A Bayesian Nonparametric Framework for Learning Disentangled Representations

Anonymous authors
Paper under double-blind review

Abstract

Disentangled representation learning aims to identify and organize the underlying sources of variation in observed data. However, learning disentangled representations 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 from observational data alone. Existing methods address this by imposing heuristic inductive biases that typically lack theoretical identifiability guarantees. They also 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 enables inference of sufficient latent capacity to represent the underlying variations without violating these 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 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 latent variable is sensitive to changes in exactly one underlying factor while being relatively invariant to changes in others Bengio (2013). Disentangled representations have been shown to improve robustness and out-of-distribution generalization (Träuble et al., 2021; Li et al., 2024), sample efficiency in few-shot learning (Van Steenkiste et al., 2019; Cheng et al., 2024), domain adaptation via separation of transferable and domain-specific features (Tran & Huang, 2019; Cai et al., 2019), controllable and interpretable generation (Zhu et al., 2021; Wang et al., 2023; Zhou et al., 2025), and causal inference and fairness through explicit separation of sensitive and task-relevant factors (Cheng et al., 2024; Locatello et al., 2019a).

However, unsupervised learning of disentangled representations is fundamentally challenged with 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 (Hyvärinen & Pajunen, 1999; Hyvärinen et al., 2019; Khemakhem et al., 2020), deep generative modeling (Wang et al., 2021; D'Amour et al., 2022), and unsupervised disentanglement (Locatello et al., 2019b) 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 mis-specification 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 piece-wise 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. 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). Critically, the factorized prior structure facilitates the orthogonal encoding of factor-specific variations. 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.

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

2 Nonparametric Bayesian Quantization for Autoencoders

Prior work (Hsu et al., 2024a;b) introduce inductive biases that encourage disentangled representation learning by structuring the latent space as a factorized Cartesian product of discrete sets, where each latent dimension is independently quantized through separate learnable codebooks. This latent quantization architecture restricts the encoder to constructing representations through combinatorial selection from small finite codebooks of scalar embeddings. This, consequently, constrains the decoder to assign consistent semantic meanings to the embeddings, such that each codebook encodes a single factor of variation with the embeddings representing specific variations within the factor. For this architectural design to serve as an effective inductive bias for learning disentangled representations, the size of each codebook C_i, and thus the support of the corresponding discrete latent variable z_i, used to index the codebook embeddings, is fixed and kept small. While this design choice encourages parsimonious representations, factors with variations larger than the size of a single codebook must necessarily be distributed across multiple codebooks, reducing the interpretability of the learned factors.

To address this limitation and ensure that each generative factor, with potentially unbounded number of variations, can be encoded in a single codebook, we propose a principled probabilistic formulation in which each codebook possesses theoretically infinite number of embeddings. Specifically, we place a nonparametric Dirichlet Process (DP) (Ferguson, 1973; Sethuraman, 1994) over each discrete codebook and use the stick-breaking construction to define a valid probability mass function with countably infinite support for the discrete latent variables z. To enable principled uncertainty quantification within this nonparametric framework, each scalar embedding is instead a stochastic variable governed by a probability distribution rather than a fixed point estimate. To realize this, we use the base distribution of the DP to generate the countably infinite set of parameters that define the distributions from which these stochastic embeddings are sampled. This formulation naturally induces a Dirichlet Process Mixture Model (DPMM) prior (A.1.1) over the embedding space, where each codebook's embeddings are modeled as samples drawn from an infinite-component mixture distribution.

To preserve the inductive biases, that makes latent quantization effective, within our nonparametric framework, we propose nested variational family-based inference for posterior approximation. During inference with this nested family, each codebook is initialized with a single component or embedding parameter. When the model encounters data requiring greater representational capacity, new components are greedily added to the codebook. This greedy expansion allows the model to gradually adapt its capacity to the complexity of the generative factor represented by the codebook, thereby providing a stronger inductive bias.

In the following sections, we first formalize the hierarchical Bayesian nonparametric prior governing the embedding space and derive the corresponding generative model. We next formalize the structured variational family, specifically designed to accurately approximate the posterior distribution with hierarchical structured priors. Finally, we present the nested extension of this structured variational family which enables principled incremental expansion of the representational capacity while preserving inductive biases of latent quantization.

2.1 Nonparametric Prior

We adopt the inductive bias of latent quantization (For a prior on vector-quantized and latent-quantized autoencoder please refer to Section A.1.2) by decomposing each datapoint's d-dimensional encoder output vector into component scalars, where each scalar is independently quantized using a separate codebook C_i. The discrete latent variable z is defined as an element of the Cartesian product of component discrete sets $\mathbf{z} \in Z_1 \times \cdots \times Z_d$ where each discrete variable $z_i \in Z_i = \{1, \ldots, \mid C_i \mid\}, \forall i \in \{1, \ldots, d\}$ indexes the embeddings of codebook C_i. Formally, we define a nonparametric prior over the parameter space Θ of the mixture components for each codebook C_i using the Dirichlet Process $\mathrm{DP}(\alpha, G_0)$. For each codebook C_i, we use the stick-breaking construction to generate an infinite sequence of stick-breaking proportions $\boldsymbol{\beta}_i = \{\beta_{i,k}\}_{k=1}^{\infty}$, with each $\beta_{i,k}$ is independently drawn from a Beta distribution with concentration parameter α controlling the expected number of active mixture components. Concurrently, the embeddings parameters $\boldsymbol{\theta}_i = \{\theta_{i,k}\}_{k=1}^{\infty}$ are independently sampled from a continuous base distribution $G_0(\lambda)$, defined over the parameter space Θ:

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

The stick-breaking proportions $\boldsymbol{\beta}_i$ are then used to define the countably infinite set of mixture weights that specify a valid probability mass function over the discrete latent variables z_i, replacing the fixed uniform prior of VQVAE and its variants. Conditional on the discrete variable $z_i = k$, the corresponding embedding e_i is sampled from the k-th mixture component distribution $p(e \mid \theta_{i,k})$. This generative process for the discrete latent variable z_i and the corresponding embedding e_i associated with codebook C_i is formalized as follows:

$$z_i = k \mid \boldsymbol{\beta}_i \sim p(z_i = k \mid \beta_i) = \beta_{i,k} \prod_{j=1}^{k-1} (1 - \beta_{i,j}), \quad e_i \sim p(e \mid z_i, \boldsymbol{\theta}_i) = \prod_{k=1}^{\infty} \left(p(e \mid \theta_{i,k})\right)^{\mathbf{1}[z_i = k]}$$

We choose the Gaussian distribution with unknown mean and precision parameters $\theta_k = \{\mu_k, s_k\}$, where μ_k denotes the mean and s_k the precision (inverse variance), to sample the embedding vectors. Following conjugacy structure, we choose the base distribution G_0 for sampling these parameters θ to be a Normal–Gamma distribution $G_0(\lambda) = \mathcal{NG}(m_0, \kappa_0, \nu_0, w_0)$ which serves as a conjugate prior to the Gaussian likelihood with unknown mean and precision. Importantly, the Normal–Gamma prior simultaneously captures uncertainty over both the mean and precision parameters, facilitating efficient joint sampling, and simplifies Bayesian inference by enabling closed-form posterior updates. We define the data-generating distribution as $p_{\theta_g}(x \mid e) \sim \mathcal{N}\big(g_{\theta_g}(e), \sigma^2\mathbf{I}\big)$ where $g_{\theta_g} : \mathcal{E} \to \mathcal{X}$ is a nonlinear mapping parameterized by θ_g that transforms the embedding vectors $\mathbf{e} \in \mathcal{E}$ into the observation space $\mathcal{X}$. The joint distribution over the observed data x, latent embeddings $\boldsymbol{e}$, discrete latent variables $\boldsymbol{z}$, stick-breaking proportions $\boldsymbol{\beta}$ and embedding distribution parameters $\boldsymbol{\theta}$ factorizes according to the hierarchical generative model as follows:

$$p\left(x, \boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \alpha, \lambda\right) = p_{\theta_g}\left(x \mid \boldsymbol{e}\right) \prod_{i=1}^{d} p\left(e_i \mid z_i, \boldsymbol{\theta}_i\right) p\left(z_i \mid \boldsymbol{\beta}_i\right) \prod_{k=1}^{\infty} p\left(\beta_{i,k} \mid \alpha\right) p\left(\theta_{i,k} \mid \lambda\right) \quad (1)$$

This hierarchical structure induces a natural partitioning of the embedding space into clusters corresponding to the mixture components and provides a principled probabilistic framework to model the underlying discrete latent structure.

2.2 Variational Inference

To enable a computationally efficient posterior approximation for nonparametric priors (for a prior on variational inference for DPMMs please refer to the Preliminary section A.1), Blei & Jordan (2006); Hoffman et al. (2013) approximate the infinite-dimensional stick-breaking process using a truncated stick-breaking variational family. This formulation introduces an explicit truncation level T by fixing the stick-breaking proportion at position T to one $q_{\nu_\beta}(\beta_T = 1) = 1$, which implicitly forces all subsequent stick lengths $\{\beta_k\}_{k>T}$ and the corresponding mixture weights to zero, thus limiting the mixture components to T. Further, this approach renders inference tractable with the use of fully factorized variational distributions; which impose strong independence constraints among latent variables, including those with hierarchical dependencies. The approach of Hoffman & Blei (2015) relaxes this mean-field assumption to allow dependencies between hierarchical latent variables, yielding more accurate posterior approximations and reducing bias; while lowering sensitivity to initialization and hyperparameters. We adapt this structured variational inference framework to preserve the hierarchical dependencies in our formulation, specifically between the stick-breaking proportions $\boldsymbol{\beta}$ and the discrete latent variables z as well as between the discrete variable z, the components parameters $\boldsymbol{\theta}$ and the embeddings e as detailed below:

$$q_\nu(e, z, \boldsymbol{\beta}, \boldsymbol{\theta}) = q(e \mid z, \boldsymbol{\theta}) q(z \mid \boldsymbol{\beta}) \prod_{k=1}^{T-1} q_{\nu_{\beta,k}}(\beta_k) \prod_{k=1}^{T} q_{\nu_{\theta,k}}(\theta_k) \quad (2)$$

Truncated variational families make inference in infinite-mixture models tractable by restricting the variational distribution to a fixed number T of mixture components. The intuition that increasing the truncation level monotonically increases the ELBO, with more mixture components allowing q to better approximate the true nonparametric posterior, motivates practitioners to choose high truncation levels. However, this intuition fails when an optimal truncation level exists, particularly for data generated by a finite mixture. Moreover, the truncated variational families are not nested; the variational family with truncation level T is not a subset of the family with truncation level $T+1$. Therefore, increasing T beyond the optimal level does not necessarily yield a better approximation and undermines the inductive bias of quantizing with a small set of latent embeddings, critical for learning disentangled representations.

