ABSTRACT

Title of Thesis: LEARNING REPRESENTATIONS FOR AUDIO

Siddhi Patil
Masters, 2026

Thesis Directed by: Professor Ramani Duraiswami
Department of Computer Science

Learning audio representations that are simultaneously useful for spatial reasoning and semantic understanding remains a persistent challenge in machine learning. Existing approaches tend to specialize: models designed for spatial audio often overfit to specific microphone array geometries or rely heavily on phase-based cues that do not generalize well across recording conditions, while large-scale audio encoders trained on diverse datasets typically prioritize semantic content and largely ignore the underlying structure of the sound field. This disconnect limits the development of unified representations that can generalize across both spatial and semantic tasks. Recent advances in self-supervised learning, particularly methods that predict latent targets in feature space rather than reconstructing raw signals, offer a promising direction for bridging this gap. However, these approaches have not yet fully reconciled the competing demands of spatial awareness and semantic richness, especially in settings with limited supervision or varying acoustic conditions.

In this work, we conduct a detailed analysis of the representations learned by modern self-supervised audio models, with a focus on understanding how different pretext tasks influence the

encoding of spatial and semantic information. Motivated by these insights, we design a set of self-supervised pretext tasks that operate directly on unstructured audio data, encouraging the model to capture motion cues, spatial consistency, and temporal dynamics without relying on explicit labels or structured inputs. These tasks are specifically crafted to introduce inductive biases that promote generalization across a wide range of downstream applications, including localization, scene understanding, and audio classification. While operating on unstructured data introduces additional computational overhead compared to pipelines that rely on pre-computed features such as log-mel spectrograms, we address this challenge by replacing standard softmax attention with a normalized attention mechanism. This modification significantly reduces the computational cost and memory footprint of transformer-based encoders while preserving their capacity to model long-range dependencies and complex temporal relationships in audio signals.

We evaluate the learned representations on a diverse set of benchmarks spanning both spatial and non-spatial tasks, including sound source localization, acoustic scene classification, and general audio tagging. Our results demonstrate that motion-informed pre-training, when combined with efficient attention mechanisms, enables models to approach—and in some cases match—the performance of larger, label-intensive baselines that rely on extensive supervised training. Notably, our approach achieves this while requiring substantially less annotated data, highlighting its effectiveness in data-scarce regimes. Furthermore, we provide qualitative and quantitative analyses showing that the learned representations capture both spatial structure and semantic content in a more balanced and interpretable manner than existing methods. These findings suggest that integrating spatial inductive biases with computationally efficient latent prediction objectives offers a promising path toward unified, data-efficient audio representations that generalize across a broad family of downstream tasks and real-world acoustic environments.

LEARNING REPRESENTATIONS FOR AUDIO

by

Siddhi Patil

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Computer Science
2026

Advisory Committee:
Professor Ramani Duraiswami, Chair/Advisor
Dr. Cornelia Fermüller
Professor Nirupam Roy

© Copyright by
Siddhi Patil
2026

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Ramani Duraiswami for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past year. It has been a pleasure to work with and learn from such an extraordinary individual.

I would like to thank Dr. Cornelia Fermüller for being a wonderful mentor to me. She has inspired me to turn my focus to audio research and for that I am grateful.

I owe my deepest thanks to my sister. She is my mentor, hero and best friend. She has inspired me to pursue excellence in everything that I do. I'd thank my mother and father who have always stood by me and guided me through my career, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Thank you for reading this!

Table of Contents

Preface ii

Dedication ii

Acknowledgements iii

Table of Contents iv

List of Tables vi

List of Figures vii

Abbreviations vii

Chapter 1: Introduction 1
- 1.1 Introduction . . . 1
- 1.2 Toward holistic audio understanding . . . 5
- 1.3 Representation learning for audio: semantic vs spatial . . . 6
- 1.4 Disentangled and geometry-preserving representations . . . 7
- 1.5 Efficient self-supervision on unstructured spatial audio . . . 9

Chapter 1: Completed Works 10
- 1.1 MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence . . . 10
  - 1.1.1 Related Works . . . 13
  - 1.1.2 The MMAU-Pro Benchmark . . . 14
  - 1.1.3 Experimental Setup . . . 18
  - 1.1.4 Results and Discussion . . . 20
  - 1.1.5 Do MLLMs Retain Skills Acquired from Text Pre-training? . . . 21
  - 1.1.6 Conclusion, Limitations and Future Work . . . 22
- 1.2 A Bayesian Non-parametric Framework for Learning Disentangled Representations . . . 23
  - 1.2.1 Introduction . . . 24
  - 1.2.2 Nonparametric Bayesian Priors for Autoencoders . . . 27
  - 1.2.3 Structured Nested Variational Family . . . 31
  - 1.2.4 The Inference Algorithm . . . 33
  - 1.2.5 Experiments . . . 37
  - 1.2.6 Related Works . . . 40
  - 1.2.7 Conclusion . . . 42
  - 1.2.8 Preliminaries . . . 43
- 1.3 Geometry-Preserving Disentanglement: Isometric Embeddings for Combinatorial Generalization 53
  - 1.3.1 Introduction . . . 56
  - 1.3.2 Latent Quantization . . . 59

1.3.3 Dirichlet Process prior for Stochastic Quantization . . . 60
1.3.4 Geometry-Preserving Isometric Manifolds . . . 61
1.3.5 Structured Latent Space . . . 62
1.3.6 Isometric Generative Map . . . 64
1.3.7 Experimental Design . . . 67
1.3.8 Evaluation Metrics . . . 68
1.3.9 Evaluation . . . 69
1.3.10 Related Works . . . 73
1.3.11 Discussion . . . 75

Appendix A: Experiments 77
A.1 Problem Definition . . . 77
A.2 Using the Indian Buffet Process for Continual Learning of Disentangled Representations 78
A.3 Related Works . . . 79

Bibliography 81

Bibliography 82

List of Tables

1.1 Disentanglement metrics measured in InfoMEC and DCI for 3DShapes dataset. For each metric a higher score is better. The scores for all the models were averaged across 5 runs with different random seeds with intervals denoting 95% confidence intervals of the mean estimated assuming a t-distribution. The results for the VQE-based and QLAE-based models are obtained using the hyperparameter settings and experimental conditions as described in [28] and [32, 65] respectively. 39
1.2 Disentanglement metrics measured in InfoMEC and DCI for MPI3D dataset. For each metric a higher score is better. The scores for all the models were averaged across 5 runs with different random seeds with intervals denoting 95% confidence intervals of the mean estimated assuming a t-distribution. 40
1.3 Fixed Hyperparameters and initializations 48
1.4 Model performance comparison across different information metrics 48
1.5 Quantitative comparison of disentanglement metrics and combinatorial generalization task performance across autoencoder architectures on Shapes3D and MPI3D benchmarks. We report the Information-theory based Modularity, Explicitness Compactness (InfoMEC), Disentanglement-Completeness-Informativeness (DCI), binary cross entropy cost for the recombination-to-range (R2R) and the Extrapolation E tasks, and classification accuracy on the recombination-to-range task (ACC). Arrows indicate desired direction: $\uparrow$ denotes higher is better, $\downarrow$ denotes lower is better. 69
1.6 Results on Shapes3D (mean $\pm$ std over 5 random seeds). 70
1.7 Results on MPI3D (mean $\pm$ std over 4 random seeds). 70
1.8 Ablation study on BayesQLAE variants for the Shapes3D dataset. DCI measures disentanglement (higher is better), while R2R and E assess recombination-to-range and extrapolation errors (lower is better), showing that the structured isometric BayesQLAE model achieves the best overall performance on both tasks. 72

List of Figures

1.1 Reconstructions and Latent traversals for the 3DShapes dataset: Reconstructions demonstrate high fidelity in capturing visual details demonstrating the model's ability to faithfully encode and decode the full range of variations in the data. Latent traversals illustrate that individual latent variables control distinct, interpretable factors of variation in the generated images. Moreover, the presence of inactive latent variables and the encoding of each factor in a single latent variables indicates that the model has learned a compact representation recovering the true generative structure 54

1.2 Evolution of Latent Traversals for 3DShapes dataset: The figure illustrates the evolution of the number of clusters associated with each generative factor demonstrating the adaptive capacity of the nonparametric formulation. Each column in each row corresponds to a factor-specific mixture component, and the distinct components within a row denote the clusters capturing the encoded variations of that factor. The vertical axis indicates cluster multiplicity, revealing how the model progressively discovers and encodes additional variations for each factor. This dynamic cluster growth exemplifies the nonparametric property of the hierarchical mixture prior, which enables data-driven inference of latent capacity without manual specification. Notably, factors with higher contribution to the reconstruction objective—such as floor hue, object hue, and wall hue—exhibit earlier cluster proliferation during training, suggesting the model prioritizes encoding variations that most significantly impact reconstruction fidelity. In contrast, geometric factors such as object orientation and shape undergo refinement in later training stages, indicating a hierarchical learning strategy wherein the model first captures high-variance attributes before refining lower-variance structural properties. This demonstrates that the nonparametric prior successfully balances model capacity across factors according to their respective complexities and contributions to data likelihood 55

List of Abbreviations

KL divergence	Kullback-Leibler divergence LALM f Large Audio-Language Model
LLM	Large Language Model
MLLM	Multimodal Large Language Model
MMAU	Multi-Modal Audio Understanding

Chapter 1: Introduction

1.1 Introduction

Learning audio representations that are useful for both spatial reasoning and semantic understanding remains a central and unresolved challenge in machine learning. Existing models tend to specialize either in content recognition or spatial processing, but rarely excel at both. Models tailored toward spatial audio often overfit to specific microphone array geometries or exploit phase-based features that do not transfer robustly across recording configurations and acoustic environments, limiting their applicability in realistic, heterogeneous scenarios. In contrast, large-scale audio encoders trained on vast collections of speech, environmental sounds, and music primarily focus on semantic content, treating audio as a one-dimensional signal and largely ignoring the rich structure of the surrounding sound field. This disconnect produces representations that are strong for classification and captioning, yet brittle when downstream tasks require precise spatial reasoning, such as localization, motion tracking, or spatialized scene understanding.

Recent self-supervised representation learning approaches offer a promising alternative to purely supervised pipelines. Instead of reconstructing raw waveforms or spectrograms, modern latent-prediction methods learn by predicting masked or transformed features in a representation space, often yielding more robust and transferable embeddings. Such approaches have demonstrated substantial gains across vision and language, and analogous techniques have begun to show benefits in audio as well, particularly for large audio–language models (LALMs) that couple audio encoders with language backbones. However, most existing self-supervised audio objectives are still optimized for semantic understanding—e.g., identifying phonetic content, categories of sounds, or musical attributes—without explicitly encoding spatial structure, motion cues, or geometry of the sound field. As a result, they have

yet to fully reconcile the dual requirements of spatial awareness and semantic richness in a unified representation.

In this thesis, we aim to close this gap by developing self-supervised methods that learn general-purpose audio representations capable of supporting both spatial and non-spatial tasks. Concretely, we study how pretext tasks, inductive biases, and architectural choices influence the extent to which learned embeddings preserve spatial information while retaining strong semantic discriminability. Building on recent insights from disentangled representation learning and geometry-preserving embeddings, we design self-supervised objectives that operate directly on unstructured spatial audio data and explicitly encourage motion-aware, geometry-sensitive representations. At the same time, we address the computational challenges introduced by operating on raw or lightly processed signals, proposing efficient attention mechanisms that reduce the cost of transformer-based encoders while preserving their capacity to model long-range temporal dependencies. Through extensive evaluation on both spatial and non-spatial benchmarks, we demonstrate that motion-informed pre-training with efficient attention can approach the performance of much larger, label-intensive baselines, while using significantly less explicit supervision.

The remainder of this chapter situates this thesis in the broader context of audio general intelligence, disentangled representation learning, and geometry-aware modeling. We first outline the emerging view of holistic audio understanding and discuss the limitations of current models and benchmarks. We then review representation learning in audio, highlighting the divide between spatial and semantic models. Next, we connect these ideas to the literature on disentanglement and geometry-preserving representations, arguing that spatial inductive biases can be seen as a domain-specific instance of more general structural and functional constraints. Finally, we summarize the main contributions of this thesis and discuss how they move toward unified, data-efficient audio representations suitable

for a broad family of downstream tasks.Learning audio representations that are useful for both spatial reasoning and semantic understanding remains a central and unresolved challenge in machine learning. Existing models tend to specialize either in content recognition or spatial processing, but rarely excel at both. Models tailored toward spatial audio often overfit to specific microphone array geometries or exploit phase-based features that do not transfer robustly across recording configurations and acoustic environments, limiting their applicability in realistic, heterogeneous scenarios. In contrast, large-scale audio encoders trained on vast collections of speech, environmental sounds, and music primarily focus on semantic content, treating audio as a one-dimensional signal and largely ignoring the rich structure of the surrounding sound field. This disconnect produces representations that are strong for classification and captioning, yet brittle when downstream tasks require precise spatial reasoning, such as localization, motion tracking, or spatialized scene understanding.

Recent self-supervised representation learning approaches offer a promising alternative to purely supervised pipelines. Instead of reconstructing raw waveforms or spectrograms, modern latent-prediction methods learn by predicting masked or transformed features in a representation space, often yielding more robust and transferable embeddings. Such approaches have demonstrated substantial gains across vision and language, and analogous techniques have begun to show benefits in audio as well, particularly for large audio–language models (LALMs) that couple audio encoders with language backbones. However, most existing self-supervised audio objectives are still optimized for semantic understanding—e.g., identifying phonetic content, categories of sounds, or musical attributes—without explicitly encoding spatial structure, motion cues, or geometry of the sound field. As a result, they have yet to fully reconcile the dual requirements of spatial awareness and semantic richness in a unified representation.

In this thesis, we aim to close this gap by developing self-supervised methods that learn general-

purpose audio representations capable of supporting both spatial and non-spatial tasks. Concretely, we study how pretext tasks, inductive biases, and architectural choices influence the extent to which learned embeddings preserve spatial information while retaining strong semantic discriminability. Building on recent insights from disentangled representation learning and geometry-preserving embeddings, we design self-supervised objectives that operate directly on unstructured spatial audio data and explicitly encourage motion-aware, geometry-sensitive representations. At the same time, we address the computational challenges introduced by operating on raw or lightly processed signals, proposing efficient attention mechanisms that reduce the cost of transformer-based encoders while preserving their capacity to model long-range temporal dependencies. Through extensive evaluation on both spatial and non-spatial benchmarks, we demonstrate that motion-informed pre-training with efficient attention can approach the performance of much larger, label-intensive baselines, while using significantly less explicit supervision.

The remainder of this chapter situates this thesis in the broader context of audio general intelligence, disentangled representation learning, and geometry-aware modeling. We first outline the emerging view of holistic audio understanding and discuss the limitations of current models and benchmarks. We then review representation learning in audio, highlighting the divide between spatial and semantic models. Next, we connect these ideas to the literature on disentanglement and geometry-preserving representations, arguing that spatial inductive biases can be seen as a domain-specific instance of more general structural and functional constraints. Finally, we summarize the main contributions of this thesis and discuss how they move toward unified, data-efficient audio representations suitable for a broad family of downstream tasks.

1.2 Toward holistic audio understanding

Comprehensive audio understanding—from spoken language and paralinguistic cues to environmental sounds and music—is increasingly recognized as a core component of machine perception and, more broadly, artificial general intelligence. Human listeners seamlessly integrate semantic content ("what is being said or sounded?") with spatial context ("where is it coming from?", "how is it moving?") and scene-level structure ("how many sources?", "what is the room like?"). For AI agents to interact effectively in the real world, they must inherit similar capabilities, using sound not only as a label-bearing modality but also as a rich source of spatial and situational cues.

Large audio–language models (LALMs) and multimodal large language models (MLLMs) have made impressive progress on high-level audio comprehension. These systems can answer questions about speech, identify sound categories, follow audio instructions, and even reason over long-form recordings that combine speech, non-speech sounds, and music. Benchmarks such as MMAU and its extension MMAU-Pro have begun to evaluate these capabilities systematically, covering a wide spectrum of skills including multi-hop reasoning, long-form audio understanding, and complex mixtures of speech, sounds, and music. Crucially, these benchmarks also introduce tasks that probe spatial audio reasoning, multi-audio understanding, and realistic "from-the-wild" scenarios, revealing that even state-of-the-art models still fall far short of human-level audio intelligence.

Despite this progress, existing evaluations highlight a critical shortcoming: most high-performing models excel at semantic recognition but struggle when questions require detailed spatial reasoning, nuanced motion understanding, or robust performance in acoustically complex environments. For instance, questions about the relative positions of speakers, the order in which a waiter takes orders in a restaurant, or the spatial layout of a sound scene expose weaknesses in representations that do

not encode spatial structure explicitly. This suggests that simply scaling up semantic encoders and training on larger datasets is insufficient; instead, we need representation learning approaches that integrate spatial inductive biases and motion-aware objectives into the audio modeling pipeline.

In this thesis, we adopt the perspective that robust audio intelligence must rest on general-purpose representations that natively encode both semantic and spatial attributes. Rather than designing narrowly specialized architectures for each task, the goal is to learn a single, versatile embedding space where downstream classifiers or decoders can easily access information about content, location, motion, and scene context. Achieving this requires revisiting the design of pretext tasks, model architectures, and training regimes in the audio domain.

1.3 Representation learning for audio: semantic vs spatial

Representation learning has become the dominant paradigm in modern audio processing, mirroring developments in computer vision and natural language processing. Self-supervised methods that learn from large corpora without manual labels have been especially influential, enabling models to discover structure in speech, environmental sounds, and music by solving proxy tasks such as masked prediction, contrastive learning, or reconstruction. In speech, for example, self-supervised encoders trained with masked prediction objectives have yielded powerful representations that support a wide variety of downstream tasks, from automatic speech recognition to speaker identification and emotion recognition.

However, most of these methods treat audio as a one-dimensional signal, or at best as a single-channel spectrogram, and therefore implicitly focus on semantic content. They deliberately discard spatial cues such as inter-channel time differences, directional energy distributions, or room impulse

responses, either by collapsing channels or by using front-end features that average over spatial structure. This design choice simplifies modeling and improves transfer to content-centric tasks, but it also erases information that is crucial for localization, spatial separation, and spatialized scene understanding.

In parallel, a separate line of work in spatial audio and acoustic signal processing has developed models that operate directly on multi-channel recordings, ambisonics, or microphone arrays. These methods exploit phase differences, array geometry, and beamforming-inspired features to estimate directions of arrival, track moving sources, and infer room acoustics. While highly effective in controlled setups, these models often bake in strong assumptions about sensor placement, array shape, and acoustic conditions, which leads to severe overfitting and poor generalization when confronted with different geometries or real-world "in-the-wild" recordings. Moreover, they are usually trained for narrow tasks, such as localization or dereverberation, and therefore produce representations that are not easily reusable for semantic understanding. This bifurcation—semantic encoders that ignore spatial structure, and spatial models that overfit to specific geometries—creates an artificial divide in the audio representation space. A central thesis of this work is that we can bridge this gap by designing self-supervised objectives that operate directly on unstructured spatial audio data and explicitly encourage the encoder to capture both spatial and semantic factors. Rather than treating spatial audio as a separate domain, I view it as an enriched version of the standard audio setting, where additional channels and motion cues provide structure that can be exploited by appropriately biased pretext tasks.

1.4 Disentangled and geometry-preserving representations

The goal of learning representations that factorize different aspects of the data is not unique to audio. In unsupervised and self-supervised learning more broadly, disentangled representations

have been proposed as a means to discover and organize latent factors of variation—such as shape, pose, lighting, or object identity—without explicit supervision. The core idea is that individual latent dimensions or subspaces should correspond to distinct, meaningful factors, thereby enabling systematic generalization to novel combinations of these factors. When successful, disentangled representations improve robustness, out-of-distribution generalization, few-shot learning, and controllable generation, as demonstrated on benchmarks such as 3DShapes and MPI3D.

However, unsupervised disentanglement faces fundamental identifiability challenges: without inductive biases, infinitely many encodings can explain the same observations, making it impossible to guarantee recovery of the true generative factors. To address this, prior work has introduced a variety of inductive biases, including information-theoretic regularizers, architectural constraints, and functional constraints on the data-generating mapping. More recent approaches leverage Bayesian nonparametric priors over discrete latent codebooks or impose geometry-preserving constraints that encourage the latent space to form approximately isometric embeddings of underlying factor manifolds. These advances suggest that carefully designed structural and geometric biases can lead to more interpretable and compositional representations.

The spatial audio setting provides a natural playground for such ideas. Many spatial attributes—such as source azimuth, elevation, distance, or trajectory—can be viewed as coordinates on smooth manifolds embedded in observation space. Geometry-preserving inductive biases that maintain local distances and angles between latent representations, as proposed in recent work on isometric embeddings for combinatorial generalization, align well with the goal of capturing spatial structure in a way that supports recombination and generalization. Similarly, viewing spatial and semantic factors as distinct but interacting generative components resonates with methods that factorize latent spaces and assign separate codebooks or subspaces to different factors.

In this thesis, I adapt and extend these ideas to the audio domain by treating spatial cues and semantic content as intertwined but partially separable factors of variation. Rather than striving for perfect disentanglement—an unrealistic goal given the complex coupling between acoustics and content—I focus on learning representations that make both kinds of information linearly or locally accessible to downstream models. Motion-informed pretext tasks, such as predicting future spatial configurations or reconstructing relative motion trajectories in latent space, serve as functional inductive biases that steer the encoder toward capturing geometry and dynamics, while latent prediction objectives ensure that the learned space remains semantically rich and robust.

1.5 Efficient self-supervision on unstructured spatial audio

Operating directly on unstructured spatial audio data introduces significant computational challenges. Unlike pipelines that rely on pre-computed features (e.g., log-mel spectrograms or handcrafted spatial descriptors), models that ingest raw or minimally processed multi-channel waveforms must process high-dimensional, long-duration inputs. Transformer-based architectures are attractive in this setting because of their ability to model long-range dependencies, but standard softmax attention incurs quadratic complexity in sequence length, making naive application to high-resolution spectrograms or long sequences prohibitively expensive.

