Beyond Flat Latent: Scale-Aware Disentanglement in Hierarchical Latent Spaces

Tingwei Liu Tian Han
Department of Computer Science, Stevens Institute of Technology
{tliu50,than6}@stevens.edu

Abstract

Existing disentanglement methods predominantly operate on a flat latent space and enforce a single global independence criterion across all dimensions, implicitly assuming that the underlying factors of variation are exchangeable and statistically independent. This assumption ignores the multi-scale structure of natural visual data, in which coarse factors such as object identity constrain but do not determine finer factors such as color and texture, leading flat methods to either over-penalize meaningful cross-scale dependencies or under-disentangle within a semantic level. We propose a scale-aware disentanglement framework built on a hierarchical VAE with two complementary regularizers: an intra-level total-correlation penalty that encourages factorized latent units within each level, and an inter-level dependence penalty that discourages redundancy across levels. We demonstrate the effectiveness of our framework through both quantitative and qualitative evidence. These results suggest that disentanglement could benefit from scale-aware hierarchical representations, rather than relying solely on a single global independence criterion.

1. Introduction

Learning structured representations of visual data is a central goal of unsupervised visual representation learning. A representation is said to be disentangled [8, 24] if each latent dimension independently captures a single, interpretable factor of variation in the data, such as object shape, color, or rotation, while remaining invariant to all other factors. The dominant framework for learning disentangled representations without supervision is the Variational Autoencoder (VAE), which encodes observations into a structured latent space through a probabilistic bottleneck. A family of methods, including β-VAE [15], FactorVAE [17], β-TCVAE [3], DAVA [11], and α-TCVAE [27], augment the standard VAE Evidence Lower Bound (ELBO) with penalties that encourage the aggregate posterior to factorize across latent dimensions, thereby promoting statistical independence among the learned factors.

Despite their empirical success on controlled benchmarks such as dSprites [26] and 3DShapes [1], these methods rest on a shared simplifying assumption: that the true generative factors underlying natural visual data are statistically independent, and that a well-disentangled model should therefore assign each factor to a single, globally independent latent dimension. This assumption becomes restrictive for natural visual data. First, real-world factors are often correlated: coarse attributes such as shape and identity may co-vary with finer attributes such as texture, color, and lighting, and such correlations can reflect meaningful structure rather than noise [31]. Second, visual variation is naturally hierarchical: global structure constrains local appearance, while local attributes still retain their own degrees of freedom [7, 20, 34]. As a result, a flat independence objective may penalize meaningful dependencies between coarse and fine factors, or fail to distinguish them from undesirable redundancy within the same semantic level. Although flat disentanglement methods can produce interpretable latent traversals, they provide limited mechanisms for organizing factors by scale, which is important for controlled recombination of global and local attributes. The core research challenge is therefore to develop a disentanglement framework that respects the multi-scale structure of visual data: encouraging independence within each semantic level while preserving structured dependencies across levels.

We propose a scale-aware disentanglement framework built on the hierarchical variational autoencoder (HVAE) [6, 25, 32], which organizes the latent space into a sequence of levels $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L$ that increase semantic abstraction. Our central empirical observation is that disentanglement can be treated as a scale-dependent property: higher-level latent variables co-vary with multiple generative factors simultaneously, governing global structure such as identity and pose. In contrast, lower-level latent variables govern finer, more localized attributes such as color and illumination. However, hierarchy alone does not guarantee this separation. Standard hierarchical VAEs can still encode redundant information across latent levels, since different layers may explain overlapping aspects of the same image. We therefore introduce a two-part training objective. First, an

intra-level total correlation [3] penalty encourages the coordinates within each level to be statistically independent, disentangling factors at the same semantic scale without imposing independence across scales. Second, an inter-level dependence penalty discourages redundancy across levels, encouraging each level to capture complementary rather than duplicated information. Together, these objectives provide an explicit scale-aware alternative to flat disentanglement: they separate factors within each semantic level while preserving structured dependencies across levels.

In summary, our contributions include: (i) We highlight scale-dependence as an important but underexplored property of disentangled representations: factors of variation operate at different semantic granularities. (ii) We propose a two-penalty training objective: intra-level total correlation for within-level disentanglement and an inter-level dependence penalty for cross-level complementarity, which formalizes scale-aware disentanglement within a principled variational framework. (iii) We demonstrate through quantitative and qualitative evaluation that our framework improves overall disentanglement while producing representations that are more factorized within levels and more semantically organized across levels.

2. Method

2.1. Hierarchical VAE

We consider a hierarchical variational autoencoder in which a high-dimensional observed sample $\mathbf{x}$ is represented by a collection of latent variables $\mathbf{z}_{1:L} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L\}$, where L denotes the number of latent levels, and different latent groups are intended to capture factors at different semantic scales. In general, hierarchical VAEs can be instantiated in two forms. The first is a *probabilistic hierarchy* [6, 25, 32], where latent variables are conditionally dependent across levels:

$$p_\theta(\mathbf{z}_{1:L}) = p_0(\mathbf{z}_L) \prod_{\ell=1}^{L-1} p_{\theta_\ell}(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}), \tag{1}$$

where $p_0(\mathbf{z}_L) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The transition prior is defined as $p_{\theta_\ell}(\mathbf{z}_\ell \mid \mathbf{z}_{\ell+1}) \sim \mathcal{N}(\boldsymbol{\mu}_{\theta_\ell}(\mathbf{z}_{\ell+1}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\theta_\ell}(\mathbf{z}_{\ell+1})))$. This formulation explicitly models top-down stochastic dependence between latent levels. While expressive, this probabilistic hierarchy can obscure the interpretation of layer-wise decomposition, since the information encoded at one level may be partially determined by variables at others.
In this work, we use another form of *architectural hierarchy* [7, 20, 34]. Instead of imposing conditional dependence among latent variables, we assume a factorized Gaussian prior across levels:

$$p(\mathbf{z}_{1:L}) = \prod_{\ell=1}^{L} p(\mathbf{z}_\ell), \text{where } p(\mathbf{z}_\ell) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

so that the prior factorizes across levels, $p(\mathbf{z}_{1:L}) = \prod_\ell p(\mathbf{z}_\ell)$, and the variational posterior is conditionally factorized given $\mathbf{x}$, $q_\phi(\mathbf{z}_{1:L} \mid \mathbf{x}) = \prod_\ell q_\phi(\mathbf{z}_\ell \mid \mathbf{x})$. Note that the corresponding *aggregate* posterior $q_\phi(\mathbf{z}_{1:L})$ generally does not factorize across levels. We follow a VLAE hierarchical autoencoding backbone [34], in which latent variables are inferred from different encoder depths and injected into the corresponding stages of a top-down decoder.

Generative Model. The decoder adopts a VLAE [34] top-down construction. The conditional distribution $p(\mathbf{x} \mid \mathbf{z}_{1:L})$ is defined implicitly through a top-down generator $\{f_L, f_{L-1}, \ldots, f_1, f_0\}$:

$$\tilde{\mathbf{z}}_L = f_L(\mathbf{z}_L), \tag{3}$$

$$\tilde{\mathbf{z}}_\ell = f_\ell(\tilde{\mathbf{z}}_{\ell+1}, \mathbf{z}_\ell), \quad \ell = L-1, \ldots, 1, \tag{4}$$

$$\mathbf{x} \sim \mathcal{N}\big(f_0(\tilde{\mathbf{z}}_1), \sigma^2 \mathbf{I}_D\big), \tag{5}$$

where $\tilde{\mathbf{z}}_\ell$ is an auxiliary hidden representation that accumulates top-down context, and each f_ℓ is instantiated as:

$$\tilde{\mathbf{z}}_\ell = u_\ell([\tilde{\mathbf{z}}_{\ell+1}; v_\ell(\mathbf{z}_\ell)]), \tag{6}$$

where $[\cdot\,;\cdot]$ denotes vector concatenation and v_ℓ, u_ℓ are shallow neural networks. The top-level variable $\mathbf{z}_L$ is decoded through the deepest pathway and is therefore encouraged to capture the coarsest, most global factors of variation; each successive level $\mathbf{z}_\ell$ refines the representation by integrating its latent code with the top-down context $\tilde{\mathbf{z}}_{\ell+1}$ inherited from above.

Inference Model. The approximate posterior $q(\mathbf{z}_{1:L} \mid \mathbf{x})$ is a bottom-up ladder encoder [34], matching the multi-level latent structure used by the decoder:

$$\mathbf{h}_\ell = g_\ell(\mathbf{h}_{\ell-1}), \tag{7}$$

$$\mathbf{z}_\ell \sim \mathcal{N}\big(\boldsymbol{\mu}_\ell(\mathbf{h}_\ell), \mathrm{diag}(\boldsymbol{\sigma}^2_\ell(\mathbf{h}_\ell))\big), \tag{8}$$

where $\ell = 1, \ldots, L$, $\mathbf{h}_0 \equiv \mathbf{x}$, and g_ℓ, $\boldsymbol{\mu}_\ell$, $\boldsymbol{\sigma}_\ell$ are neural networks. Each $\mathbf{z}_\ell$ is obtained from the feature map at the corresponding encoder depth, so that shallower latents capture low-level features and deeper latents capture high-level semantic structure.

2.2. Scale-Aware Disentanglement

Intra-Level Total Correlation. Although the factorized prior encourages each $\mathbf{z}_\ell$ to be marginally close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, it does not enforce independence among the individual latent units within a level in the aggregate posterior. Following the TC decomposition of [3], we decompose the per-level KL in expectation over the data as:

$$\mathbb{E}_{p(\mathbf{x})}[\mathrm{KL}(q_\phi(\mathbf{z}_\ell \mid \mathbf{x}) \,\|\, p(\mathbf{z}_\ell))] = \underbrace{I_q(\mathbf{x}; \mathbf{z}_\ell)}_{\text{index-code MI}} + \underbrace{\mathrm{TC}(q_\phi(\mathbf{z}_\ell))}_{\text{total correlation}} + \underbrace{\sum_j \mathrm{KL}\Big(q_\phi(z_\ell^{(j)}) \,\|\, p(z_\ell^{(j)})\Big)}_{\text{dim-wise KL}}, \quad (9)$$

where $\mathrm{TC}(q_\phi(\mathbf{z}_\ell)) = \mathrm{KL}(q_\phi(\mathbf{z}_\ell) \,\|\, \prod_j q_\phi(z_\ell^{(j)}))$ measures the statistical dependency among latent units within level ℓ, and the aggregate posterior is defined as $q_\phi(\mathbf{z}_\ell) = \mathbb{E}_{p(\mathbf{x})}[q_\phi(\mathbf{z}_\ell \mid \mathbf{x})]$. We penalize the total-correlation term with a level-specific weight $\beta_\ell > 1$ to encourage within-level disentanglement:

$$\mathcal{L}_{\text{intra}} = \sum_{\ell=1}^{L} \beta_\ell \, \mathrm{TC}(q_\phi(\mathbf{z}_\ell)). \quad (10)$$

Inter-Level Dependence Penalty. The hierarchical VAE architecture organizes latent variables into L distinct groups, where each group $\mathbf{z}_\ell$ is associated with a dedicated decoding pathway of different depth and capacity. This design imposes a natural group structure on the latent space: higher-level groups are decoded through deeper pathways and are intended to capture global semantic factors, while lower-level groups are decoded through shallower pathways and capture finer-grained variation. Critically, this group structure directly generalizes and aligns with the assumptions underlying the partial disentanglement framework [10, 14, 19], which relaxes the requirement of global factor independence in favor of independence *between* semantically meaningful groups of latent variables.