Therefore, we employ the nested variational family framework of Kurihara et al. (2006), which defines an infinite-dimensional variational parameter space $\{\nu_{\beta,k}, \nu_{\theta,k}\}_{k=1}^{\infty}$, to support an unbounded number of mixture components using parameter tying.

$$q_\nu(\boldsymbol{\beta}) = \prod_{k=1}^{T} q_{\nu_{\beta,k}}(\beta_k) \prod_{k=T}^{\infty} p(\beta_k \mid \alpha), \qquad q_\nu(\boldsymbol{\theta}) = \prod_{k=1}^{T} q_{\nu_{\theta,k}}(\theta_k) \prod_{k=T}^{\infty} p(\theta_k \mid \lambda) \quad (3)$$

When data are associated only with the first T components, the variational parameters for all components $k > T$ are tied to their corresponding prior values, such that $q_{\nu_{\beta,k}}(\beta_k) = p(\beta \mid \alpha)$ and

$q_{\nu_{\theta,k}}(\theta_k) = p(\theta \mid \lambda)$. This parameter tying effectively constrains the variational distributions beyond the implicit truncation level T to the prior distribution. This constraint ensures that although the variational distribution theoretically includes an unbounded number of mixture components, we only need to represent and optimize parameters up to the implicit truncation level T. Crucially, data belonging to components beyond the truncation level T, can be assigned to the infinite components, with their parameters tied to the prior. This enables our inference algorithm to proceed greedily, starting with $T = 1$ and incrementally adding components only when they yield a significant improvement in the empirical ELBO. This process of incrementally adding components is continued until all data is assigned to components within truncation level T and no data is assigned to the prior. Notably, under this nested formulation, codebook components that fail to encode meaningful variations collapse back to their prior distribution during training.

For a single data point, the ELBO, estimated using Monte Carlo samples, under this hierarchical, structured, nested variational family q_ν, given the generative model in Equation equation 1, can be expressed as:

$$\mathcal{L} = \frac{1}{N}\mathbb{E}_{q_{\nu_\beta}}\left[\log\frac{p(\boldsymbol{\beta} \mid \alpha)}{q_{\nu_\beta}(\boldsymbol{\beta})}\right] + \frac{1}{N}\mathbb{E}_{q_{\nu_\theta}}\left[\log\frac{p(\boldsymbol{\theta} \mid \lambda)}{q_{\nu_\theta}(\boldsymbol{\theta})}\right] + \mathbb{E}_{q_\nu}\left[\log\frac{p_{\theta_g}(x \mid e)\, p(e \mid z, \boldsymbol{\theta})\, p(z \mid \boldsymbol{\beta})}{q(e \mid z, \boldsymbol{\theta})\, q(z \mid \boldsymbol{\beta})}\right]$$

2.3 The algorithm

Efficient inference algorithms for hierarchical models rely on the use of conjugate exponential family data likelihoods, which preserve tractable structure. Specifically, hierarchical models where the latent variables follow distributions from the exponential family and the generative model is conjugate to the prior, the resulting conditional posterior distributions remain within the same exponential family as the prior, thereby facilitating such efficient inference. However, for general neural network observation likelihoods, such as p_{θ_g}, the absence of such conjugacy structure significantly increases the computational complexity of inferring the latent variables requiring multiple passes through the generative model. To ensure computational tractability while using general non-conjugate observation likelihoods with structured latent variable priors, we use deep amortized recognition networks $h(x;\phi)$ of Johnson et al. (2016). For each datapoint, these networks output local conjugate likelihood potentials $\hat{p}_\phi$, as defined in Equation 4 unlike standard variational autoencoder encoders that directly output variational distribution parameters. These conjugate potentials replace the original non-conjugate observation likelihoods during inference, and are combined with the structured latent variable prior using efficient message-passing algorithms, thereby preserving the tractability of conjugate graphical model inference.

$$\hat{p}_\phi(\boldsymbol{e} \mid x) = \prod_{i=1}^{d} \hat{p}_\phi(e_i \mid x) = \prod_{i=1}^{d} \exp\{\langle h_i(x;\phi), t_e(e_i)\rangle\} \tag{4}$$