To make self-supervised learning on unstructured spatial audio tractable, I combine motion-aware pretext tasks with computationally efficient attention mechanisms. Specifically, I replace the standard softmax attention with a normalized attention formulation that reduces complexity while preserving the ability to aggregate information over long time horizons. This modification reduces memory footprint and computational cost, enabling training on longer inputs and more channels

without sacrificing the expressiveness of the encoder. By integrating this efficient attention into latent prediction frameworks, the model can learn to capture both fine-grained temporal details and global spatial structure. This architectural choice is tightly coupled with the design of pretext tasks. Latent prediction objectives that operate on spectrogram-like representations can be adapted to multi-channel settings by jointly predicting masked or transformed features across channels, encouraging the encoder to align spatially corresponding regions and to exploit inter-channel coherence. Motion-informed tasks, such as predicting future latent states conditioned on past spatial configurations, further bias the representations toward capturing temporal dynamics and source trajectories. Together, these design elements pave the way for general-purpose spatial audio representations that maintain semantic fidelity while "seeing" the sound field as a structured, dynamic object.

Chapter 1: Completed Works

1.1 MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence

Comprehensive audio understanding-from spoken language to environmental sounds and music-is fundamental to human general intelligence. Correspondingly, AI systems must possess comparable capabilities for effective real-world interaction (Sakshi et al. 2025). Recent advancements in multimodal large language models (MLLMs) have led to the emergence of Large Audio-Language Models (LALMs), demonstrating notable audio comprehension skills (Ghosh et al. 2024, 2025b; Goel et al. 2025;

Gong et al. 2024; Deshmukh et al. 2023; KimiTeam et al. 2025; Xie et al. 2025a; Xu et al. 2025; Chu et al. 2024). Despite numerous benchmarks assessing progress toward Artificial General Intelligence (AGI) through text, audio intelligence evaluation remains notably underserved. Given audio's inherent diversity and complexity, we contend that progress toward AGI is incomplete without strong audio intelligence capabilities-and that their rigorous evaluation remains an open challenge. Recently, several benchmarks have emerged to evaluate LALMs. MMAU (Sakshi et al. 2025), a pioneering comprehensive benchmark, comprises 10,000 carefully selected audio clips across speech, sounds, and music, with single- turn, single-audio questions requiring knowledge and reasoning. Following MMAU, MMAR (Ma et al. 2025) introduced more challenging queries, while MMSU (Wang et al. 2025b) expanded spoken language understanding assessments. Domain-specific benchmarks like Speech-IFEval (Lu, Kuan, and Lee 2025) focus on instruction-following and CMM (Leng et al. 2024) focuses on hallucinations. Nevertheless, existing benchmarks inadequately represent the complexity of realistic auditory scenarios - such as multiple and overlapping audios, long-duration inputs, open-ended answers, and culturally varied content-which demand deeper comprehension and multi-hop reasoning beyond basic recognition.

To this end, we present MMAU-Pro, a novel benchmark consisting of 5,305 expert-annotated instances designed to evaluate 49 distinct auditory intelligence skills spanning speech, environmental sounds, and music. MMAU-Pro presents challenges overlooked by prior benchmarks, including long-form audio understanding (up to 10 minutes), reasoning across multiple clips, spatial audio perception, multicultural music interpretation, instruction-following abilities, etc. All questions are crafted to require deliberate multi-hop reasoning and include a balanced mix of multiple-choice and open-ended formats. To address the shortcomings of existing evaluation methodologies, we further propose a retrieval-based evaluation framework that enables more robust and reliable assessment. By

emphasizing realistic and demanding auditory tasks, MMAU-Pro provides a comprehensive testbed to accelerate the development of auditory intelligence in multimodal AI systems. To summarize, our main contributions are:

1. We introduce MMAU-Pro, the most comprehensive benchmark to date for evaluating auditory intelligence. It comprises 5,305 expert-annotated question–answer pairs spanning 49 distinct skills across speech, environmental sounds, music, and their mixtures. MMAU-Pro introduces novel challenges, including spatial audio reasoning, multi-clip audio reasoning, voice–chat comprehension, and tasks requiring prosodic, world-knowledge, and STEM-based reasoning. All audio samples are drawn from the wild, with durations up to ten minutes, significantly surpassing the short clips typical of prior benchmarks where current models are near-saturated.

2. We benchmark over 15 open-source and proprietary multimodal LLMs on MMAU-Pro, finding that even the strongest models face substantial challenges. Gemini 2.5 Flash achieves only 59.2% accuracy; the best-performing fully open-source model, Audio Flamingo 3, reaches 51.7%; and the strongest open-weights omni model, Qwen2.5-Omni-7B-Instruct, achieves just 52.2%.

3. We provide an in-depth analysis of model responses, uncovering key failure modes in auditory perception and reasoning. These include shallow audio grounding, degradation in text-only and STEM reasoning, poor performance in multi-audio and spatial reasoning, and limited understanding of multicultural music.

1.1.1 Related Works

Recent advances in multimodal modeling have led to (L)ALMs-models that pair audio perception with (L)LMs to tackle complex audio tasks. Early systems such as Whisper [152] [160] and CLAP [170] [137]; [138] focused on foundational tasks like transcription, captioning, and retrieval, but struggled with reasoning-centric challenges. More recent models-GAMA [142], Audio Flamingo [141], Mellow [134], Phi-4MM [129], Qwen2-Audio [133], and AudioPALM [163] proposed improved architectures and large-scale training, targeting deeper understanding. These efforts have culminated in large audio reasoning models (LARMs), including Audio-Reasoner [171], SoundMind [135], R1-AQA [153], and Audio-CoT [171] which explicitly model step-by-step reasoning. In parallel, general purpose Omni-Language Models (OLMs) such as Qwen2.5-Omni [172], Baichuan-Omni [154], and Ming-Omni [144] - though not tailored for audio have demonstrated surprising proficiency on audio tasks. While progress is promising, robust benchmarks remain essential to evaluate audio intelligence.

Existing benchmarks provide strong foundations but fall short in evaluating holistic audio intelligence. MMAU [164] introduced 10,000 QA pairs across 27 skills for speech, sounds, and music, but used existing datasets and short, single-source clips-achieving only 52–60% accuracy. MMAR [156] added 1,000 real-world QA triplets with hierarchical reasoning layers and rationales, yet remained limited in scale and scope. AudioBench [166] unifies 26 datasets across 8 tasks, while MuChoMusic [169] probes 1,100 MCQs on culturally diverse music, exposing models' over-reliance on text. MMSU [59] tests 5,000 spoken language QA pairs across 47 speech skills. Beyond Single-Audio [132] evaluates multi-audio reasoning across 20 datasets, showing that most ALLMs struggle when reasoning over more than one audio stream. Dynamic SUPERB Phase-2 [145] expands to 180 tasks. While some recent models, such as Mellow [134] and BAT [175], begin to address multi- and spatial-audio tasks,

benchmark evaluations remain shallow and fragmented. Moreover, no existing benchmark systematically tests instruction following or jointly evaluates long-form (up to 10 minutes), multi-audio, spatial, open-ended, and multicultural scenarios. Addressing these gaps, MMAU-Pro offers the most comprehensive benchmark to date, targeting underexplored dimensions critical to advancing real-world audio general intelligence.

1.1.2 The MMAU-Pro Benchmark

MMAU-Pro is designed to holistically evaluate audio intelligence in AI systems. It comprises 5,305 expert-annotated question–answer pairs covering 49 distinct skills. Questions require multi-step, multi-hop reasoning and were authored and validated by domain experts to ensure high quality. To avoid data leakage, all audio (except our spatial subset) is sourced from in-the-wild recordings. For spatial audio reasoning, we reuse high-fidelity multi-channel recordings from the EasyCom dataset [136]. While prior benchmarks such as MMAU and MMAR primarily evaluate models using multiple-choice questions, MMAU-Pro extends evaluation to include open-ended responses and MCQs with up to 10 options, thereby substantially reducing the likelihood of models succeeding by random guessing. It also categorizes audio clips by duration: short ($\leq$ 30s), medium (30s–3min), long (3–8min), and ultra-long (8–10min), enabling characterization and analysis across varying temporal contexts.

Inspired by prior benchmarks in this space, we design a specialized multi-stage pipeline with more human involvement in the process to construct high-quality data for MMAU-Pro.

1. **Domain and Task Design:** We define a diverse set of reasoning tasks across speech, sound, music, and their mixtures, including long-context comprehension, spatial reasoning, multi-audio analysis, and multicultural music understanding.

2. **Task Allocation:** Domain experts (authors) are assigned tasks based on specialization, guided by detailed instructions to ensure comprehensive domain coverage.

3. **QA Generation:** The experts then manually collect audio and craft QA pairs. The QA pairs are created with an emphasis on multi-hop reasoning and real-world use cases. We include both MCQs and open-ended questions in the benchmark, with explanatory answers authored for the latter. Annotators determine the appropriate format based on factors such as susceptibility to elimination via language cues and the added value of open-endedness for deeper evaluation.

4. **Distractor Construction:** For MCQs, experts create challenging distractors that discourage superficial pattern matching. Distractors are crafted to avoid trivial elimination and encourage careful audio grounding. Unlike other benchmarks, which resort to LLMs for the generation of distractors, experts carefully create distractors for each question to pose a higher level of challenge to the models being evaluated.

5. **Annotation Verification:** A second expert independently verifies each QA instance for accuracy, clarity, and reasoning validity. Discrepancies are resolved iteratively, followed by grammar and style checks using both experts and LLMs.

6. **Expert Review:** Final review ensures cultural sensitivity, task appropriateness, and explanatory quality, particularly for open-ended responses.

7. **Benchmark Finalization:** The finalized dataset balances domain, task type, and audio length to ensure diverse and representative coverage.

Over 25 individuals were involved in this process of data collection, QA categorization and design, validation, curation, and evaluation.

MMAU-Pro introduces several key advancements, including multi-audio reasoning, spatial audio understanding, and STEM-based evaluation. In the following subsections, we describe these core innovations in detail.

Long-Audio Understanding Previous benchmarks such as MMAU (avg. 10.1 sec), MMAR (avg. 19.4 sec), and MMSU (7.01 sec) primarily focus on short audio clips, limiting evaluation to brief segments. Audio Flamingo 2 introduced long-audio perception with LALMs, motivated by real-world applications such as video, podcast, and movie analysis. This was followed by models like Qwen2.5-Omni and Audio Flamingo 3. MMAU-Pro builds on recent efforts such as LongAudioBench [141] and BLAB [130] by incorporating long-form audio inputs, categorized into four duration bins: short ($\leq$ 30s), medium (30s–3min), long (3–8min), and ultralong (8–10min), comprising 2,589; 1,897; 1,307; and 348 QA instances respectively. Long-form comprehension poses unique challenges-such as locating sparse events ("needle in a haystack") and understanding narrative or temporal structure, which is explicitly tested in MMAU-Pro through specialized QA designs.

Multi-Audio Understanding While multi-image understanding has been extensively studied [146] [174] [151], multi-audio understanding remains largely underexplored. Although many real-world use cases require understanding and reasoning over multiple audio inputs, most frontier MLLMs with audio perception capabilities do not natively support multi-audio processing. Chen et al. [132] make an initial attempt, and Audio Flamingo 3 supports multi-audio multi-turn dialogue, but lacks explicit multi-audio analysis support. MMAU-Pro addresses this gap by extending beyond single-audio QA. It includes 430 and 26 QA instances with two and three audios, respectively, each requiring understanding all individual audios for answering the QA correctly.

Multicultural Music Understanding Most existing benchmarks evaluating music understanding focus predominantly on Western music, overlooking the rich diversity of global musical traditions.

MMAU-Pro expands this scope by incorporating music from eight culturally distinct regions: African (21), Chinese (496), European (54), Indian (112), Latin American (11), Middle Eastern (7), Western (901), and Other Asian cultures (16). In Appendix B, we show that models trained primarily on Western music struggle with non-Western musical reasoning, highlighting he need to diversify training datasets for more inclusive music understanding.

Spatial Audio Understanding Understanding properties such as directionality, reverberation, and acoustic environment is a critical component of spatial awareness in auditory intelligence. Unlike visual spatial reasoning, spatial cues in audio often require multi-channel input. MMAU-Pro includes 325 expertly curated QA pairs paired with binaural recordings, designed to assess models' ability to perceive spatial relationships, such as sound direction and room characteristics, requiring fine-grained spatial awareness.

Voice QA As AI agents become more capable and widely adopted, voice-to-voice interaction is poised to become the default interface [165]. However, enabling faithful voice based interaction requires more than just spoken language understanding. It demands robust paralinguistic comprehension, including age, emotion, demographic cues, and urgency [161]. Moreover, models must process spoken content that extends beyond natural language-such as mathematical expressions and STEM-related queries.

To evaluate these capabilities, MMAU-Pro introduces questions that assess paralinguistic understanding, including age, emotion, and urgency. Additionally, we convert STEM questions into spoken form using GPT-4oTTS to test voice-based comprehension of mathematical expressions. These tasks also allow us to probe the model's STEM reasoning abilities, a known challenge for MLLMs.

Instruction Following Enabling foundation models to follow human instructions is essential for building controllable and reliable AI assistants [159] [131] [177]. However, evaluating instruction-

following remains challenging due to the open-ended nature of many prompts (e.g., "Write a short poem based on the sound you hear") (Chen et al. 2024a; Wang et al. 2024). To enable objective evaluation, we adopt the constraint-based approach of Zhou et al. (2023b) [12], framing instruction-following as a verifiable subtask within MMAU-Pro. Our design is further inspired by IFEval-Audio [139], which introduced structured spoken instruction evaluation for audio-language models. We construct a dedicated subset with 87 constraint in stances drawn from 28 instruction types, grouped into six categories (e.g., Length Constraints, Keyword Usage, Format). Each instruction is paired with one of seven open-ended prompt templates (e.g., "Describe the audio") and instantiated with variations to test robustness across prompt styles. Final inputs are synthesized using ChatterboxTTS (Resemble AI 2025), combining spoken instructions with audio segments from the MMAU dataset (e.g., speech, music, ambient sounds). We provide deterministic regex-based evaluation scripts for each constraint, enabling scalable, reproducible scoring in realistic conditions.

1.1.3 Experimental Setup

LALMs We evaluate a wide range of Large Audio-Language Models on the MMAU-Pro benchmark to assess their capabilities in long and short form reasoning, spatial understanding, multicultural music interpretation, and multi-audio comparisons.

Cascaded Systems To evaluate the robustness of our benchmark MMAU-Pro, we also conduct assessments on cascaded systems. In this approach, we first obtain captions for sound and music, and transcripts for speech-based questions. Subsequently, we combine these captions and questions and pass them to text-only open and closed-source Large Language Models (LLMs). These LLMs include GPT-4o [158], one closed-source, state-of-the-art LLM, and Qwen3-235B-A22B-Instruct [173], an

open-source, instruction-tuned model. For obtaining the captions of sound and music audios, we resort to Audio Flamingo 3, and for obtaining the speech transcripts, we use Whisper-Large-v3 [162].

Evaluation Strategy To evaluate MCQs, we compute the embedding of each answer choice using a pretrained transformer model, i.e, NV-Embed-v2 [148] [157] in our case, and compare it to the model's output embedding for the question context. Rather than computing the question embedding directly, the model generates an output vector representing its predicted response. This output embedding is compared against the embeddings of all available answer choices using cosine similarity. The choice with the highest similarity is selected as the predicted answer. The evaluation is then conducted by comparing this prediction to the ground truth label. This embedding-based selection strategy allows for semantically meaningful predictions even when explicit answer tokens are not generated, and avoids reliance on string-based pattern matching. For evaluating open-ended responses, we employ Qwen2.5-7B-Instruct as a judge and provide it with the ground truth answer and the prediction. We evaluate each model's response on 5 fronts - (i) Correctness: How factually accurate is the response compared to the reference? (ii) Relevance: How well does the response address the specific question asked? (iii) Completeness: Does the response cover all important aspects mentioned in the reference? and (iv) Clarity: How clear and well-structured is the response? and we also ask the LLM to assign an overall assessment score, which we report. For evaluating open-ended evaluations, we first obtain scores on a scale of 1 to 5. Then, we convert these scores into percentage values to ensure that all reported scores remain on the same scale. We also evaluate LLM as a judge vs Human annotation score, and find a high correlation value to validate the strength of our LLM-as-a-judge framework. We show these correlations on the MMAU-test-mini and MMAR dataset. For evaluation of multi-audio, for models that do not support multi-audio analysis, we concatenate the audios with a silence of 2 seconds and feed it to the model, and mention it in the prompt, whereas for the models that support multiple audios,

we feed them sequentially.

1.1.4 Results and Discussion

We present a comprehensive breakdown of model performance across twelve main modalities. Several clear patterns emerge. First, on the core single-modality tasks (Sound, Music, Speech), most end-to-end LALMs achieve only moderate accuracy (30–60%), with smaller models such as SALMONN-7B and Phi4-MM-Instruct often below 50%. Even the strongest open-source models Qwen2.5-Omni-7B rarely exceed 65% on Music or Speech, indicating that foundational audio understanding remains challenging.

Performance degrades further as tasks grow more complex. On mixed modalities (Sound–Music, Speech–Music, Speech–Sound, and three-way mixtures), accuracies typically fall into the 20–50% range. This suggests that models cannot yet reliably fuse information across multiple audio streams. A similar drop can be seen in the spatial audio understanding performance, where even the top models rarely surpass 40%.

Voice-chat reasoning, which tests conversational and world-knowledge, and STEM knowledge inference, also exposes weaknesses, with most models scoring between 25% and 60%. Notably, Qwen2.5-Omni-7B and Gemini-2.5 Flash perform decently well on these tasks, scoring 60% and 71.7% respectively, but smaller or less instruction-tuned models often languish below 50%.

Multi-audio reasoning and open-ended question answering remain the most challenging tasks: no model surpasses 30% accuracy on the "Multi-Audio" subset, and open-ended QA tops out at only 45% even for the largest models.

For multiple-choice questions with four options, performance may be inflated because models

can rely on elimination strategies or benefit from the higher probability of guessing correctly (25%). To further stress-test this, we expand certain questions to include 10 options and observe a substantial drop in accuracy as the number of options increases. Models like Mellow, Qwen2.5-Omni, and AF3, which support multi-audio, still do not exceed 30%. GPT4o-Audio achieves slightly higher than 30%. Finally, instruction-following ("IF") is dominated by large closed source models.

In summary, while most closed-source LALMs can handle many single-domain input tasks well, they still struggle with nuances in audio like temporal understanding, prosodic, and emotional reasoning. In addition, they struggle much more with multi-audio analysis, spatial reasoning, free-form answering, and instruction following. These areas represent clear directions for future model and benchmark development.

1.1.5 Do MLLMs Retain Skills Acquired from Text Pre-training?

Performance on STEM QA. To examine whether models can effectively link audio understanding with text-based knowledge and reasoning skills, we compare the STEM focused QA performance of AF3 (in "Think" mode) with its base LLM, Qwen2.5-7B-Instruct. In this setup, Qwen2.5-7B-Instruct is evaluated on the original text-only STEM questions from the source dataset, while AF3+Think is evaluated on the corresponding audio-based MMAU-Pro Voice STEM subset. Qwen2.5-7B-Instruct achieves 36.17% accuracy, whereas AF3+Think reaches only 31.91%. We identify two possible causes: (i) AF3 may lose part of its text-based math reasoning ability during audio fine-tuning—a gap that could potentially be mitigated with high-quality instruction tuning data; or (ii) an auditory perception gap, where the model correctly interprets the audio and retains the necessary reasoning skills, but fails to connect perception with knowledge. Similar issues have been observed in LVLMs

(Ghosh et al. 2025a), where models demonstrate sufficient reasoning ability in text form but struggle to bridge perception with understanding, a phenomenon described as the visual perception gap.

Instruction Following. We compare the performance of AF3 and Qwen-2.5-Omni-7B on the instruction-following subset of MMAU-Pro. For the Change Cases task, AF3 attains 35.4% accuracy, whereas Qwen2.5-Omni-7B reaches 75.2%. On Detectable Format, AF3 fails to produce many correct responses (8.6%), while Qwen2.5-Omni-7B correctly formats 40.3% of cases. In Length Constraints, AF3 scores 30.7% compared to 68.5% for Qwen2.5-Omni-7B. AF3's only relative strength appears on Detectable Content, where it achieves 67.8% accuracy versus Qwen2.5-Omni-7B's 60.4%. The Keywords task again highlights the gap AF3 manages just 5.9% while Qwen2.5-Omni-7B succeeds on 65.1%. Finally, for Multi-Part Response, AF3 records 55.6% accuracy compared with 60.8% for Qwen2.5-Omni-7B. This consistent advantage for Qwen2.5 Omni-7B on five of six subtasks underscores the crucial role of extensive text-only pretraining and instruction-tuning: without a robust textual instruction corpus, AF3's performance on language- centric directives remains significantly weaker despite its strong audio understanding and reasoning ability.

1.1.6 Conclusion, Limitations and Future Work

In this paper, we introduced MMAU-Pro, a comprehensive benchmark designed to holistically evaluate general audio intelligence in multimodal language models. MMAU-Pro advances prior efforts by incorporating 5,305 expert-annotated QA pairs spanning 49 diverse skills across speech, sounds, music, and their combinations. The benchmark introduces several key innovations, including long-audio understanding, multi-audio reasoning, spatial audio comprehension, multicultural music understanding, instruction following, etc. These tasks mirror real-world challenges and require advanced perception,

contextual understanding, and complex reasoning. Our evaluation across open and proprietary LALMs demonstrates that even the strongest models struggle across several categories. Of course, humans are able to do amazing feats with the sound they hear, and to truly benchmark an audio model's ability to do all of these will always be a work in progress, and we do not claim to explore all dimensions of audio processing/reasoning ability in the benchmark.

As part of future work, we plan to: (i) further expand the scale of MMAU-Pro to include more languages and low-resource acoustic environments; (ii) introduce dynamic and interactive audio tasks, such as real-time reasoning over streaming audio; (iii) refine instruction-following evaluation with free-form generation and adversarial constraints; and (iv) develop better metrics for evaluating paralinguistic understanding and culturally-grounded reasoning. We hope MMAU-Pro serves as a stepping stone toward developing more capable and general-purpose audio-language models.