This alignment motivates a direct integration of group-level total correlation as an inter-level regularizer. Specifically, the partial disentanglement objective identifies between-group TC as the principled measure of cross-group statistical dependence in the aggregate posterior. Applied to our ladder hierarchy, this yields:

$$\begin{aligned}\mathcal{L}_{\text{inter}} &= \mathrm{KL}\left(q_\phi(\mathbf{z}_{1:L}) \,\middle\|\, \prod_{\ell=1}^{L} q_\phi(\mathbf{z}_\ell)\right) \\ &= \mathbb{E}_{q_\phi(\mathbf{z}_{1:L})}\left[\log q_\phi(\mathbf{z}_{1:L}) - \sum_{\ell=1}^{L} \log q_\phi(\mathbf{z}_\ell)\right],\end{aligned} \quad (11)$$

which measures how much the joint aggregate posterior across all levels deviates from a product of independent per-level marginals. When $\mathcal{L}_{\text{inter}} = 0$, the levels are statistically independent in the aggregate posterior, so that each level contributes complementary information.

Crucially, $\mathcal{L}_{\text{inter}}$ is distinct from the intra-level TC $\mathcal{L}_{\text{intra}}$: the former penalizes dependence *between* levels while the latter penalizes entanglement *within* each level. Together, they target complementary failure modes: within-level entanglement and cross-level redundancy. It is worth noting that the inter-level penalty alone is insufficient to produce meaningful hierarchical disentanglement. If the model architecture does not provide a strong hierarchical inductive bias, different latent groups may still collapse or learn arbitrary partitions of information. In other words, the structural hierarchy determines the semantic roles of different latent groups, and the proposed penalties refine this decomposition by encouraging independence within each level and non-redundancy across levels.

Aggregate posterior estimator. Both $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$ require evaluating log-densities of the aggregate posterior $q_\phi(\cdot)$, which is intractable because it marginalizes over the empirical data distribution. We adopt the mini-batch importance-sampling estimators introduced in β-TCVAE [3], applied at two distinct granularities: across coordinates within a level for $\mathcal{L}_{\text{intra}}$, and across levels for $\mathcal{L}_{\text{inter}}$. Given a minibatch of size B, $\{\mathbf{x}_b\}_{b=1}^{B}$, drawn from the entire dataset of size M and the corresponding latent samples $\mathbf{z}_{\ell,b} \sim q_\phi(\mathbf{z}_\ell \mid \mathbf{x}_b)$, the log-density of the aggregate posterior at level ℓ is approximated by

$$\log q_\phi(\mathbf{z}_\ell) \;\approx\; \log\left(\frac{1}{BM}\sum_{b'=1}^{B} q_\phi(\mathbf{z}_\ell \mid \mathbf{x}_{b'})\right), \quad (12)$$

following the importance-weighting argument of [3]. The estimator is biased in $\log q_\phi$, and the bias grows as the batch size shrinks [3]. More expressive estimators are a natural direction for future work, including adversarial density-ratio estimation as used in FactorVAE [17] and DAVA [11], and the α-divergence formulation of α-TCVAE [27], which interpolates between TC and other dependence measures and may better balance disentanglement against sample diversity in the hierarchical setting.

Training Objective. The full training objective comprises a reconstruction term, index-code mutual information, dimension-wise KL, intra-level total correlation, and inter-level dependence:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \sum_{\ell=1}^{L}\left[I_q(\mathbf{x}; \mathbf{z}_\ell) + \sum_j \mathrm{KL}\Big(q_\phi(z_\ell^{(j)}) \,\|\, p(z_\ell^{(j)})\Big)\right] + \mathcal{L}_{\text{intra}} + \gamma\mathcal{L}_{\text{inter}}. \quad (13)$$

where $\mathcal{L}_{\text{rec}} = \mathbb{E}_{q_\phi}[-\log p_\theta(\mathbf{x} \mid \mathbf{z}_{1:L})]$ is the reconstruction loss. The coefficient β_ℓ in Eq. 10 controls the final weight of

intra-level TC at level ℓ, encouraging factorization among latent dimensions within the same level. The coefficient γ controls the final weight of inter-level dependence, used as a soft redundancy penalty across latent levels. Thus, the objective preserves the variational regularization of the base VLAE while explicitly reweighting the dependence terms most relevant to scale-aware disentanglement.

3. Experiment

In this section, we conduct empirical experiments to evaluate the proposed framework and understand its behavior under different conditions by exploring the following questions: (*Q1*) Does our hierarchical framework achieve stronger global disentanglement than flat and hierarchical baselines? (*Q2*) Do the intra-level and inter-level regularization terms each contribute meaningfully to the overall disentanglement quality, and what is the effect of removing either component? (*Q3*) Does the learned hierarchy support more selective latent traversals that better preserve unrelated attributes?

3.1. Experiment Setup

Dataset. We conduct quantitative evaluation on three factorized benchmarks and qualitative traversal analysis on CelebA [22]. **3DShapes** [1] contains 480,000 images with six ground-truth factors: object shape, object color, floor color, wall color, object scale, and azimuth. **MPI3D-Real** [13] contains 103,680 images of objects mounted on a robot arm, with seven factors: object color, size, shape, camera height, background color, azimuth, and robot arm altitude; it is considered one of the most challenging factorized benchmarks due to its photorealistic rendering. **Cars3D** [28] contains 16,185 images spanning three factors: car type, elevation, and azimuth. **CelebA** [22] contains over 200,000 face images with 40 binary attribute annotations across a broad range of poses, expressions, and lighting conditions; as a real-world dataset with correlated and partially observed factors, it represents a realistic and challenging setting. All datasets use resolution 64×64 RGB images.