This independence structure of the recognition network, enables the inference of local latent variables $\{e_i, z_i\}$ associated with each codebook C_i independently of the other codebooks. With this structural constraint, the data-likelihood third term of the ELBO (2.2) with local latent variables decomposes into a sum over individual dimensions:

$$\mathcal{L}_i = \mathbb{E}_{q_{\nu_\beta}(\boldsymbol{\beta})q(z_i|\boldsymbol{\beta})}\left[\log\frac{p(z_i \mid \boldsymbol{\beta})}{q(z_i \mid \boldsymbol{\beta})} + \mathbb{E}_{q_{\nu_\theta}(\boldsymbol{\theta})q(e_i|z_i,\boldsymbol{\theta})}\left[\log\frac{p(e_i \mid z_i, \boldsymbol{\theta})}{q(e_i \mid z_i, \boldsymbol{\theta})} + \log\hat{p}_\phi(e_i \mid x)\right]\right] \tag{5}$$

Similar to the work of Hoffman & Blei (2015), we observe that, with the global latent variables $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and the local latent variable z_i held fixed, the second term—dependent on the latent variable e_i—can be expressed as a variational lower bound on the conditional marginal likelihood associated with the conjugate potential of that dimension.

$$\begin{aligned}&\mathbb{E}_{q(e_i|z_i,\theta)}\left[\log\frac{p(e_i \mid z_i, \boldsymbol{\theta})}{q(e_i \mid z_i, \boldsymbol{\theta})} + \hat{p}_\phi(e_i \mid x)\right]\\ &\quad = -D_{\mathrm{KL}}\left(q(e_i \mid z_i, \boldsymbol{\theta}) \| \hat{p}_\phi(e_i \mid x_i, z_i, \boldsymbol{\theta})\right) + \log\hat{p}_\phi(x_i, \mid z_i, \boldsymbol{\theta}) \leq \log\hat{p}_\phi(x_i, \mid z_i, \boldsymbol{\theta})\end{aligned} \tag{6}$$

with the local posterior distribution of e_i conditioned on the data and the latent variables, defined as

$$\hat{p}_\phi(e_i \mid x_i, z_i, \boldsymbol{\theta}) = \frac{\hat{p}_\phi(e_i \mid x_i)p(e_i \mid z_i, \boldsymbol{\theta})}{\hat{p}_\phi(x_i \mid z_i, \boldsymbol{\theta})} \tag{7}$$

Here, the marginal likelihood of the conjugate potential defined as $\hat{p}_\phi(x_i \mid z_i, \boldsymbol{\theta}) = \int \hat{p}_\phi(e_i \mid x_i)\, p(e_i \mid z_i, \boldsymbol{\theta}) de_i$ to ensure that the posterior $\hat{p}_\phi(e_i \mid x_i, z_i, \boldsymbol{\theta})$ is a valid probability distribution. Since the Kullback-Leibler (KL) divergence is non-negative, choosing the variational distribution $q(e_i \mid z_i, \boldsymbol{\theta})$ to be exactly equal to the local posterior $\hat{p}_\phi(e_i \mid x_i, z_i, \boldsymbol{\theta})$ minimizes the KL divergence to zero and yields the tightest possible lower bound on the local ELBO.

$$\begin{aligned} q(e_i \mid z_i, \boldsymbol{\theta}) &= \hat{p}_\phi(e_i \mid x_i, z_i, \boldsymbol{\theta}) \\ &= \exp\left\{\langle \eta_e(z_i, \eta_\theta(\boldsymbol{\theta}), \phi), t_e(e)\rangle - A_e(\eta_e(z_i, \eta_\theta(\boldsymbol{\theta}), \phi))\right\} \end{aligned} \tag{8}$$

where the natural parameters $\eta_e(z_i, \eta_\theta(\boldsymbol{\theta}), \phi)$, defined as

$$\eta_e(\eta_\theta(\boldsymbol{\theta}), \phi) = \sum_{k=1}^{T} \mathbf{1}_{[z_i=k]}\, \eta_\theta(\theta_k) + h_i(x_i; \phi)$$

This formulation explicitly expresses the variational distribution $q(e_i \mid z_i, \boldsymbol{\theta})$ as an exponential family distribution resulting from combining the conjugate observation likelihood with the structured latent prior distribution. It is worth noting that for effective partitioning of the data through quantization, instead of propagating embeddings to the decoder which are sampled from the variational distribution $q(e_i \mid z_i, \boldsymbol{\theta})$, we propagate embeddings sampled from their prior distribution $p(e_i \mid z_i, \boldsymbol{\theta})$. This forces the representations to cluster around the prior and encoding variations common to all data belonging to the same cluster.