1.2 A Bayesian Non-parametric Framework for Learning Disentangled Representations

Disentangled representation learning aims to identify and organize the underlying sources of variation in observed data. However, learning disentangled representations from observational data alone without any additional supervision necessitates inductive biases to solve the fundamental identifiability problem of uniquely recovering the true latent structure and parameters of the data-generating process. Existing methods address this by imposing heuristic inductive biases that typically lack these theoretical identifiability guarantees. Additionally, these methods rely on strong regularization to impose these inductive biases, creating an inherent trade-off in which stronger regularization improves disentanglement but limits the latent capacity to represent underlying variations. To address both challenges, we propose a principled generative model with a Bayesian nonparametric hierarchical mixture prior that

embeds inductive biases within a provably identifiable framework for unsupervised disentanglement. Specifically, the hierarchical mixture prior imposes the structural constraints necessary for identifiability guarantees, while the nonparametric formulation allows the latent representation to scale with infinite capacity to faithfully represent the complete set of underlying variations without violating these structural constraints. To enable tractable inference under this nonparametric hierarchical prior, we develop a structured variational inference framework with a nested variational family that both preserves the hierarchical structure of the identifiable generative model and approximates the expressiveness of the nonparametric prior. We evaluate our proposed probabilistic model on standard disentanglement benchmarks, 3DShapes and MPI3D datasets characterized by diverse source variation distributions, to demonstrate that our method consistently outperforms strong baseline models through structural biases and a unified objective function, obviating the need for auxiliary regularization constraints or careful hyperparameter tuning.

1.2.1 Introduction

A primary objective of representation learning is not merely to perform density estimation or generate realistic samples, but to discover and characterize the latent structure inherent in observational data. This notion is formalized by disentangled representations that aim to separate the distinct, independent, and informative generative factors of variation in the data such that each variable of the representation is sensitive to changes in exactly one underlying factor while being relatively invariant to changes in others [1]. Disentangled representations have been shown to improve robustness and out-of-distribution generalization [3] [4], sample efficiency in few-shot learning [5] [6], domain adaptation via separation of transferable and domain-specific features [7] [8] controllable and interpretable generation

[9] [11] [12], and causal inference and fairness through explicit separation of sensitive and task-relevant factors [6] [28]. However, unsupervised learning of disentangled representations is fundamentally challenged by identifiability which refers to whether the true generative factors and their structure can be uniquely inferred from observed data alone. Without identifiability, different parameterizations of the generative model can produce identical distributions over observed data, making it theoretically impossible to recover the true generative factors. Prior work in nonlinear independent component analysis [25] [26], deep generative modeling [49], and unsupervised disentanglement [28] has shown that enforcing the commonly used simple isotropic Gaussian prior in combination with a nonlinear generative function is generally insufficient to recover the true sources of variation. Without additional inductive biases, the model can learn infinitely many, potentially entangled representations that satisfy the marginal prior distribution yet fail to align with the true data-generating factors.

Moreover, prior works primarily impose heuristic inductive biases and typically rely on strong regularization to enforce them inducing an inherent trade-off whereby stronger regularization enhances disentanglement but simultaneously restricts the representation capacity. Consequently, this misspecification of the latent capacity either under-represents all relevant modes of variation or forces encoding of the data in a manner that conflicts with the natural structure, leading to systematic violation of the disentanglement-inducing constraints.

To address both these limitations we build upon the theoretical framework of Kivva et al. (2022), who prove that mixture priors provide sufficient inductive bias for identifiability in deep generative models with piecewise affine data-generating functions. We propose a Bayesian nonparametric hierarchical mixture prior that inherits these identifiability guarantees lacking in simple Gaussian priors while simultaneously addressing the representation capacity mis-specification problem through its nonparametric formulation. The nonparametric formulation allows the complexity of factor-specific mixtures to

remain unspecified a priori, a characteristic analogous to species discovery in unexplored ecosystems, where the number and types of species present cannot be predicted in advance. This formulation endows our model with universal approximation capabilities, ensuring that the identifiable architecture is, in principle, expressive enough to recover the natural underlying structure of the data. To specifically learn disentangled representations, we define a factorized prior structure under which a nonparametric hierarchical mixture prior is placed over the space of each generative factor independently, such that mixture components correspond to discrete variations of the respective factor. Consistent with the principles of classical factor analysis, this factorized structure entails that observations are generated through the combinatorial composition of factor-specific mixture components, with each observation determined by a unique combination of components across all generative factors Hsu et al. (2024a) [32].

For tractable inference under this nonparametric hierarchical prior, we develop a structured variational inference framework with a nested variational family. The structured inference framework preserves the hierarchical structure of the identifiable generative model thereby enabling joint optimization of the prior and deep generative model parameters within a unified objective function. The nested formulation enables the variational distribution to approximate the expressiveness of the nonparametric prior while maintaining computational tractability.

Empirically, we show that this hierarchical mixture prior provides substantially stronger inductive biases enabling the learning of modular and compact disentangled representations that enhance interpretability. Our results on two image datasets with distinct factor distributions further demonstrate that the nonparametric hierarchical mixture prior and the corresponding inference framework provide sufficient inductive bias without additional computationally expensive auxiliary inductive biases or careful manual tuning of regularization hyperparameters.

1.2.2 Nonparametric Bayesian Priors for Autoencoders

The theoretical framework of Kivva et al. (2022) establishes identifiability of deep latent variable models by rigorously delineating the conditions under which such models are uniquely recoverable from data. Specifically, identifiability of the latent variable distribution up to affine transformation requires two critical conditions: (i) the marginal distribution of the latent variables follows a Gaussian mixture model (GMM), and (ii) the corresponding data-generating function is piecewise affine and satisfies weak injectivity. Building on these conditions, the authors further demonstrate that stronger identifiability, up to permutation, scaling, and translation, necessitates the introduction of additional discrete latent variables which index the mixture components of the GMM, inducing a hierarchical structure in which the continuous latent variables are conditionally independent given these discrete variables. Furthermore, imposing an additional maximality condition (P3 in Kivva et al. (2022)) on this hierarchical structure yields identifiability of the complete latent representation, including the latent dimensionality, the cardinality and distribution of the discrete variables, and the conditional distribution of the continuous latent variables.

To further bias this identifiable framework toward learning disentangled representations, we impose an additional structural constraint on the discrete latent space. Specifically, we factorize the multivariate discrete latent variable into statistically independent components, where each component encodes a distinct factor of variation and indexes a set of mixture components representing the discrete modes of variations of its respective factor. Critically, this factorized prior enforces an orthogonal partitioning of the latent space, such that factor-specific variations are encoded in non-overlapping subspaces of the continuous latent representation, thereby promoting the independent encoding of factor-specific variations. Moreover, this factorized latent space induces a generative process in which

observations are generated through the combinatorial composition of factor-specific mixture components, with each observation generated by a unique combination of component assignments.

This inductive bias of combinatorial composition toward learning disentangled representations is closely related to the works of Hsu et al. (2024a;b) [32], which similarly structure the latent space as a product of independent discrete sets to promote disentangled representations. In their framework, the discrete latent space is formed from continuous latent representations that are discretized via vector quantization, wherein each latent dimension is independently mapped to discrete codes through distinct, learnable codebooks. For this architecture to serve as an effective inductive bias toward disentanglement, the size of each codebook, and thus the cardinality of the corresponding discrete latent variable indexing its embeddings, is kept small and fixed, constraining each dimension to capture only a limited set of discrete modes. While this design choice encourages parsimonious representations, it imposes a rigid upper bound on the expressive capacity of each factor: factors whose discrete modes of variations exceed the cardinality of a single codebook must either be distributed across multiple independent codebooks or be only partially captured, in both cases reducing the interpretability and semantic coherence of the learned latent factors.

To address the expressive limitations of fixed-cardinality discrete variables, we introduce a nonparametric Dirichlet Process mixture [94] [96] [95] prior over the latent space, allowing the support of each latent variable to grow adaptively with the intrinsic complexity of its associated factor of variation. By placing nonzero probability mass over a countably infinite support, the DP prior equips each discrete latent variable with a universal approximation property over discrete distributions, enabling it to represent an arbitrary discrete distribution over the modes of its associated factor. Consequently, this universal approximation property reinforces the statistical independence of the joint latent prior, ensuring that the complete set of discrete modes of variation associated with each independent factor is

represented entirely by a discrete latent variable. Critically, because the DP prior is applied independently to each discrete latent variable, each factor can adaptively expand its variation space without disrupting the orthogonal partition structure induced by the remaining factors, thereby preserving statistical independence of the representation.

In the following sections, we formalize this framework by deriving a Bayesian nonparametric hierarchical generative model that is jointly identifiable, universally expressive through the nonparametric prior, and structurally biased toward disentangled representations with a particular emphasis on completeness and interpretability. We then formalize the structured variational family, which is specifically designed to accurately approximate the posterior distribution under the hierarchical factorized structure of the prior. Finally, we present a nested extension of this structured variational family that enables principled incremental expansion of representational capacity, with the structured component preserving the inductive biases of identifiability and factorial independence, and the nested component approximating the universal approximation capacity of the nonparametric prior, jointly ensuring that the learning algorithm remains faithful to the theoretical properties of the generative model throughout inference.

We consider a hierarchical latent variable generative model in which each observation $x \in \mathcal{X}$ is generated via a structured latent representation comprising a continuous latent variable $e \in R^d$ and a multivariate discrete latent variable $z = (z_1, \ldots, z_d) \in \mathcal{Z}$ with $\mathcal{Z}$ denoting the corresponding discrete latent space. To impose an inductive bias toward disentangled representations, we assume a factorial structure on the discrete space $\mathcal{Z}$, defined as the Cartesian product $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_d$, where each $z_i \in \mathcal{Z}_i = \{1, \ldots\}$ is mutually independent and captures a distinct factor of variation. To ensure identifiability of the continuous latent variable e up to affine transformations, we follow the conditions of [**?**] and specify the marginal distribution of e as a Gaussian mixture model (Assumption P1). To further satisfy the conditions for stronger identifiability (Assumption P2), we require that, conditioned

on the discrete latent variable z, the components of e are mutually independent, $e_i \perp\!\!\!\perp e_j \mid z$ for all $i \neq j$ enforcing an orthogonal partitioning of the continuous latent space into d non-overlapping factor-specific subspaces.

To allow an unbounded number of mixture components, we model each e_i under an independent Dirichlet Process mixture prior, inducing a countably infinite Gaussian mixture over the marginal $p(e_i)$. Formally, we place a DP prior $G_i \sim \mathrm{DP}(\alpha, G_0)$ over the component parameters, where $\alpha > 0$ is the concentration parameter controlling the richness of the induced partition, and G_0 is the base distribution over the parameter space Θ_i. Under the stick-breaking representation of the Dirichlet Process, we define an infinite sequence of stick breaking proportions $\beta_i = \{\beta_{i,k}\}_{k=1}^{\infty}$, with $\beta_{i,k} \sim \mathrm{Beta}(1, \alpha)$ and draw the corresponding component parameters independently from the base distribution, $\theta_{i,k} \sim G_0(\lambda)$ as follows,

$$\beta_{i,k} \mid \alpha \sim p(\beta \mid \alpha) = Beta\,(1, \alpha)\,, \quad \theta_{i,k} \mid \lambda \sim p(\theta \mid \lambda) = G_0(\lambda), \quad \forall i \in \{1, \dots\, d\},\, \forall k \in N$$

From the stick-breaking proportions β_i, we construct an infinite sequence of mixing weights $\pi_{i,k} = \beta_{i,k} \prod_{j=1}^{k-1}(1-\beta_{i,j})$, satisfying $\sum_{k=1}^{\infty} \pi_{i,k} = 1$ almost surely, yielding the random discrete measure $G_i = \sum_{k=1}^{\infty} \pi_{i,k}\, \delta_{\theta_{i,k}}$. The discrete latent variable z_i then selects a component index according to $z_i \sim \mathrm{Categorical}(\pi_i)$, upon which the continuous component e_i is drawn from the corresponding component-specific distribution parameterized by $\theta_{i,k} \in \Theta_i$, i.e. $e_i \mid z_i = k, \theta_i \sim p(e_i \mid \theta_{i,k})$.

Completing the generative hierarchy, the observation x is then generated via a piecewise affine deep neural network (with Rectified Linear Unit (ReLU) nonlinearities) data-generating function $g_{\theta_g} : R^d \to \mathcal{X}$ satisfying the weak injectivity conditions (F1 and F2 of kivva2022identifiability), with additive observation noise: $x \mid e \sim p_{\theta_g}(x \mid e) = \mathcal{P}(g_{\theta_g}(e))$ where $\mathcal{P}$ denotes an appropriate likelihood family.

Together, these modeling choices yield a generative model with three key properties: identifiability

of the latent space up to permutation, scaling, and translation; nonparametric expressiveness through the DP prior; and an explicit factorial structure promoting disentangled, factor-specific representations. The joint distribution over observations x, continuous variables e, discrete variables z, stick-breaking proportions β, and component parameters θ factorizes according to the hierarchical generative process as $\mathrm{p}(x, e, z, \beta, \theta \mid \alpha, \lambda) = p_{\theta_g}(x \mid e) \prod_{i=1}^{d} p(e_i \mid z_i, \theta_i)\, p(z_i \mid \beta_i) \prod_{k=1}^{\infty} p(\beta_{i,k} \mid \alpha)\, p(\theta_{i,k} \mid \lambda)$ where the outer product over i encodes the factorial independence across latent dimensions, and the inner product over k reflects the infinite mixture structure induced by the Dirichlet Process prior.

1.2.3 Structured Nested Variational Family

To enable tractable posterior approximation under the nonparametric DP prior (see Section 1.2.8 for a primer on variational inference for DPMMs), existing methods blei2006variational, hoffman2013stochastic rely on two simplifying assumptions. First, they approximate the infinite-dimensional stick-breaking process using a truncated variational family by introducing an explicit truncation level T, fixing $q_{\nu_{\beta_i}}(\beta_{i,T} = 1) = 1$ for each dimension i, thereby forcing all subsequent stick-breaking weights $\{\beta_{i,k}\}_{k>T}$ and their corresponding mixing weights to zero, limiting the mixture components to T. Second, they adopt a mean-field variational family, imposing full factorization across all latent variables, thereby severing the hierarchical dependencies induced by the generative model.

We address the second limitation by adopting the structured variational framework of hoffman2015structured, which relaxes the mean-field assumption and retains the conditional dependencies induced by the generative model, yielding more accurate posterior approximations with reduced bias and lower sensitivity to initialization and hyperparameters. Specifically, for each dimension i, we preserve the hierarchical dependencies between the stick-breaking proportions β_i and the discrete variable z_i along with the

component parameters θ_i, and the continuous latents e_i with the following structured variational family,

$$q_\nu(e_i, z_i, \beta_i, \theta_i) = q(e_i \mid z_i, \theta_i) q(z_i \mid \beta_i) \prod_{k=1}^{T-1} q_{\nu_{\beta_{i,k}}}(\beta_{i,k}) \prod_{k=1}^{T} q_{\nu_{\theta_{i,k}}}(\theta_{i,k}) \tag{1.1}$$

Truncated variational families render inference in infinite-mixture models tractable by restricting the variational distribution to a fixed number T of mixture components. Intuitively, one might expect the ELBO to increase monotonically with T, as additional components allow the variational posterior to better approximate the true nonparametric posterior, thereby motivating the use of large truncation levels. However, when the data are generated by a finite mixture, there can exist an optimal truncation level beyond which additional components do not improve the approximation, because the truncated families are not nested: the variational family with truncation level T is not a subset of the family with truncation level $T + 1$. Consequently, increasing T beyond the optimal level does not necessarily improve the approximation, and crucially, undermines the inductive bias toward parsimonious representations [32] that is essential for learning disentangled factors.

To address this, we employ the nested variational family of kurihara2006accelerated, which supports an unbounded number of components in the variational family by tying the variational parameters of all components beyond an implicit truncation level T to their prior values:

$$q_{\nu_\beta}(\beta_i) = \prod_{k=1}^{T} q_{\nu_{\beta_{i,k}}}(\beta_{i,k}) \prod_{k=T}^{\infty} p(\beta_{i,k} \mid \alpha), \qquad q_{\nu_\theta}(\theta_i) = \prod_{k=1}^{T} q_{\nu_{\theta_{i,k}}}(\theta_{i,k}) \prod_{k=T}^{\infty} p(\theta_{i,k} \mid \lambda) \tag{1.2}$$

Under this formulation, only the first T components have free variational parameters, ensuring that despite the variational distribution supporting an unbounded number of components only T sets of parameters need to be represented and optimized, rendering inference computationally tractable.

Crucially, under this nested variational family, data points that are not explained by the first T components with free variational parameters can still be assigned to components beyond T, whose parameters remain tied to the prior. This enables a greedy inference procedure that begins with $T = 1$ and incrementally adds components only when they yield a significant improvement in the ELBO, continuing until all data are assigned to components within the active truncation level and no mass remains on prior-tied components. Notably, under this nested formulation, components that fail to encode meaningful variations collapse back to their prior distribution during training.

Given the generative model in eq:generative and the hierarchical structured nested variational family q_ν, the evidence lower bound (ELBO) for a single data point, estimated via Monte Carlo, can be written as $\mathrm{L} = {}_{q_{\nu_\beta}}\left[\log \frac{p(\beta|\alpha)}{q_{\nu_\beta}(\beta)}\right] + {}_{q_{\nu_\theta}}\left[\log \frac{p(\theta|\lambda)}{q_{\nu_\theta}(\theta)}\right] + {}_{q_\nu}\left[\log \frac{p_{\theta_g}(x|e)\, p(e|z,\theta)\, p(z|\beta)}{q(e|z,\theta)\, q(z|\beta)}\right]$

1.2.4 The Inference Algorithm

Efficient inference in hierarchical models relies on conjugate exponential family likelihoods, which preserve tractable conditional posterior distributions within the same exponential family as the prior admitting closed-form updates. For general neural network observation models such as p_{θ_g}, the absence of conjugacy makes inference over latent variables substantially more expensive and often requires multiple passes through the generative network for a single update. To retain the computational advantages of conjugate structured inference while using flexible non-conjugate likelihoods, we follow johnson2016composing and use deep amortized recognition networks $h(x;\phi)$ that output local conjugate likelihood potentials $\hat{p}_\phi$ (Eq. 1.3), rather than directly parameterizing the variational distributions as in standard variational autoencoders. These conjugate potentials serve as amortized, data-dependent surrogates for the intractable likelihood terms during inference and are

combined with the structured latent prior via message passing, leaving the generative model itself unchanged.

$$\hat{p}_\phi(e \mid x) = \prod_{i=1}^{d} \hat{p}_\phi(e_i \mid x) = \prod_{i=1}^{d} \exp\{\langle h_i(x;\phi), t_e(e_i)\rangle\} \tag{1.3}$$

This factored structure of the recognition network mirrors the factorial prior over e_i, allowing inference for the pair of local latent variables (e_i, z_i) associated with each dimension i to be inferred independently. Consequently, the data-dependent term of the ELBO in Eq. 1.2.3 decomposes into a sum of local contributions $\mathcal{L}_i$, one per dimension:

$$\mathcal{L}_i = {}_{q_{\nu_\beta}(\beta_i)q(z_i|\beta_i)}\left[\log \frac{p(z_i \mid \beta_i)}{q(z_i \mid \beta_i)} + {}_{q_{\nu_\theta}(\theta_i)q(e_i|z_i,\theta_i)}\left[\log \frac{p(e_i \mid z_i, \theta_i)}{q(e_i \mid z_i, \theta_i)} + \log \hat{p}_\phi(e_i \mid x)\right]\right] \tag{1.4}$$

Following hoffman2015structured, conditioning on the global variables β_i and θ_i, for any given value of the local variable $z_i = k$, the inner expectation over e_i admits the form of a variational lower bound on the conditional marginal likelihood of the conjugate potential: ${}_{q(e_i|z_i,\theta_i)}\left[\log \frac{p(e_i|z_i,\theta_i)}{q(e_i|z_i,\theta_i)} + \hat{p}_\phi(e_i \mid x)\right]$ $= -\left(q(e_i \mid z_i, \theta_i) \| \hat{p}_\phi(e_i \mid x_i, z_i, \theta_i)\right) + \log \hat{p}_\phi(x_i \mid z_i, \theta_i) \leq \log \hat{p}_\phi(x_i \mid z_i, \theta_i)$ where the local posterior and conjugate marginal likelihood are defined as

$$\hat{p}_\phi(e_i \mid x_i, z_i, \theta_i) = \frac{\hat{p}_\phi(e_i \mid x_i) p(e_i \mid z_i, \theta_i)}{\hat{p}_\phi(x_i \mid z_i, \theta_i)},\ \hat{p}_\phi(x_i \mid z_i, \theta_i) = \int \hat{p}_\phi(e_i \mid x_i)\, p(e_i \mid z_i, \theta_i) de_i \tag{1.5}$$

Since the Kullback-Leibler (KL) divergence is non-negative, setting $q(e_i \mid z_i, \theta)$ equal to the local posterior $\hat{p}_\phi(e_i \mid x_i, z_i, \theta)$ minimizes the KL divergence to zero and yields the tightest lower bound on the local ELBO. Under conjugacy, this optimal variational distribution takes the exponential

family form:

$$q(e_i \mid z_i, \theta_i) = \hat{p}_\phi(e_i \mid x_i, z_i, \theta_i) = \exp\left\{\langle \eta_e(z_i, \eta_\theta(\theta_i), \phi), t_e(e)\rangle - A_e(\eta_e(z_i, \eta_\theta(\theta_i), \phi))\right\} \eta_e(z_i, \eta_\theta(\theta_i), \phi) = \sum_{k=1}^{T} {}_{[z_i=k}$$

(1.6)

with natural parameters which explicitly combines the structured prior through $\eta_\theta(\theta_{i,k})$ with the data-dependent signal from the recognition network $h_i(x; \phi)$. With this optimal choice of $q(e_i \mid z_i, \theta_i)$, the local ELBO reduces to a variational lower bound on the marginal likelihood of the data conditioned on the global variable β_i as follows:

$$\mathcal{L}_i = {}_{q_{\nu_\beta}(\beta_i) q(z_i \mid \beta_i)} \left[\log \frac{p(z_i \mid \beta_i)}{q(z_i \mid \beta_i)} + {}_{q_{\nu_\theta}(\theta_i)} \left[\log \hat{p}_\phi(x \mid z_i, \theta_i)\right]\right] = -(q(z_i \mid \beta_i) \| \hat{p}_\phi(z_i \mid x_i, \beta_i)) + \log \hat{p}_\phi(x_i \mid \beta_i) \leq \log \hat{p}$$

(1.7)

and similar to the variational distribution of e_i, the locally optimal variational distribution over z_i for $k \leq T$ is given by:

$$\log q(z_i = k \mid \beta_i) \propto \log p(z_i = k \mid \beta_i) + {}_{q_{\nu_\theta}(\theta_i)} \log \hat{p}_\phi(x_i \mid \theta_{i,k}) = \log \pi_{i,k} + {}_{q_{\nu_\theta}(\theta_{i,k})} [A_e\left(\eta_\theta(\theta_{i,k}) + h_i(x_i; \phi)\right) - A_e\left(\eta_\theta($$

(1.8)

While the nested variational family supports an unbounded number of components, naively computing the variational assignment $q(z_i = k \mid \beta_i)$ for all $k > T$ would require sampling infinitely many stick-breaking proportions to evaluate the corresponding mixing weights $\pi_{i,k}$. However, two properties of the nested family together enable a tractable closed-form estimate. First, since the parameters of all components $k > T$ are tied to the prior $p(\theta_{i,k} \mid \lambda)$, the expected log-likelihood $E_{p(\theta_{i,k} \mid \lambda)}[\log \hat{p}_\phi(x_i \mid \theta_{i,k})]$ is identical across all components beyond T and can be factored out of the infinite sum. Second, since the mixing weights are normalized and satisfy $\sum_{k=1}^{\infty} \pi_{i,k} = 1$ almost surely, the total weight

assigned to all components beyond T reduces to $(1 - \sum_{k=1}^{T} \pi_{i,k})$, which depends only on the T free variational parameters. Together, these two properties yield the following tractable expression for the total assignment probability to all prior-tied components:

$$\log q(z_i = k \mid \beta_i) \propto \log p(z_i = k \mid \beta_i) + {}_{p(\theta_{i,k}|\lambda)} \log \hat{p}_\phi(x_i \mid \theta_{i,k}) q(z_i = k \mid \beta_i) \propto \pi_{i,k} \cdot \exp\Big\{ {}_{p(\theta|\lambda)} \log \hat{p}_\phi(x_i \mid \theta) \Big\} q(z_i > \tag{1.9}$$

This quantity serves as the stopping criterion for the greedy component-expansion procedure where a new component is added to the active truncation level when $q(z_i > T \mid \beta_i)$ exceeds a threshold, indicating that residual data cannot be explained by the current T components.