Disentanglement Metrics. We evaluate disentanglement using three complementary metrics: **DCI-D**, **DCI-C** [9], and **MIG** [3]. Following [9], we train a predictor for each factor using the learned representations and collect the resulting feature-importance weights into a factor-dimension importance matrix $R \in \mathbb{R}^{d \times K}$, where d is the latent dimensionality and K is the number of ground-truth factors.

DCI-D measures whether each latent dimension is specialized to at most one factor. For each latent dimension j, we normalize the j-th row of R into a probability distribution $\hat{R}_{jk} = R_{jk} / \sum_{k'} R_{jk'}$ and compute its entropy $H(\hat{R}_{j\cdot})$ in base K, so that $H(\hat{R}_{j\cdot}) \in [0, 1]$. The overall score is:

$$\text{DCI-D} = 1 - \sum_{j=1}^{d} \frac{\sum_k R_{jk}}{\sum_{j,k} R_{jk}} H\Big(\hat{R}_{j\cdot}\Big), \tag{14}$$

where lower entropy indicates that a latent dimension is associated with fewer factors.

DCI-C measures whether each ground-truth factor is captured compactly by a small number of latent dimensions. For each factor k, we normalize the k-th column of R into a probability distribution $\hat{R}_{jk} = R_{jk} / \sum_{j'} R_{j'k}$ and compute its entropy $H(\hat{R}_{\cdot k})$ in base d, so that $H(\hat{R}_{\cdot k}) \in [0, 1]$:

$$\text{DCI-C} = 1 - \frac{1}{K} \sum_{k=1}^{K} H\Big(\hat{R}_{\cdot k}\Big). \tag{15}$$

A high DCI-D score indicates that each latent unit is factor-specific, whereas a high DCI-C score indicates that each factor is concentrated in only a few latent units.

Mutual Information Gap (MIG) [3] measures disentanglement by comparing the top two latent dimensions in terms of mutual information with each ground-truth factor. For factor k, let $I(z^{(j)}; v_k)$ denote the mutual information between the j-th latent unit and factor v_k, and let $z^{(j_1)}$ and $z^{(j_2)}$ be the top two dimensions ranked by this quantity. The MIG score is

$$\begin{aligned} \text{MIG} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{H(v_k)} \\ \times \Big[I(z^{(j_1)}; v_k) - I(z^{(j_2)}; v_k) \Big], \end{aligned} \tag{16}$$

where $H(v_k)$ normalizes by the entropy of the factor. A high MIG score indicates that one latent dimension captures substantially more information about a factor than the second-best dimension.

Together, DCI-D, DCI-C, and MIG provide complementary views of representation quality: DCI-D evaluates whether each latent unit is factor-specific, DCI-C evaluates whether each factor is compactly represented, and MIG evaluates the gap between the most and second-most informative latent units for each factor.

Baseline Methods. We compare against a suite of representative disentanglement methods: β-VAE [15], β-TCVAE [3], FactorVAE [17], DAVA [11], Hausdorff Factorized Support (HFS) [29], and α-TCVAE [27]. These baselines operate on a single flat latent space and apply a global independence criterion across all latent dimensions, without any hierarchical organization of factors by semantic scale. We additionally include methods that operate on hierarchical latent spaces to isolate the contribution of our regularization from the benefits of the hierarchical architecture or latent space alone, including NVAE [32] and VLAE [34].

Table 1. **Disentanglement comparison on all latent units.** We compare flat and hierarchical baselines using DCI-D, DCI-C, and MIG. For hierarchical latents, we gather all level groups and evaluate all latent units. Higher values indicate better performance. We evaluate with five random seeds and report the standard deviation. Best results for each metric are **bold**, second-best underlined.

Method	3DShapes			MPI3D-Real			Cars3D		
	DCI-D↑	DCI-C↑	MIG↑	DCI-D↑	DCI-C↑	MIG↑	DCI-D↑	DCI-C↑	MIG↑
Flat Latent									
β-VAE	0.58 ± 0.11	0.50 ± 0.06	0.28 ± 0.06	0.26 ± 0.02	0.28 ± 0.04	0.16 ± 0.01	0.16 ± 0.01	0.08 ± 0.01	0.03 ± 0.02
β-VAE+HFS	0.71 ± 0.03	0.62 ± 0.02	0.36 ± 0.02	0.22 ± 0.02	0.33 ± 0.04	0.18 ± 0.05	0.19 ± 0.02	0.17 ± 0.02	0.04 ± 0.02
β-TCVAE	0.69 ± 0.02	0.59 ± 0.03	0.35 ± 0.03	0.10 ± 0.01	0.18 ± 0.01	0.06 ± 0.01	0.19 ± 0.02	0.17 ± 0.01	0.07 ± 0.02
FactorVAE	0.73 ± 0.05	0.63 ± 0.03	0.29 ± 0.08	0.27 ± 0.02	0.30 ± 0.05	0.13 ± 0.06	0.10 ± 0.02	0.10 ± 0.02	0.03 ± 0.01
DAVA	0.77 ± 0.02	0.74 ± 0.03	0.53 ± 0.03	0.12 ± 0.02	0.16 ± 0.02	0.18 ± 0.02	0.33 ± 0.02	0.28 ± 0.02	$\mathbf{0.13 \pm 0.05}$
α-TCVAE	0.72 ± 0.09	0.62 ± 0.03	0.37 ± 0.03	0.38 ± 0.03	0.37 ± 0.02	$\mathbf{0.24 \pm 0.03}$	0.25 ± 0.03	0.20 ± 0.03	0.08 ± 0.03
Hierarchical Latent									
NVAE	0.10 ± 0.04	0.07 ± 0.04	0.02 ± 0.02	0.10 ± 0.02	0.03 ± 0.01	0.02 ± 0.01	0.11 ± 0.03	0.06 ± 0.02	0.02 ± 0.01
VLAE	0.86 ± 0.02	0.72 ± 0.02	0.40 ± 0.03	0.14 ± 0.02	0.16 ± 0.02	0.16 ± 0.02	0.17 ± 0.03	0.09 ± 0.02	0.05 ± 0.01
Ours	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.56 \pm 0.02}$	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.42 \pm 0.03}$	0.20 ± 0.01	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.30 \pm 0.01}$	0.11 ± 0.02