It is worth noting that for effective data partitioning through representation quantization, we propagate embeddings sampled from the prior distribution $p(e_i \mid z_i, \boldsymbol{\theta})$ to the decoder, rather than from the variational distribution $q(e_i \mid z_i, \boldsymbol{\theta})$. This approach forces the representations to encode variations common to all data belonging to the same cluster, rather than datapoint-specific information. With this choice of variational distribution for e_i, it is crucial to note that the local ELBO can be further expressed as a variational lower bound on the marginal likelihood of the data, conditioned on the global variable $\boldsymbol{\beta}$ with respect to the latent variable z_i.

$$\begin{aligned} \mathcal{L}_i &= \mathbb{E}_{q_{\nu_\beta}(\boldsymbol{\beta}) q(z_i \mid \boldsymbol{\beta})}\left[\log \frac{p(z_i \mid \boldsymbol{\beta})}{q(z_i \mid \boldsymbol{\beta})} + \mathbb{E}_{q_{\nu_\theta}(\boldsymbol{\theta})}\left[\log \hat{p}_\phi(x \mid z_i, \boldsymbol{\theta})\right]\right] \\ &= -D_{\mathrm{KL}}(q(z_i \mid \boldsymbol{\beta}) \| \hat{p}_\phi(z_i \mid x_i, \boldsymbol{\beta})) + \log \hat{p}_\phi(x_i \mid \boldsymbol{\beta}) \le \log \hat{p}_\phi(x_i \mid \boldsymbol{\beta}) \end{aligned} \tag{9}$$

Therefore, similar to the variational distribution of e_i, the variational distribution of z_i can be set to the local optimal value given by

$$\begin{aligned} \log q(z_i = k) &\propto E_{q_{\nu_\theta}(\boldsymbol{\theta})} \log \hat{p}_\phi(x_i \mid z_i = k, \boldsymbol{\theta}) \\ &= A_e\left(\eta_\theta(\theta_k) + h_i(x_i; \phi)\right) - A_e\left(\eta_\theta(\theta_k)\right) \end{aligned} \tag{10}$$

The use of structured variational inference with deep amortized recognition networks enables the variational distributions over local latent variables to be set to their locally optimal values, thereby ensuring tractable and efficient inference. At each iteration, we first sample the global parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ from their variational distributions $q_\nu(\boldsymbol{\beta})$ and $q_\nu(\boldsymbol{\theta})$ respectively and use the samples to compute the local variational distributions.

However, in the context of structured variational inference, as noted by Hoffman & Blei (2015) exact inference over the global variables is generally intractable due to the restored dependencies between global and local variables. Moreover, the use of recognition networks in place of the non-conjugate likelihood generative functions necessitates gradient-based estimation of global parameters. As a result, we use low-variance Monte Carlo estimators to approximate the required expectations and efficiently implement this using the reparameterization trick, which enables gradient-based optimization of the variational parameters. For approximate posterior inference of the global variables, we use stochastic gradient-based optimization methods equipped with adaptive preconditioning matrices, such as RMSProp (Graves, 2013) and Adam Adam et al. (2014). These optimizers facilitate efficient updates by scaling the gradients according to the geometry of the parameter space. To further enhance convergence and stability, we select the step-size following the recommendations of Mandt et al. (2017), which provide principled guidance for optimal learning rates for posterior inference.

3 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 (Burgess & Kim, 2018) 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 (Gondal et al., 2019), collected from a real-world robotic environment, which exhibits a power-law distribution across the number of variations of different factors. Specifically, a few factors contain extensive variations (e.g., 40 discrete values for each rotational degrees of freedom), while the majority possess substantially fewer variations (e.g., 2-6 values for object properties).

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 Higgins et al. (2017) and β-TCVAE Chen et al. (2018) which enforces disentanglement through information-theoretic regularization encouraging independence across latent dimensions. We further consider BioAE Whittington et al. (2022), 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 Hsu et al. (2024a), which introduces an architectural bias based on latent quantization, and subsequently Tripod Hsu et al. (2024a), 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. For a concise prior on the disentanglement metrics and the different properties measured please refer to Section A.3.

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 and Table 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 A.2. [1]

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

Table 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 Locatello et al. (2019b) and Hsu et al. (2024a;b) respectively.

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

Table 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.

4 Related Works

The challenge of separating mutually independent sources in data traces back to the classical statistical problem of Independent Component Analysis (ICA) Comon (1994); Hyvärinen & Oja (2000). This core problem was later reinterpreted in the context of modern machine learning as disentanglement, formally articulated by Bengio Bengio (2013) and formalized by Higgins et al. (2018). When the data generating process is governed by nonlinear transformations Hyvärinen & Pajunen (1999), the task of learning disentangled representations becomes theoretically unidentifiable Hyvärinen & Oja (2000); Khemakhem et al. (2020); Locatello et al. (2019b). Consequently, the incorporation of auxiliary data Hyvärinen & Pajunen (1999); Hyvärinen et al. (2019); Khemakhem et al. (2020) or weak supervision Shu et al. (2019); Locatello et al. (2020) 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 (Locatello et al., 2019b) 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 Higgins et al. (2017); Burgess et al. (2018) 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 Kim & Mnih (2018) and

[1] The code can be found here

β-TCVAE Chen et al. (2018) further refine these constraints by explicitly penalizing total correlation to enforce statistical independence between dimensions. BioAE (Whittington et al., 2022) 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 (Sprekeler et al., 2014; Klindt et al., 2020).

Architectural and Structural Biases. Structural inductive biases embedded directly into model architectures have proved powerful. Vector quantization in models like QLAE (Hsu et al., 2024a) and the recent Tripod framework (Hsu et al., 2024b) induce grid-like latent spaces that simplify factor separation. FactorQLAE (Baykal et al., 2024) combine scalar quantization of the latent variables with a total correlation term in the optimization as an inductive bias. On the theoretical front, Barin-Pacela et al. (2024) establish identifiability for quantized factors under nonlinear mappings. Further, Leeb et al. (2020) 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. Yang et al. (2023) 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, Yang et al. (2024) 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 Jun et al. (2025) 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 (Jung et al., 2025).