As in structured variational inference more broadly, restoring dependencies between global variables and local variables along with amortized recognition networks enables the local variational distributions to be set to their locally optimal values in closed form. At each iteration, the global parameters β_i and θ_i are sampled from $q_{\nu_\beta}(\beta_i)$ and $q_{\nu_\theta}(\theta_i)$ respectively using the reparameterization trick, and used to compute the local variational distributions. Since exact inference over the global variables remains intractable due to the hierarchical dependencies hoffman2015structured, and since the use of recognition networks precludes analytic gradient computation, we optimize the ELBO using low-variance Monte Carlo estimators, yielding unbiased gradient estimates for both global and recognition network parameters. Step sizes are selected following the principled stability and convergence guidelines of mandt2017stochastic, with optimization carried out using adaptive preconditioned gradient methods, specifically Adam [179].

1.2.5 Experiments

In this section, we present experiments designed to empirically assess whether the hierarchical Bayesian nonparametric approach to latent quantization provides effective inductive biases for learning interpretable disentangled representations. Specifically, we evaluate whether our approach achieves comparable performance relative to prior work which impose equivalent inductive biases through multiple, distinct regularization terms.

Datasets. Our experimental framework systematically addresses these questions through comprehensive quantitative evaluations conducted on two benchmark datasets labeled with ground-truth source information. Each dataset is constructed from mutually independent sources through a deterministic data generation process. In particular, we use the 3DShapes 3dshapes18 dataset of 3D shapes generated from six ground-truth independent latent factors with approximately uniform and small number of variations. Additionally, we use the MPI3D dataset NEURIPS2019$_d$97d404$b, collectedfromareal-worldroboticenvironmen$
$lawdistributionacrossthenumberofvariationsofdifferentfactors.Specifically, afewfactorscontainexten$
$6valuesforobjectproperties)$.

Prior Methods. We evaluate our proposed approach against several state-of-the-art methods that incorporate distinct inductive biases for unsupervised disentanglement. Specifically, we compare to β-VAE higgins2017beta and β-TCVAE [24] which enforce disentanglement through information-theoretic regularization encouraging independence across latent dimensions. We further consider BioAE whittington2022disentanglement, which introduces biologically inspired constraints—namely nonnegativity and energy efficiency—to promote compact representations enforcing neurons to become selective for single factors of task variation, together with a grid-like structural constraint as an architectural inductive bias. In addition, we examine QLAE hsu2024disentanglement, which introduces an architectural

bias based on latent quantization, and subsequently Tripod hsu2024disentanglement, which combines latent quantization with additional inductive biases enforcing independence among latent variables as well as constraining the functional mapping from latent representations to the data space.

Quantitative Comparison with Prior Methods. The experimental evaluation demonstrates that the proposed Bayes-QLAE consistently outperforms most baseline methods across both datasets in terms of modularity metrics (InfoM and D), with the notable exception of achieving competitive performance relative to QLAE and Tripod (Table 1.1 and Table 1.2). The observed improvements in compactness (InfoC) are particularly pronounced when compared to the baseline QLAE, demonstrating the effectiveness of the nonparametric prior in adapting to the complexity of underlying generative factors while maintaining consistency in modularity. Contrary to the position advanced by the authors of QLAE, who prioritize modularity/disentanglement over compactness/completeness through specific architectural design choices, we argue that achieving interpretable representations that faithfully capture mutually independent generative factors requires balanced weighting of both modularity and compactness metrics. With competitive explicitness and informativeness measures, Bayes-QLAE demonstrates performance consistent with QLAE and Tripod while substantially outperforming alternative approaches, reinforcing the efficacy of latent quantization for disentangled representation learning.

It is worth noting that Tripod achieves its superior modularity and compactness performance through the application of a Normalized Hessian Penalty, which necessitates multiple forward passes through the generative network, thereby incurring additional computational overhead. In contrast, Bayes-QLAE achieves competitive performance through architectural inductive biases alone, without requiring additional regularization terms. Furthermore, Tripod's disentanglement performance, particularly in modularity and compactness dimensions, exhibits sensitivity to quantization level hyperparameters, which must be specified a priori. Conversely, Bayes-QLAE demonstrates adaptive behavior that

automatically learns quantization levels from the data while maintaining robustness across evaluation metrics.

We observe that the performance improvement of Bayes-QLAE is notably more pronounced on the 3DShapes dataset compared to the MPI3D dataset, where factor variations are characterized by a power-law distribution. This differential performance suggests that the underlying distributional properties of the generative factors significantly influence the efficacy of the nonparametric prior. We hypothesize that replacing the Dirichlet Process prior with a more flexible, generalized prior such as the Pitman–Yor process—which allows for a richer clustering structure and can model power-law behaviors—may yield further performance gains. We perform detailed ablation studies to systematically isolate and quantify the inductive biases contributed by each component of our hierarchical Bayesian nonparametric framework in Section **??**.

model	InfoM	InfoC	InfoE	D	C	I
β-VAE	0.62 $\pm$ .02	0.44 $\pm$ .03	0.93 $\pm$ .02	0.58 $\pm$ .02	0.42 $\pm$ .02	**0.97** $\pm$ .02
β-TCVAE	0.65 $\pm$ .03	**0.56** $\pm$ .02	0.91 $\pm$ .02	0.56 $\pm$ .02	0.46 $\pm$ .02	**0.95** $\pm$ .02
BioAE	0.58 $\pm$ .02	0.42 $\pm$ .02	0.90 $\pm$ .01	0.48 $\pm$ .01	0.39 $\pm$ .02	0.91 $\pm$ .02
QLAE	0.84 $\pm$ .02	0.49 $\pm$ .01	**0.97** $\pm$ .01	**0.79** $\pm$ .01	0.56 $\pm$ .01	**0.97** $\pm$ .01
Tripod	**0.91** $\pm$ .03	**0.58** $\pm$ .03	**0.96** $\pm$ .02	**0.80** $\pm$ .03	**0.63** $\pm$ .03	**0.97** $\pm$ .02
Bayes-QLAE	**0.91** $\pm$ .03	**0.61** $\pm$ .02	**0.95** $\pm$ .02	**0.84** $\pm$ .03	**0.65** $\pm$ .03	**0.97** $\pm$ .02

Table 1.1: Disentanglement metrics measured in InfoMEC and DCI for 3DShapes dataset. For each metric a higher score is better. The scores for all the models were averaged across 5 runs with different random seeds with intervals denoting 95% confidence intervals of the mean estimated assuming a t-distribution. The results for the VQE-based and QLAE-based models are obtained using the hyperparameter settings and experimental conditions as described in [28] and [32, 65] respectively.

Qualitative Comparisons. We conduct qualitative assessments of our proposed method on each dataset to evaluate both sample reconstructions and latent traversals in Section **??**. Furthermore, we also assess the evolution of the latent traversals for the 3DShapes dataset with an increase in the number of components in Figure 1.2.

model	InfoM	InfoC	InfoE	D	C	I
β-VAE	0.41 $\pm$ .03	0.40$\pm$ .03	**0.68** $\pm$ .03	0.24 $\pm$ .03	0.19 $\pm$ .03	**0.80** $\pm$ .03
β-TCVAE	0.48 $\pm$ .03	0.46 $\pm$ .03	0.62 $\pm$ .03	0.27 $\pm$ .03	0.24 $\pm$ .03	**0.79** $\pm$ .03
BioAE	0.44 $\pm$ .03	0.38 $\pm$ .02	0.61 $\pm$ .03	0.26 $\pm$ .02	0.14 $\pm$ .02	0.77 $\pm$ .02
QLAE	0.52 $\pm$ .02	0.43 $\pm$ .02	**0.68** $\pm$ .04	0.38 $\pm$ .04	0.34 $\pm$ .04	0.81 $\pm$ .04
Tripod	**0.59** $\pm$ .05	**0.54** $\pm$ .05	**0.74** $\pm$ .06	**0.47** $\pm$ .04	**0.45** $\pm$ .05	**0.84** $\pm$ .05
Bayes-QLAE	**0.60** $\pm$ .03	**0.56** $\pm$ .03	**0.71** $\pm$ .04	**0.48** $\pm$ .03	**0.47** $\pm$ .03	**0.81** $\pm$ .03

Table 1.2: Disentanglement metrics measured in InfoMEC and DCI for MPI3D dataset. For each metric a higher score is better. The scores for all the models were averaged across 5 runs with different random seeds with intervals denoting 95% confidence intervals of the mean estimated assuming a t-distribution.

1.2.6 Related Works

The challenge of separating mutually independent sources in data traces back to the classical statistical problem of Independent Component Analysis (ICA). This core problem was later reinterpreted in the context of modern machine learning as disentanglement, formally articulated by Bengio [**?**] and formalized by higgins2018towards. When the data generating process is governed by nonlinear transformations [**?**], the task of learning disentangled representations becomes theoretically unidentifiable [**?**, 26, 28]. Consequently, the incorporation of auxiliary data [**?**, 25, 26] or weak supervision [**?**, **?**] is necessary to achieve identifiability in disentanglement. A distinct line of research focuses on the incorporation of inductive biases either in the model, training objective, or the data locatello2019challenging for identifiability.

Information-theoretic Regularization Biases. Many early and influential works leverage information-theoretic constraints on the latent space to encourage factorization. The β-VAE variants [22, 34] introduces a scalar multiplicative factor on the KL divergence penalty with isotropic Gaussian priors, forming an information bottleneck, limiting the amount of information each latent can capture. Extensions like FactorVAE [23] and β-TCVAE [24] further refine these constraints by explicitly penalizing total

correlation to enforce statistical independence between dimensions. BioAE whittington2022disentanglement demonstrates that biologically inspired constraints, specifically, minimizing latent activity and weight energy while promoting latent non-negativity encourage more factorized representations. In a similar vein, temporal sparsity is used to encourage the learning of factors varying independently across sequences sprekeler2014extension, klindt2020towards.

Architectural and Structural Biases. Structural inductive biases embedded directly into model architectures have proved powerful. Vector quantization in models like QLAE hsu2024disentanglement and the recent Tripod framework hsu2024tripod induce grid-like latent spaces that simplify factor separation. FactorQLAE baykal2024disentanglement combine scalar quantization of the latent variables with a total correlation term in the optimization as an inductive bias. On the theoretical front, barin2024identifiability establish identifiability for quantized factors under nonlinear mappings. Further, leeb2020structure demonstrate that restricting different latents to enter the decoding computation graph at different points can enable disentanglement. Diffusion-based architectural biases have emerged as particularly effective inductive structures. yang2023disdiff introduce the first unsupervised framework for disentangling pre-trained diffusion models by automatically discovering latent factors and decomposing gradient fields into factor-conditioned sub-gradients. Further, [**?**] show that diffusion models with cross-attention mechanisms serve as strong inductive biases, relying on the inherent information bottlenecks in the diffusion process and cross-attention mechanisms. Dynamic Gaussian Anchoring in [**?**] bias towards a cluster structure in the latent space of diffusion models with cross-attention mechanisms for better separability between factor variations. Compositional constraints offer another structural approach, where maximizing the validity of composite images generated through stochastic mixing operators between latent representations enforces meaningful factor recombination without factor-specific architectural biases jung2025disentangled. Recent work emphasizes the incorporation of multiple, complementary

inductive biases, for example, Tripod integrates quantization, statistical independence, and inter-latent influence minimization into a single framework [65]. Similarly, our work combines complementary inductive biases derived from nonparametric priors, structured variational inference, and stochastic quantization, within a principled Bayesian framework with a unified objective that provides theoretical grounding for their integration.

1.2.7 Conclusion

In this paper, we introduce a novel approach that incorporates Bayesian nonparametric priors into the latent space of autoencoders. By leveraging the flexibility of nonparametric Bayesian methods, our approach enables the model to adaptively partition the latent space in accordance with the underlying data complexity, promoting more interpretable and structured latent encodings. This prior biases the learned representations toward capturing the underlying structure inherent in the data, thereby facilitating the learning of disentangled representations.

To enable accurate posterior inference under this flexible and hierarchical prior, we introduce a tailored nested and structured variational family. This variational family is specifically designed to preserve both the hierarchical structure of the prior and the inductive bias imposed by a discrete latent space, ensuring that the inference procedure remains expressive enough to capture complex dependencies while maintaining the structural properties essential for effective representation learning.

Our ablation studies systematically isolate and quantify the inductive biases contributed by each component of our hierarchical Bayesian nonparametric framework—namely, the nested variational family, structured variational inference, and stochastic quantization. Bayes-QLAE consistently outperforms all ablated variants across disentanglement metrics, demonstrating that each component provides complementary

inductive biases that, when combined, enhance distinct aspects of disentanglement. Our empirical results demonstrate the effectiveness and generalizability of the proposed approach across image datasets characterized by diverse factor variation distributions. Bayes-QLAE consistently achieves superior or competitive performance relative to baseline methods on both 3DShapes and MPI3D, particularly in terms of both modularity and compactness-based disentanglement metrics. Importantly, this performance is attained solely through architectural inductive biases, without reliance on additional computationally expensive regularization.

The differential in performance on the two datasets suggests that the underlying distributional properties of generative factors significantly influence the efficacy of the nonparametric prior. We hypothesize that replacing the Dirichlet Process prior with a more flexible prior such as the Pitman–Yor process—which allows for a richer clustering structure and can model power-law behaviors—may yield further performance improvements. These findings highlight the potential of our framework for interpretable and structured representation learning in varied settings.

1.2.8 Preliminaries

For a theoretical foundation for our proposed approach, we first review the Dirichlet Process and its role as a nonparametric prior in mixture models, with particular emphasis on the natural clustering structure that emerges from this formulation. Subsequently, we discuss the quantization-based generative model, the VQ-VAE, which utilizes discrete latent representation spaces and review prior work leveraging this quantization-structure for learning disentangled representations. **Dirichlet Process Mixture Models** We begin by considering nonparametric models, defined by an infinite-dimensional parameter space that fundamentally allows the model's complexity to adapt and grow

with the data. These models are typically used as priors over distributions with broad support that encompasses the entire space of all possible distributions. The Dirichlet Process (DP), in particular, is a stochastic process whose realizations are discrete probability distributions, thereby defining a valid nonparametric prior probability distribution over the space of discrete probability measures. sethuraman1994constructive constructive definition of the DP represents each discrete distribution drawn from the DP as a weighted sum of countably infinite atomic measures sampled from a continuous base distribution. This definition uses the stick-breaking construction, where the infinite sequence of weights for the atomic measures of the discrete distribution is generated by iteratively partitioning a unit-length stick. In the first step, a segment of length β_1 is broken off the stick, where β_1 is drawn from a Beta distribution, $\beta_1 \sim \mathrm{Beta}(1, \alpha)$, parameterized by α, ensuring $\beta_1 \in (0, 1)$. This segment is assigned as the weight of the first atomic measure θ_1, which is independently sampled from a base distribution $G_0(\lambda)$ with parameters λ. The remaining portion of the stick, with length $1 - \beta_1$, is then recursively partitioned in the same manner: at each step, a segment of length $\beta_i \sim \mathrm{Beta}(1, \alpha)$, scaled to the length of the remaining stick given by $\prod_{j=1}^{i-1}(1 - \beta_j)$, is broken off and assigned as the weight of an atomic measure θ_i, drawn independently from the base distribution. This explicit stick-breaking construction generates a random discrete distribution $G \sim \mathrm{DP}(\alpha, G_0)$ over the countably infinite atomic measures θ drawn from the base distribution G_0. Building on this, the DP serves as the nonparametric prior over the mixture components in the Dirichlet Process Mixture Models (DPMMs) [95]. DPMMs generate data by first sampling a discrete distribution G from the prior $\mathrm{DP}(\alpha, G_0)$ using the stick-breaking construction and then using the set of atomic measures θ sampled from the base distribution G_0 to parameterize a data-generating distribution F. To generate each data point, we first sample a latent variable z from the discrete distribution defined by the stick-breaking weights, then use the corresponding atomic measure θ_z to parameterize the data-generating distribution

F, which is further used to draw the observation x: $\beta_k \mid \alpha \sim p(\beta_k \mid \alpha) = Beta\,(1, \alpha)\,, \quad \theta_k \mid \lambda \sim p(\theta_k \mid \lambda) = G_0(\lambda), \quad \forall k \in N$

$z = k \mid \beta \sim p(z = k \mid \beta) = \beta_k \prod_{j=1}^{k-1}(1 - \beta_j), \quad x \mid \{z, \theta\} \sim p(x \mid z, \theta) = F(\theta_z)$ Because each realization G drawn from the DP is a discrete distribution over the atomic measures θ, the above data-generating process results in repeated parameter values for the data-generating function F. This effectively induces a partitioning of the data, where each partition or component corresponds to the data points generated with identical parameter values, allowing the generative process to be interpreted as a mixture model. Consequently, this results in a hierarchical Bayesian framework, where the parameters of the data-generating distribution F are sampled from a discrete probability distribution drawn from the DP. The joint distribution of the data $\{x_1, \ldots, x_N\}$ and the latent variables: stick-breaking lengths $\beta = \{\beta_1, \beta_2, \ldots\}$, component parameters $\theta = \{\theta_1, \theta_2, \ldots\}$ and assignment variables $\{z_1, \ldots, z_N\}$, factorizes hierarchically as follows: $\mathrm{p}(x, z, \beta, \theta \mid \alpha, \lambda) = p\,(\beta \mid \alpha)\,p\,(\theta \mid \lambda)\,p\,(z \mid \beta)\,p\,(x \mid z, \theta)$ The primary objective of the learning process is to infer the posterior distribution of the latent variables β, θ and z conditioned on the observed data x and the hyperparameters α, λ, denoted by $p(\beta, \theta, z \mid x, \alpha, \lambda)$.

Computing the exact posterior over the latent variables given the observed data introduces dependencies among the variables. As a consequence evaluating the marginal likelihood of the data requires integrating over every possible latent configuration, making it intractable. In the nonparametric setting, such as under a Dirichlet Process (DP) prior, the posterior cannot be computed exactly and must be approximated.wainwright2008graphical introduce a deterministic approach to approximate the intractable posterior with a simpler, tractable family of distributions by breaking certain dependencies among latent variables.They define a variational family q_ν, parameterized by free parameters ν, and optimize ν to minimize the Kullback–Leibler divergence between q_ν and the true posterior. Equivalently, this corresponds to maximizing the evidence lower bound (ELBO) on the log marginal likelihood of

the data, as defined below: $\log p(x \mid \alpha, \lambda) \geq_{q_\nu} [\log p(x, e, z, \beta, \theta \mid \alpha, \lambda) - \log q_\nu(e, z, \beta, \theta)]$

Next, we discuss the Vector Quantized-Variational AutoEncoder (VQ-VAE) [31], a generative model which learns a discrete latent representation via vector quantisation (VQ). The VQ-VAE model discretizes the continuous encoder outputs $z_e(x)$ by mapping them to a discrete latent space consisting of a codebook with a finite set of K embedding vectors $\{e_k\}_{k=1}^{K}$. The posterior distribution $q(z \mid x)$ of the latent variable z is categorical over the embedding space, with probabilities determined by the Euclidean distances between the encoder output and the embedding vectors in the codebook. Samples drawn from this distribution index the set of embedding vectors, which are then passed as input to the decoder z_q as follows: q(z = k | x) = { 1 $fork = \arg\min_j \|z_e(x) - e_j\|_2$,

$0otherwise$,

$z \sim q(z \mid x), \quad z_q(x) = e_z = e_k$ To enable gradient propagation through the non-differentiable quantization step, a straight-through estimator is used, wherein gradients from the decoder are directly propagated back to the encoder output $z_e(x)$. The loss function used to train the VQ-VAE, defined in eq:vqloss, consists of the reconstruction loss, jointly optimizing the encoder and decoder to maximize the evidence lower bound (ELBO) on the data log-likelihood. Assuming a uniform prior $p(z)$ and a deterministic posterior as in equation 1.2.8, the KL divergence of the ELBO simplifies to the constant $\log K$ and is ignored. The second term corresponds to the vector quantization loss, which updates the embedding vectors by moving them toward the encoder outputs. The third term is the commitment loss, encouraging the encoder outputs to remain close to the selected embeddings and thereby ensuring alignment between the encoder space and the embedding space.

$$L = -\log p(x \mid z_q(x)) + \|sg[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - sg[e]\|_2^2, \tag{1.10}$$

where sg stands for the stop-gradient operator which blocks the gradient from propagating through the computational branch of the operand, treating it as a constant. While standard VQ-VAE approaches discretize the latent representations using a single codebook of high-dimensional embedding vector and optimize primarily for reconstruction fidelity, learning disentangled representations necessitates strong inductive biases [28]. To structure the latent space such that distinct dimensions capture independent generative factors, the approach of hsu2024disentanglement instead propose latent quantization, which enforces structural regularity in the latent space by quantizing each latent dimension using separate learnable scalar codebooks. Specifically, the proposed quantized latent autoencoder (QLAE) parameterizes the latent space as the Cartesian product $Z = C_1 \times \cdots \times C_d$, where each codebook C_j contains scalar embeddings. This element-wise quantization enforces a combinatorial factorized encoding, allowing the decoder to learn consistent interpretations for each latent dimension. Furthermore, a higher weight decay is used to regularize the model to encourage reliance on the discrete codebook structure. Collectively, these design choices promote disentangled representations through explicit architectural and regularization biases.

