that assess the health status of a patient. Even without labels that describe the medical condition, our methods and insights can be applied to infer disease characteristics that are expressed in different modalities.

Methodological constraints To address the research questions, we provided theoretical justifications and conducted a wide range of experiments. Yet, there are still possible methodological limitations that we identified in our research.

In the context of generative learning with multimodal VAEs, we discussed the limitations in a dedicated chapter (Chapter 7), but further points need to be addressed. To evaluate the learned representations, we mainly used linear probing and measures of generative quality and coherence. However, we did not formally analyze the identifiability of the latent variables and only provided empirical results to assess the quality of learned representations.

Conversely, in Chapter 7, we did not establish theoretical limitations for the learned representations but merely for the generative capabilities of multimodal VAEs. For future work, it would be interesting to formally analyze the (non-)identifiability of latent variables for multimodal VAEs. Moreover, it might be worthwhile to examine whether the limitations we have established also apply to other types of generative models. Specifically, in light of the promising developments in conditional generative learning with transformers [Vas+17] and diffusion models [Soh+15; SE19; HJA20; Son+21] that led to remarkable improvements in image-to-text generation (e.g., [Ram+21; Ram+22; Rom+22]). Thus, it might be worth investigating whether different types of generative models could be combined or substituted.

In the context of contrastive learning, our theoretical results from Chapter 8 rest on some non-trivial but justifiable assumptions that we discuss in Section 8.5. Specifically, we would like to highlight that our results only consider the case of *two* modalities. While it might be sufficient to use pairs of modalities in many applications, it would be interesting to investigate whether the theoretical results could be extended beyond pairs of modalities. Additionally, our results assumed content invariance (Assumption 4), which might not be satisfied in real-world settings, e.g., due to measurement errors or other inaccuracies in the data collection. For future work, it might be interesting to consider relaxations of the assumption; e.g., in the setup of sequential decision making, as described in Appendix A.8.

Overall, despite the potential limitations regarding real-world applications and the methodological constraints of our work, we can reasonably infer that our methods and findings address the research questions posed in Section 1.2. Thus, we believe that our work provides a contribution to the overarching goal of discovering latent structure from low-level observations by leveraging weak supervision in the form of multiple modalities.

10.3 Conclusion

In this thesis, we investigated the topic of multimodal representation learning in the setup of weak supervision. We developed machine learning methods that leverage statistical dependencies between observations of different modalities to draw inferences about the latent factors of variation, particularly those shared between modalities. Employing both generative and discriminative approaches, we analyzed existing and proposed novel methods that transform observations into meaningful representations of shared and modality-specific information without explicit supervision by labels. Specifically, we devised methods that can encode and disentangle shared and modality-specific information and in some cases provably recover latent factors up to acceptable ambiguities. Throughout this work, we demonstrated that learned representations satisfying these properties can improve the performance on downstream tasks, as showcased by the generation of missing modalities based on the learned representations.

Thus, we established a theoretical basis for multimodal representation learning and explained in which settings generative and discriminative approaches can be effective in practice. More broadly, we contributed to the discovery of latent structure from low-level observations by leveraging weak supervision in the form of multiple modalities.

Appendix

A.1 Derivation of Equation (6.16)

In the following, we explain the individual steps used to derive Equation (6.16).

We start from Equation (6.15), to which we first apply the definition of the KL-divergence and then use the product rule for logarithms to group the corresponding terms as follows:

$$D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \,||\, p(\mathbf{c}) \prod_{j=1}^{M} p(\mathbf{m}_j)\Big) \tag{A.1}$$

$$= \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)}{p(\mathbf{c}) \prod_{j=1}^{M} p(\mathbf{m}_j)} \right] \tag{A.2}$$

$$= \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} + \sum_{j=1}^{M} \log \frac{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)}{p(\mathbf{m}_j)} \right] . \tag{A.3}$$

Then, we use the linearity of expectation and simplify the individual terms as follows:

For the first term, we have

$$\mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} \right] \tag{A.4}$$

$$= \int \int \cdots \int q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} \, d\mathbf{c} \, d\mathbf{m}_1 \cdots d\mathbf{m}_M \tag{A.5}$$

$$= \int q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} \underbrace{\int \cdots \int \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \, d\mathbf{m}_1 \cdots d\mathbf{m}_M}_{=1} \, d\mathbf{c} \tag{A.6}$$

$$= \mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})} \left[\log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} \right] . \tag{A.7}$$

Analogously, we apply the same procedure for each of the remaining terms, so that

$$\mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \prod_{j=1}^{M} q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)}{p(\mathbf{m}_j)} \right] = \mathbb{E}_{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)}{p(\mathbf{m}_j)} \right] \tag{A.8}$$

for each $j \in \{1, \ldots, M\}$.

Thus, Equation (A.3) can be expressed as sum of KL-divergences, i.e.,

$$\mathbb{E}_{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})} \left[\log \frac{q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}})}{p(\mathbf{c})} \right] + \sum_{j=1}^{M} \mathbb{E}_{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)} \left[\log \frac{q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j)}{p(\mathbf{m}_j)} \right] \tag{A.9}$$

$$= D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{c}}}(\mathbf{c} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c})\Big) + \sum_{j=1}^{M} D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_j}}(\mathbf{m}_j \mid \mathbf{x}_j) \,||\, p(\mathbf{m}_j)\Big) , \tag{A.10}$$

which completes the derivation of Equation (6.16).

A.2 Proof of Lemma 4

Lemma 4. *Objective* $\mathcal{L}^{MMVAE+}_{ELBO\text{-}Part.}(\bar{\mathbf{x}}; \phi, \theta, \psi)$ *(Equation 6.22) forms a lower bound on* $\log p(\bar{\mathbf{x}})$*, i.e.,*

$$\log p(\bar{\mathbf{x}}) \geq \mathcal{L}^{MMVAE+}_{ELBO\text{-}Part.}(\bar{\mathbf{x}}; \phi, \theta, \psi) \ . \tag{6.23}$$

Proof. We start from the definition of the ELBO for the multimodal VAE with a joint latent space (Definition 5), i.e., the inequality

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z} \mid \bar{\mathbf{x}})}\Big[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{z})\Big] - D_{\mathrm{KL}}\Big(q_\phi(\mathbf{z} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{z})\Big) \tag{A.11}$$

and first adapt the notation to describe a model with a partitioned latent space. Based on the similarity of the graphical models in terms of their conditional independence assumptions (see Figures 6.1a and 6.1b), we rename the latent variable $\mathbf{z}$ to $\mathbf{c}$ and the encoder parameters ϕ_i to $\phi_{\mathbf{c}_i}$. Thus, the right-hand side of Equation (A.11) equals

$$\mathbb{E}_{q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}})}\Big[\log p_\theta(\bar{\mathbf{x}} \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c})\Big) \ . \tag{A.12}$$

Then, we use the conditional independence assumption $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{c}$ to rewrite the first term from Equation (A.12) as a sum of likelihood terms. Thus, we have

$$\sum_{i=1}^{M} \mathbb{E}_{q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}})}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(q_{\phi_\mathbf{c}}(\mathbf{c} \mid \bar{\mathbf{x}}) \,||\, p(\mathbf{c})\Big) \ . \tag{A.13}$$

We plug in the mixture of experts (MoE) decomposition of the variational joint posterior (i.e., Equation 6.18) into Equation (A.13) to obtain

$$\sum_{i=1}^{M} \mathbb{E}_{\frac{1}{M}\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(\tfrac{1}{M}\textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c})\Big) \ , \tag{A.14}$$

which is equal to the objective of the MMVAE [Shi+19] using a slightly different notation.

Next, we use the linearity of expectation to take the weighted sum from the variational joint posterior out of the expectation. Thus, we have

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(\tfrac{1}{M}\textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c})\Big) \ , \tag{A.15}$$

for which we again notice the grouping of likelihood terms into self- and crossmodal-reconstructions, as described in Section 6.3.2.

For self-reconstructions (i.e, likelihood terms for which $i = j$), we use the lower-bound

$$\mathbb{E}_{q_{\phi_{\mathbf{c}_i}}(\mathbf{c}\,|\,\mathbf{x}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] \tag{A.16}$$

$$\geq \mathbb{E}_{q_{\phi_{\mathbf{c}_i}}(\mathbf{c}\,|\,\mathbf{x}_i)}\Big[\mathbb{E}_{q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i \mid \mathbf{c})\Big)\Big] \tag{A.17}$$

$$= \mathbb{E}_{q_{\phi_{\mathbf{c}_i}}(\mathbf{c}\,|\,\mathbf{x}_i)q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] - D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i)\Big)\,. \tag{A.18}$$

Concretely, in Equation (A.17), we use a variational approximation, for which we introduce the variational posterior $q_{\phi_i}(\mathbf{m}_i \mid \mathbf{x}_i)$. Equation (A.18) follows from the assumed independence of shared and modality-specific information (c.f., Figure 6.1), whereby $p(\mathbf{m}_i \mid \mathbf{c}) = p(\mathbf{m}_i)$; consequently, the second term is constant with respect to $\mathbf{c}$.

For cross-reconstructions (i.e., likelihood terms for which $i \neq j$), we instead use the following approximation, for which we introduce the auxiliary prior $q_{\psi_i}(\mathbf{m}_i)$:

$$\mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big] = \mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)}\Big[\log \mathbb{E}_{q_{\psi_i}(\mathbf{m}_i)}[p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})]\Big] \tag{A.19}$$

$$\geq \mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)q_{\psi_i}(\mathbf{m}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c})\Big]\,, \tag{A.20}$$

where the lower bound follows from Jensen's inequality because the logarithm is a concave function. Notably, the first equation holds for any distribution $q_{\psi_i}(\mathbf{m}_i)$ for which $\mathbf{x}_i \perp\!\!\!\perp \tilde{\mathbf{m}}_i | \mathbf{c}$, where $\tilde{\mathbf{m}}_i \sim q_{\psi_i}(\mathbf{m}_i)$ and $\mathbf{c} \sim q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j)$.

Finally, we apply the approximations from Equations (A.18) and (A.20) to the respective likelihood terms in Equation (A.15) and thus obtain

$$\begin{aligned}
&\frac{1}{M}\sum_{i=1}^{M}\Big\{\mathbb{E}_{q_{\phi_{\mathbf{c}_i}}(\mathbf{c}\,|\,\mathbf{x}_i)q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i\,|\,\mathbf{x}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)\Big] \\
&+ \sum_{j=1}^{M} \mathbb{1}_{\{i\neq j\}}\mathbb{E}_{q_{\phi_{\mathbf{c}_j}}(\mathbf{c}\,|\,\mathbf{x}_j)q_{\psi_i}(\mathbf{m}_i)}\Big[\log p_{\theta_i}(\mathbf{x}_i \mid \mathbf{c}, \mathbf{m}_i)\Big] \\
&- D_{\mathrm{KL}}\Big(q_{\phi_{\mathbf{m}_i}}(\mathbf{m}_i \mid \mathbf{x}_i) \,||\, p(\mathbf{m}_i)\Big)\Big\} \\
&- D_{\mathrm{KL}}\Big(\tfrac{1}{M}\textstyle\sum_{j=1}^{M} q_{\phi_{\mathbf{c}_j}}(\mathbf{c} \mid \mathbf{x}_j) \,||\, p(\mathbf{c})\Big)\,,
\end{aligned} \tag{A.21}$$

which corresponds to the MMVAE+ objective from Equation (6.22).