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 Hsu et al. (2024b). 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.

5 Conclusion

In this paper, we introduce a novel approach that incorporates Bayesian nonparametric priors into the embedding space of latent quantizing 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 latent quantization, 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 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.

REFERENCES

Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412(6), 2014.

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pp. 1152–1174, 1974.

Vitória Barin-Pacela, Kartik Ahuja, Simon Lacoste-Julien, and Pascal Vincent. On the identifiability of quantized factors. In *Causal Learning and Reasoning*, pp. 384–422. PMLR, 2024.

Gulcin Baykal, Melih Kandemir, and Gozde Unal. Disentanglement with factor quantized variational autoencoders. *arXiv preprint arXiv:2409.14851*, 2024.

Yoshua Bengio. Deep learning of representations: Looking forward. In *International conference on statistical language and speech processing*, pp. 1–37. Springer, 2013.

David M Blei and Michael I Jordan. Variational inference for dirichlet process mixtures. 2006.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, pp. 2060, 2019.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23595–23604, 2024.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations*, 2018.

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and

R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/d97d404b6119214e4a7018391195240a-Paper.pdf.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, pp. 361–369, 2015.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36, 2024a.

Kyle Hsu, Jubayer Ibn Hamid, Kaylee Burns, Chelsea Finn, and Jiajun Wu. Tripod: Three complementary inductive biases for disentangled representation learning. *arXiv preprint arXiv:2404.10282*, 2024b.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pp. 859–868, 2019.

Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29, 2016.

Youngjun Jun, Jiwoo Park, Kyobin Choo, Tae Eun Choi, and Seong Jae Hwang. Disentangling disentangled representations: Towards improved latent units via diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3559–3569. IEEE, 2025.

Whie Jung, Dong Hoon Lee, and Seunghoon Hong. Disentangled representation learning via modular compositional bias. *arXiv preprint arXiv:2510.21402*, 2025.

Ilyes Khemakhem, Diederik Kingma, Ricardo Pio Monti, and Aapo 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. URL https://proceedings.mlr.press/v119/khemakhem20a.html.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pp. 2649–2658. PMLR, 2018.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

Kenichi Kurihara, Max Welling, and Nikos Vlassis. Accelerated variational dirichlet process mixtures. *Advances in neural information processing systems*, 19, 2006.

Felix Leeb, Guilia Lanzillotta, Yashas Annadani, Michel Besserve, Stefan Bauer, and Bernhard Schölkopf. Structure by architecture: Structured representations without regularization. *arXiv preprint arXiv:2006.07796*, 2020.

Haoyang Li, Xin Wang, Zeyang Zhang, Haibo Chen, Ziwei Zhang, and Wenwu Zhu. Disentangled graph self-supervised learning for out-of-distribution generalization. In *Forty-first International Conference on Machine Learning*, 2024.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124, 2019b.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pp. 6348–6359. PMLR, 2020.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pp. 639–650, 1994.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

Vu-Hoang Tran and Ching-Chun 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.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *International conference on machine learning*, pp. 10401–10412. PMLR, 2021.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Sjoerd Van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? *Advances in neural information processing systems*, 32, 2019.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Tsun-Hsuan Wang, Wei Xiao, Tim Seyde, Ramin Hasani, and Daniela Rus. Measuring interpretability of neural policies of robots with disentangled representation. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 602–641. PMLR, 06–09 Nov 2023.

Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5443–5455. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2b6921f2c64dee16ba21ebf17f3c2c92-Paper.pdf.

James CR Whittington, Will Dorrell, Surya Ganguli, and Timothy EJ Behrens. Disentanglement with biological constraints: A theory of functional cell types. *arXiv preprint arXiv:2210.01768*, 2022.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion. In *Advances in Neural Information Processing Systems (NeurIPS) 37*, 2023. URL https://proceedings.neurips.cc/paper/2023/file/da47bfaf3f3a8d5bbab0d60c5195dc18-Paper-Conference.pdf.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. Diffusion model with cross attention as an inductive bias for disentanglement. *NeurIPS*, 2024.

Yuan Zhou, Richang Hong, Yanrong Guo, Lin Liu, Shijie Hao, and Hanwang Zhang. Controllable relation disentanglement for few-shot class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Xinqi Zhu, Chang Xu, and Dacheng 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.

A Appendix

A.1 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.