Recognition Network (or Encoder).[1] The encoder is a feedforward convolutional network. Each convolutional block consists of a downsampling convolution (kernel size 4, stride 2) followed by batch normalization, and each block is followed by a Leaky ReLU activation (slope 0.3). Four such blocks with channel widths 32, 64, 128, and 256 produce a $256 \times 4 \times 4$ feature map. This feature map is flattened and passed through two fully connected layers of width 256 with Leaky ReLU activations. A final linear layer outputs the mean and precision parameters of the approximate conjugate likelihood, using a Softplus transformation on the precision to enforce positivity.

Data-Generating Function (or Decoder). The sampled latent representation is first passed

[1]The implementation is available https://github.com/vspatil-npbayes/bayes-qlae.git

through two fully connected layers of width 256 with ReLU activations and no normalization, preserving the piecewise affine structure of the mapping. The resulting feature vector is reshaped to $256 \times 4 \times 4$ and processed by four transposed-convolution blocks. Each block applies an upsampling transposed convolution (kernel size 4, stride 2) followed by a Leaky ReLU activation (slope 0.3). The blocks use channel widths 128, 64, 32, and 3, yielding a final observation of shape $3 \times 64 \times 64$. For a binary cross entropy loss function the output is followed by a sigmoid activation function.

hyperparameter	value
d	10
α	1 (with a Gamma(1, 1) prior)
β (scaling of KL div.)	0.005
$\lambda = (mu_0, \kappa_0, \nu_0, w_0)$	$(0., 1., 2., 0.1)$
Autoencoder learning rate	1×10^{-4}
Latent learning rate	1×10^{-3}
Adam β_1	0.9
Adam β_2	0.99
Adam updates	2×10^5
Grad clip norm	5.0
batch size	64

Table 1.3: Fixed Hyperparameters and initializations

Model	Info M	Info C	Info E
Bayes-QLAE	$0.58 \pm .04$	$0.51 \pm .03$	$0.71 \pm .04$
T-QLAE (k=10)	$0.54 \pm .04$	$0.40 \pm .03$	$0.68 \pm .04$
T-QLAE (k=50)	$0.51 \pm .06$	$0.48 \pm .05$	$0.62 \pm .06$
MF-QLAE	$0.49 \pm .04$	$0.49 \pm .04$	$0.76 \pm .04$
DQ-QLAE	$0.52 \pm .02$	$0.43 \pm .02$	$0.71 \pm .02$

Table 1.4: Model performance comparison across different information metrics

We structure the experiments in this section to isolate and quantify the specific inductive biases derived from each component of our hierarchical Bayesian nonparametric framework: the nested variational family, structured variational inference, and stochastic quantization. Specifically, we perform an ablation by replacing the nested variational family with a truncated one (T-QLAE) with different

truncation levels K. Similarly, to evaluate the role of the structured variational family, we substitute it with a mean-field variational family (MF-QLAE), as detailed in johnson2016composing. Finally, to isolate the effect of stochastic quantization, we replace with a deterministic nearest-neighbor quantization, with the straight-through estimator used to propagate gradients through the quantization step.

From our experiments (as detailed in Table 1.4) Bayes-QLAE consistently outperforms its ablated variants across all disentanglement metrics, confirming that each component contributes an inductive bias which, when combined, enhances performance. For models based on truncated variational families, we observe a negative correlation between modularity and truncation level, while a positive correlation with compactness. This aligns with the intuition that representations obtained with fewer quantized values are biased toward modularity. Notably, the truncated model with $k = 10$ surpasses QLAE in modularity due to the benefits of stochastic quantization and structured variational inference, though with a slight reduction in compactness as a consequence of stochasticity. Removing structured variational inference and defaulting to deterministic quantization degrades both modularity and compactness, with the mean-field family exhibiting a more severe decline in modularity, suggesting a bias towards representation which minimize the reconstruction cost over representations adhering to structured prior distribution. Finally, deterministic quantization exacerbates posterior collapse, leading to representations with a lower compactness metric.

We empirically demonstrate that a small codebook is not a prerequisite for disentanglement and therefore need not be fixed a priori. Critically, across all three axes of disentanglement assessment—informativeness (reconstruction fidelity), modularity (independence), and compactness (one-to-one factor-dimension correspondence)—we observe that disentanglement quality remains stable or improves as the codebook size expands adaptively in response to data complexity. These findings directly challenge the common assumption that small, fixed codebooks are necessary for learning disentangled representations.

Rather, the critical factors enabling disentanglement are two structural properties: (1) the implicit regularization effect induced by discrete latent encodings, and (2) the combinatorial composition of factor-specific codes to encode representations. During early training stages, when the codebook size is small, the encoder operates under a representational bottleneck that necessitates the construction of latent representations through combinatorial composition of the restricted set of available codes. This bottleneck implicitly regularizes the learning process, strongly biasing the encoder toward allocating disjoint, factor-specific codes to each factor-specific codebook. Consequently, the learned compositional structure mirrors the underlying generative process of the dataset, wherein the set of observations arise from the Cartesian product of discrete factor instantiations. This early-stage regularization effect establishes a foundation for disentanglement by enforcing a modular, compositional encoding scheme that respects the factorial structure of the data-generating distribution.

In our approach we initialize the nonparametric prior with a single code per codebook. The nested variational family provides a principled mechanism to increase the number of codes: new codes are instantiated if and only if their inclusion yields an improvement in the variational lower bound. This criterion ensures that capacity expansion occurs only when statistically justified by the data. Consequently, the model inherits the inductive bias of sparse codebooks while avoiding any explicit hard constraint on the upper bound of the cardinality of the codebooks. This adaptive regularization mechanism resolves the tension between early-stage structural learning and asymptotic expressiveness.

We validate this hypothesis through ablation studies comparing our nested variational inference framework against a truncated variant. In the truncated approach, we fix the number of mixture components at a predetermined upper bound for each factor, effectively eliminating the adaptive capacity of the nonparametric formulation. This modification results in measurable degradation across all disentanglement metrics relative to the nested variational inference approach. These results demonstrate

that the adaptive, data-driven discovery of codebook size—rather than absolute codebook cardinality—is the essential mechanism underlying successful disentanglement.

Moreover, the tendency toward cluster expansion is explicitly governed by the concentration parameter α of the nonparametric prior, which is itself assigned a Gamma hyperprior. This hierarchical Bayesian formulation provides regularization of the cluster proliferation rate, enabling the model to infer from data the appropriate balance between model parsimony and representational capacity without manual specification.

For quantitative evaluation, we compute two complementary disentanglement metrics which comprehensively measure disentanglement properties using different computational approaches. The InfoMEC metric hsu2024disentanglement relies on information-theoretic mutual information estimation computed from the empirical joint distribution between latent representations and ground-truth factors. In contrast, the DCI metric eastwood2018framework trains predictive models to map the learned representations to the underlying factors of variation. Both metrics evaluate disentanglement quality across three fundamental dimensions, though with different terminology: InfoMEC measures InfoModularity (InfoM) while DCI measures Disentanglement (D) to quantify the extent to which sources are encoded in mutually disjoint subsets of representations, InfoExplicitness (InfoE) and Informativeness (I) measure the degree to which the relationship between the sources and representations can be characterized by a simple functional or statistical dependency, and InfoCompactness (InfoC) and Completeness (C) quantifies the degree to which latent variables encode information exclusively about mutually disjoint subsets of the sources. We train our proposed approach in a completely unsupervised manner on the entire dataset and evaluate the learned representations on a subset of samples, using the open-source implementations of disentanglement metrics by locatello2019challenging, hsu2024disentanglement.

We conduct qualitative assessments of our proposed method on each dataset to evaluate both

sample reconstructions and latent traversals. For latent traversals, we encode a single image into the latent space and systematically visualize the effects of intervening on individual latent dimensions. Specifically, for each latent variable we vary its value across the range of values encoded in the representations (sample with replacement) while holding all other latent dimensions fixed, then decode the resulting latent vectors to observe their effects in the data space. In the visualization, each row corresponds to interventions on a single latent variable, while columns represent different values sampled from the empirical distribution of that dimension. Well-disentangled representations should exhibit smooth, semantically meaningful changes along individual latent dimensions, with each dimension controlling a distinct generative factor independently of others.

Reconstruction Fidelity. We assess the informativeness of learned representations by examining reconstruction quality. High-fidelity reconstructions that faithfully preserve visual details of the original images indicate that the latent representations are sufficiently informative to capture the full range of variations present in the data. Conversely, poor reconstructions suggest that certain factors of variation have been inadequately encoded or lost during the encoding process.

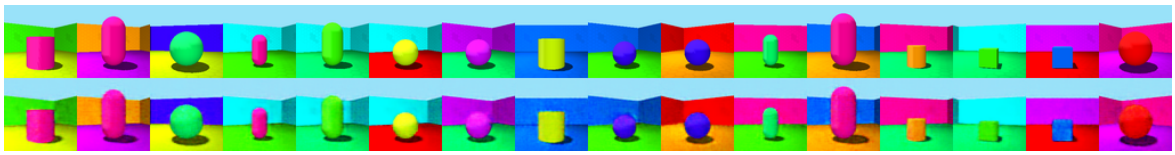
Modularity. We evaluate the modularity, or disentanglement, properties of learned representations through latent traversal analysis. A representation exhibits modularity when each latent dimension independently controls a single underlying generative factor while remaining invariant to variations in other factors. Operationally, this is assessed by examining whether each row in the traversal visualization demonstrates isolated semantic changes corresponding to a single factor of variation without coupling to other factors. Such independence in the latent space reflects successful recovery of the true compositional structure of the underlying independent generative factors.

Compactness. We further assess the compactness, or completeness, of the learned representations by determining whether all variations of a single generative factor are captured within a single latent

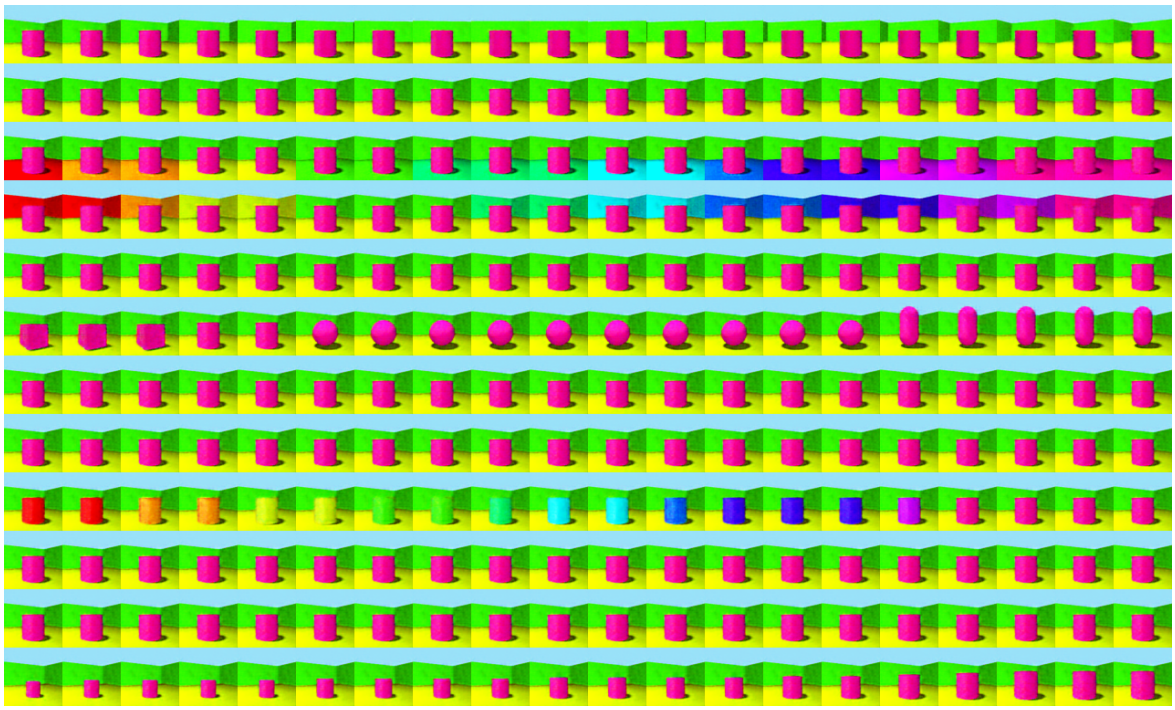
dimension. Compact representations, wherein each factor is encoded by exactly one latent variable, are crucial for interpretability as they establish a one-to-one correspondence between latent dimensions and semantically meaningful generative factors. This property enables intuitive understanding and manipulation of specific attributes in the generated outputs.

1.3 Geometry-Preserving Disentanglement: Isometric Embeddings for Combinatorial Generalization

Real-world data distributions exhibit inherent compositional structure that enables human cognitive abilities such as combinatorial generalization which is defined as the capacity to systematically understand and generate novel compositions of familiar attributes beyond observed examples. Learning disentangled representations offers a promising approach to this challenge by discovering and encoding the natural structure underlying data distributions. Recent vector quantization methods introduce structural inductive biases through discrete latent codes to capture compositional structure; however, these biases alone are insufficient for robust combinatorial generalization. We address this limitation through a principled geometric framework grounded in identifiable representation learning theory. Our approach assumes each generative factor admits a smooth Riemannian manifold structure and seeks to construct Euclidean latent representations that induce approximately isometric embeddings. Critically, we enforce local isometry on the data-generating function, ensuring that transformations in latent space induce geometrically consistent changes in observation space. This functional inductive bias, combined with structural constraints in latent space, enables the model to infer factor manifolds whose geometric properties, local distances, angles, and curvature, remain intact under learned transformations, thereby supporting compositional recombination of latent components. Experiments on 3DShapes and MPI3D benchmarks

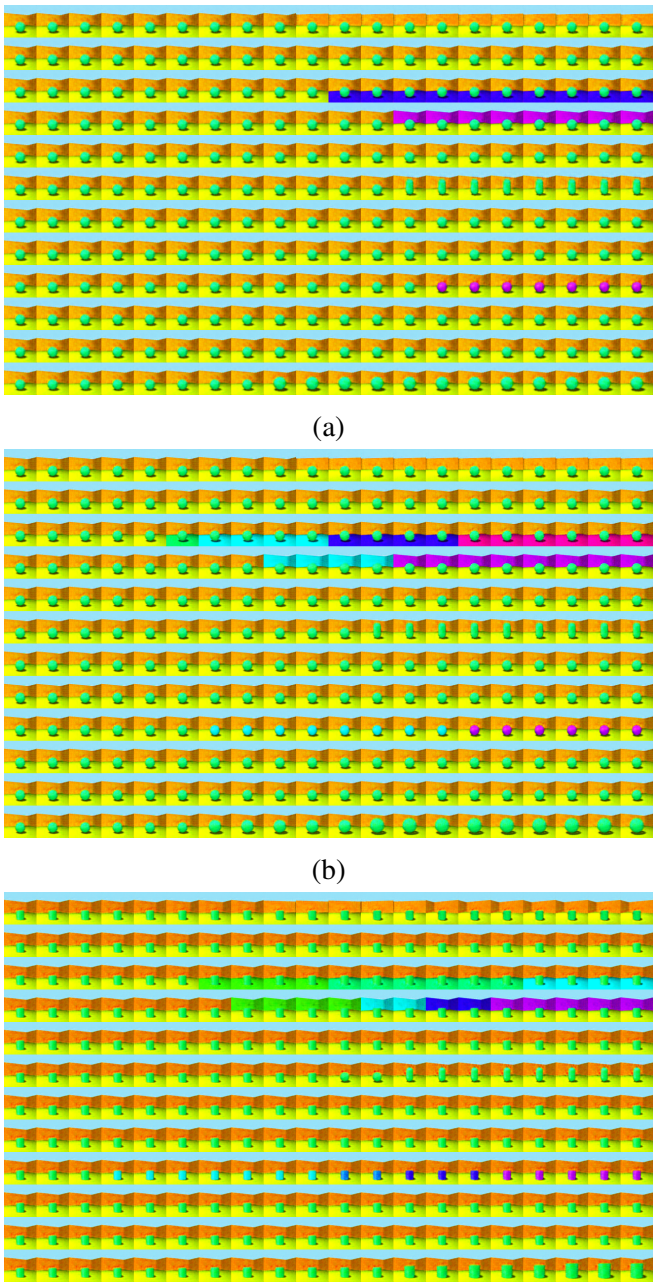


(a) Sample reconstructions: Original images (top row) and corresponding reconstructions (bottom row)



(b) Latent traversals: Each row shows the effect of systematically varying a single latent dimension while holding all other dimensions fixed. Columns represent different values sampled from the distribution of that dimension. The model successfully disentangles six ground-truth factors of variation: object orientation (row 1), floor hue (row 3), wall hue (row 4), object shape (row 6), object hue (row 9), and object scale (row 12). Rows 2, 5, 7, 8, 10, and 11 correspond to inactive latent dimensions that do not encode interpretable factors.

Figure 1.1: Reconstructions and Latent traversals for the 3DShapes dataset: Reconstructions demonstrate high fidelity in capturing visual details demonstrating the model's ability to faithfully encode and decode the full range of variations in the data. Latent traversals illustrate that individual latent variables control distinct, interpretable factors of variation in the generated images. Moreover, the presence of inactive latent variables and the encoding of each factor in a single latent variables indicates that the model has learned a compact representation recovering the true generative structure



(a)

(b)

(c)

Figure 1.2: Evolution of Latent Traversals for 3DShapes dataset: The figure illustrates the evolution of the number of clusters associated with each generative factor demonstrating the adaptive capacity of the nonparametric formulation. Each column in each row corresponds to a factor-specific mixture component, and the distinct components within a row denote the clusters capturing the encoded variations of that factor. The vertical axis indicates cluster multiplicity, revealing how the model progressively discovers and encodes additional variations for each factor. This dynamic cluster growth exemplifies the nonparametric property of the hierarchical mixture prior, which enables data-driven inference of latent capacity without manual specification. Notably, factors with higher contribution to the reconstruction objective—such as floor hue, object hue, and wall hue—exhibit earlier cluster proliferation during training, suggesting the model prioritizes encoding variations that most significantly impact reconstruction fidelity. In contrast, geometric factors such as object orientation and shape undergo refinement in later training stages, indicating a hierarchical learning strategy wherein the model first captures high-variance attributes before refining lower-variance structural properties. This demonstrates

demonstrate that our geometric framework consistently achieves superior or competitive combinatorial generalization compared to strong baselines, providing empirical evidence that geometry-preserving inductive biases are essential for robust disentanglement.

1.3.1 Introduction

The ability to systematically compose known elements into novel configurations lies at the core of human intelligence. Cognitive science has long argued that this combinatorial generalization, the capacity to understand and generate new combinations of familiar components, arises from the structured, compositional nature of human representations of knowledge. In human perceptual and conceptual systems, such compositionality relies on the assumption that the world can be described as combinations of relatively independent factors of variation, which interact through consistent generative principles. Formalizing this intuition in machine learning frameworks remains a central goal for achieving models that generalize systematically when confronted with novel compositions of known factors.

One prominent approach to addressing this challenge involves learning disentangled representations, which aim to discover and explicitly encode the natural factorized structure present in data distributions. The core hypothesis underlying this paradigm is that by decomposing observations into their constituent factors of variation, each corresponding to a distinct, interpretable dimension in the latent space, models can better capture the underlying generative process and thereby facilitate more robust generalization. When successful, disentangled representations enable models to manipulate individual factors independently, potentially allowing systematic recombination to form novel configurations.

Recent work has explored vector quantization-based methods that introduce explicit structural

inductive biases to encourage disentanglement. These approaches leverage discrete latent codes to represent compositional structure, with the intuition that discrete representations may better align with the categorical nature of many real-world factors. By constraining the latent space through architectural choices and training objectives, these methods have shown promise in learning factorized representations. However, despite these advances, a critical question remains: are structural inductive biases alone sufficient to guarantee robust combinatorial generalization?

Moreover, recent empirical investigations have revealed a critical disconnect between disentanglement and combinatorial generalization. montero2022lost demonstrated that models achieving highly disentangled representations nevertheless fail systematically when tested on unseen combinations of factor values, particularly in settings where novel compositions lie within the individual ranges of each factor. This finding suggests that inductive biases in latent space alone are insufficient for robust compositional reasoning. Even when factors are correctly identified and separated, the learned decoder may fail to produce semantically coherent outputs for novel latent configurations, indicating that the functional properties of the generative mapping are equally critical.

Through systematic empirical evaluation, we demonstrate that existing structural inductive biases, while beneficial, are insufficient to consistently achieve robust combinatorial generalization across diverse scenarios. Moreover, many current approaches lack principled theoretical foundations that would provide guarantees about their generalization behavior. This gap between empirical success in constrained settings and theoretical understanding limits our ability to predict when and why these methods will succeed or fail, hindering their reliable deployment in practice.

To move beyond heuristic biases, a more principled approach must situate disentanglement within a formal probabilistic and geometric framework. From this perspective, compositional generalization depends on the existence of an isometric correspondence between the true latent manifold of generative

factors and its learned counterpart within the model. An isometric or geometrically consistent representation ensures that relations among latent variables, such as independence and local linearity, are preserved under the learned mapping. This condition is necessary for recombining latent components in novel configurations to yield meaningful outputs in the observation space.