3.2. Quantitative Results

Latent Unit Disentanglement. We follow the evaluation protocol [27] and measure global disentanglement over all latent units. For a fair comparison, we follow the baseline methods by using 10 latent units in total, allocated across the hierarchy in a top-down configuration of [4,2,2,2]. To ensure that the performance gain is not simply due to increased model capacity, we also control the total number of parameters. The baseline models contain approximately 0.8M–1.2M parameters, while our model has 1.2M parameters. As shown in Tab. 1, our method achieves the strongest DCI-D and DCI-C performance across three datasets, while remaining competitive on MIG.

A further observation is that not all hierarchical latent models are equally suitable for disentanglement. NVAE [32] performs poorly across all datasets, suggesting that NVAE's probabilistic top-down conditional hierarchy is a less suitable inductive bias for learning disentangled representations. In models such as NVAE, lower-level latents are conditionally generated from higher-level latents and are optimized to improve likelihood and reconstruction. This design encourages different latent levels to cooperate in explaining the image, but it does not require them to capture independent semantic factors or distinct levels of abstraction. As a result, information can be redundantly distributed across levels, and the hierarchy may represent conditional residual details rather than an interpretable factor structure.

In contrast, structure-wise hierarchical models such as VLAE [34] and our method can provide a useful inductive bias for disentanglement by explicitly organizing the representation into separate latent groups associated with different levels of variation. However, the gap between VLAE and our method indicates that architectural hierarchy alone remains insufficient. This is consistent with the qualitative observation in Fig. 1, where the unregularized hierarchy already shows a coarse-to-fine factor stratification but exhibits both within-level entanglement and cross-level redundancy that the architecture alone cannot resolve. Our intra-level and inter-level constraints further encourage each level to learn factorized information while reducing redundancy across levels. Therefore, the results support our first experimental goal *Q1*: a structure-wise hierarchical representation, when combined with suitable disentanglement regularization, provides an effective framework that improves over hierarchical baselines and achieves strong performance relative to flat disentanglement methods.

3.3. Qualitative Analysis

To examine how the proposed regularizers shape the organization of the latent hierarchy, we visualize unit-wise latent traversals under progressively adding training objectives on the 3DShapes [1] dataset, which comprises six ground-truth factors: object shape, object color, floor color, wall color, object scale, and azimuth. Fig. 1 visualizes per-dimension traversals of every latent unit $\mathbf{z}_{\ell j}$ in a three-level $\ell = 3$ hierarchy $\mathbf{Z}_\ell$, where each latent level contains $j = 3$ units. We visualize four training regimes: the base ELBO without either regularizer, with the intra-level penalty $\mathcal{L}_{\text{intra}}$ only, with the inter-level penalty $\mathcal{L}_{\text{inter}}$ alone, and with the full objective combining $\mathcal{L}_{\text{inter}}$ and $\mathcal{L}_{\text{intra}}$ from left to right. Each row sweeps a single unit across seven values while holding all others fixed. A well-organized hierarchical disentangled representation should exhibit factor-specific variation within each row while avoiding redundant encoding of the same factor across multiple levels.

Without either penalty, the architectural hierarchy alone induces a partial level-wise organization of factors: the up-