Since Equation (A.15) is a lower-bound on the log-evidence $\log p(\bar{\mathbf{x}})$ and the used approximations (Equations A.18 and A.20) are also lower bounds, it follows that the MMVAE+ objective is a lower bound on the log-evidence, which concludes the proof.

□

A.3 Proof of Corollary 2

Corollary 2. *Let $\mathbf{x}_{M+1}$ be a random vector that describes an additional modality and let $\bar{\mathbf{x}}^+ \coloneqq \bar{\mathbf{x}} \cup \{\mathbf{x}_{M+1}\}$ denote the extended set of modalities. Further, let $\mathcal{S}^+$ denote the model-specific set of subsets of modalities and corresponding mixture coefficients for $\bar{\mathbf{x}}^+$.*

For the MMVAE and MoPoE-VAE, the discrepancy $\Delta(\bar{\mathbf{x}}, \mathcal{S})$ increases, i.e.,

$$\Delta(\bar{\mathbf{x}}^+, \mathcal{S}^+) > \Delta(\bar{\mathbf{x}}, \mathcal{S}) \,, \tag{7.27}$$

if $\mathbf{x}_{M+1}$ is sufficiently diverse in the following sense:

$$\begin{aligned}\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A \in \mathcal{S}} I(\mathbf{x}_{\{1,\ldots,M\}\setminus A}; \mathbf{x}_{M+1} \mid \mathbf{x}_A) \; < \; & \frac{1}{|\mathcal{S}^+||\mathcal{S}|} \sum_{A \in \mathcal{S}} H(\mathbf{x}_A \mid \mathbf{x}_{M+1}) \\ & + \frac{1}{|\mathcal{S}^+|} \sum_{A \in \mathcal{S}} H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \,.\end{aligned} \tag{7.28}$$

Proof. First, note that all subsets from $\mathcal{S}$ are still contained in $\mathcal{S}^+$ but that $\mathcal{S}^+$ contains new subsets in addition to those in $\mathcal{S}$. Further, due to the re-weighting of mixture coefficients, $\mathcal{S}^+$ can have different mixture coefficients for the subsets it shares with $\mathcal{S}$. Let $\mathcal{S}^- \coloneqq \{(A, \omega_A^+) \in \mathcal{S}^+ : A \notin \mathcal{S}\}$. In the following, ω_A^+ denotes the new mixture coefficients, for which typically $\omega_A \neq \omega_A^+$ due to the re-weighting.

We are interested in the change of the discrepancy when we add modality $M+1$, i.e.,

$$\Delta(\bar{\mathbf{x}}^+, \mathcal{S}^+) - \Delta(\bar{\mathbf{x}}, \mathcal{S}) \tag{A.22}$$

$$= \sum_{B \in \mathcal{S}^+} \omega_B^+ \, H(\mathbf{x}_{\{1,\ldots,M+1\}\setminus B} \mid \mathbf{x}_B) - \sum_{A \in \mathcal{S}} \omega_A \, H(\mathbf{x}_{\{1,\ldots,M\}\setminus A} \mid \mathbf{x}_A) \,. \tag{A.23}$$

Rewrite the right-hand side in terms of subsets that are contained in both $\mathcal{S}$ and $\mathcal{S}^+$ and subsets that are only contained in $\mathcal{S}^+$. For this, we decompose the first term as follows:

$$\sum_{B \in \mathcal{S}^+} \omega_B^+ \, H(\mathbf{x}_{\{1,\ldots,M+1\}\setminus B} \mid \mathbf{x}_B) \tag{A.24}$$

$$= \sum_{A \in \mathcal{S}} \omega_A^+ \, H(\mathbf{x}_{\{1,\ldots,M+1\}\setminus A} \mid \mathbf{x}_A) + \sum_{B \in \mathcal{S}^-} \omega_B^+ \, H(\mathbf{x}_{\{1,\ldots,M+1\}\setminus B} \mid \mathbf{x}_B) \tag{A.25}$$

$$\begin{aligned}= & \sum_{A \in \mathcal{S}} \omega_A^+ \, H(\mathbf{x}_{\{1,\ldots,M\}\setminus A} \mid \mathbf{x}_A) + \sum_{A \in \mathcal{S}} \omega_A^+ \, H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \\ & + \sum_{B \in \mathcal{S}^-} \omega_B^+ \, H(\mathbf{x}_{\{1,\ldots,M+1\}\setminus B} \mid \mathbf{x}_B) \,,\end{aligned} \tag{A.26}$$

where the last equation follows from

$$H(\mathbf{x}_{\{1,\dots,M+1\}\setminus A} \mid \mathbf{x}_A) = H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) + H(\mathbf{x}_{M+1} \mid \mathbf{x}_A, \mathbf{x}_{\{1,\dots,M\}\setminus A}) \tag{A.27}$$

$$= H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) + H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \; . \tag{A.28}$$

We can use the decomposition from Equation (A.26) to rewrite the right-hand side of Equation (A.23) by collecting the corresponding terms for $H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A)$:

$$\begin{aligned} &\sum_{A\in\mathcal{S}} (\omega_A^+ - \omega_A) H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) + \sum_{A\in\mathcal{S}} \omega_A^+ \, H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \\ &+ \sum_{B\in\mathcal{S}^-} \omega_B^+ \, H(\mathbf{x}_{\{1,\dots,M+1\}\setminus B} \mid \mathbf{x}_B) \; . \end{aligned} \tag{A.29}$$

Notice that in Equation (A.29) only the first term can be negative, due to the re-weighting of mixture coefficients for terms that do not contain $\mathbf{x}_{M+1}$. Hence, the generative discrepancy can only decrease if the mixture coefficients change in such a way that the first term in Equation (A.29) dominates the other two terms.

For the relevant special case of uniform mixture weights, which applies to both the MMVAE and MoPoE-VAE, we can further decompose Equation (A.29) into (i) information shared between $\bar{\mathbf{x}}$ and $\mathbf{x}_{M+1}$, and (ii) information that is specific to $\bar{\mathbf{x}}$ or $\mathbf{x}_{M+1}$.

Using uniform mixture coefficients $\omega_A = \frac{1}{|\mathcal{S}|}$ and $\omega_A^+ = \frac{1}{|\mathcal{S}^+|}$ for all subsets, we can factor out the coefficients and rewrite Equation (A.29) as follows:

$$\begin{aligned} &\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A\in\mathcal{S}} H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) + \frac{1}{|\mathcal{S}^+|} \sum_{A\in\mathcal{S}} H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \\ &+ \frac{1}{|\mathcal{S}^+|} \sum_{B\in\mathcal{S}^-} H(\mathbf{x}_{\{1,\dots,M+1\}\setminus B} \mid \mathbf{x}_B) \; , \end{aligned} \tag{A.30}$$

where the second term already denotes information that is specific to $\mathbf{x}_{M+1}$. Hence, we decompose the first and last terms corresponding to the criteria (i) and (ii).

For the first term from Equation (A.30), we have

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A\in\mathcal{S}} H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A) \tag{A.31}$$

$$= \left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right) \sum_{A\in\mathcal{S}} \Big\{ H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{x}_{M+1} \mid \mathbf{x}_A) \Big\} \; . \tag{A.32}$$

For the last term from Equation (A.30), we have

$$\frac{1}{|\mathcal{S}^+|} \sum_{B\in\mathcal{S}^-} H(\mathbf{x}_{\{1,\dots,M+1\}\setminus B} \mid \mathbf{x}_B) \tag{A.33}$$

$$= \frac{1}{|\mathcal{S}^+|} \Big\{ H(\bar{\mathbf{x}} \mid \mathbf{x}_{M+1}) + \sum_{A\in\mathcal{S}} \mathbb{1}_{\{(A\cup\{M+1\})\in\mathcal{S}^-\}} H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) \Big\} \; , \tag{A.34}$$

where we can further decompose

$$\frac{1}{|\mathcal{S}^+|}H(\bar{\mathbf{x}} \mid \mathbf{x}_{M+1}) = \frac{1}{|\mathcal{S}^+|}\Big\{H(\bar{\mathbf{x}} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + I(\bar{\mathbf{x}}; \mathbf{x}_A \mid \mathbf{x}_{M+1})\Big\} \tag{A.35}$$

$$= \frac{1}{|\mathcal{S}^+|}\Big\{H(\bar{\mathbf{x}} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + H(\mathbf{x}_A \mid \mathbf{x}_{M+1})\Big\} \tag{A.36}$$

$$= \frac{1}{|\mathcal{S}^+||\mathcal{S}|}\sum_{A\in\mathcal{S}}\Big\{H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + H(\mathbf{x}_A \mid \mathbf{x}_{M+1})\Big\}\,. \tag{A.37}$$

Collecting all corresponding terms from Equations (A.32), (A.34) and (A.37), we can rewrite Equation (A.30) as follows:

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|} + \frac{1}{|\mathcal{S}^+||\mathcal{S}|}\right)\sum_{A\in\mathcal{S}} H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + \tag{A.38}$$

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right)\sum_{A\in\mathcal{S}} I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{x}_{M+1} \mid \mathbf{x}_A) + \tag{A.39}$$

$$\frac{1}{|\mathcal{S}^+|}\sum_{A\in\mathcal{S}} \mathbb{1}_{\{(A\cup\{M+1\})\in\mathcal{S}^-\}} H(\mathbf{x}_{\{1,\dots,M\}\setminus A} \mid \mathbf{x}_A, \mathbf{x}_{M+1}) + \tag{A.40}$$

$$\frac{1}{|\mathcal{S}^+||\mathcal{S}|}\sum_{A\in\mathcal{S}} H(\mathbf{x}_A \mid \mathbf{x}_{M+1}) + \tag{A.41}$$

$$\frac{1}{|\mathcal{S}^+|}\sum_{A\in\mathcal{S}} H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}})\,. \tag{A.42}$$

For both the MMVAE and MoPoE, the first and last terms cancel out, which can see by plugging in the respective definitions of $\mathcal{S}$ into the above equation. Recall that for the MMVAE, $\mathcal{S}$ is comprised of the set of unimodal subsets $\{\{1\},\dots,\{M\}\}$ and thus $\mathcal{S}^+$ is comprised of $\{\{1\},\dots,\{M+1\}\}$. For the MoPoE-VAE, $\mathcal{S}$ is comprised of the powerset $\mathcal{P}(M)$ and thus $\mathcal{S}^+$ is comprised of the powerset $\mathcal{P}(M+1)$.

