MVU-AE: Minimum Variance Unbiased Autoencoder

Anonymous Authors[1]

Abstract

Variational Autoencoders (VAEs) play a crucial role in deep generative modeling that involves latent variables. The choice of prior distribution in the latent space significantly impacts reconstruction quality and performance in downstream tasks. Motivated by the concept of the Minimum Variance Unbiased Estimator from classical statistics, we propose a novel approach for evaluating latent representation in deep variable learning. We cast the problem of latent representation learning as finding an unbiased representation with lower variance, leading to the development of a Minimum Variance Unbiased Autoencoder (MVU-AE). The MVU-AE incorporates a flexible prior distribution, ensuring a more accurate, informative and complete latent space for image representation. Extensive experiments demonstrate the effectiveness of MVU-AE for image reconstruction and generation tasks on natural and human face datasets, achieving significant improvements in performance. Furthermore, MVU-AE exhibits superior generalization across multiple datasets and excels in unconditional generation when integrated with diffusion models.

1. Introduction

Variational Autoencoders (VAEs) (Kingma, 2013; Kingma et al., 2014) have emerged as a prominent class of generative models that learn probabilistic latent representations. By effectively mapping high-dimensional data into a compact latent space, VAEs facilitate efficient representation learning while preserving essential information about the underlying data distribution. This fundamental capability has led to their widespread adoption across diverse domains, including image and video generation (Razavi et al., 2019; Hu et al., 2022), audio processing (Mao et al., 2023; Li et al., 2021), natural language processing (Chien, 2019; Yang et al., 2020), and unsupervised learning (Nasiri & Bepler, 2022; Lee et al., 2020). A critical aspect of VAE applications lies in their ability to learn well-structured latent representations, which not only enhance the quality of generated samples but also significantly improve reconstruction fidelity - a crucial requirement for many downstream tasks.

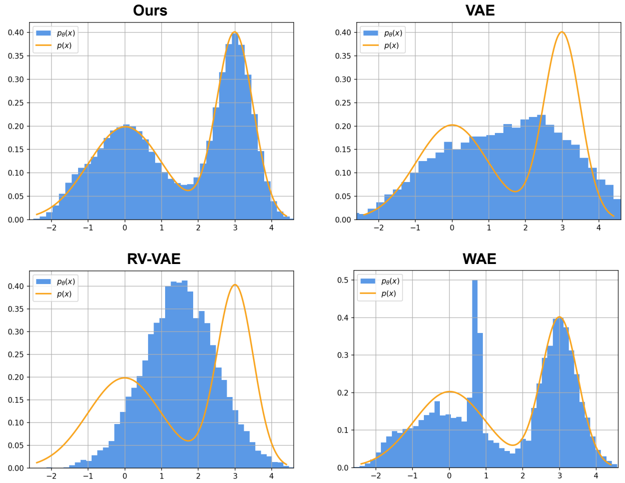

Figure 1. Latent Alignment Bias. The data distribution $p(X)$ (orange line) and the learned latent distribution (blue area) show a misalignment in VAE, RV-VAE and WAE, especially multimodal patterns, which are omitted in VAE and RV-VAE. In contrast, our proposed MVU-AE achieves a well-aligned, expressive latent representation that accurately models $p(X)$.

Despite notable progress, latent representation learning of VAEs continues to grapple with significant challenges. A prominent issue is posterior collapse, which has spurred the development of various strategies, such as introducing hyperparameters (Higgins et al., 2017), utilizing optimal transport (Hao & Shafto, 2023), adopting mixed priors (Bai et al., 2022), discretizing the latent space (Van Den Oord et al., 2017), *etc.* Nevertheless, these approaches still fall short of exploring the underlying structural inefficiencies in VAEs' latent representation learning. A key limitation is that VAEs impose a predefined prior on the latent space, which often misaligns with the true data manifold. While techniques like posterior regularization and discretization reshape the latent space, they introduce additional constraints that distort rather than preserve the underlying structure.

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

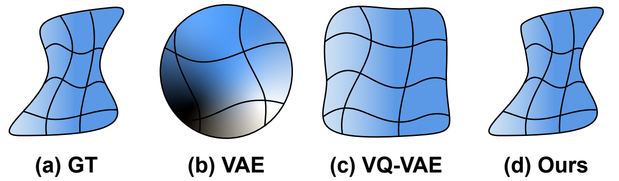

Figure 2. Structural Inductive Bias. The true latent distribution is depicted in (a), while the latent distributions learned by the standard VAE and VQ-VAE exhibit deviations from the true data manifold, as shown in (b) and (c). In contrast, our proposed MVU-AE better preserves the true latent geometry, as illustrated in (d).