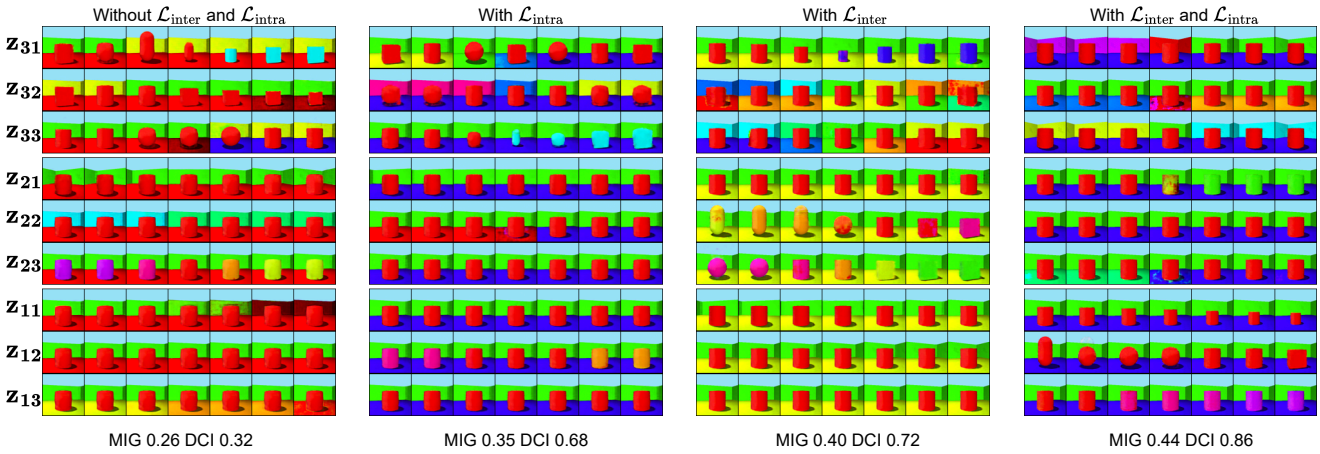

Figure 1. **Latent traversals of each unit $\mathbf{z}_{\ell j}$ across all three hierarchical levels.** Comparing the base ELBO, with $\mathcal{L}_{\text{intra}}$ only, with $\mathcal{L}_{\text{inter}}$ only, and with the full objective $\mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{intra}}$. The traversals suggest progressively cleaner disentanglement both across levels and within each level. Each row shows the effect of traversing a single latent unit over seven evenly spaced values from -3 to 3, while keeping all other latent units fixed. We also report MIG and DCI-D results for comparison.

per level $\mathbf{Z}_3$, whose latent units are decoded through the deepest top-down pathway, predominantly captures five semantic factors. The middle level $\mathbf{Z}_2$ responds to factors such as wall color, object color, and azimuth, while the lowest level $\mathbf{Z}_1$ encodes some variations in wall and floor color. Despite this partial level-wise organization, the unregularized model still exhibits two co-occurring failure modes. First, there is substantial within-level entanglement: a single unit can jointly control multiple factors, as $\mathbf{z}_{31}$ simultaneously modulates object color, shape, scale, and wall color rather than isolating fewer factors. Second, the same factor can be redundantly represented across levels: wall color is modulated by $\mathbf{z}_{31}$, $\mathbf{z}_{22}$, and $\mathbf{z}_{11}$, indicating that architectural hierarchy alone does not prevent repeated use of the same information throughout the latent space.

Adding $\mathcal{L}_{\text{intra}}$ appears to improve unit-level specialization primarily. Compared with the base model, the traversals become more concentrated within each row, and the highly mixed variations in the upper level are reduced. For example, the broad multi-factor changes previously expressed by $\mathbf{z}_{31}$ and $\mathbf{z}_{32}$ become noticeably simpler after applying the intra-level constraint, reducing the number of correlated factors from four to two. Additionally, although $\mathcal{L}_{\text{intra}}$ is defined within each latent group, its effect is not restricted to local row-wise cleanup. Comparing the base and the $\mathcal{L}_{\text{intra}}$-only model, we also observe a different allocation of active factors across levels. For example, object color variation expressed by $\mathbf{z}_{23}$ is no longer carried by the same middle-level group. This suggests that because all groups are optimized jointly, enforcing within-level independence can indirectly reshape the distribution of information across the hierarchy.

Conversely, $\mathcal{L}_{\text{inter}}$ appears to reduce cross-level redundancy. Relative to the unregularized model, factors such as wall and floor color become more concentrated at the higher level $\mathbf{Z}_3$, rather than being expressed repeatedly across all three levels. This suggests that the inter-level constraint encourages different latent groups to carry more complementary information. However, because it does not directly regularize the dimensions inside each group, active units can remain entangled; for instance, the traversals of $\mathbf{z}_{32}$ and $\mathbf{z}_{33}$ continue to mix multiple semantic factors.

The full objective combines these complementary effects. Compared with either regularizer alone, it yields a visually better-organized latent hierarchy, with more factor-specific traversals overall, less factor mixing within individual units, and less redundant reuse of factors across levels. Although some residual factor mixing remains, the comparison suggests that there is potential for combining structural hierarchy with targeted regularization to improve hierarchical disentanglement. This qualitative evidence supports our second experimental question in *Q2*: under hierarchical modeling, the two regularizers play distinct yet complementary roles in improving disentanglement.

Layer-Wise Semantic Isolation Traversal. A practical benefit of the hierarchy is that it separates editable factors across semantic scales. Higher levels capture coarser attributes, while lower levels model increasingly fine residual variations after the higher-level semantics are fixed. Consequently, traversing a unit at a finer level should produce a more selective edit with less disturbance to coarser at-

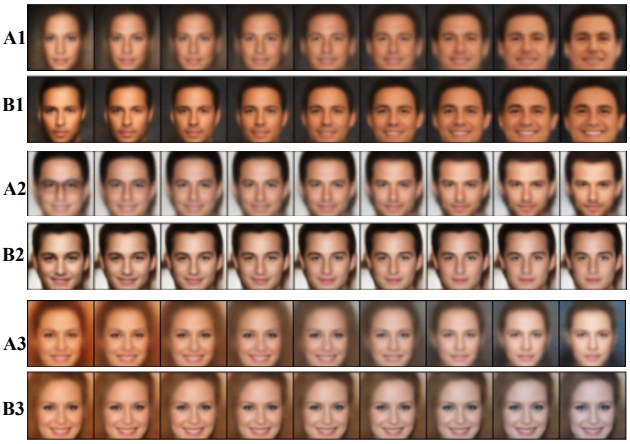

Figure 2. **Scale-aware attribute editing on CelebA via traversals.** Each row sweeps a single latent unit across nine values with all other units fixed. Rows A1, A2, and A3 show traversals from α-TCVAE [27]; rows B1, B2, and B3 show traversals from our method with the same source input images. *(A1, B1)* A unit regulates facial width. *(A2, B2)* A unit controls smiling. *(A3, B3)* A unit controls illumination and ambient tone.

tributes. Fig. 2 illustrates this behavior on CelebA by comparing latent traversals of α-TCVAE [27] and our four-level hierarchical model from the same source images.