A.1.1 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. Sethuraman (1994) 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) Antoniak (1974). 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:

$$\beta_k \mid \alpha \sim p(\beta_k \mid \alpha) = \text{Beta}(1, \alpha), \quad \theta_k \mid \lambda \sim p(\theta_k \mid \lambda) = G_0(\lambda), \quad \forall k \in \mathbb{N} \tag{11}$$

$$z = k \mid \beta \sim p(z = k \mid \beta) = \beta_k \prod_{j=1}^{k-1}(1 - \beta_j), \quad x \mid \{z, \theta\} \sim p(x \mid z, \theta) = F(\theta_z) \tag{12}$$

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 $\{x_1, \dots, x_N\}$ and the latent variables: stick-breaking lengths $\beta = \{\beta_1, \beta_2, \dots\}$, component parameters $\theta = \{\theta_1, \theta_2, \dots\}$ and assignment variables $\{z_1, \dots, z_N\}$, factorizes hierarchically as follows:

$$p(x, z, \beta, \theta \mid \alpha, \lambda) = p(\beta \mid \alpha)\, p(\theta \mid \lambda)\, p(z \mid \beta)\, p(x \mid z, \theta) \tag{13}$$

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.Wainwright et al. (2008) 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:

$$\log p(x \mid \alpha, \lambda) \geq \mathbb{E}_{q_\nu}\left[\log p(x, \boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\theta} \mid \alpha, \lambda) - \log q_\nu(\boldsymbol{e}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\theta})\right] \tag{14}$$

A.1.2 Vector Quantization for Disentanglement

Next, we discuss the Vector Quantized-Variational AutoEncoder (VQ-VAE) Van Den Oord et al. (2017), 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:

$$q(z = k \mid x) = \begin{cases} 1 & \text{for } k = \arg\min_j \|z_e(x) - e_j\|_2, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

$$z \sim q(z \mid x), \quad z_q(x) = e_z = e_k \tag{16}$$

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 equation 17, 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 15, 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.

$$L = -\log p(x \mid z_q(x)) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2, \tag{17}$$

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 Locatello et al. (2019b). To structure the latent space such that distinct dimensions capture independent generative factors, the approach of Hsu et al. (2024a) 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.

A.2 Ablation Studies

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 3: 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 (Johnson et al., 2016). 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 3) 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 constrained. 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.

A.3 Disentanglement Metrics

For quantitative evaluation, we compute two complementary disentanglement metrics which comprehensively measure disentanglement properties using different computational approaches. The InfoMEC metric (Hsu et al., 2024a) 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 (Eastwood & Williams, 2018) 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 Locatello et al. (2019b); Hsu et al. (2024a).

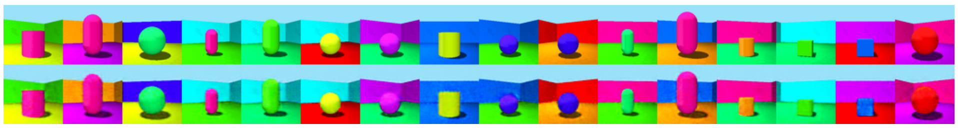
A.4 Qualitative Evaluation

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.

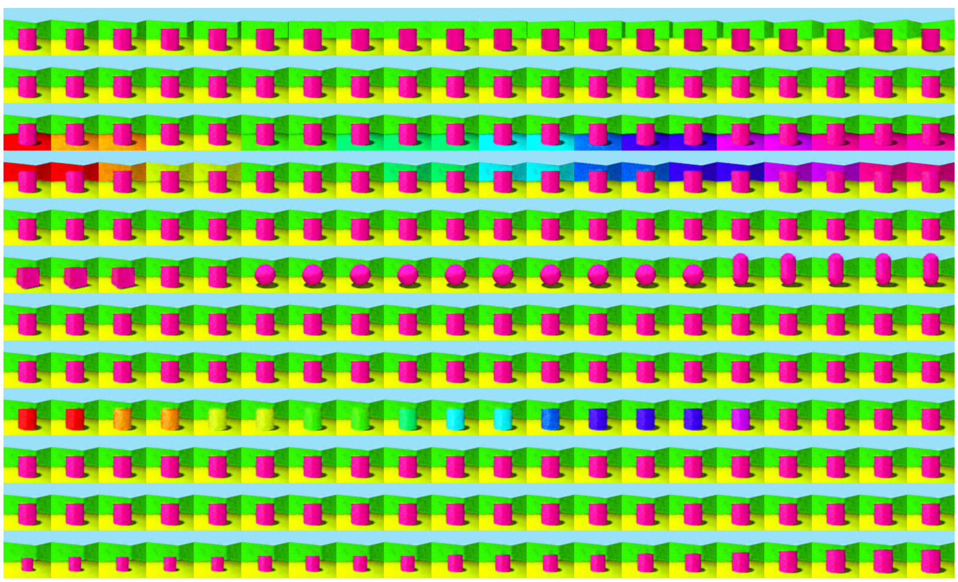
Reconstruction Fidelity. We assess the informativeness of learned representations by examining reconstruction quality. High-fidelity reconstructions that faithfully preserve visual details of the original images indicate that the latent representations are sufficiently informative to capture the full 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.