Hence, for the MMVAE and MoPoE-VAE, we have shown that $\Delta(\bar{\mathbf{x}}^+, \mathcal{S}^+) - \Delta(\bar{\mathbf{x}}, \mathcal{S})$ is equal to the following expression:

$$\left(\frac{1}{|\mathcal{S}^+|} - \frac{1}{|\mathcal{S}|}\right)\sum_{A\in\mathcal{S}} I(\mathbf{x}_{\{1,\dots,M\}\setminus A}; \mathbf{x}_{M+1} \mid \mathbf{x}_A) + \tag{A.43}$$

$$\frac{1}{|\mathcal{S}^+||\mathcal{S}|}\sum_{A\in\mathcal{S}} H(\mathbf{x}_A \mid \mathbf{x}_{M+1}) + \frac{1}{|\mathcal{S}^+|}\sum_{A\in\mathcal{S}} H(\mathbf{x}_{M+1} \mid \bar{\mathbf{x}}) \tag{A.44}$$

where the information is decomposed into (i) information shared between $\mathbf{x}$ and $\mathbf{x}_{M+1}$ (Equation A.43) and (ii) information that is specific to $\mathbf{x}$ or $\mathbf{x}_{M+1}$ (the first and second terms in Equation (A.44), respectively), where only (i) can be negative since $|\mathcal{S}^+| > |\mathcal{S}|$.

This concludes the proof of Corollary 2, showing that $\Delta(\bar{\mathbf{x}}^+, \mathcal{S}^+) - \Delta(\bar{\mathbf{x}}, \mathcal{S}) > 0$, if $\mathbf{x}_{M+1}$ is sufficiently diverse in the sense that $(ii) > (i)$.

□

white	yellow	blue	red		green	gray	brown	black
$[0, 0, 120]$	$[25, 50, 70]$	$[90, 50, 70]$	$[0, 50, 70]$	$[159, 50, 70]$	$[36, 50, 70]$	$[0, 0, 50]$	$[24, 255, 255]$	$[0, 0, 0]$
$[180, 18, 255]$	$[35, 255, 255]$	$[158, 255, 255]$	$[15, 255, 255]$	$[180, 255, 255]$	$[89, 255, 255]$	$[180, 18, 120]$	$[16, 50, 70]$	$[180, 255, 50]$

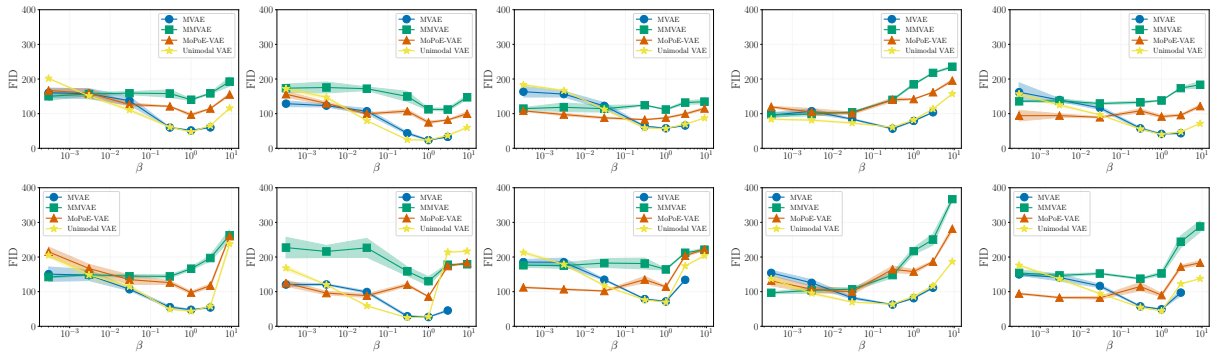
Table A.1: HSV color ranges used to assign pixels to color classes. For each color, we report the lower and upper limits. To include all tonalities of *red*, we consider two distinct ranges.

A.4 Coherence Estimation for CUB Image-Captions

To estimate the semantic coherence for the CUB Image-Captions dataset, we use a proxy to evaluate the coherence for caption-to-image generation. We construct eight captions of the form "*this bird is completely* [*color*]", where [*color*] takes values in the set {*white, yellow, red, blue, green, grey, brown, black*}. We divide the HSV color range according to these eight color classes (see Table A.1). For each of the constructed captions, we generate ten images and count how many pixels belong to each color class. We label an image as *coherent* if the color for the given caption is among the two classes with the highest pixel count. We consider *two* color classes because the highest count for some images might be the background color rather than the color of the bird. Finally, the ratio of coherent images over the total number of generated images is our coherence metric.

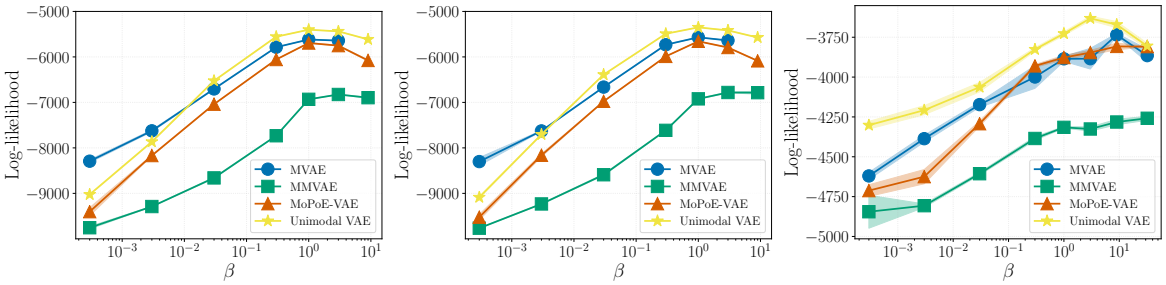
A.5 Additional Results for Chapter 7

Log-likelihoods and qualitative results Figure A.2 shows the generative quality in terms of joint log-likelihoods. We observe a similar ranking of models as with FID, but we notice that the gap between MVAE and MoPoE-VAE appears less pronounced. The reason for this discrepancy is that, to be consistent with Chapter 5, we estimate joint log-likelihoods given *all* modalities—a procedure that resembles reconstruction more than it does unconditional generation. It can be of independent interest that log-likelihoods might overestimate the generative quality for unconditional generation for certain types of models. Qualitative results for unconditional generation (Figure A.3) support the hypothesis that the presented log-likelihoods do not reflect the visible lack of generative quality for the MoPoE-VAE. Further, qualitative results for conditional generation (Figure A.4) indicate a lack of diversity for both the MMVAE and MoPoE-VAE—even though we draw different samples from the posterior, the respective conditionally generated images (i.e., the ten images along each column) show little diversity in terms of backgrounds or writing styles.



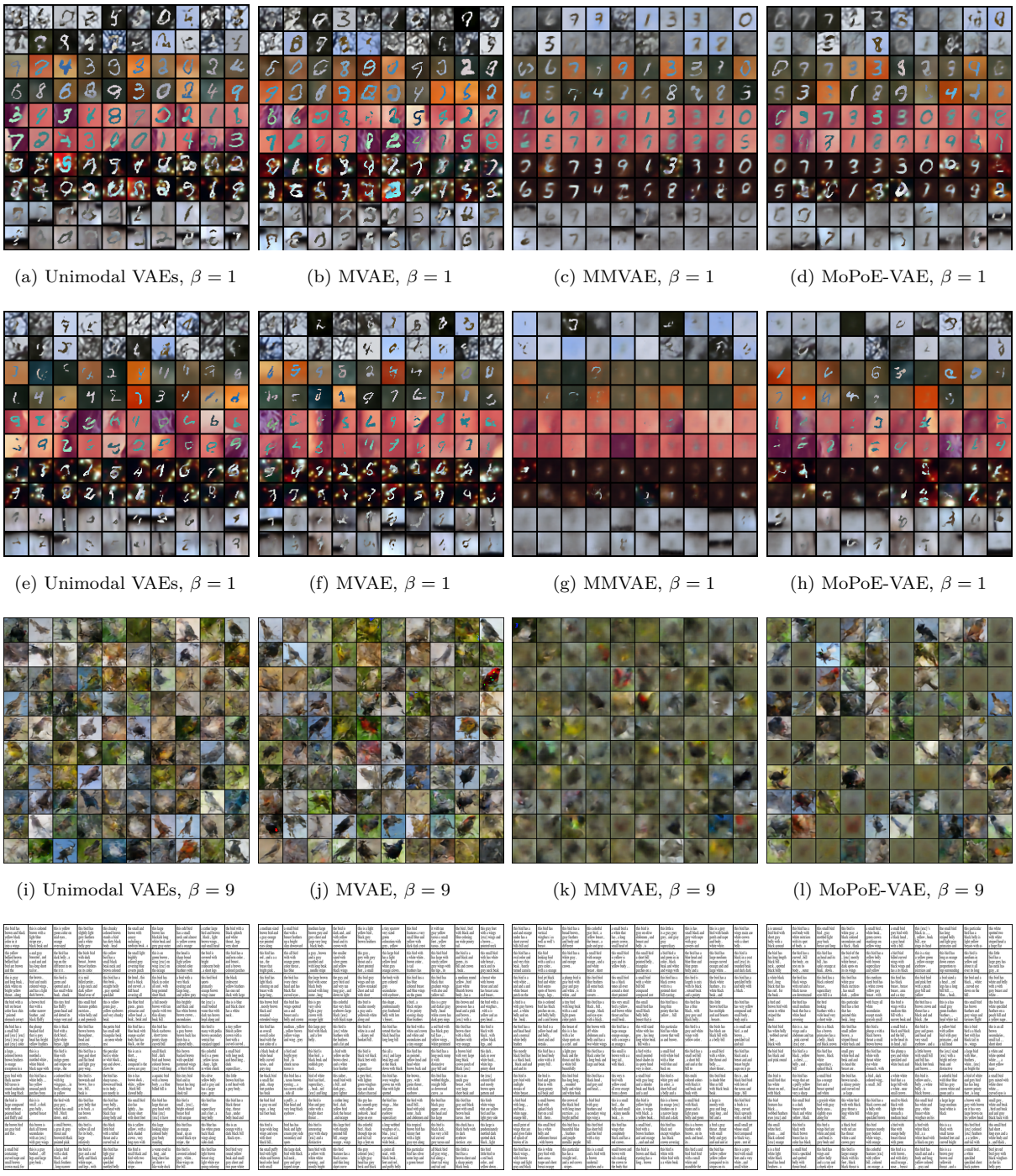
(a) Modality 1 (b) Modality 2 (c) Modality 3 (d) Modality 4 (e) Modality 5

Figure A.1: FID values for each of the five modalities. The top row shows the FIDs for PolyMNIST and the bottom row for Translated-PolyMNIST respectively. Points denote the FID averaged over three seeds and bands show one standard deviation respectively.



(a) PolyMNIST (b) Translated-PolyMNIST (c) CUB Image-Captions

Figure A.2: Joint log-likelihoods over a range of β values. Each point denotes the joint log-likelihood of the respective model on the test set as an average over three seeds and the bands show one standard deviation respectively.