Building on this conceptual foundation, we propose a probabilistic framework that integrates structural inductive biases with the formal machinery of identifiable deep generative modeling. The central objective is to ensure that the model not only factorizes observed variability into interpretable latent dimensions but also preserves the intrinsic geometry of the underlying manifold through an isometric data-generating function. This formulation yields a representation in which local transformations in latent space correspond to consistent, compositional changes in observation space which is an essential precondition for combinatorial generalization. Theoretically, our framework can be viewed as imposing a two-level constraint: (1) a probabilistic factorization constraint that aligns the learned latent space with the true intrinsic manifold of the generative factors, and (2) a geometric constraint ensuring that the composition of these factors through the data-generating function preserves isometries. By encouraging inference of the intrinsic manifold structure of generative factors, the model is theoretically positioned to generalize systematically to novel combinations, those unobserved during training but valid within the compositional algebra of the learned generative space.

We demonstrate through systematic experiments on standard benchmarks (3DShapes and MPI3D) that our geometric framework achieves superior or competitive combinatorial generalization compared to strong baselines. Our results provide empirical evidence that geometry-preserving inductive biases are essential for bridging the gap between disentanglement and compositional reasoning, advancing both the theoretical understanding and practical capabilities of representation learning systems. We begin by reviewing the theoretical foundations underlying our approach. First, we introduce latent-

quantization-based disentanglement methods followed by the stochastic latent quantization framework guaranteeing identifiability of the learned representations. We build upon these foundations by incorporating additional inductive biases, specifically, structural and isometry constraints, to better support combinatorial generalization.

1.3.2 Latent Quantization

We begin with the framework proposed by hsu2024disenta nglement which introduces a structural inductive bias on the latent space by learning factorized discrete latent representations via vector quantization [31] to promote disentanglement. The central idea is to factorize the latent space into independent scalar subspaces and quantize each dimension of the latent representation independently using learnable codebooks with scalar codes. Formally, given an input x, the encoder outputs a continuous latent representation vector $\mathbf{e}(x) = (e_1(x), \ldots, e_d(x)) \in R^d$. Each scalar component $e(x)_i$ is independently quantized using a corresponding codebook $C_i = \{e_{i,1}, \ldots, e_{i,|C_i|}\}$ where each $e_{i,k} \in R$ represents a learnable code embedding. For each dimension i, a discrete latent variable $z_i \in Z_i = \{1, \ldots, |C_i|\}$ indexes the codes of the corresponding codebook C_i, i.e.,

$$z_i =_k |e(x)_i - e_{i,k}|, \quad k = 1, \ldots, |C_i| \tag{1.11}$$

where the discrete latent variable $\mathbf{z} = (z_1, \ldots, z_d)$ is formed from the Cartesian product $\mathbf{z} \in Z_1 \times \ldots \times Z_d$. Each z_i indexes the quantized embedding e_{i,z_i} for dimension i, and the decoder operates on the set of quantized embeddings $\{e_{i,z_i}\}_{i=1}^d$ in place of the continuous encoder outputs. This element-wise independent quantization enforces a combinatorial structure over the latent representations and provides an explicit inductive bias towards factorized representations in which distinct latent variables

capture distinct generative factors.

1.3.3 Dirichlet Process prior for Stochastic Quantization

Building on the theoretical framework of kivva2022identifiability for establishing identifiability in latent generative models, anonymous2026a introduce stochastic latent quantization as a principled mechanism for identifiable learning of disentangled representations. In this formulation, a mixture prior is defined independently for each latent dimension, governing the distribution of embeddings within its corresponding codebook. To endow this framework with universal approximation capabilities, a Dirichlet Process (DP) mixture prior is defined, allowing each codebook to represent a countably infinite set of embeddings.

Formally, under this construction, each embedding corresponding to dimension i is drawn from an infinite mixture governed by a random probability measure $G_i \sim \mathrm{DP}(\alpha_i, G_0) = \sum_{k=1}^{\infty} \pi_{i,k}\, \delta_{\theta_{i,k}}$ where α_i is the concentration parameter and G_0 denotes the base distribution over component parameters. The mixture weights $\pi_{i,k}$ are generated via the stick-breaking process $\pi_{i,k} = \beta_{i,k} \prod_{t<k}(1-\beta_{i,t})$ with $\beta_{i,k} \sim \mathrm{Beta}(1, \alpha_i)$, while each component parameter $\theta_{i,k} \sim G_0$ specifies the distribution from which an embedding in the codebook is generated. An embedding e_i is then generated by first sampling an index $z_i \sim \mathrm{Categorical}(\pi_{i,1:\infty})$ followed by drawing the scalar embedding;

$$e_i \mid z_i \sim p(e \mid z_i, \theta_i) = \prod_{k=1}^{\infty} p(e \mid \theta_{i,k})^{I[z_i=k]} \tag{1.12}$$

Following the identifiability framework, the data likelihood is modeled through a piecewise affine data-generating function g implemented as a deep neural network with ReLU activations, which maps

the latent representations to the observation space,

$$p(x \mid \mathbf{e}) = \mathcal{N}(x \mid g(\mathbf{e}), \sigma^2 \mathbf{I}), \tag{1.13}$$

where σ^2 denotes a fixed observation noise variance. The hierarchical Bayesian structure of this stochastic latent quantization framework provably ensures identifiability of the generative process while enabling principled learning of disentangled representations.

1.3.4 Geometry-Preserving Isometric Manifolds

We begin with the assumption that each true generative factor lies (approximately) on a smooth Riemannian manifold endowed with an intrinsic metric (geodesic distance). By the Nash embedding theorem, any smooth Riemannian manifold can be isometrically embedded into a Euclidean space of sufficiently high dimension. For instance, while a 2-sphere S^2 intrinsically has dimension 2, it requires embedding into R^3 to preserve distances isometrically. Our objective is to learn a Euclidean latent representation whose dimension is close to the intrinsic dimension of each factor manifold while approximately preserving its Riemannian metric, that is, to recover an embedding that incurs minimal distortion of geodesic distances, thereby aligning the learned representations with the true geometric structure of the underlying generative factors.

In turn, a central requirement for such geometrically faithful, disentangled representations that support compositional generalization is that the data-generating function preserve this geometric structure when mapping latent representations back to observation space. As observed by horan2021unsupervised, if the mapping from representations to the data manifold is locally isometric, then the latent parameterization faithfully encodes the manifold's intrinsic geometry, preserving local distances, angles, and curvatures

associated with each factor. This geometric consistency ensures that relations among latent variables, such as independence and local linearity, remain intact under the learned transformation. Models that respect such invariances, ensuring that transformations in latent space induce well-defined, bijective correspondences in data space, provably recover the true generative structure up to isometric equivalence. Critically, this isometry constraint is necessary for compositional generalization: only when geodesic structure is approximately preserved can latent components be recombined in novel configurations to produce semantically coherent outputs in observation space

To satisfy these geometric requirements of faithful latent embeddings of factor manifolds and isometry-preserving generative mappings we propose a probabilistic framework that combines the identifiable structural inductive biases in the latent space with the functional inductive biases on the data-generating map, enabling recovery of the true manifold of each generative factor. In the sections that follow, we formalize this framework by first describing the structure of the latent space, which factorizes observed variability into identifiable dimensions while preserving the intrinsic geometry of each underlying factor manifold. We then introduce the functional inductive biases, specifically, local isometry regularizations on the data-generating function that ensure local transformations in latent space induce consistent, compositional changes in observation space, thereby establishing the conditions necessary for combinatorial generalization.

1.3.5 Structured Latent Space

We structure the latent space to learn a higher-dimensional Euclidean embedding for each factor that approximately isometrically embeds its true manifold, while preserving the structural constraints required for identifiability. Specifically, the embedding space is structured as $\mathbf{e} \in R^{d \times l}$, decomposed

into d vector components where each $\mathbf{e}_i \in R^l$ encodes a generative factor in an l-dimensional Euclidean subspace. Extending the stochastic latent quantization framework, each embedding vector e_i conditioned on the discrete latent variable z_i is generated according to the mixture model

$$p(\mathbf{e}_i \mid z_i, \theta_i) = \prod_{k=1}^{K} p(\mathbf{e}_i \mid z_i = k, \theta_{i,k})^{I[z_i=k]} \tag{1.14}$$

Crucially, we impose conditional independence within each subspace: the l scalar components $e_{i,1}, \ldots, e_{i,l}$ of e_i are independent given the discrete latent variable $z_i = k$ and the component-specific parameters. This yields the factorized likelihood,

$$p(\mathbf{e}_i \mid z_i = k, \theta_{i,k}) = \prod_{\ell=1}^{l} p(e_{i,\ell} \mid \theta_{i,k,\ell}), \tag{1.15}$$

where $\theta_{i,k,\ell}$ parameterizes the scalar distribution of ℓ-th coordinate. Importantly, this factorization satisfies the structural constraints required for identifiability under the framework of [**?**], specifically the (P2) assumption that continuous latent variables must be conditionally independent given their corresponding discrete latent variable.

Consequently, for each subspace i we introduce l separate scalar codebooks, one per coordinate, where each codebook contains the parameters $\theta_{i,k,\ell}$ governing the distribution of scalar coordinate $e_{i,\ell}$ under discrete assignment $z_i = k$. This coordinate-specific parameterization allows dimensions that encode no information to collapse to their prior distribution, effectively pruning uninformative coordinates. In our implementation, we choose a univariate Gaussian likelihood for each scalar embedding:

$$e_{i,\ell} \sim \mathcal{N}(\mu_{i,k,\ell}, s_{i,k,\ell}^{-1}), \tag{1.16}$$

where $\theta_{i,k,\ell} = (\mu_{i,k,\ell}, s_{i,k,\ell})$. To enable principled Bayesian inference with a conjugate structure, we use a Normal–Gamma prior as the base distribution of the corresponding DP prior.

1.3.6 Isometric Generative Map

We motivate isometric mapping as a functional inductive bias based on theoretical and empirical advances in geometry-preserving representation learning. lee2022regularized regularize autoencoders to ensure approximate isometry between latent space and data manifold, preserving intrinsic geometric structure to improve downstream performance. Extending this principle to unsupervised disentanglement, [73] prove that local isometry combined with non-Gaussian priors enables provably identifiable disentanglement, establishing that geometric constraints on the decoder suffice to recover true generative factors.

Intuitively an isometry is a mapping between two spaces that preserves distances and angles, preserving the inner product everywhere. [Isometry] Let $f : \mathcal{M} \to \mathcal{N}$ be a mapping between Riemannian manifolds $(\mathcal{M}, G)$ and $(\mathcal{N}, H)$. We say that f is an isometry if, for all $z \in \mathcal{M}$,

$$G(z) \;=\; J_f(z)^\top \, H(f(z)) \, J_f(z), \tag{1.17}$$

where $J_f(z)$ denotes the Jacobian of f at z. Requiring f to be an isometry can be overly restrictive whereas preserving only the angles and scaled distances may suffice. A conformal map instead preserves angles but not necessarily distances [**?**].

We impose isometry constraints independently on each subspace e_i, where each subspace admits a Gaussian mixture prior. The data-generating function must map approximately isometrically from the parameter space, endowed with the Fisher information metric, to Euclidean space. This requires preserving Fisher-Rao distances between component distributions rather than naïve Euclidean distances

between parameter vectors, as the Fisher metric defines the intrinsic geometry of the statistical manifold.

To achieve this, we impose two complementary constraints. First, a global isometry loss ensures Euclidean distances between mapped component centers match Fisher-Rao distances between their underlying distributions. Second, a local isometry loss enforces that small perturbations around each component preserve Fisher-metric norms in their Euclidean embedding. Together, these losses align the learned representation with the Fisher geometry of the mixture manifold.

To ensure that the data-generating function preserves the global geometric structure of the latent mixture distribution for each embedding space e_i, we impose a constraint that Fisher-Rao distances between mixture components are approximately preserved under the piece-wise linear transformation $f(z) = Wz+b$. The Fisher information metric $G(\theta)$ on the parameter manifold of Gaussian components induces a geodesic distance $d_{FR}(\theta_{i,k}, \theta_{i,k'})$ between components k and k' of subspace i. For univariate Gaussians with parameters $\theta_{i,k} = (\mu_{i,k}, s_{i,k})$, the Fisher-Rao metric has a closed form, and for multivariate diagonal-covariance Gaussians, the total squared distance decomposes as the sum of coordinate-wise squared distances.

Given representative samples $\{e_i \mid z_i = k\}$ from each component, we define the target Fisher-Rao distance as

$$D^{FR}_{kk'} = d_{FR}(e_i \mid z_i = k,\, e_i \mid z_i = k')^2. \tag{1.18}$$

The data-generating function induces Euclidean distances in the transformed space:

$$D^{W}_{kk'} = \|W(e_i \mid z_i = k) - W(e_i \mid z_i = k')\|_2^2. \tag{1.19}$$

Approximate isometry is then enforced by minimizing the distance-matching objective

$$L_{global}(W) = \sum_{(k,k')\in\mathcal{N}} (D^W_{kk'} - D^{FR}_{kk'})^2, \tag{1.20}$$

This objective corresponds to the standard isometric embedding problem, wherein a linear transformation is constrained to preserve geodesic distances defined by the Fisher-Rao metric, with $\mathcal{N}$ denoting pairs of components to emphasize preservation of local manifold structure.

To enforce approximate local-isometry between the latent manifold and a layer's output space, we constrain the local metric structure at sampled embeddings. The Fisher information matrix $G(e_i)$ induces a Riemannian metric on the latent manifold, defining the norm of tangent vectors v as $\|v\|^2_G = v^\top G(e_i)v$. Local isometry requires that this intrinsic norm be preserved under the functional mapping for tangent directions at each point e_i.

For a piecewise affine decoder with weight matrix W, the Jacobian is constant within each linear region: $J = W$. The local isometry 1.3.6 condition then becomes

$$v^\top G(e_i)v \approx \|Wv\|^2_2 \quad for tangent vectors vate_i. \tag{1.21}$$

We approximate tangent directions via finite differences between samples from the same mixture component. This yields the regularization objective

$$L_{local}(W) = E_{e_i,v}[\|Wv\|_2 - \sqrt{v^\top G(e_i)v}]^2, \tag{1.22}$$

which penalizes deviations from isometry by minimizing the discrepancy between the Euclidean norm of transformed tangent vectors and their Fisher-metric norm in the latent space.

In the context of geometry-preserving representation learning, a conformal mapping of degree zero [**?**], also referred to as scaled isometry, represents a relaxation of strict isometric mappings that permits global scaling while preserving local geometric structure. Unlike strict isometries which preserve exact geodesic distances, scaled isometries preserve angles and relative distances up to a constant scaling factor, thereby offering greater flexibility in optimization while maintaining the essential geometric structure of the manifold. We instead adopt this conformal regularization framework of lee2022regularized, which preserves the angular structure of the data manifold while permitting local metric scaling, thereby relaxing the strict distance-preservation requirement of isometric embeddings.

Further, to ensure numerical stability and well-conditioned transformations, we apply spectral normalization to the weight matrices constraining their singular values. Combined with the isometry objectives defined above, this regularization aligns the learned embedding with the Fisher-Rao geometry of the mixture distribution yielding an approximately isometric linear embedding of the Gaussian mixture latent manifold.

1.3.7 Experimental Design

Datasets To evaluate combinatorial generalization we employ two benchmark datasets with distinct compositional structures: 3D Shapes [85] and MPI3D-Real [89]. The 3D Shapes dataset comprises rendered images varying systematically along six independent generative factors (floor hue, wall hue, object hue, shape, scale, and orientation), providing a controlled testbed where each factor spans uniformly across its range. In contrast, MPI3D-Real consists of real-world robotic images exhibiting highly imbalanced factor variability: certain dimensions, such as object rotation, span up to 40 discrete values, while most object-specific properties exhibit only 2–6 distinct states. This

imbalance creates a challenging combinatorial regime where models must generalize across factors with vastly different degrees of training coverage, testing whether learned manifold structure enables extrapolation to under-sampled regions of the compositional space.

Generalization Tasks To evaluate combinatorial generalization, we adopt the compositional holdout task following montero2021role, systematically excluding factor combinations from training and evaluating reconstruction on these unseen configurations. We test two generalization regimes of increasing difficulty: (i) recombination-to-range (R2R), where entire ranges of combinations involving factor subsets are withheld, requiring generalization over broader compositional gaps and (ii) extrapolation (E), where models must generate outputs for factor values outside the training distribution, representing the most challenging regime. We exclude the recombination-to-element regime, where all factor values appear in training but certain complete combinations are withheld, due to its limited difficulty. This protocol isolates compositional reasoning from memorization, providing a principled measure of whether learned representations support systematic recombination of disentangled factors to synthesize novel configurations.

1.3.8 Evaluation Metrics

Disentanglement Metrics. We evaluate disentanglement using two complementary metrics. InfoMEC [32] quantifies factor-specific information captured by each latent dimension via mutual information estimation. DCI [64] employs predictive modeling to assess three properties: Disentanglement (one-to-one factor-latent alignment or Modularity), Completeness (factor information preservation or Compactness), and Informativeness (overall encoding quality or Explicitness). Together, these provide information-theoretic and prediction-based perspectives on representation quality.

Table 1.5: Quantitative comparison of disentanglement metrics and combinatorial generalization task performance across autoencoder architectures on Shapes3D and MPI3D benchmarks. We report the Information-theory based Modularity, Explicitness Compactness (Info-MEC), Disentanglement-Completeness-Informativeness (DCI), binary cross entropy cost for the recombination-to-range (R2R) and the Extrapolation E tasks, and classification accuracy on the recombination-to-range task (ACC). Arrows indicate desired direction: ↑ denotes higher is better, ↓ denotes lower is better.

	SHAPES3D					MPI3D				
(LR)2-6 (LR)7-11 MODEL	INFOMEC↑	DCI↑	R2R↓	E↓	ACC↑	INFOMEC↑	DCI↑	R2R↓	E↓	AC
β-VAE	0.66	0.65	0.41	0.52	0.34	0.49	0.41	0.72	0.72	0.3
β-TCVAE	0.70	0.66	0.32	0.48	0.47	0.52	0.43	0.61	0.68	0.3
SAE	**0.83**	**0.80**	0.12	0.24	**0.97**	0.59	0.56	0.29	0.34	**0.**
QLAE	0.76	0.77	0.18	0.35	0.87	0.56	0.51	0.41	0.47	0.6
BAYES QLAE	**0.81**	**0.80**	0.14	0.30	0.90	0.62	0.58	0.37	0.43	0.7
ST-ISO-BAYES QLAE	**0.83**	**0.82**	**0.09**	**0.18**	**0.98**	**0.64**	**0.62**	**0.24**	**0.38**	**0.**

Measuring Combinatorial Generalization We assess combinatorial generalization using the reconstruction cost (R2R and E) and classification accuracy (ACC) [108]. The reconstruction cost measures reconstruction fidelity on unseen combinations of generative factors, capturing how well the model reproduces variations beyond the training distribution. Accuracy evaluates the model's ability to correctly identify discrete factor configurations, reflecting categorical generalization performance. These complementary metrics jointly quantify a model's capacity to generalize systematically to novel factor combinations.

1.3.9 Evaluation

Quantitative Evaluation. Table 1.5 presents a comprehensive comparison of disentanglement performance and combinatorial generalization performance across six autoencoder variants, with distinct inductive biases, on the Shapes3D and MPI3D benchmarks. Specifically, we compare against β-VAE higgins2017beta and β-TCVAE chen2018isolating, which enforce disentanglement through information-

Table 1.6: Results on Shapes3D (mean ± std over 5 random seeds).

MODEL	INFOMEC↑	DCI↑	R2R↓	E↓	ACC↑
β-VAE	0.65 ± 0.02	0.64 ± 0.02	0.43 ± 0.04	0.55 ± 0.06	0.31 ± 0.03
β-TCVAE	0.71 ± 0.03	0.67 ± 0.02	0.35 ± 0.05	0.53 ± 0.07	0.44 ± 0.04
SAE	$\mathbf{0.84 \pm 0.03}$	$\mathbf{0.81 \pm 0.03}$	$\mathbf{0.14 \pm 0.05}$	0.22 ± 0.06	$\mathbf{0.96 \pm 0.03}$
QLAE	0.76 ± 0.01	0.75 ± 0.02	0.20 ± 0.03	0.40 ± 0.05	0.85 ± 0.03
BAYES QLAE	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.80 \pm 0.02}$	$\mathbf{0.15 \pm 0.03}$	0.34 ± 0.05	0.89 ± 0.06
ST-ISO-BAYES QLAE	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.83 \pm 0.03}$	$\mathbf{0.10 \pm 0.04}$	$\mathbf{0.24 \pm 0.07}$	$\mathbf{0.94 \pm 0.05}$

Table 1.7: Results on MPI3D (mean ± std over 4 random seeds).

MODEL	INFOMEC↑	DCI↑	R2R↓	E↓	ACC↑
β-VAE	0.45 ± 0.04	0.39 ± 0.03	0.75 ± 0.05	0.76 ± 0.07	0.26 ± 0.06
β-TCVAE	0.54 ± 0.04	0.44 ± 0.04	0.58 ± 0.05	0.66 ± 0.08	0.29 ± 0.06
SAE	0.62 ± 0.05	0.59 ± 0.04	$\mathbf{0.32 \pm 0.06}$	$\mathbf{0.36 \pm 0.08}$	$\mathbf{0.80 \pm 0.07}$
QLAE	0.56 ± 0.03	0.52 ± 0.03	0.43 ± 0.05	0.49 ± 0.06	0.58 ± 0.05
BAYES QLAE	0.59 ± 0.04	0.58 ± 0.03	0.41 ± 0.05	0.48 ± 0.07	0.68 ± 0.06
ST-ISO-BAYES QLAE	$\mathbf{0.65 \pm 0.04}$	$\mathbf{0.63 \pm 0.04}$	$\mathbf{0.28 \pm 0.05}$	$\mathbf{0.41 \pm 0.06}$	$\mathbf{0.75 \pm 0.06}$

theoretic regularization that encourages independence across latent dimensions. We further consider SAE [68], which learn structured disentangled representations by introducing latent variables hierarchically in the decoder architecture without relying on explicit regularization terms. In addition, we examine QLAE [32], which introduces an architectural bias based on latent quantization, and subsequently BayesQLAE, which use stochastic quantization with identifiable generative models to learn disentangled representations.