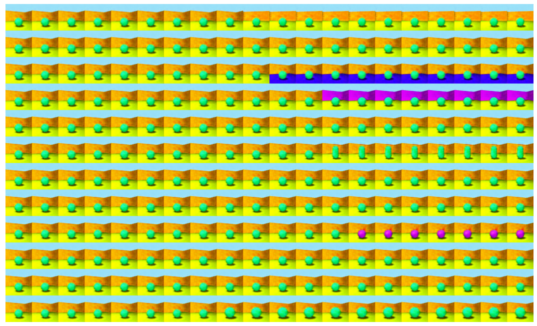


(a) Sample reconstructions: Original images (top row) and corresponding reconstructions (bottom row)

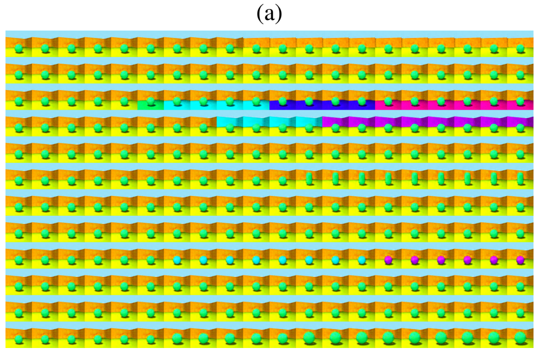


(b) Latent traversals: Each row shows the effect of systematically varying a single latent dimension while holding all other dimensions fixed. Columns represent different values sampled from the distribution of that dimension. The model successfully disentangles six ground-truth factors of variation: 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.

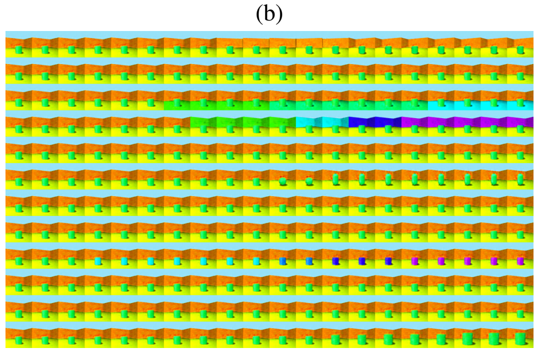
Figure 1: Reconstructions and Latent traversals for the 3DShapes dataset: Reconstructions demonstrate high fidelity in capturing visual details demonstrating the model's ability to faithfully encode and decode the full range of variations in the data. Latent traversals illustrate that individual latent 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)

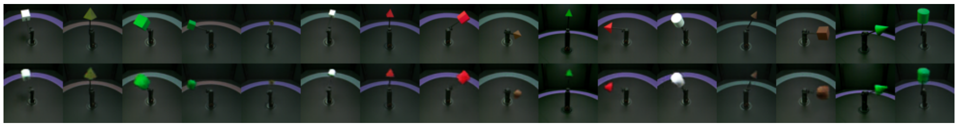


(b)

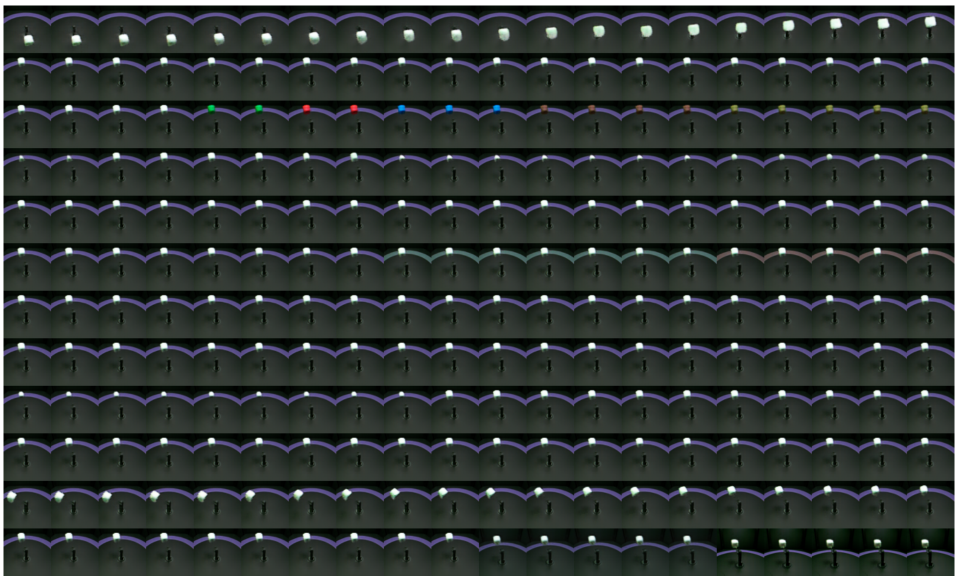


(c)

Figure 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



(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 the following ground-truth factors of variation: vertical axis (row 1), object color (row 3), object shape (row 4), background color (row 6), object size (row 9), horizontal axis (row 11) and camera height (row 12). Rows 2, 5, 7, 8, 10 correspond to inactive latent dimensions that do not encode interpretable factors.

Figure 3: Reconstructions and Latent traversals for the MPI3D real dataset: Reconstructions demonstrate high fidelity in capturing visual details demonstrating the model's ability to faithfully encode and decode the an extensive range of variations in the data. Latent traversals illustrate that individual latent variables control distinct, interpretable factors of variation in the generated images. It is worth noting that, although the representations do not capture the full set of underlying variations, they remain both modular and compact, closely reflecting the true underlying structure of the generative process.