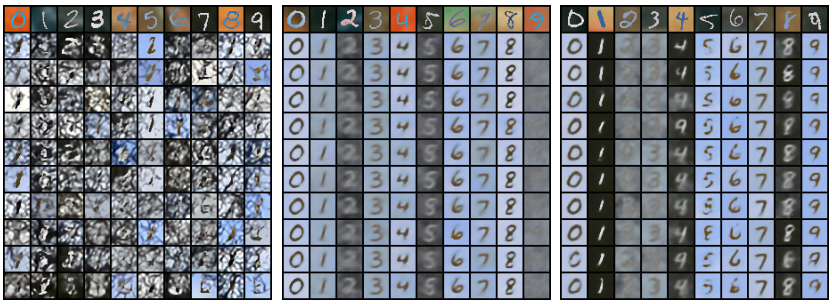
(a) Unimodal VAEs, $\beta = 1$ (b) MVAE, $\beta = 1$ (c) MMVAE, $\beta = 1$ (d) MoPoE-VAE, $\beta = 1$

(e) Unimodal VAEs, $\beta = 1$ (f) MVAE, $\beta = 1$ (g) MMVAE, $\beta = 1$ (h) MoPoE-VAE, $\beta = 1$

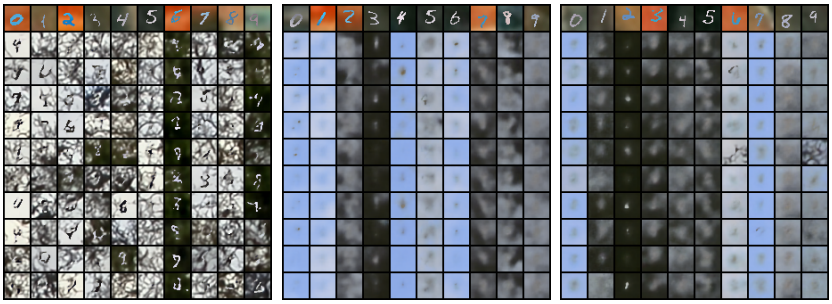
(i) Unimodal VAEs, $\beta = 9$ (j) MVAE, $\beta = 9$ (k) MMVAE, $\beta = 9$ (l) MoPoE-VAE, $\beta = 9$

(m) Unimodal VAEs, $\beta = 9$ (n) MVAE, $\beta = 9$ (o) MMVAE, $\beta = 9$ (p) MoPoE-VAE, $\beta = 9$

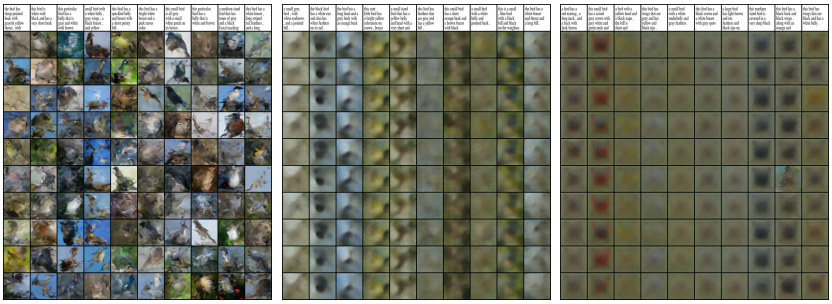
Figure A.3: Qualitative results for the unconditional generation using prior samples. For PolyMNIST (Figures (a) to (d)) and Translated-PolyMNIST (Figures (e) to (h)), we show 20 samples for each modality. For CUB Image-Captions, we show 100 generated images (Figures (i) to (l)) and 100 generated captions (Figures (m) to (p)) respectively.



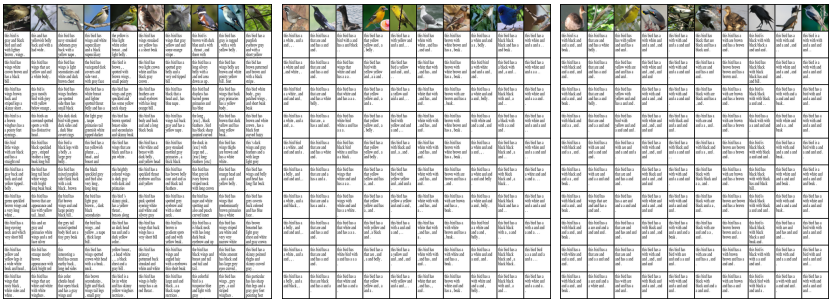
(a) MVAE, $\beta = 1$ (b) MMVAE, $\beta = 1$ (c) MoPoE-VAE, $\beta = 1$



(d) MVAE, $\beta = 1$ (e) MMVAE, $\beta = 1$ (f) MoPoE-VAE, $\beta = 1$



(g) MVAE, $\beta = 9.0$ (h) MMVAE, $\beta = 9.0$ (i) MoPoE-VAE, $\beta = 9.0$



(j) MVAE, $\beta = 9.0$ (k) MMVAE, $\beta = 9.0$ (l) MoPoE-VAE, $\beta = 9.0$

Figure A.4: Qualitative results for the conditional generation of missing modalities. For PolyMNIST (Figures (a) to (c)) and Translated-PolyMNIST (Figures (d) to (f)), we show 10 conditionally generated samples of the first modality given the second modality. Inputs are shown in the first row of the respective subfigure. For CUB Image-Captions, we show the generation of images given captions (Figures (g) to (i)) as well as the generation of captions given images (Figures (j) to (l)).

A.6 Additional Results for Chapter 8

Parameter	Value
Generating function	3-layer MLP
Encoder	7-layer MLP
Optimizer	Adam
Cond. threshold ratio	1e-3
Dimensionality d	15
Batch size	6144
Learning rate	1e-4
Temperature τ	1.0
# Seeds	3
# Iterations	300 000
Similarity metric	Euclidean
Gradient clipping	2-norm; max value 2

(a) Numerical simulation

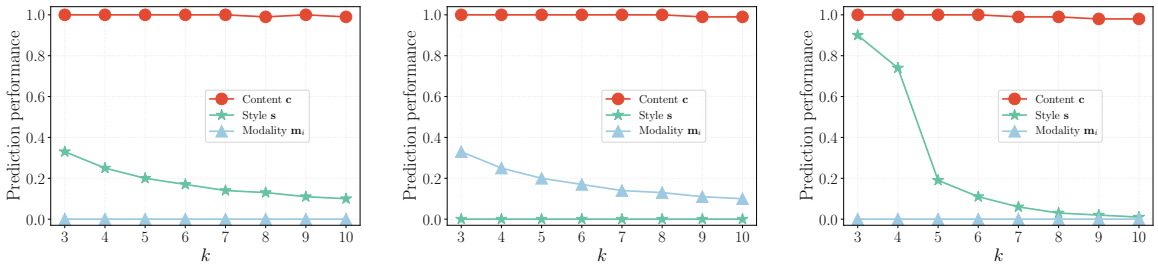
Parameter	Value
Generating function	Image and text rendering
Image encoder	ResNet-18
Text encoder	4-layer ConvNet
Optimizer	Adam
Batch size	256
Learning rate	1e-5
Temperature τ	1.0
# Seeds	3
# Iterations	100 000
# Samples (train / val / test)	125 000 / 10 000 / 10 000
Similarity metric	Cosine similarity
Gradient clipping	2-norm; max value 2

(b) Multimodal3DIdent

Table A.2: Hyperparameter values used for the two experiments in Chapter 8.

Numerical simulation with discrete latent factors Extending the numerical simulation from Section 8.4.1, we test block-identifiability of content information when observations are generated from a mixture of continuous and discrete latent variables, thus violating one of the assumptions from Theorem 2. In this setting, content, style and modality-specific information are random vectors with 5 components sampled from either a continuous normal distribution or a discrete multinomial distribution with k classes, for which we experiment with different $k \in \{3, 4, \dots, 10\}$. For all settings, we train an encoder with the InfoNCE objective and set the encoding size to 5 dimensions. The other hyperparameters used in this set of experiments are detailed in Table A.2a. To ensure convergence of the models, we extended the number of training iterations to 600 000 and 3 000 000 for experiments with discrete style/modality-specific and discrete content variables respectively.

With discrete style or modality-specific variables and continuous content (Figures A.5a and A.5b), the results suggest that content is block-identified, since the prediction of style and modality-specific information is at chance level (i.e., $accuracy = 1/k$) while content is consistently fully recovered ($R^2 \geq 0.99$). In the opposite setting, with continuous style and modality-specific variables and discrete content (Figure A.5c), the number of content classes appears to be a critical factor for block-identifiability of content: while content is always encoded well, style information is also encoded to a significant extent when the



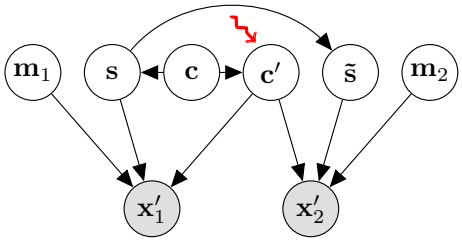
(a) Only $\mathbf{s}$ is discrete (b) Only $\mathbf{m}_1, \mathbf{m}_2$ are discrete (c) Only $\mathbf{c}$ is discrete

Figure A.5: Numerical simulations with discrete latent factors. The results show three settings in each of which one group of latent variables is discrete while the remaining groups are continuous. Continuous variables are normally distributed, whereas discrete variables are sampled from a multinomial distribution with k distinct classes. We evaluate the prediction performance using nonlinear probing and measure the R^2 coefficient of determination for continuous factors and classification accuracy for discrete factors respectively. Each point denotes the average across three seeds and error bars show the standard deviation.

number of content classes is small but significantly less style information can be recovered when the number of content classes increases.

Through this set of experiments, we challenge the assumption that *all* generative factors should be continuous (c.f., Section 8.2) and show that block-identifiability of content can still be satisfied when content is continuous while style or modality-specific variables are discrete. On the other hand, style is encoded to a significant extent when content is discrete, which might explain our observation for the Multimodal3DIdent dataset (Section 8.4.2), where we saw that, in the presence of discrete content factors, some style information can be encoded. However, the additional experiments do not explain our observation that modality-specific information was encoded as well.