The first comparison *(A1 vs. B1)* targets face width, which is represented at the second top level of our hierarchy. Although this traversal still shows some coupling between face width and smiling, our model better preserves the subject identity, whereas the closest-matching α-TCVAE unit introduces greater unintended changes, including a shift in gender. The second comparison *(A2 vs. B2)* targets smiling, represented at the third top level. At this finer level, our traversal becomes more attribute-specific: it changes the smile while largely preserving facial identity and other surrounding attributes, whereas α-TCVAE still entangles the edit with broader facial changes. Taken together, these two examples suggest a useful role for the hierarchy: although smiling remains partially mixed with face width at a coarser level, a separate unit at a finer level captures expression more selectively, changing the smile while preserving facial identity and face geometry. The final comparison *(A3 vs. B3)* targets illumination and ambient tone, which are assigned to the bottom level. This finest-level traversal appears more selective in our model, modifying lighting and color cast while leaving higher-level attributes such as identity, expression, gaze, and hairstyle intact.

Together, these traversals offer preliminary support for *Q3*. Beyond encouraging statistical independence among latent coordinates, allocating variations of different semantic scales to different levels of the hierarchy may help isolate individual factors. This suggests a hierarchical structure as one possible route toward more selective editing on real-world images, where purely flat latent factorization may still leave semantically distinct attributes entangled.

Table 2. Ablation on hierarchy depth and latent allocation on 3DShapes. All settings keep the total latent dimensionality fixed at 10 and use the same model capacity.

Layers	Allocation	MIG ↑		DCI-D ↑		DCI-C ↑	
		VLAE	Ours	VLAE	Ours	VLAE	Ours
2	$(4, 6)$	0.24	0.20	0.43	0.56	0.42	0.57
	$(5, 5)$	0.19	0.20	0.44	0.50	0.41	0.46
	$(6, 4)$	0.10	**0.36**	0.40	**0.80**	0.34	**0.70**
3	$(2, 4, 4)$	0.20	0.41	0.65	0.84	0.57	**0.74**
	$(3, 3, 4)$	0.30	**0.50**	0.78	**0.86**	0.64	0.72
	$(4, 4, 2)$	0.24	0.36	0.50	0.78	0.45	0.66
4	$(4, 3, 2, 1)$	0.29	0.33	0.84	0.87	0.70	0.72
	$(4, 2, 2, 2)$	0.38	**0.56**	0.86	**0.92**	0.72	**0.80**
	$(3, 3, 2, 2)$	0.27	0.44	0.80	0.88	0.71	0.75

3.4. Ablation Study

Hierarchical Latent Allocation. We first examine how a fixed latent budget might be distributed across hierarchy levels. Keeping the total latent dimensionality and model capacity fixed, we vary the per-level allocation, denoted from the top to the bottom level, and report the mean over five seeds for the plain VLAE [34] and our full model in Tab. 2. In our experiments, allocation appears to matter even at fixed depth and capacity: neither a uniform nor a strictly tapering split is consistently optimal, suggesting that hierarchical structure alone may not be sufficient and that the way capacity is partitioned across levels can influence the learned factorization. The performance of both models varies with this choice, and each tends to perform best under particular allocations. For a given allocation, our method generally matches or improves upon VLAE, with the larger differences appearing where the baseline is weakest.

Effect of Hierarchy Depth. We next vary the number of levels while keeping the latent budget and capacity constant. When we compare the best allocation found at each depth, disentanglement increases with depth for both models across all three metrics, indicating that additional levels can support a more expressive factorization of the generative factors. However, when considering all configurations, a deeper but poorly partitioned hierarchy can underperform a shallower, well-allocated one. Overall, these ablations are consistent with the view that depth and per-level allocation jointly influence disentanglement, and that our objective remains effective across the evaluated configurations.

4. Related Work

4.1. Disentangled Representation Learning

Unsupervised disentanglement has developed along a single dominant trajectory: starting from the VAE ELBO and progressively refining a regularizer that drives the aggregate posterior toward factorization across latent dimensions. The earliest formulations [2, 15] achieve this implicitly through KL upweighting; subsequent work makes the target explicit by isolating and penalizing total correlation [3, 12, 17, 18]; and the most recent generation refines either the estimator, the auxiliary geometry, or the training dynamics, for example, through adversarial schemes [11], explicit support factorization for correlated factors [29], latent quantization [16], or generalized divergences [27]. Despite this methodological diversity, the trajectory shares a single structural commitment: the latent space is flat, all dimensions are exchangeable, and a global independence criterion is applied uniformly across them. Both theoretical analyses [24] and empirical studies on naturally correlated data [31] have exposed the limitations of this commitment, and a small line of work [10, 19] has explored partial disentanglement that enforces independence between *groups* of latent variables rather than individual dimensions. However, these group-based formulations still treat the partition as flat and leave the relationship between group structure and the multi-scale nature of visual variation unaddressed. Our work departs from this trajectory by treating the latent partition as fundamentally hierarchical and tying it to the semantic scale of the underlying generative factors.

4.2. Hierarchical Latent-Variable Modeling

Hierarchical VAEs have followed two parallel paths, distinguished by how dependencies between latent groups are modeled. The first emphasizes *probabilistic* hierarchy, in which lower-level latents condition on higher-level ones through a top-down generative process [6, 25, 30, 32]. This design has produced state-of-the-art likelihoods, but analyses of these models [5, 33] show that conditional dependence encourages levels to explain the data cooperatively rather than capture independent factors, leading to opaque per-level semantics and, as observed in our experiments, weaker disentanglement. The second path emphasizes *architectural* hierarchy with a factorized prior, separating latent groups by the depth of the decoding pathway so that scale is encoded structurally rather than probabilistically [7, 20, 34]. This second path is naturally aligned with disentanglement, since architectural depth induces a coarse-to-fine specialization across levels, and it forms the structural basis of our work.