Across both Shapes3D and MPI3D, the proposed Iso-Bayes QLAE consistently outperforms most baselines on both the disentanglement metrics along with the combinatorial generalization performance metrics. On Shapes3D, Iso-Bayes QLAE obtains the comparable InfoMEC and DCI scores while also yielding the lowest R2R, E error, indicating the isometric constraint effectively preserves the manifold structure necessary for both disentanglement and generalization tasks. The performance improvement is most pronounced in the extrapolation regime, where generalization capability depends critically on

recovering the true manifold structure of generative factors. Our results are most comparable to SAE, which similarly constrains the decoder to learn structured representations, corroborating that functional constraints on the data-generating function are essential for combinatorial generalization. The MPI3D dataset presents a more challenging scenario, as evidenced by generally lower scores across all models, however, a similar pattern holds on the dataset. Iso-Bayes QLAE outperforms most baselines on both disentanglement metrics and generalization tasks underscoring the robustness of the approach under increased visual complexity.

Ablation Studies. We conduct systematic ablations to isolate the contribution of each inductive bias in our framework.

Representation structure. We first evaluate the necessity of vector-valued latent representations by applying isometry constraints to the data-generating function of the baseline BayesQLAE with scalar representations (Iso-BayesQLAE). Additionally, we assess the role of conditional independence by replacing the diagonal covariance structure with full covariance matrices (C-BayesQLAE), thereby removing the factorization constraint on within-factor variability.

Isometry regularization. We ablate the geometric constraints on the data-generating function through three configurations: (i) no isometry regularization (St-BayesQLAE), (ii) global isometry alone (component-level distance matching) (GIso-BayesQLAE), and (iii) local isometry alone (tangent-space preservation) (LIso-BayesQLAE). This isolates the contribution of each geometric constraint to the overall performance.

Target metric. Finally, we evaluate whether preserving Fisher-Rao geometry is essential by replacing the Fisher information metric with standard Euclidean distances on sampled embeddings from the mixture prior (EIso-BayesQLAE). This tests whether the statistical manifold structure, rather than simple embedding-space distances, is necessary for effective disentanglement and generalization.

Table 1.8: Ablation study on BayesQLAE variants for the Shapes3D dataset. DCI measures disentanglement (higher is better), while R2R and E assess recombination-to-range and extrapolation errors (lower is better), showing that the structured isometric BayesQLAE model achieves the best overall performance on both tasks.

Model	DCI ↑	R2R ↓	E ↓
St-Iso-BayesQLAE	0.82	0.09	0.18
Iso-BayesQLAE	0.84	0.14	0.31
C-BayesQLAE	0.71	0.24	0.38
St-BayesQLAE	0.74	0.27	0.35
G-Iso-BayesQLAE	0.78	0.19	0.21
L-Iso-BayesQLAE	0.79	0.13	0.22
E-Iso-BayesQLAE	0.80	0.16	0.23

From our experiments (as detailed in Table 1.8) **Discussion.** Across all settings, the full model consistently outperforms its ablated variants on both disentanglement and combinatorial generalization metrics, indicating that each inductive bias provides a complementary contribution and that their combination is necessary to achieve the best performance overall. In particular, the C-BayesQLAE variant exhibits a pronounced drop in disentanglement metrics, suggesting that replacing scalar with vector quantization without appropriate structural constraints destroys the compositional structure of the codebook and leads to multiple generative factors being entangled within individual codes. For the Iso-BayesQLAE variant, disentanglement metrics are slightly higher than those of the proposed model, but combinatorial generalization performance, especially in the extrapolation regime, is substantially worse, indicating that improved factor separation alone is insufficient without the full set of geometric constraints.

For the isometry-regularization ablations, we observe a marked reduction in generalization performance whenever either global or local isometry loss is removed. The global isometry term contributes the largest single improvement, yet both global and local terms are required to achieve the strongest results. Notably, when all isometry regularization is removed, disentanglement scores also decline

in the vector-representation setting, indicating insufficient regularization to enforce a factorized latent structure.

Finally, the target-metric ablation shows that defining isometry regularization in the naïve Euclidean space of sampled embeddings, instead of using the Fisher information metric on the mixture components, leads to reduced performance on both recombination-to-range and extrapolation regimes. This highlights the importance of respecting the statistical manifold geometry of the prior, rather than relying solely on Euclidean distances in the latent embedding space.

1.3.10 Related Works

Disentangled Representations. Disentanglement learning builds on the classical problem of Independent Component Analysis (ICA) , later reframed through the lens of representation learning as the challenge of identifying statistically independent generative factors [20]. While linear ICA achieves identifiability under certain assumptions, nonlinear variants render disentanglement theoretically unidentifiable [26]; [13] without additional information. To address this, researchers have introduced auxiliary data [25]; [26], weak supervision or inductive biases [13] that constrain model behavior. Information-theoretic approaches encourage factorization via bottleneck formulations [22]; [34] and explicit penalties on total correlation [23]; [24], while recent biologically [76] and temporally inspired methods leverage activity sparsity and independent temporal variation. Complementarily, architectural biases such as quantized latent structures [65]; [67], constrained decoding circuits [68], modular sub-network designs [48], flow-based invertible layers [75], causal factorization, and language-informed or diffusion-based anchoring have demonstrated empirical and theoretical advantages. Contemporary work increasingly integrates multiple complementary biases spanning quantization, independence

enforcement, and structured inference within unified probabilistic frameworks [32] to achieve more identifiable and robust disentangled representations.

Combinatorial Generalization. While the methods above primarily target disentanglement quality itself, a parallel line of inquiry examines whether and how disentangled representations enable compositional generalization—the ability to recombine learned factors in novel configurations. Empirical studies reveal that standard disentanglement metrics correlate weakly with compositional generalization performance, with models achieving high disentanglement scores often failing to generalize to out-of-distribution combinations. Recent work has begun identifying specific mechanisms that close this gap. yang2023vector demonstrate that vector-based representations which use vectors rather than scalars to encode individual generative factors are crucial for achieving both strong disentanglement and compositional generalization, suggesting that representation granularity fundamentally impacts systematic generalization capabilities. Taking a group-theoretic perspective, they propose MAGANet, which learns latent group actions and transformations between data pairs rather than encoding individual examples independently, achieving superior combinatorial generalization under distribution shift through group-equivariant decoder constraints. Complementing these representational insights, liang2025compositional argue that the decoder architecture is equally critical, showing that forcing disentangled latent factors to be directly rendered into observation space through architectural constraints or regularization dramatically improves data efficiency and out-of-distribution compositional robustness. Most recently, jung2025disentangled introduce a modular compositional bias framework enabling flexible disentanglement of both attributes and objects through factor-specific latent mixing rules, demonstrating that unified compositional priors can handle diverse generative structures without modifying base architectures or training objectives.

1.3.11 Discussion

This work establishes that geometric structure preservation is essential for bridging disentanglement and combinatorial generalization. While prior research has achieved strong disentanglement through structural inductive biases in latent space, we demonstrate that these constraints alone are insufficient.

Our framework introduces a principled geometric approach grounded in identifiable representation learning theory, combining two complementary inductive biases. First, we structure the latent space as factorized vector-valued subspaces that can express higher-dimensional Euclidean embeddings of each generative factor's intrinsic manifold. Second, and crucially, we enforce approximately isometric mappings from latent space to observations through Fisher-Rao distance preservation ensuring that the data-generating function respects the geometric relationships encoded in latent representations.

Our empirical results on 3DShapes and MPI3D benchmarks validate this theoretical framework, demonstrating that isometry-regularized models consistently achieve superior or competitive performance on both disentanglement metrics and combinatorial generalization tasks, particularly in challenging extrapolation regimes. Systematic ablations confirm that each component contributes meaningfully: vector representations with conditional independence structure enable richer geometric embeddings, while both global and local isometry constraints are necessary to preserve manifold structure across novel factor combinations

Our findings challenge the prevailing focus on latent-space regularization alone and underscore that successful compositional reasoning requires explicit geometric constraints on how representations map to observations. This perspective aligns with recent evidence that decoder architecture fundamentally shapes generalization capabilities and suggests a shift toward viewing disentanglement not merely as statistical independence in latent space, but as geometric consistency across the full generative

pathway.

The capacity for combinatorial generalization represents a cornerstone of human intelligence. By demonstrating that geometry-preserving inductive biases enable machine learning systems to systematically recombine learned factors beyond observed examples, this work advances both the understanding and practical capabilities of disentangled representation learning. We believe that honoring the intrinsic geometric structure of generative processes, rather than imposing arbitrary statistical constraints, offers a principled path toward more robust, interpretable, and generalizable AI systems capable of human-like compositional reasoning.

Appendix A: Experiments

A.1 Problem Definition

The problem of continual learning of disentangled representations can be mathematically formulated as the capacity of a model to sequentially learn and maintain disentangled encodings of independent latent factors across a stream of domains, while preserving previously acquired factor knowledge and avoiding catastrophic forgetting.

Consider a sequence of tasks or domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$, where each domain is characterized by data $\mathcal{D}_t = \{x_i^{(t)}\}_{i=1}^{N_t}$ generated from a potentially evolving set of underlying factors $\mathcal{F}^{(t)} = \{f_1^{(t)}, f_2^{(t)}, \ldots, f_{k_t}^{(t)}\}$. Each factor $f_j^{(t)}$ takes values in a domain-specific set $\mathcal{F}_j^{(t)}$, and the generative process for task t is given by:

$$x^{(t)} = g^{(t)}(f_1^{(t)}, f_2^{(t)}, \ldots, f_{k_t}^{(t)}) \tag{A.1}$$

where $g^{(t)}$ is a task-specific composition function that shares structural similarities across tasks. The continual disentangled representation learning objective is to sequentially learn encoder functions $\Phi_t : \mathcal{X} \to R^{d_t}$ for each task t such that:

$$\Phi_t(x^{(t)}) = [\phi_1^{(t)}(f_1^{(t)}), \phi_2^{(t)}(f_2^{(t)}), \ldots, \phi_{k_t}^{(t)}(f_{k_t}^{(t)})] \tag{A.2}$$

where each $\phi_j^{(t)}$ encodes only factor $f_j^{(t)}$ and is invariant to other factors within the task. Key constraints in continual learning of disentangled representations include the following. First, factor preservation requires that for factors $f_j^{(s)}$ encountered in earlier tasks $s < t$, the representations remain consistent

across tasks, such that $\phi_j^{(t)}(f_j^{(s)}) \approx \phi_j^{(s)}(f_j^{(s)})$, provided $f_j^{(s)} \in \mathcal{F}_j^{(t)}$, thereby preventing catastrophic forgetting. Second, factor alignment requires that shared factors across tasks, $f_j^{(t)} = f_j^{(s)}$ for some j, s, t, are encoded using identical functions, that is, $\phi_j^{(t)} = \phi_j^{(s)}$, ensuring consistent semantics in the latent space. Third, the model must possess the capacity to dynamically expand and adapt its representational space to accommodate novel variations within existing factors or to incorporate entirely new factors that emerge across different domains or tasks.

A.2 Using the Indian Buffet Process for Continual Learning of Disentangled Representations

The Indian Buffet Process (IBP) is a stochastic process that defines a probability distribution over sparse binary matrices Z with a finite number of rows and an infinite number of columns. Let Z be an $N \times K$ binary matrix where $z_{nk} = 1$ indicates that object n possesses feature k. The IBP intuitively follows a sequential generative process analogous to customers entering an Indian restaurant buffet where the first customer samples $Poisson(\alpha)$ dishes. Similarly, the i-th customer samples each previously chosen dish k with probability $\frac{m_k}{i}$, where m_k is the number of previous customers who chose dish k. Further, the i-th customer then samples $Poisson(\alpha/i)$ new dishes. The probability of generating a particular binary matrix Z is modeled under the IBP.

We propose leveraging the IBP as a nonparametric prior for continual learning of disentangled representations, where different domains or tasks are generated from subsets of the total learned factors. In this proposed approach we use the IBP as a prior distribution over the latent factors and the binary matrix indicating which factors are active for a specific task or domain. The IBP prior naturally determines which subset of factors $\mathcal{F}^{(t)}$ is active for each task, with sparsity encouraged by the process structure. Further, previously discovered factors can be reactivated in new tasks with

probability proportional to their historical usage $\frac{m_k}{t}$, promoting positive transfer learning. New factors are introduced with rate $\frac{\alpha}{t}$, allowing the model to grow its representational capacity as needed. The probabilistic framework naturally handles uncertainty about factor relevance and improves robustness.

This IBP-based framework provides a principled, nonparametric approach to continual disentangled representation learning that automatically adapts to the complexity and factor requirements of each task while promoting knowledge transfer and preventing catastrophic forgetting.

A.3 Related Works

Continual learning, or lifelong learning, addresses the challenge of learning from sequential data streams without catastrophic forgetting of previously acquired knowledge [113, 114]. Recently, substantial efforts have focused on integrating disentangled representations into continual learning frameworks to leverage modularity and interpretability for improved knowledge retention and transfer. achille2018life introduced continual disentangled representation learning across multiple datasets, assuming each dataset forms a cluster characterized by a local set of generative factors, with some factors shared across clusters. Their method processes data sequentially by cluster and employs AnnealVAE (Burgess et al., 2018), which heuristically increases latent dimension capacity. Relevant latent factors for each cluster are identified by thresholding KL divergence values, providing a heuristic to determine cluster-specific latent usage. [116] improve upon [115] and model the latent representations with a topologically-connected mixture of distributions via Bayesian self-organizing maps (SOM) to find the relations between continually-arrived data based on their underlying semantic factors. The learned mixture automatically discovers the active semantic factors underlying each data environment, and accordingly accumulate their relational structure in the representations. They do so by encouraging

sparse coding to discover latent dimensions explaining active semantic factors in VAE by using a mixture of spike and slab variables (a set of bernoulli variables) along with the normal Gaussian variables in the latent space.

Bibliography

[1] Y. Bengio, A. Courville, and P. Vincent, *Representation learning: A review and new perspectives*, IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(8), 1798–1828 (2013).

[2] R. G. Millikan, *Beyond Concepts: Unicepts, Language, and Natural Information* (Oxford University Press, 2017).

[3] F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer, *On disentangled representations learned from correlated data*, in *International Conference on Machine Learning*, pp. 10401–10412 (PMLR, 2021).

[4] H. Li, X. Wang, Z. Zhang, H. Chen, Z. Zhang, and W. Zhu, *Disentangled graph self-supervised learning for out-of-distribution generalization*, in *Forty-first International Conference on Machine Learning* (2024).

[5] S. Van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem, *Are disentangled representations helpful for abstract visual reasoning?*, Advances in Neural Information Processing Systems **32** (2019).

[6] D. Cheng, Z. Xu, X. Jiang, N. Wang, D. Li, and X. Gao, *Disentangled prompt representation for domain generalization*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23595–23604 (2024).

[7] V.-H. Tran and C.-C. Huang, *Domain adaptation meets disentangled representation learning and style transfer*, in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 2998–3005 (IEEE, 2019).

[8] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, *Learning disentangled semantic representation for domain adaptation*, in *IJCAI: Proceedings of the Conference*, Vol. 2019, p. 2060 (2019).

[9] X. Zhu, C. Xu, and D. Tao, *Where and what? Examining interpretable disentangled representations*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5861–5870 (2021).

[10] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, *Toward causal representation learning*, Proceedings of the IEEE **109**(5), 612–634 (2021).

[11] T.-H. Wang, W. Xiao, T. Seyde, R. Hasani, and D. Rus, *Measuring interpretability of neural policies of robots with disentangled representation*, in *Proceedings of The 7th Conference on Robot Learning*, edited by J. Tan, M. Toussaint, and K. Darvish, Proceedings of Machine Learning Research, Vol. 229, pp. 602–641 (PMLR, November 6–9, 2023).

[12] Y. Zhou, R. Hong, Y. Guo, L. Liu, S. Hao, and H. Zhang, *Controllable relation disentanglement for few-shot class-incremental learning*, IEEE Transactions on Circuits and Systems for Video Technology (2025).

[13] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, *On the fairness of disentangled representations*, Advances in Neural Information Processing Systems **32** (2019).

[14] Y. Li, J. Hestness, M. Elhoseiny, L. Zhao, and K. Church, *Efficiently disentangle causal representations*, in *Conference on Parsimony and Learning*, pp. 54–71 (PMLR, 2024).

[15] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, *Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness*, in *International Conference on Machine Learning*, pp. 6056–6065 (PMLR, 2019).

[16] A. Bird and E. Tobin, *Natural Kinds*, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta and U. Nodelman, Spring 2024 ed. (Metaphysics Research Lab, Stanford University, 2024), `https://plato.stanford.edu/archives/spr2024/entries/natural-kinds/`.

[17] J. R. Saffran, R. N. Aslin, and E. L. Newport, *Statistical learning by 8-month-old infants*, Science **274**(5294), 1926–1928 (1996).

[18] S. Dehaene, F. Meyniel, C. Wacongne, L. Wang, and C. Pallier, *The neural representation of sequences: From transition probabilities to algebraic patterns and linguistic trees*, Neuron **88**(1), 2–19 (2015).

[19] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, *Building machines that learn and think like people*, Behavioral and Brain Sciences **40**, e253 (2017).

[20] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, *Towards a definition of disentangled representations*, arXiv:1812.02230 (2018).

[21] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, arXiv:1312.6114 (2013).

[22] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, *β-VAE: Learning basic visual concepts with a constrained variational framework*, in *International Conference on Learning Representations* (2017).

[23] H. Kim and A. Mnih, *Disentangling by factorising*, in *International Conference on Machine Learning*, pp. 2649–2658 (PMLR, 2018).

[24] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, *Isolating sources of disentanglement in variational autoencoders*, Advances in Neural Information Processing Systems **31** (2018).

[25] A. Hyvärinen, H. Sasaki, and R. E. Turner, *Nonlinear ICA using auxiliary variables and generalized contrastive learning*, in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 89, pp. 859–868 (2019).

[26] I. Khemakhem, D. Kingma, R. P. Monti, and A. Hyvärinen, *Variational autoencoders and nonlinear ICA: A unifying framework*, in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217 (2020), `https://proceedings.mlr.press/v119/khemakhem20a.html`.

[27] A. Hyvärinen, I. Khemakhem, and H. Morioka, *Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning*, Patterns **4**(10), 100765 (2023), doi: `https://doi.org/10.1016/j.patter.2023.100765`.

[28] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, *Challenging common assumptions in the unsupervised learning of disentangled representations*, in *International Conference on Machine Learning*, pp. 4114–4124 (PMLR, 2019).

[29] H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvärinen, *Disentangling identifiable features from noisy data with structured nonlinear ICA*, Advances in Neural Information Processing Systems **34**, 1624–1633 (2021).

[30] A. Kumar, P. Sattigeri, and A. Balakrishnan, *Variational inference of disentangled latent concepts from unlabeled observations*, arXiv:1711.00848 (2017).

[31] A. Van Den Oord, O. Vinyals, et al., *Neural discrete representation learning*, Advances in Neural Information Processing Systems **30** (2017).

[32] K. Hsu, W. Dorrell, J. Whittington, J. Wu, and C. Finn, *Disentanglement via latent quantization*, Advances in Neural Information Processing Systems **36** (2024).

[33] D. J. Rezende, S. Mohamed, and D. Wierstra, *Stochastic backpropagation and approximate inference in deep generative models*, in *International Conference on Machine Learning*, pp. 1278–1286 (PMLR, 2014).

[34] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, *Understanding disentangling in β-VAE*, arXiv:1804.03599 (2018).

[35] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, *InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets*, Advances in Neural Information Processing Systems **29** (2016).

[36] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh, *InfoGAN-CR and model centrality: Self-supervised model training and selection for disentangling GANs*, in *International Conference on Machine Learning*, pp. 6127–6139 (PMLR, 2020).

[37] U. Ojha, K. K. Singh, C.-J. Hsieh, and Y. J. Lee, *Elastic-InfoGAN: Unsupervised disentangled representation learning in class-imbalanced data*, Advances in Neural Information Processing Systems **33**, 18063–18075 (2020).

[38] K. Liao, C. Lin, Y. Zhao, and M. Gabbouj, *DR-GAN: Automatic radial distortion rectification using conditional GAN in real-time*, IEEE Transactions on Circuits and Systems for Video Technology **30**(3), 725–733 (2019).

[39] I. Jeon, W. Lee, M. Pyeon, and G. Kim, *IB-GAN: Disentangled representation learning with information bottleneck generative adversarial networks*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(9), pp. 7926–7934 (2021).

[40] J. Zhao, Z. Liu, X. Guo, and L. Pan, *CoDeGAN: Contrastive disentanglement for generative adversarial network*, arXiv:2103.03636 (2021).

[41] W. Lee, D. Kim, S. Hong, and H. Lee, *High-fidelity synthesis with disentangled representation*, in *European Conference on Computer Vision*, pp. 157–174 (Springer, 2020).

[42] W. Nie, T. Karras, A. Garg, S. Debhath, A. Patney, A. B. Patel, and A. Anandkumar, *Disentangled GANs for controllable generation of high-resolution images* (2020).

[43] Y. Shen, C. Yang, X. Tang, and B. Zhou, *InterfaceGAN: Interpreting the disentangled face representation learned by GANs*, IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(4), 2004–2018 (2020).

[44] Z. Wu, D. Lischinski, and E. Shechtman, *StyleSpace analysis: Disentangled controls for StyleGAN image generation*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12863–12872 (2021).

[45] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, *StyleCLIP: Text-driven manipulation of StyleGAN imagery*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094 (2021).

[46] Y. Gao, J.-X. Zhuang, S. Lin, H. Cheng, X. Sun, K. Li, and C. Shen, *DiSCO: Remedying self-supervised learning on lightweight models with distilled contrastive learning*, in *European Conference on Computer Vision*, pp. 237–253 (Springer, 2022).

[47] R. Gal, A. Bermano, H. Zhang, and D. Cohen-Or, *MRGAN: Multi-rooted 3D shape representation learning with unsupervised part disentanglement*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2039–2048 (2021).

[48] T. Yang, Y. Wang, Y. Lu, and N. Zheng, *DisDiff: Unsupervised disentanglement of diffusion*, in *Advances in Neural Information Processing Systems (NeurIPS) 37* (2023), `https://proceedings.neurips.cc/paper/2023/file/da47bfaf3f3a8d5bbab0d60c5195dc18-Paper-Conference.pdf`.

[49] Y. Wang, Y. Schiff, A. Gokaslan, W. Pan, F. Wang, C. De Sa, and V. Kuleshov, *InfoDiffusion: Representation learning using information maximizing diffusion models*, in *International Conference on Machine Learning*, pp. 36336–36354 (PMLR, 2023).