Evaluation with test-time interventions In Chapter 8, we observed that style information can be predicted to some degree when there are causal dependencies from content to style (Table 8.1), which can be attributed to style information being partially predictable from the encoded content information in the causal setup. To verify that the encoders only depend on content information (i.e., that content is block-identified), we assess the trained models using a novel, more rigorous empirical evaluation for the numerical simulation. We test the effect of *interventions* $\mathbf{c} \to \mathbf{c}'$, which perturb the content information at *test time* via batch-wise permutations of content, before generating $\mathbf{x}'_1 = \mathbf{f}_1(\mathbf{c}', \mathbf{s}, \mathbf{m}_1)$ and



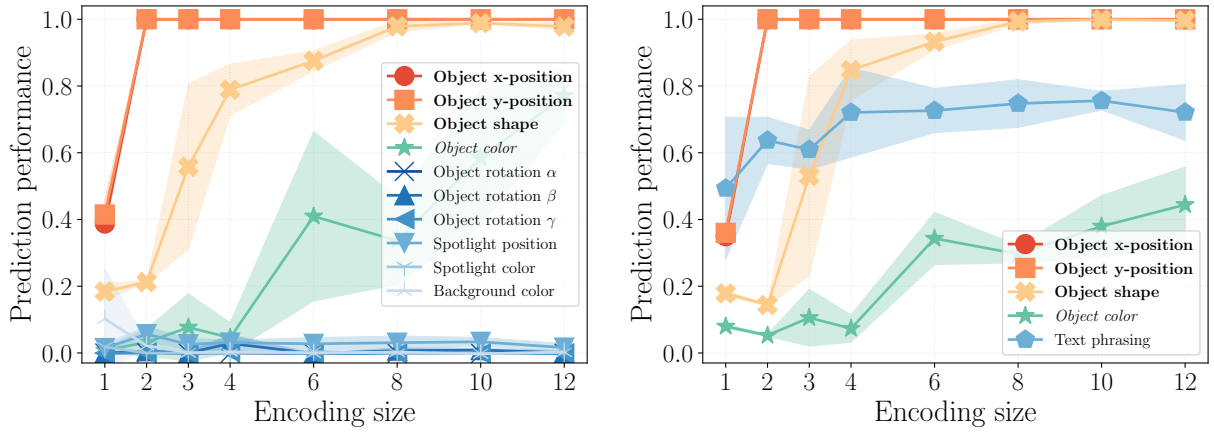
Generative process			R^2 (nonlinear)			
p(chg.)	Stat.	Cau.	Content c	Content c′	Style s	Modality $\mathbf{m}_i$
1.0	✗	✗	0.00 (±0.00)	**1.00** (±0.00)	0.00 (±0.00)	0.00 (±0.00)
0.75	✗	✗	0.00 (±0.00)	**1.00** (±0.00)	0.00 (±0.00)	0.00 (±0.00)
0.75	✓	✗	0.00 (±0.00)	**1.00** (±0.00)	<u>0.50</u> (±0.19)	0.00 (±0.00)
0.75	✗	✓	0.01 (±0.00)	**0.98** (±0.00)	0.03 (±0.01)	0.00 (±0.00)
0.75	✓	✓	<u>0.28</u> (±0.14)	**0.91** (±0.03)	<u>0.39</u> (±0.20)	0.00 (±0.00)

Figure A.6: Evaluation with test-time interventions. We use the interventional setup that is illustrated on the left, i.e., perturbed samples $\mathbf{x}'_1, \mathbf{x}'_2$ that are generated from the intervened content $\mathbf{c}'$, which is a copy of the original content $\mathbf{c}$ with an intervention, i.e., a batch-wise permutation ($\rightsquigarrow$) that makes $\mathbf{c}'$ independent of $\mathbf{s}$. Each row presents the results of a different setup with varying style-change probability p(chg.) and possible statistical (Stat.) and/or causal (Caus.) dependencies. Each value denotes the R^2 coefficient of determination (averaged across 3 seeds) for a nonlinear regression model that predicts the respective ground truth factor ($\mathbf{c}, \mathbf{c}', \mathbf{s}$, or $\mathbf{m}_i$) from the learned representation.

$\mathbf{x}'_2 = \mathbf{f}_1(\mathbf{c}', \tilde{\mathbf{s}}, \mathbf{m}_1)$. Hence, we break the causal dependence between content and style (see illustration in Figure A.6), which allows us to better assess whether the trained encoders depend on content or style information. Specifically, we train the encoders for 3 000 000 iterations to ensure convergence and then use nonlinear probing to predict both the original and the intervened content variables from the learned representations.

Figure A.6 presents our results using the interventional setup, showing that in most cases only content information can be recovered. We observe an exception (underlined values) in the two cases with statistical dependencies, where some style information can be recovered, which is expected because statistical dependencies reduce the effective dimensionality of content [von+21]. Analogously, in the case of statistical and causal dependencies, some of the original content information can be recovered via the encoded style information. In summary, the evaluation with interventions provides a more rigorous assessment of block-identifiability in the causal setup, showing that neither style nor modality-specific information can be recovered when the encoding size matches the true number of content factors.

Multimodal3DIdent with mutually independent factors For the results of the experiments in Section 8.4.2, we used the Multimodal3DIdent dataset, which was designed such that object color is causally dependent on the x-position of the object to impose a



(a) Prediction of image factors

(b) Prediction of text factors

Figure A.7: Result on the Multimodal3DIdent dataset with mutually independent factors. As a function of the encoding size of the model, show the results for nonlinear probing with respect to the ground truth image factors (Figure A.7a) and text factors (Figure A.7b) to quantify how well the embeddings encode the respective information. Content factors are denoted in bold and style factors in italic. Along the x-axis, we vary the encoding size, i.e., the output dimensionality of the model. We measure the prediction performance in terms of the R^2 coefficient of determination for continuous factors and classification accuracy for discrete factors respectively. Each point denotes the average across three seeds and bands show one standard deviation.

causal dependence of style on content. In Figure A.7, we provide a similar analysis using a version of the dataset *without* the causal dependence, i.e., with mutually independent factors. For both modalities, we observe that object color can only be recovered when the encoding size is larger than four, i.e., when there is excess capacity beyond the capacity needed to encode all content factors. Hence, the results in Figure A.7 corroborate that contrastive learning can block-identify content factors in the Multimodal3DIdent dataset.

A.7 Outlook: Symmetric Generative Process

In Chapter 8, we described an asymmetric generative process, where $\mathbf{z}_2$ is a perturbed version of $\mathbf{z}_1$. In this section, we sketch out how our model and results can be adapted to a symmetric setting, where $\mathbf{z}_1$ and $\mathbf{z}_2$ are generated as perturbations of $\mathbf{z}$.

Concretely, we would need to make small adjustments to Assumptions 4 and 5 as follows. We start with the content invariance in Assumption 4 to specify how $\mathbf{z}_1 = (\tilde{\mathbf{c}}_1, \tilde{\mathbf{s}}_1, \tilde{\mathbf{m}}_1)$ and $\mathbf{z}_2 = (\tilde{\mathbf{c}}_2, \tilde{\mathbf{s}}_2, \tilde{\mathbf{m}}_2)$ are generated.

Let $i \in \{1, 2\}$. The conditional density $p(\mathbf{z}_i \mid \mathbf{z})$ over $\mathcal{Z}_i \times \mathcal{Z}$ takes the form

$$p(\mathbf{z}_i \mid \mathbf{z}) = \delta(\tilde{\mathbf{c}}_i - \mathbf{c})\delta(\tilde{\mathbf{m}}_i - \mathbf{m}_i)p(\tilde{\mathbf{s}}_i \mid \mathbf{s}) \ , \tag{A.45}$$

where $\delta(\cdot)$ is the Dirac delta function, used to specify that $\tilde{\mathbf{c}}_i = \mathbf{c}$ almost everywhere as well as $\tilde{\mathbf{m}}_i = \mathbf{m}_i$ almost everywhere. Since $\tilde{\mathbf{c}}_1 = \mathbf{c}$ a.e. and $\mathbf{c} = \tilde{\mathbf{c}}_2$ a.e., it follows that $\tilde{\mathbf{c}}_1 = \tilde{\mathbf{c}}_2$ almost everywhere, which is a property that is needed in Step 1 of the proof of Theorem 2. In addition, it still holds that $\tilde{\mathbf{m}}_i \perp\!\!\!\perp \mathbf{z}_j$, for $i, j \in \{1, 2\}$ and $i \neq j$, which is needed in Step 2 of the proof to show that modality-specific information is not encoded.

Additionally, we need to revisit Assumption 5, for which both $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ would be generated through perturbations of $\mathbf{s}$ via the conditional distribution $p(\tilde{\mathbf{s}}_i \mid \mathbf{s})$ on $\mathcal{S} \times \mathcal{S}$, as described in Assumption 5, for each $i \in \{1, 2\}$ individually. As a technical nuance, we would need to specify the conditional generation of the perturbed style variables $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ such that they are not perturbed in an identical manner w.r.t. $\mathbf{s}$. To ensure this, one could for example constrain p_A appropriately to exclude the degenerate case where dimensions in $\tilde{\mathbf{s}}_1$ and $\tilde{\mathbf{s}}_2$ are perfectly aligned. This would be necessary for Step 2 of the proof of Theorem 2.

A.8 Outlook: Sequential Decision Making

In this section, we describe an extension of our formulation of the multimodal generative process from Chapter 2 to a setup with sequential decision making, i.e., a dynamical system where an agent interacts with its environment. Instead of a fixed dataset, we consider a *time series* of observations and subsequent *actions* that affect the generative process on the level of latent variables or *hidden states*, as they are typically called in the context of partially observable Markov decision processes (POMDPs).

A.8.1 Motivation

For illustration, consider the following example from everyday life. Imagine that you intend to cross a busy street and suddenly hear a car approaching. As you turn your head into the perceived direction of sound, your visual experience matches your auditory perception—you *see* the car approaching. Hence, the car, which is initially perceived only through the auditory sensory system, comes into visual perception given a suitable course of actions. Analogously, machines can learn to *seek* sensory inputs that carry shared information about external objects by minimizing the prediction error across modalities through enhanced perception and adaptive behavior.

In this thesis, we have only investigated the former mechanism (i.e., enhanced perception) by learning representations that encode shared information in a suitable format. We believe that adaptive behavior, which represents the flipside of *active inference* [e.g., PPF22], offers rich opportunities for future work on multimodal representation learning. In the following, we sketch out how adaptive behavior in the form of sequential decision making can be combined with our framework.

A.8.2 Formulation

Consider a *policy*, i.e., a function $\boldsymbol{\pi}_\eta : \hat{\mathcal{Z}} \to \mathcal{A}$ that maps from the representation space $\hat{\mathcal{Z}}$ to a set of possible actions $\mathcal{A}$ and that is parameterized by η. Let $\mathbf{a}^{(t)} = \boldsymbol{\pi}(\hat{\mathbf{z}}^{(t)})$ be an action at time t that feeds back into a generative process that evolves over time. Specifically, consider a partially observable Markov decision process (POMDP), for which the hidden states at time $t+1$ depend on the previous states and actions, i.e.,

$$\mathbf{z}^{(t+1)} = \mathbf{f}_\mathbf{z}(\mathbf{z}^{(t)}, \mathbf{a}^{(t)}) \,, \tag{A.46}$$

where $\mathbf{f}_\mathbf{z}$ is an unknown function. The observations at time $t+1$ are generated as before (i.e., following Definition 1) but with an additional temporal dependence,

$$\mathbf{x}_i^{(t+1)} = \mathbf{f}_i(\mathbf{z}^{(t+1)}), \quad \text{for each modality } i \in \{1, \dots, M\} \,. \tag{A.47}$$

Within the framework of sequential decision making, one could reason about a set of invariant factors given a suitable course of actions and therefore consider a *joint optimization* with respect to the parameters of the encoder and policy network.