A handful of recent methods have explicitly combined architectural hierarchy with disentanglement objectives [4, 14, 21, 23]. These works establish that hierarchical structure can induce a coarse-to-fine factor stratification, but they leave two failure modes largely unaddressed: within-level entanglement and cross-level redundancy. Our work targets exactly this gap. Building on the architectural hierarchy with factorized priors, we add a dual regularizer: an intra-level total correlation penalty that enforces independence among coordinates within each level, and an inter-level dependence penalty that suppresses redundancy across levels. Together, these terms formalize disentanglement as a scale-dependent property, within a single principled variational objective.

5. Conclusion and Future Work

We introduced a scale-aware framework for disentangled representation learning that treats disentanglement as a property of semantic granularity rather than a single global independence objective. Built on a hierarchical VAE, our method combines an intra-level TC penalty to promote within-level factorization with an inter-level dependence penalty to reduce cross-level redundancy. Across our experiments, this objective achieves competitive performance relative to existing disentanglement methods, while the traversal analysis suggests that the two regularizers play complementary roles in organizing the latent hierarchy. On CelebA, the traversals suggest that hierarchy may provide an additional route for factor isolation by mapping variations across different semantic scales to distinct levels, thereby supporting more selective, fine-grained edits. Overall, architectural hierarchy, paired with within- and cross-level regularization, is a useful inductive bias for disentanglement.

A natural next step is to extend our objective with the progressive training scheme of proVLAE [20], which introduces latent levels from coarse to fine. This temporal inductive bias is complementary to ours: progressive training encourages appropriate level-wise allocation, while our regularizers promote factorization within and across levels. In this setting, the intra-level penalty can be applied to each active level, and the inter-level penalty can be restricted to the active hierarchy at each stage. More broadly, we will explore coupling diffusion models with our hierarchical encoder to improve generation quality while preserving latent disentanglement, and conduct more extensive ablation studies on hierarchy depth, per-level latent capacity, and datasets of increasing complexity.

Acknowledgment

This work is supported in part by NSF 2339604. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018. 1, 4, 5

[2] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018. 8

[3] 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. 1, 2, 3, 4, 8

[4] Xi Chen and Shaoyi Li. PH-VAE: A polynomial hierarchical variational autoencoder towards disentangled representation learning. *arXiv preprint arXiv:2502.02856*, 2025. 8

[5] Yixuan Chen, Yubin Shi, Dongsheng Li, Yujiang Wang, Mingzhi Dong, Yingying Zhao, Robert P Dick, Qin Lv, Fan Yang, and Li Shang. Recursive disentanglement network. In *International Conference on Learning Representations*, 2021. 8

[6] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021. 1, 2, 8

[7] Jiali Cui, Ying Nian Wu, and Tian Han. Learning hierarchical features with joint latent space energy-based prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2218–2227, 2023. 1, 2, 8

[8] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012. 1

[9] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018. 4

[10] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem Meent. Structured disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2525–2534. PMLR, 2019. 3, 8

[11] Benjamin Estermann and Roger Wattenhofer. DAVA: Disentangling Adversarial Variational Autoencoder. In *International Conference on Learning Representations*, 2023. 1, 3, 4, 8

[12] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Auto-encoding total correlation explanation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1157–1166. PMLR, 2019. 8

[13] 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 *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 4

[14] Mohammad Reza Hasanabadi and Davood Gharavian. The hierarchical disentangled information framework enables the discovery of meaningful structures. *Scientific Reports*, 2025. 3, 8

[15] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 1, 4, 8

[16] Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[17] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658, 2018. 1, 3, 4, 8

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 8

[19] Chengrui Li, Yunmiao Wang, Yule Wang, Weihan Li, Dieter Jaeger, and Anqi Wu. A revisit of total correlation in disentangled variational auto-encoder with partial disentanglement. *arXiv preprint arXiv:2502.02279*, 2025. 3, 8

[20] Zhiyuan Li, Jaideep Vitthal Murkute, Prashnna Kumar Gyawali, and Linwei Wang. Progressive learning and disentanglement of hierarchical representations. In *International Conference on Learning Representations*, 2020. 1, 2, 8

[21] F. Lin, X. Yuan, L. Peng, and N.-F. Tzeng. Cascade variational auto-encoder for hierarchical disentanglement. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1248–1257, 2022. 8

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 4

[23] Ziwei Liu, Mingsheng Li, and Cheng Han. Blocked and hierarchical disentangled representation from information theory perspective. *arXiv preprint arXiv:2101.08408*, 2021. 8

[24] 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*, pages 4114–4124. PMLR, 2019. 1, 8

[25] Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 8

[26] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. 1

[27] Cristian Meo, Louis Mahon, Anirudh Goyal, and Justin Dauwels. Alpha tc-vae: On the relationship between disentanglement and diversity. In *International Conference on Learning Representations*, 2024. 1, 3, 4, 5, 7, 8

[28] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28, 2015. 4

[29] Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. In *The Eleventh International Conference on Learning Representations*, 2023. 4, 8

[30] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016. 8

[31] 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 *Proceedings of the 38th International Conference on Machine Learning*, pages 10401–10412, 2021. 1, 8

[32] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 1, 2, 4, 5, 8

[33] Tim Z. Xiao and Robert Bamler. Trading Information between Latents in Hierarchical Variational Autoencoders. In *International Conference on Learning Representations*, 2023. 8

[34] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from deep generative models. In *International Conference on Machine Learning*, pages 4091–4099, 2017. 1, 2, 4, 5, 7, 8