[50] B. M. G. Nielsen, A. Christensen, A. Dittadi, and O. Winther, *DiffEnc: Variational diffusion with a learned encoder*, arXiv:2310.19789 (2023).

[51] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, *Diffusion autoencoders: Toward a meaningful and decodable representation*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629 (2022).

[52] X. Jin, B. Li, B. Xie, W. Zhang, J. Liu, Z. Li, T. Yang, and W. Zeng, *Closed-loop unsupervised representation disentanglement with β-VAE distillation and diffusion probabilistic feedback*, in *European Conference on Computer Vision*, pp. 270–289 (Springer, 2024).

[53] Z. Lu, C. Wu, X. Chen, Y. Wang, L. Bai, Y. Qiao, and X. Liu, *Hierarchical diffusion autoencoders and disentangled image manipulation*, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5374–5383 (2024).

[54] A. Uppal, Y. Takida, C.-H. Lai, and Y. Mitsufuji, *Denoising multi-beta VAE: Representation learning for disentanglement and generation*, arXiv:2507.06613 (2025).

[55] S. Gupta, R. Gadde, R. Chen, and A. M. Martinez, *Disentanglement in T-space for faster and distributed training of diffusion models with fewer latent-states*, arXiv:2508.14413 (2025).

[56] Y. Kim, D. Kim, H. Lee, and I.-C. Moon, *Unsupervised controllable generation with score-based diffusion models: Disentangled latent code guidance*, in *Score-Based Models in Machine Learning Workshop, Advances in Neural Information Processing Systems (NeurIPS)* (2022), `https://openreview.net/forum?id=9X_AZydQfir`.

[57] G. Lozupone, A. Bria, F. Fontanella, F. J. A. Meijer, C. De Stefano, and H. Huisman, *Latent diffusion autoencoders: Toward efficient and meaningful unsupervised representation learning in medical imaging*, arXiv:2504.08635 (2025), https://arxiv.org/abs/2504.08635.

[58] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, *DDDM-VC: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion*, arXiv:2305.15816 (2023), https://arxiv.org/abs/2305.15816.

[59] Y. Wang et al., *Exploring behavior-relevant and disentangled neural dynamics with generative diffusion models*, arXiv:2410.09614 (2024), https://arxiv.org/abs/2410.09614.

[60] G. Kwon and J. C. Ye, *Diffusion-based image translation using disentangled style and content representation*, in *International Conference on Learning Representations (ICLR)* (2023), https://openreview.net/forum?id=Nayau9fwXU.

[61] H. Jia, Y. Li, H. Cui, D. Xu, C. Yang, Y. Wang, and T. Yu, *DisControlFace: Disentangled control for personalized facial image editing*, arXiv:2312.06193 (2023), https://arxiv.org/abs/2312.06193.

[62] T. Qi, S. Fang, Y. Wu, H. Xie, J. Liu, L. Chen, Q. He, and Y. Zhang, *DEADiff: An efficient stylization diffusion model with disentangled representations*, arXiv:2403.06951 (2024), https://arxiv.org/abs/2403.06951.

[63] K. Ridgeway and M. C. Mozer, *Learning deep disentangled embeddings with the F-statistic loss*, Advances in Neural Information Processing Systems **31** (2018).

[64] C. Eastwood and C. K. I. Williams, *A framework for the quantitative evaluation of disentangled representations*, in *6th International Conference on Learning Representations* (2018).

[65] K. Hsu, J. I. Hamid, K. Burns, C. Finn, and J. Wu, *Tripod: Three complementary inductive biases for disentangled representation learning*, arXiv:2404.10282 (2024).

[66] D. Bouchacourt, R. Tomioka, and S. Nowozin, *Multi-level variational autoencoder: Learning disentangled representations from grouped observations*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1) (2018).

[67] G. Baykal, M. Kandemir, and G. Unal, *Disentanglement with factor quantized variational autoencoders*, arXiv:2409.14851 (2024).

[68] F. Leeb, G. Lanzillotta, Y. Annadani, M. Besserve, S. Bauer, and B. Schölkopf, *Structure by architecture: Disentangled representations without regularization* (2020).

[69] H. Caselles-Dupré, M. Garcia Ortiz, and D. Filliat, *Symmetry-based disentangled representation learning requires interaction with environments*, Advances in Neural Information Processing Systems **32** (2019).

[70] X. Zhu, C. Xu, and D. Tao, *Commutative Lie group VAE for disentanglement learning*, in *International Conference on Machine Learning*, pp. 12924–12934 (PMLR, 2021).

[71] T. Yang, X. Ren, Y. Wang, W. Zeng, and N. Zheng, *Towards building a group-based unsupervised representation disentanglement framework*, arXiv:2102.10303 (2021).

[72] T. Yang, C. Lan, Y. Lu, and N. Zheng, *Diffusion model with cross attention as an inductive bias for disentanglement*, Advances in Neural Information Processing Systems **37**, 82465–82492 (2024).

[73] D. Horan, E. Richardson, and Y. Weiss, *When is unsupervised disentanglement possible?*, Advances in Neural Information Processing Systems **34**, 5150–5161 (2021).

[74] J. Kahana and Y. Hoshen, *A contrastive objective for learning disentangled representations*, in *European Conference on Computer Vision*, pp. 579–595 (Springer, 2022).

[75] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma, *VideoFlow: A flow-based generative model for video*, arXiv:1903.01434 **2**(5), 3 (2019).

[76] J. C. R. Whittington, W. Dorrell, S. Ganguli, and T. E. J. Behrens, *Disentanglement with biological constraints: A theory of functional cell types*, arXiv:2210.01768 (2022).

[77] W. Peebles, J. Peebles, J.-Y. Zhu, A. A. Efros, and A. Torralba, *The Hessian penalty: A weak prior for unsupervised disentanglement*, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 581–597 (2020), doi: https://doi.org/10.1007/978-3-030-58539-6_35.

[78] Y. Wei, Y. Shi, X. Liu, Z. Ji, Y. Gao, Z. Wu, and W. Zuo, *Orthogonal Jacobian regularization for unsupervised disentanglement in image generation*, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1171–1180 (2021), https://arxiv.org/abs/2108.07668.

[79] A. Gresele, D. Angles, I. Khemakhem, A. Hyvärinen, and A. Gresele, *On the identifiability of nonlinear ICA: Sparsity and beyond*, arXiv:2206.07751 (2021), https://arxiv.org/abs/2206.07751.

[80] P. Lopez, J. Regier, D. Blei, and M. Jordan, *Information constraints on auto-encoding variational Bayes*, arXiv:1804.03599 (2020), https://arxiv.org/abs/1804.03599.

[81] P. Rößler, D. Wagner, and P. Eisert, *Sparsity-enforcing variational autoencoders*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 921–922 (2020), https://openaccess.thecvf.com/content_CVPRW_2020/html/w10/Roessler_Sparsity-Enforcing_Variational_Autoencoders_CVPRW_2020_paper.html.

[82] F. Yang, T. Zhang, S. Xu, H. Yang, J. Zhou, and J. Huang, *CausalVAE: Disentangled representation learning via neural structural causal models*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12205–12214 (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

[83] Z. Ke, M. Zhu, Y. Zhang, M. Jian, and Q. Zhao, *DEAR: Bidirectional disentangled effect assignment for causal representation learning*, in *Advances in Neural Information Processing Systems (NeurIPS)* (2021), https://arxiv.org/abs/2106.01116.

[84] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, *The "wake-sleep" algorithm for unsupervised neural networks*, Science **268**(5214), 1158–1161 (1995).

[85] C. Burgess and H. Kim, *3D Shapes Dataset* (2018), https://github.com/deepmind/3dshapes-dataset/.

[86] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, *dSprites: Disentanglement Testing Sprites Dataset* (2017), https://github.com/deepmind/dsprites-dataset/.

[87] Z. Liu, P. Luo, X. Wang, and X. Tang, *Deep learning face attributes in the wild*, in *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015).

[88] J. Zaidi, J. Boilard, G. Gagnon, and M.-A. Carbonneau, *Measuring disentanglement: A review of metrics*, arXiv:2012.09276 (2021).

[89] M. W. Gondal, M. Wuthrich, D. Miladinovic, F. Locatello, M. Breidt, V. Volchkov, J. Akpo, O. Bachem, B. Schölkopf, and S. Bauer, *On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset*, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Vol. 32 (Curran Associates, Inc., 2019), https://proceedings.neurips.cc/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

[90] D. L. Donoho and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, Proceedings of the National Academy of Sciences **100**(10), 5591–5596 (2003), doi: https://doi.org/10.1073/pnas.1031596100.

[91] Y. Bengio and Y. LeCun, *Scaling learning algorithms towards AI*, in *Large Scale Kernel Machines* (MIT Press, 2007).

[92] G. E. Hinton, S. Osindero, and Y. W. Teh, *A fast learning algorithm for deep belief nets*, Neural Computation **18**, 1527–1554 (2006).

[93] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, Vol. 1 (MIT Press, 2016).

[94] T. S. Ferguson, *A Bayesian analysis of some nonparametric problems*, The Annals of Statistics, 209–230 (1973).

[95] C. E. Antoniak, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, The Annals of Statistics, 1152–1174 (1974).

[96] J. Sethuraman, *A constructive definition of Dirichlet priors*, Statistica Sinica, 639–650 (1994).

[97] D. M. Blei and M. I. Jordan, *Variational inference for Dirichlet process mixtures* (2006).

[98] K. Kurihara, M. Welling, and N. Vlassis, *Accelerated variational Dirichlet process mixtures*, Advances in Neural Information Processing Systems **19** (2006).

[99] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, *Stochastic variational inference*, Journal of Machine Learning Research **14**(1), 1303–1347 (2013).

[100] M. D. Hoffman and D. M. Blei, *Structured stochastic variational inference*, in *Artificial Intelligence and Statistics*, pp. 361–369 (2015).

[101] M. J. Wainwright, M. I. Jordan, et al., *Graphical models, exponential families, and variational inference*, Foundations and Trends in Machine Learning **1**(1–2), 1–305 (2008).

[102] L. Saul and M. Jordan, *Exploiting tractable substructures in intractable networks*, Advances in Neural Information Processing Systems **8** (1995).

[103] H. Ishwaran and L. F. James, *Gibbs sampling methods for stick-breaking priors*, Journal of the American Statistical Association **96**(453), 161–173 (2001).

[104] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta, *Composing graphical models with neural networks for structured representations and fast inference*, Advances in Neural Information Processing Systems **29** (2016).

[105] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, *Toward controlled generation of text*, in *International Conference on Machine Learning*, pp. 1587–1596 (PMLR, 2017).

[106] T. Yang, Y. Wang, C. Lan, Y. Lu, and N. Zheng, *Vector-based representation is the key: A study on disentanglement and compositional generalization*, arXiv:2305.18063 (2023), doi: `https://doi.org/10.48550/arXiv.2305.18063`.

[107] Q. Liang, D. Qian, L. Ziyin, and I. R. Fiete, *Compositional generalization via forced rendering of disentangled latents*, arXiv:2501.18797 (2025), ICML 2025.

[108] A. Yang et al., *A study on disentanglement and compositional generalization*, arXiv:2305.18063 (2023).

[109] H. Zheng and M. Lapata, *Disentangled sequence-to-sequence learning for compositional generalization*, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4256–4268 (Association for Computational Linguistics, 2022).

[110] A. Xu et al., *A study on disentanglement and emergent language*, in *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[111] A. Hwang et al., *Achieving combinatorial generalization by modeling a group action*, in *Proceedings of the 40th International Conference on Machine Learning (ICML)* (2023).

[112] T. Ito et al., *Compositional generalization through abstract representations in reinforcement learning*, in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[113] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro, *Learning to learn without forgetting by maximizing transfer and minimizing interference*, arXiv:1810.11910 (2018).

[114] R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, *Measuring catastrophic forgetting in neural networks* (AAAI Press, 2018).

[115] A. Achille, T. Eccles, L. Matthey, C. Burgess, N. Watters, A. Lerchner, and I. Higgins, *Life-long disentangled representation learning with cross-domain latent homologies*, Advances in Neural Information Processing Systems **31** (2018).

[116] Z. Li, X. Jiang, R. Missel, P. K. Gyawali, N. Kumar, and L. Wang, *Continual unsupervised disentangling of self-organizing representations*, in *The Eleventh International Conference on Learning Representations*.

[117] D. Rao, F. Visin, A. Rusu, R. Pascanu, Y. W. Teh, and R. Hadsell, *Continual unsupervised representation learning*, Advances in Neural Information Processing Systems **32** (2019).

[118] J. Ramapuram, M. Gregorova, and A. Kalousis, *Lifelong generative modeling*, Neurocomputing **404**, 381–400 (2020).

[119] D. Madaan, J. Yoon, Y. Li, Y. Liu, and S. J. Hwang, *Representational continuity for unsupervised continual learning*, in *International Conference on Learning Representations* (2022).

[120] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola, *A kernel statistical test of independence*, Advances in Neural Information Processing Systems **20** (2007).

[121] L. B. Almeida, *MISEP—linear and nonlinear ICA based on mutual information*, Journal of Machine Learning Research **4**(Dec), 1297–1318 (2003).

[122] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, Advances in Neural Information Processing Systems **33**, 6840–6851 (2020).

[123] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al., *Conditional image generation with PixelCNN decoders*, Advances in Neural Information Processing Systems **29** (2016).

[124] H. Larochelle and I. Murray, *The neural autoregressive distribution estimator*, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 29–37 (JMLR Workshop and Conference Proceedings, 2011).

[125] L. Dinh, J. Sohl-Dickstein, and S. Bengio, *Density estimation using Real NVP*, arXiv:1605.08803 (2016).

[126] A. Vahdat and J. Kautz, *NVAE: A deep hierarchical variational autoencoder*, Advances in Neural Information Processing Systems **33**, 19667–19679 (2020).

[127] R. Salakhutdinov, A. Mnih, and G. Hinton, *Restricted Boltzmann machines for collaborative filtering*, in *Proceedings of the 24th International Conference on Machine Learning*, pp. 791–798 (2007).

[128] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, *Zero-shot text-to-image generation*, in *International Conference on Machine Learning*, pp. 8821–8831 (PMLR, 2021).

[129] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen, et al., *Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs*, arXiv:2503.01743 (2025).

[130] O. Ahia, M. Bartelds, K. Ahuja, H. Gonen, V. Hofmann, S. Arora, S. S. Li, V. Puttagunta, M. Adeyemi, C. Buchireddy, et al., *BLAB: Brutally long audio bench*, arXiv:2505.03054 (2025).

[131] L. Chen, C. Zhu, D. Soselia, J. Chen, T. Zhou, T. Goldstein, H. Huang, M. Shoeybi, and B. Catanzaro, *Odin: Disentangled reward mitigates hacking in RLHF*, arXiv:2402.07319 (2024).

[132] Y. Chen, X. Yue, X. Gao, C. Zhang, L. F. D'Haro, R. T. Tan, and H. Li, *Beyond single-audio: Advancing multi-audio processing in audio large language models*, in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10917–10930 (2024).

[133] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, et al., *Qwen2-Audio technical report*, arXiv:2407.10759 (2024).

[134] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, *Pengi: An audio language model for audio tasks*, Advances in Neural Information Processing Systems **36**, 18090–18108 (2023).

[135] X. Diao, C. Zhang, K. Kong, W. Wu, C. Ma, Z. Ouyang, P. Qing, S. Vosoughi, and J. Gui, *SoundMind: RL-incentivized logic reasoning for audio-language models*, in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 528–540 (2025).

[136] J. Donley, V. Tourbabin, J.-S. Lee, M. Broyles, H. Jiang, J. Shen, M. Pantic, V. K. Ithapu, and R. Mehra, *EasyCom: An augmented reality dataset to support algorithms for easy communication in noisy environments*, arXiv:2107.04174 (2021).

[137] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, *CLAP: Learning audio concepts from natural language supervision*, in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5 (IEEE, 2023).

[138] B. Elizalde, S. Deshmukh, and H. Wang, *Natural language supervision for general-purpose audio representations*, in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 336–340 (IEEE, 2024).

[139] Y. Gao, B. Wang, C. Wei, S. Sun, and A. Aw, *IFEval-Audio: Benchmarking instruction-following capability in audio-based large language models*, in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 120–127 (2025).

[140] S. Ghosh, C. K. R. Evuru, S. Kumar, U. Tyagi, O. Nieto, Z. Jin, and D. Manocha, *Visual description grounding reduces hallucinations and boosts reasoning in LVLMs*, arXiv:2405.15683 (2024).

[141] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, *Audio Flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities*, arXiv:2503.03983 (2025).

[142] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, *GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities*, in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313 (2024).

[143] Inclusion AI, B. Gong, C. Zou, C. Zheng, C. Zhou, C. Yan, C. Jin, C. Shen, D. Zheng, F. Wang, et al., *Ming-Omni: A unified multimodal model for perception and generation*, arXiv:2506.09344 (2025).

[144] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, *Listen, think, and understand*, in *The Twelfth International Conference on Learning Representations* (2023).

[145] C.-Y. Huang, W.-C. Chen, S.-W. Yang, A. T. Liu, C.-A. Li, Y.-X. Lin, W.-C. Tseng, A. Diwan, Y.-J. Shih, J. Shi, et al., *Dynamic-SUPERB Phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks*, arXiv:2411.05361 (2024).

[146] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, *Mantis: Interleaved multi-image instruction tuning*, arXiv:2405.01483 (2024).

[147] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang, et al., *Kimi-Audio technical report*, arXiv:2504.18425 (2025).

[148] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, *NV-Embed: Improved techniques for training LLMs as generalist embedding models*, arXiv:2405.17428 (2024).

[149] S. Leng, Y. Xing, Z. Cheng, Y. Zhou, H. Zhang, X. Li, D. Zhao, S. Lu, C. Miao, and L. Bing, *The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio*, arXiv:2410.12787 (2024).

[150] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al., *LLaVA-OneVision: Easy visual task transfer*, arXiv:2408.03326 (2024).

[151] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, *Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering*, arXiv:2503.11197 (2025).

[152] M. Li, C.-T. Do, S. Keizer, Y. Farag, S. Stoyanchev, and R. Doddipatla, *WHISMA: A speech-LLM to perform zero-shot spoken language understanding*, in *2024 IEEE Spoken Language Technology Workshop*, pp. 1115–1122 (IEEE, 2024).

[153] Y. Li, J. Liu, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, L. Ming, G. Dong, D. Pan, et al., *Baichuan-Omni-1.5 technical report*, arXiv:2501.15368 (2025).

[154] Y. Li, H. Sun, M. Lin, T. Li, G. Dong, T. Zhang, B. Ding, W. Song, Z. Cheng, Y. Huo, et al., *Baichuan-Omni technical report*, arXiv:2410.08565 (2024).

[155] K.-H. Lu, C.-Y. Kuan, and H.-Y. Lee, *Speech-IFEval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models*, arXiv:2505.19037 (2025).

[156] Z. Ma, Y. Ma, Y. Zhu, C. Yang, Y.-W. Chao, R. Xu, W. Chen, Y. Chen, Z. Chen, J. Cong, et al., *MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix*, arXiv:2505.13032 (2025).

[157] G. de Souza P. Moreira, R. Osmulski, M. Xu, R. Ak, B. Schifferer, and E. Oldridge, *NV-Retriever: Improving text embedding models with effective hard-negative mining*, arXiv:2407.15831 (2024).

[158] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., *GPT-4o system card*, arXiv:2410.21276 (2024).

[159] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., *Training language models to follow instructions with human feedback*, Advances in Neural Information Processing Systems **35**, 27730–27744 (2022).

[160] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li, J. Shi, S. Arora, W. Chen, R. Sharma, et al., *Reproducing Whisper-style training using an open-source toolkit and publicly available data*, in *2023 IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–8 (IEEE, 2023).

[161] S. B. H. Pias, R. Huang, D. S. Williamson, M. Kim, and A. Kapadia, *The impact of perceived tone, age, and gender on voice assistant persuasiveness in the context of product recommendations*, in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pp. 1–15 (2024).

[162] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, *Robust speech recognition via large-scale weak supervision*, in *International Conference on Machine Learning*, pp. 28492–28518 (PMLR, 2023).

[163] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, et al., *AudioPaLM: A large language model that can speak and listen*, arXiv:2306.12925 (2023).

[164] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, *MMAU: A massive multi-task audio understanding and reasoning benchmark*, arXiv:2410.19168 (2024).

[165] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, *Voice in human–agent interaction: A survey*, ACM Computing Surveys **54**(4), 1–43 (2021).

[166] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. Chen, *AudioBench: A universal benchmark for audio large language models*, in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316 (2025).

[167] D. Wang, J. Li, J. Wu, D. Yang, X. Chen, T. Zhang, and H. Meng, *MMSU: A massive multi-task spoken language understanding and reasoning benchmark*, arXiv:2506.04779 (2025).

[168] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, L. Kong, Q. Liu, T. Liu, et al., *Large language models are not fair evaluators*, in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450 (2024).

[169] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, *MuChoMusic: Evaluating music understanding in multimodal audio-language models*, arXiv:2408.01337 (2024).

[170] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, *Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation*, in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5 (IEEE, 2023).

[171] Z. Xie, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, *Audio-Reasoner: Improving reasoning capability in large audio language models*, arXiv:2503.02318 (2025).

[172] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al., *Qwen3-Omni technical report*, arXiv:2509.17765 (2025).

[173] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., *Qwen3 technical report*, arXiv:2505.09388 (2025).

[174] H. Zhao, Z. Cai, S. Si, X. Ma, K. An, L. Chen, Z. Liu, S. Wang, W. Han, and B. Chang, *MMICL: Empowering vision-language model with multi-modal in-context learning*, arXiv:2309.07915 (2023).

[175] Z. Zheng, P. Peng, Z. Ma, X. Chen, E. Choi, and D. Harwath, *BAT: Learning to reason about spatial sounds with large language models*, arXiv:2402.01591 (2024).

[176] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al., *LIMA: Less is more for alignment*, Advances in Neural Information Processing Systems **36**, 55006–55021 (2023).

[177] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, *Instruction-following evaluation for large language models*, arXiv:2311.07911 (2023).

[178] S. Mandt, M. D. Hoffman, and D. M. Blei, *Stochastic gradient descent as approximate Bayesian inference*, Journal of Machine Learning Research **18**(134), 1–35 (2017).

[179] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, arXiv:1412.6980 (2014).