The goal of sequential decision making is to find an optimal behavior subject to some optimality criterion. Often, the objective is to choose actions at each time step to maximize the expected future discounted reward (or to minimize the cost, respectively), i.e.,

$$\max_{\eta} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], \tag{A.48}$$

where r_t is the reward at time t and $\gamma \in [0, 1)$ is a discount factor. In the context of multimodal representation learning, we could for example use the InfoNCE loss (Equation 3.58) as a cost function and formulate the objective as a joint optimization over the space of actions and encoder parameters, namely

$$\min_{\phi_1, \phi_2, \eta} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{L}_{\text{InfoNCE}}\left(\{\mathbf{x}_1^{(k,t)}, \mathbf{x}_2^{(k,t)}\}_{k=1}^{K}; \phi_1, \phi_2\right)\right]. \tag{A.49}$$

Crucially, the setup of sequential decision making can allow for relaxations of content invariance (e.g., Assumption 3 or 4). For instance, when the assumption is violated initially at time $t = 0$, there can still be set of actions $\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(T)}\}$, such that content invariance is satisfied at time $T > 0$. This is analogous to our previous example of the approaching car, for which it takes some time to turn our head in the right direction to confirm with our eyes what we initially perceive only with our ears.

In Chapter 8, we have shown that contrastive learning tends to encode the invariant factors over modality-specific and shared but variable factors when content invariance is satisfied. We conjecture that similar results could be obtained in the described setup of sequential decision making, even under violations of content invariance.

Bibliography

[AB16] Guillaume Alain and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes". *arXiv preprint arXiv:1610.01644* (2016) (cit. on pp. 19, 20).

[Ale+17] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin P. Murphy. "Deep Variational Information Bottleneck". *International Conference on Learning Representations*. 2017 (cit. on p. 36).

[BAM19] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A survey and Taxonomy". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443 (cit. on p. 13).

[Bay12] Tim Bayne. *The unity of consciousness*. Oxford University Press, 2012 (cit. on p. 3).

[BCV13] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828 (cit. on pp. 2, 24, 25, 117).

[BH92] Suzanna Becker and Geoffrey E. Hinton. "Self-organizing neural network that discovers surfaces in random-dot stereograms". *Nature* 355.6356 (1992), pp. 161–163 (cit. on p. 41).

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006 (cit. on p. 5).

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians". *Journal of the American statistical Association* 112.518 (2017), pp. 859–877 (cit. on p. 32).

[BPL22] Adrien Bardes, Jean Ponce, and Yann LeCun. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". *International Conference on Learning Representations*. 2022 (cit. on p. 42).

[Bre+22] Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. "Weakly supervised causal representation learning". *Advances in Neural Information Processing Systems*. 2022 (cit. on p. 18).

[Bro+93] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. "Signature verification using a siamese time delay neural network". 1993 (cit. on p. 41).

[BTN18] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. "Multi-level variational autoencoder: Learning disentangled representations from grouped observations". *AAAI Conference on Artificial Intelligence*. 2018 (cit. on pp. 14, 18).

[Car+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. "Emerging Properties in Self-Supervised Vision Transformers". *International Conference on Computer Vision*. 2021 (cit. on p. 42).

[CB20] Junxiang Chen and Kayhan Batmanghelich. "Weakly Supervised Disentanglement by Pairwise Similarities". *AAAI Conference on Artificial Intelligence*. 2020 (cit. on p. 18).

[CF14] Yanshuai Cao and David J. Fleet. "Generalized product of experts for automatic and principled fusion of Gaussian process predictions". *arXiv preprint arXiv:1410.7827* (2014) (cit. on pp. 40, 60).

[CF21] Tim Crane and Craig French. "The Problem of Perception". *The Stanford Encyclopedia of Philosophy*. 2021 (cit. on p. 16).

[CH21] Xinlei Chen and Kaiming He. "Exploring Simple Siamese Representation Learning". *Conference on Computer Vision and Pattern Recognition*. 2021 (cit. on p. 42).

[Che+10] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. "Large Scale Online Learning of Image Similarity Through Ranking". *Journal of Machine Learning Research* 11 (2010), pp. 1109–1135 (cit. on p. 41).

[Che+20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". *International Conference on Machine Learning*. 2020 (cit. on p. 41).

[CHL05] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a Similarity Metric Discriminatively, with Application to Face Verification". *Computer Vision and Pattern Recognition*. 2005 (cit. on p. 41).

[CJ10] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press, 2010 (cit. on p. 27).

[CT12] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012 (cit. on pp. 35, 97, 134).

[CW00] Michael A. Casey and Alex Westner. "Separation of Mixed Audio Sources By Independent Subspace Analysis". *International Computer Music Conference*. 2000 (cit. on p. 27).

[CW08] Ronan Collobert and Jason Weston. "A unified architecture for natural language processing: deep neural networks with multitask learning". *International Conference on Machine Learning*. 2008 (cit. on p. 41).

[Dar51] George Darmois. "Analyse des liaisons de probabilité". *Proc. Int. Stat. Conferences 1947*. 1951, p. 231 (cit. on p. 119).

[Dau+20] Imant Daunhawer, Thomas M. Sutter, Ricards Marcinkevics, and Julia E. Vogt. "Self-supervised Disentanglement of Modality-Specific and Shared Factors Improves Multimodal Generative Models". *German Conference on Pattern Recognition*. 2020 (cit. on p. 64).

[Dau+22] Imant Daunhawer, Thomas M. Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E. Vogt. "On the Limitations of Multimodal VAEs". *International Conference on Learning Representations*. 2022 (cit. on p. 100).

[DC04] Adele Diederich and Hans Colonius. "Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time". *Perception & psychophysics* 66.8 (2004), pp. 1388–1404 (cit. on p. 3).

[Den+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: a large-scale hierarchical image database". *Conference on Computer Vision and Pattern Recognition*. 2009 (cit. on p. 20).

[Doe16] Carl Doersch. "Tutorial on variational autoencoders". *arXiv preprint arXiv:1606.05908* (2016) (cit. on p. 96).

[Dor+19] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. "Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation". *Medical Image Computing and Computer Assisted Intervention*. 2019 (cit. on pp. 90, 130, 147).

[Doy+07] Kenji Doya, Shin Ishii, Alexandre Pouget, and Rajesh P. N. Rao. *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007 (cit. on p. 3).

[EB04] Marc O. Ernst and Heinrich H. Bülthoff. "Merging the senses into a robust percept". *Trends in cognitive sciences* 8.4 (2004), pp. 162–169 (cit. on p. 3).

[EL04] Ahmed Elgammal and Chan-Su Lee. "Separating style and content on a nonlinear manifold". *Conference on Computer Vision and Pattern Recognition*. 2004 (cit. on p. 14).

[Eve95] H. R. Everett. *Sensors for mobile robots*. CRC Press, 1995 (cit. on p. 3).

[Fed+20] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. "Learning Robust Representations via Multi-View Information Bottleneck". *International Conference on Learning Representations*. 2020 (cit. on p. 44).

[FTF21] Marco Federici, Ryota Tomioka, and Patrick Forré. "An Information-theoretic Approach to Distribution Shifts". *Advances in Neural Information Processing Systems*. 2021 (cit. on p. 127).

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016 (cit. on p. 29).

[GEB16] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "Image style transfer using convolutional neural networks". *Conference on Computer Vision and Pattern Recognition*. 2016 (cit. on p. 14).

[GH10] Michael Gutmann and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models". *International Conference on Artificial Intelligence and Statistics*. 2010 (cit. on pp. 6, 41, 42, 112).

[GH12] Michael Gutmann and Aapo Hyvärinen. "Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics". *Journal of Machine Learning Research* 13 (2012), pp. 307–361 (cit. on p. 41).

[Gho+20] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael Black, and Bernhard Schölkopf. "From Variational to Deterministic Autoencoders". *International Conference on Learning Representations*. 2020 (cit. on p. 137).

[Goo] *Google Scholar: Top venues*. https://scholar.google.com/citations?view_op=top_venues. Accessed: 2023-04-27 (cit. on p. 25).

[Goo+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". *Advances in Neural Information Processing Systems*. 2014 (cit. on p. 106).

[Gre+19] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. "The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA". *Conference on Uncertainty in Artificial Intelligence*. 2019 (cit. on pp. 26, 28, 110, 112, 127).

[Gri+20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo A. Pires, Zhaohan Guo, Mohammad G. Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning". *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 42).

[GS06] Asif A. Ghazanfar and Charles E. Schroeder. "Is neocortex essentially multisensory?" *Trends in cognitive sciences* 10.6 (2006), pp. 278–285 (cit. on p. 2).

[GZP21] Boying Gong, Yun Zhou, and Elizabeth Purdom. "Cobolt: integrative analysis of multimodal single-cell sequencing data". *Genome biology* 22.1 (2021), pp. 1–21 (cit. on pp. 130, 147).

[Har+03] Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, and Klaus-Robert Müller. "Kernel-Based Nonlinear Blind Source Separation". *Neural Computation* 15.5 (2003), pp. 1089–1124 (cit. on p. 26).

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. "Dimensionality reduction by learning an invariant mapping". *Computer Vision and Pattern Recognition*. 2006 (cit. on p. 41).

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". *Conference on Computer Vision and Pattern Recognition*. 2016 (cit. on pp. 50, 101, 123, 136).

[He+20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. "Momentum Contrast for Unsupervised Visual Representation Learning". *Conference on Computer Vision and Pattern Recognition*. 2020 (cit. on pp. 41, 44).

[Heu+17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. "GANs trained by a two time-scale update rule converge to a local nash equilibrium". *Advances in Neural Information Processing Systems*. 2017 (cit. on pp. 20, 21, 81, 102, 140).

[HG18] Wei-Ning Hsu and James Glass. "Disentangling by Partitioning: A Representation Learning Framework for Multimodal Sensory Data". *arXiv preprint arXiv:1805.11264* (2018) (cit. on pp. 39, 64).

[HGIL16] Jerónimo Hernández-González, Inaki Inza, and Jose A. Lozano. "Weak supervision and other non-standard classification problems: a taxonomy". *Pattern Recognition Letters* 69 (2016), pp. 49–55 (cit. on p. 18).

[HH00] Aapo Hyvärinen and Patrik Hoyer. "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces". *Neural computation* 12.7 (2000), pp. 1705–1720 (cit. on p. 27).

[Hig+17] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". *International Conference on Learning Representations*. 2017 (cit. on pp. 82, 101).

[Hig+18] Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loïc Matthey, Danilo J. Rezende, and Alexander Lerchner. "Towards a Definition of Disentangled Representations". *arXiv preprint arXiv:1812.02230* (2018) (cit. on pp. 20, 25, 117).

[Hin02] Geoffrey E. Hinton. "Training products of experts by minimizing contrastive divergence". *Neural Computation* 14.8 (2002), pp. 1771–1800 (cit. on pp. 40, 58, 131).

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". *Advances in Neural Information Processing Systems*. 2020 (cit. on p. 148).

[Hje+19] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. "Learning deep representations by mutual information estimation and maximization". *International Conference on Learning Representations*. 2019 (cit. on p. 41).

[HKM23] Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. "Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning". *Patterns* 4.10 (2023) (cit. on pp. 25, 27, 28).

[HM16] Aapo Hyvärinen and Hiroshi Morioka. "Unsupervised feature extraction by time-contrastive learning and nonlinear ICA". *Advances in Neural Information Processing Systems*. 2016 (cit. on pp. 26, 122, 127).

[HM17] Aapo Hyvärinen and Hiroshi Morioka. "Nonlinear ICA of Temporally Dependent Stationary Sources". *International Conference on Artificial Intelligence and Statistics*. 2017 (cit. on pp. 26, 122).

[HO00] Aapo Hyvärinen and Erkki Oja. "Independent component analysis: algorithms and applications". *Neural networks* 13.4-5 (2000), pp. 411–430 (cit. on p. 26).

[HP99] Aapo Hyvärinen and Petteri Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results". *Neural networks* 12.3 (1999), pp. 429–439 (cit. on pp. 2, 25, 28, 111, 119, 127).

[HRR22] Irina Higgins, Sébastien Racanière, and Danilo J. Rezende. "Symmetry-based representations for artificial and biological general intelligence". *Frontiers in Computational Neuroscience* (2022), p. 28 (cit. on pp. 20, 25).

[HST19] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. "Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning". *International Conference on Artificial Intelligence and Statistics*. 2019 (cit. on pp. 26, 127).

[Hua+18] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. "Multimodal Unsupervised Image-to-Image Translation". *European Conference on Computer Vision*. 2018 (cit. on p. 14).

[Hug+94] Howard C. Hughes, Patricia A. Reuter-Lorenz, George Nozawa, and Robert Fendrich. "Visual-auditory interactions in sensorimotor processing: saccades versus manual responses." *Journal of Experimental Psychology: Human Perception and Performance* 20.1 (1994), p. 131 (cit. on p. 3).

[Hén+20] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aäron van den Oord. "Data-efficient image recognition with contrastive predictive coding". *International Conference on Machine Learning*. 2020 (cit. on p. 41).

[Jam+13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013 (cit. on p. 125).

[Jeb04] Tony Jebara. *Machine Learning: Discriminative and Generative*. Springer, 2004 (cit. on p. 6).

[Jin+22] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. "Understanding Dimensional Collapse in Contrastive Self-supervised Learning". *International Conference on Learning Representations*. 2022 (cit. on p. 42).

[Joy+22] Tom Joy, Yuge Shi, Philip H. S. Torr, Tom Rainforth, Sebastian M. Schmon, and Siddharth N. "Learning Multimodal VAEs through Mutual Supervision". *International Conference on Learning Representations*. 2022 (cit. on p. 81).

[KB15] Diederik P. Kingma and Jimmy Ba. "Adam: A method for stochastic gradient descent". *International Conference on Learning Representations*. 2015 (cit. on p. 101).

[KGS19] Richard Kurle, Stephan Guennemann, and Patrick van der Smagt. "Multi-Source Neural Variational Inference". *AAAI Conference on Artificial Intelligence*. 2019 (cit. on pp. 58, 99).

[Khe+20] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". *International Conference on Artificial Intelligence and Statistics*. 2020 (cit. on p. 26).

[Kla+14] Arto Klami, Seppo Virtanen, Eemeli Leppäaho, and Samuel Kaski. "Group factor analysis". *IEEE Transactions on Neural Networks and Learning Systems* 26.9 (2014), pp. 2136–2147 (cit. on p. 14).

[Kli+21] David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. "Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding". *International Conference on Learning Representations*. 2021 (cit. on p. 127).

[KM18] Hyunjik Kim and Andriy Mnih. "Disentangling by factorising". *International Conference on Machine Learning*. 2018 (cit. on p. 134).

[Kon+22] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. "Partial disentanglement for domain adaptation". *International Conference on Machine Learning*. 2022 (cit. on p. 112).

[KP04] David C. Knill and Alexandre Pouget. "The Bayesian brain: the role of uncertainty in neural coding and computation". *Trends in Neurosciences* 27.12 (2004), pp. 712–719 (cit. on p. 3).

[KW14] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". *International Conference on Learning Representations*. 2014 (cit. on pp. 6, 29, 33).

[KW19] Diederik P. Kingma and Max Welling. "An Introduction to Variational Autoencoders". *Found. Trends Mach. Learn.* 12.4 (2019), pp. 307–392 (cit. on p. 31).

[KW52] Jack Kiefer and Jacob Wolfowitz. "Stochastic estimation of the maximum of a regression function". *The Annals of Mathematical Statistics* (1952), pp. 462–466 (cit. on p. 29).

[Kör+07] Konrad P. Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams. "Causal inference in multisensory perception". *PLoS one* 2.9 (2007), e943 (cit. on pp. 3, 14).

[LC06] Erich L. Lehmann and George Casella. *Theory of point estimation*. Springer, 2006 (cit. on p. 111).

[LCB98] Yann LeCun, Corinna Cortes, and Christopher J. C. Burges. "The MNIST database of handwritten digits". *http://yann.lecun.com/exdb/mnist/* (1998) (cit. on p. 47).

[Le+11] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis". *Conference on Computer Vision and Pattern Recognition*. 2011 (cit. on p. 27).

[LeC+98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 47).

[LF20] Qi Lyu and Xiao Fu. "Nonlinear Multiview Analysis: Identifiability and Neural Network-Assisted Implementation". *IEEE Trans. Signal Process.* 68 (2020), pp. 2697–2712 (cit. on pp. 28, 112).

[LHS20] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. "Contrastive Representation Learning: a Framework and Review". *IEEE Access* 8 (2020), pp. 193907–193934 (cit. on p. 41).

[Lin76] Seppo Linnainmaa. "Taylor expansion of the accumulated rounding error". *BIT Numerical Mathematics* 16.2 (1976), pp. 146–160 (cit. on p. 29).

[Liu+15] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". *International Conference on Computer Vision*. 2015 (cit. on p. 49).

[Liu+23] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. "Self-Supervised Learning: Generative or Contrastive". *IEEE Trans. Knowl. Data Eng.* 35.1 (2023), pp. 857–876 (cit. on pp. 6, 29, 41).

[Loc+19] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". *International Conference on Machine Learning*. 2019 (cit. on pp. 25, 127).

[Loc+20a] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. "Disentangling Factors of Variations Using Few Labels". *International Conference on Learning Representations*. 2020 (cit. on p. 26).

[Loc+20b] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. "Weakly-Supervised Disentanglement Without Compromises". *International Conference on Machine Learning*. 2020 (cit. on pp. 18, 26, 28, 112, 125, 127).

[LS21] Changhee Lee and Mihaela van der Schaar. "A Variational Information Bottleneck Approach to Multi-Omics Data Integration". *International Conference on Artificial Intelligence and Statistics*. 2021 (cit. on pp. 90, 130, 147).

[Lyu+22] Qi Lyu, Xiao Fu, Weiran Wang, and Songtao Lu. "Understanding Latent Correlation-Based Multiview Learning and Self-Supervision: An Identifiability Perspective". *International Conference on Learning Representations*. 2022 (cit. on pp. 28, 110, 112, 127).

[Mat] *Matplotlib v3.7.1 colors API*. `https://matplotlib.org/stable/api/colors_api.html`. Accessed: 2023-11-05 (cit. on p. 52).

[Mer02] M. Alex Meredith. "On the neuronal basis for multisensory convergence: a brief overview". *Cognitive brain research* 14.1 (2002), pp. 31–40 (cit. on p. 2).

[Mil91] Jeff Miller. "Channel interaction and the redundant-targets effect in bimodal divided attention." *Journal of Experimental Psychology: Human Perception and Performance* 17.1 (1991), p. 160 (cit. on p. 3).

[Min+21] Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. "A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data". *Cell Reports Methods* (2021) (cit. on pp. 90, 130, 147).

[Mit+21] Jovana Mitrovic, Brian McWilliams, Jacob C. Walker, Lars H. Buesing, and Charles Blundell. "Representation Learning via Invariant Causal Mechanisms". *International Conference on Learning Representations*. 2021 (cit. on pp. 41, 127).

[Mon+21] Milton L. Montero, Casimir J. H. Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. "The role of Disentanglement in Generalisation". *International Conference on Learning Representations*. 2021 (cit. on p. 106).

[MS09a] Micah M. Murray and Lucas Spierer. "Auditory spatio-temporal brain dynamics and their consequences for multisensory interactions in humans". *Hearing research* 258.1-2 (2009), pp. 121–133 (cit. on p. 2).

[MS09b] Gabriella Musacchia and Charles E. Schroeder. "Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex". *Hearing research* 258.1-2 (2009), pp. 72–79 (cit. on p. 2).

[MS20] David McAllester and Karl Stratos. "Formal Limitations on the Measurement of Mutual Information". *International Conference on Artificial Intelligence and Statistics*. 2020 (cit. on p. 43).

[Mur22] Kevin P. Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022 (cit. on pp. 29, 31).

[Mur23] Kevin P. Murphy. *Probabilistic machine learning: advanced topics*. MIT press, 2023 (cit. on pp. 33, 42).

[Nat+13] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. "Learning with Noisy Labels". *Advances in Neural Information Processing Systems*. 2013 (cit. on p. 18).

[NK10] Marcus J. Naumer and Jochen Kaiser. *Multisensory object perception in the primate brain*. Springer, 2010 (cit. on p. 3).

[NWJ10] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. "Estimating divergence functionals and the likelihood ratio by convex risk minimization". *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861 (cit. on p. 134).

[O12] Casey O'Callaghan. "Perception and Multimodality". *The Oxford Handbook of Philosophy of Cognitive Science*. Oxford University Press, 2012 (cit. on p. 3).

[OED] *multimodal, adj. OED Online*. Oxford University Press, Mar. 2023. URL: https://www.oed.com/view/Entry/123567 (visited on 04/07/2023) (cit. on p. 1).

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". *arXiv preprint arXiv:1807.03748* (2018) (cit. on pp. 6, 41–43, 112).

[Pea00] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000 (cit. on p. 26).

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017 (cit. on pp. 14, 15).

[Poo+19] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. "On Variational Bounds of Mutual Information". *International Conference on Machine Learning*. 2019 (cit. on p. 43).

[PPF22] Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022 (cit. on p. 168).

[Pre+19] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. "Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer". *International Conference on Learning Representations*. 2019 (cit. on p. 14).

[Qui+05] Rodrigo Q. Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. "Invariant visual representation by single neurons in the human brain". *Nature* 435.7045 (2005), pp. 1102–1107 (cit. on p. 3).

[Qui+09] Rodrigo Q. Quiroga, Alexander Kraskov, Christof Koch, and Itzhak Fried. "Explicit encoding of multimodal percepts by single neurons in the human brain". *Current Biology* 19.15 (2009), pp. 1308–1313 (cit. on p. 3).

[Rad+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". *International Conference on Machine Learning*. 2021 (cit. on pp. 45, 110, 147).

[Ram+21] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. "Zero-Shot Text-to-Image Generation". *International Conference on Machine Learning*. 2021 (cit. on pp. 110, 148).

[Ram+22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. "Hierarchical Text-Conditional Image Generation with CLIP Latents". *arXiv preprint arXiv:2204.06125* (2022) (cit. on pp. 110, 148).

[Rei56] Hans Reichenbach. *The direction of time*. University of California Press, 1956 (cit. on pp. 3, 14).

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". *Nature* 323.6088 (1986), pp. 533–536 (cit. on p. 29).

[RM51] Herbert Robbins and Sutton Monro. “A stochastic approximation method”. *The annals of mathematical statistics* (1951), pp. 400–407 (cit. on p. 29).

[RMW14] Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. “Stochastic backpropagation and approximate inference in deep generative models”. *International Conference on Machine Learning*. 2014 (cit. on pp. 6, 29).

[Roj+18] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard E. Turner, and Jonas Peters. “Invariant Models for Causal Transfer Learning”. *Journal of Machine Learning Research* 19 (2018), pp. 1309–1342 (cit. on p. 127).

[Rom+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models”. *Conference on Computer Vision and Pattern Recognition*. 2022 (cit. on pp. 110, 148).

[RW17] Brendan van Rooyen and Robert C. Williamson. “A Theory of Learning with Corrupted Labels”. *Journal of Machine Learning Research* 18 (2017) (cit. on p. 18).

[Sah+22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Kamyar Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha G. Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”. *Advances in Neural Information Processing Systems*. 2022 (cit. on p. 110).

[Sch+16] Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. “Modeling confounding by half-sibling regression”. *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7391–7398 (cit. on p. 127).

[Sch+21] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. “Toward Causal Representation Learning”. *Proceedings of the IEEE* 109.5 (2021), pp. 612–634 (cit. on pp. 2, 25, 26).

[SDV20] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. “Multimodal Generative Learning Utilizing Jensen-Shannon-Divergence”. *Advances in Neural Information Processing Systems*. 2020 (cit. on pp. 49, 61, 64, 90).

[SDV21] Thomas M. Sutter, Imant Daunhawer, and Julia E. Vogt. "Generalized Multimodal ELBO". *International Conference on Learning Representations*. 2021 (cit. on pp. 82, 90, 92).

[SE19] Yang Song and Stefano Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution". *Advances in Neural Information Processing Systems*. 2019 (cit. on p. 148).

[Shi+19] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". *Advances in Neural Information Processing Systems*. 2019 (cit. on pp. 21, 40, 50, 55, 58, 61, 62, 64, 78, 81, 82, 86, 90–92, 101, 103, 106, 136, 154).

[Shi+21] Yuge Shi, Brooks Paige, Philip H. S. Torr, and Siddharth N. "Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models". *International Conference on Learning Representations*. 2021 (cit. on pp. 50, 81, 101, 103).

[Shu+20] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Weakly Supervised Disentanglement with Guarantees". *International Conference on Learning Representations*. 2020 (cit. on pp. 18, 117).

[SK19] Connor Shorten and Taghi M. Khoshgoftaar. "A survey on image data augmentation for deep learning". *Journal of big data* 6.1 (2019), pp. 1–48 (cit. on p. 41).

[SLY15] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". *Advances in Neural Information Processing Systems*. 2015 (cit. on p. 37).

[SM93] Barry E. Stein and M. Alex Meredith. *The merging of the senses*. MIT press, 1993 (cit. on p. 2).

[SMB05] Ladan Shams, Wei Ji Ma, and Ulrik Beierholm. "Sound-induced flash illusion as an optimal percept". *Neuroreport* 16.17 (2005), pp. 1923–1927 (cit. on p. 3).

[SNM16] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. "Joint Multimodal Learning with Deep Generative Models". *arXiv preprint arXiv:1611.01891* (2016) (cit. on p. 39).

[SNS11] Roland Siegwart, Illah R. Nourbakhsh, and Davide Scaramuzza. *Introduction to autonomous mobile robots*. MIT press, 2011 (cit. on p. 3).

[Soh+15] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". *International Conference on Machine Learning*. 2015 (cit. on p. 148).

[Son+14] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. "Nonparametric Estimation of Multi-View Latent Variable Models". *International Conference on Machine Learning*. 2014 (cit. on p. 28).

[Son+21] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. "Score-Based Generative Modeling through Stochastic Differential Equations". *International Conference on Learning Representations*. 2021 (cit. on p. 148).

[Spe11] Charles Spence. "Crossmodal correspondences: A tutorial review". *Attention, Perception, & Psychophysics* 73 (2011), pp. 971–995 (cit. on p. 3).

[Spi+00] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000 (cit. on p. 26).

[SS08] Barry E. Stein and Terrence R. Stanford. "Multisensory integration: current issues from the perspective of the single neuron". *Nature Reviews Neuroscience* 9.4 (2008), pp. 255–266 (cit. on p. 2).

[SSK12] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. "Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation". *Annals of the Institute of Statistical Mathematics* 64.5 (2012), pp. 1009–1044 (cit. on p. 134).

[SSR14] Barry E. Stein, Terrence R. Stanford, and Benjamin A. Rowland. "Development of multisensory integration from the perspective of the individual neuron". *Nature Reviews Neuroscience* 15.8 (2014), pp. 520–535 (cit. on p. 2).

[Ste12] Barry E. Stein. *The new handbook of multisensory processing*. MIT Press, 2012 (cit. on p. 2).

[Sug+22] Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine learning from weak supervision: An empirical risk minimization approach*. MIT Press, 2022 (cit. on p. 18).

[Sze+15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". *Conference on Computer Vision and Pattern Recognition*. 2015 (cit. on p. 20).

[SZW14] Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. "An extension of slow feature analysis for nonlinear blind source separation". *Journal of Machine Learning Research* 15.1 (2014), pp. 921–947 (cit. on p. 26).

[TF00] Joshua B. Tenenbaum and William T. Freeman. "Separating style and content with bilinear models". *Neural computation* 12.6 (2000), pp. 1247–1283 (cit. on p. 14).

[TF96] Joshua B. Tenenbaum and William T. Freeman. "Separating style and content". *Advances in Neural Information Processing Systems* (1996) (cit. on p. 14).

[The06] Fabian J. Theis. "Towards a general independent subspace analysis". *Advances in Neural Information Processing Systems*. 2006 (cit. on p. 27).

[Tia+20] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. "What Makes for Good Views for Contrastive Learning?" *Advances in Neural Information Processing Systems*. 2020 (cit. on pp. 41, 44).

[TKI20] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive Multiview Coding". *European Conference on Computer Vision*. 2020 (cit. on p. 127).

[TOB16] Lucas Theis, Aäron van den Oord, and Matthias Bethge. "A note on the evaluation of generative models". *International Conference on Learning Representations*. 2016 (cit. on pp. 21, 140).

[Tsa+21] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. "Self-supervised Learning from a Multi-view Perspective". *International Conference on Learning Representations*. 2021 (cit. on p. 44).

[Tsc+20] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. "On Mutual Information Maximization for Representation Learning". *International Conference on Learning Representations*. 2020 (cit. on p. 44).

[Tye03] Michael Tye. *Consciousness and persons: Unity and identity*. MIT Press, 2003 (cit. on p. 3).

[Uex92] Jakob von Uexküll. "A stroll through the worlds of animals and men: A picture book of invisible worlds" (1992) (cit. on p. 3).

[Vas+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". *Advances in Neural Information Processing Systems*. 2017 (cit. on p. 148).

[Ved+18] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin P. Murphy. "Generative Models of Visually Grounded Imagination". *International Conference on Learning Representations*. 2018 (cit. on p. 39).

[Vir+12] Seppo Virtanen, Arto Klami, Suleiman Khan, and Samuel Kaski. "Bayesian group factor analysis". *Artificial Intelligence and Statistics*. 2012 (cit. on p. 14).

[Wah+11] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. Tech. rep. CNS-TR-2011-001. California Institute of Technology, 2011 (cit. on p. 50).

[Wan+14] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. "Learning fine-grained image similarity with deep ranking". *Conference on Computer Vision and Pattern Recognition*. 2014 (cit. on p. 41).

[WG18] Mike Wu and Noah D. Goodman. "Multimodal Generative Models for Scalable Weakly-Supervised Learning". *Advances in Neural Information Processing Systems*. 2018 (cit. on pp. 5, 39, 40, 55, 56, 58, 60, 62, 82, 90–92, 106, 131, 136).

[WG19] Mike Wu and Noah D. Goodman. "Multimodal Generative Models for Compositional Representation Learning". *arXiv preprint arXiv:1912.05075* (2019) (cit. on p. 56).

[WI20] Tongzhou Wang and Phillip Isola. "Understanding contrastive representation learning through alignment and uniformity on the hypersphere". *International Conference on Machine Learning*. 2020 (cit. on pp. 43, 44, 112, 113).

[Wu+18] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. "Unsupervised feature learning via non-parametric instance discrimination". *Conference on Computer Vision and Pattern Recognition*. 2018 (cit. on p. 41).

[Yil14] Ilker Yildirim. "From perception to conception: learning multisensory representations". PhD thesis. University of Rochester, 2014 (cit. on p. 3).

[Yon22] Ed Yong. *An immense world: How animal senses reveal the hidden realms around us*. Random House, 2022 (cit. on p. 3).

[Zbo+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". *International Conference on Machine Learning*. 2021 (cit. on p. 42).

[Zha+22] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. "Contrastive Learning of Medical Visual Representations from Paired Images and Text". *Machine Learning for Healthcare*. 2022 (cit. on p. 45).

[Zho17] Zhi-Hua Zhou. "A brief introduction to weakly supervised learning". *National Science Review* 5.1 (2017), pp. 44–53 (cit. on p. 18).

[Zim+21] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. "Contrastive Learning Inverts the Data Generating Process". *International Conference on Machine Learning*. 2021 (cit. on pp. 26, 51, 110, 121, 122, 127).

[ZZC18] Yexun Zhang, Ya Zhang, and Wenbin Cai. "Separating style and content for generalized style transfer". *Conference on Computer Vision and Pattern Recognition*. 2018 (cit. on p. 14).

[Ble18] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. 2018. URL: http://www.blender.org (cit. on p. 51).

[von+21] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. "Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style". *Advances in Neural Information Processing Systems*. 2021 (cit. on pp. 28, 44, 51, 110, 112, 113, 115–118, 120–122, 127, 165).

[von67] Hermann von Helmholtz. *Handbuch der physiologischen Optik*. Leopold Voss, 1867 (cit. on p. 3).