In this work, we systematically identify two fundamental issues in VAEs' latent representation learning. First, *latent alignment bias*: A misalignment arises between the latent representation z and the input data X. Due to suboptimal encoding mechanisms, the learned latent representation may either omit essential information from X or encode irrelevant details, as depicted in Fig. 1. As a result, the model struggles to capture complex structures such as multimodal distributions, ultimately reducing its generative and reconstructive capabilities. Second, *structural inductive bias*: The learned latent space may be structurally misaligned with the intrinsic geometry of the data manifold. As illustrated in Fig. 2, imposing a fixed prior or discretizing the latent space can artificially constrain its shape, preventing it from flexibly adapting to the true data distribution.

We propose to learn a flexible and informative latent representation, which should satisfy three conditions, termed *d*eep Minimum Variance Unbiased (d-MVU) conditions. These conditions are inspired by the statistical concept of the minimum variance unbiased estimator (Blackwell & Girshick, 1947; Lee, 1990), establishing a theoretical connection between latent representation learning and the problem of finding an optimal, unbiased estimator with minimal variance. d-MVU conditions include three fundamental conditions, which are rigorously derived to correspond to the classical statistical properties of unbiasedness, sufficiency, and completeness. This formulation provides a principled extension of statistical estimation theory to the domain of deep learning, offering a feasible approach for evaluating and developing high-quality latent representations.

To enforce these conditions, we introduce the Minimum Variance Unbiased Autoencoder (MVU-AE), a novel architecture designed to constrain the learned latent representation to adhere to the d-MVU conditions. Rather than utilizing KL divergence to constrain the latent representation to be close to a predefined prior distribution that leads to latent alignment bias and structural inductive bias, our framework incorporates a specialized form of latent-space regularization that penalizes non-informative subspace in z while capturing all essential information from the dataset X. This allows the latent representation learning to avoid latent alignment bias and structural inductive bias, thus improving the ability of reconstruction and generation. We extensively evaluate our approach through comprehensive experiments on three benchmark datasets: ADE20K, CelebA-HQ, and FFHQ, demonstrating superior performance in both image reconstruction and generation tasks. We further conduct cross-domain experiments on three additional datasets to demonstrate the generalization capability of our model.

In a nutshell, our contributions are summarized below:

- Following the concept of minimum variance unbiased estimator, we propose a novel term the d-MVU conditions in deep learning and combine latent representation training with finding a representation satisfying d-MVU conditions problem. We also use d-MVU conditions to evaluate existing VAEs.
- We introduce the MVU-AE, which features a flexible and informative latent space and achieves accurate latent representation by devising a regularization term to ensure d-MVU conditions. d-MVU conditions prevent structural inductive bias as well as latent alignment bias altogether.

2. Related Work

2.1. Variational Autoencoder

VAEs were first introduced in (Kingma, 2013), where the latent variables are sampled from a probability distribution instead of being deterministic. This approach enables VAEs to model latent distributions and generate diverse samples, making them highly suitable for tasks like image synthesis and semi-supervised learning.

One of the critical challenges in VAEs is balancing the reconstruction quality and the quality of the learned latent space. The original formulation often suffers from issues like posterior collapse (Chen et al., 2016; Lucas et al., 2019; Dai et al., 2020), where the encoder learns to ignore the latent variables, leading to poor generative capabilities. To address this, several techniques have been proposed, such as β-VAEs (Higgins et al., 2017), which introduce a weighting factor to control the trade-off between reconstruction loss and KL divergence. Similarly, VQ-VAEs (Van Den Oord et al., 2017) replace the KL divergence with a discrete codebook and quantization loss, yielding more interpretable latent representations.

Another prominent line of research focuses on hierarchical VAEs (Sønderby et al., 2016; Maaløe et al., 2019), which introduce multi-scale latent spaces to capture complex data distributions more effectively. These models have shown promise in generating high-resolution images and capturing fine-grained details. However, their increased complexity often requires careful tuning and computational resources.

In contrast to adding complexity to network architectures, our work concentrates on refining the latent representation directly. We aim to heighten reconstruction quality by ensuring the unbiasedness, sufficiency and completeness of the latent representation and attaining a more accurate, flexible and informative distribution.

2.2. Latent Representation Optimization

There have been several methods to improve latent representations by various ideas. VQ-VAE (Van Den Oord et al., 2017) and VQGAN (Esser et al., 2021) employ discrete latent representation. SVAE (Bendekgey et al., 2023) allows for unbiased learning to improve the fidelity and expressiveness of latent space and (Li et al., 2014) proposed learning unbiased features to mitigate bias in learned representations, enhancing the robustness of latent space for downstream tasks. Apart from exploring better representation in discrete latent space, HVAE (Caterini et al., 2018) augments the villina VAE with Hamiltonian dynamics to improve posterior inference; (Salimans et al., 2015) and (Wolf et al., 2016) introduced a hybrid framework combining Markov Chain Monte Carlo (MCMC) and variational methods, enabling a more accurate approximation of posterior distributions. However, these methods do not answer what a good latent representation looks like. Following the concept of minimum variance unbiased estimator, we propose a solution involving an accurate, flexible and informative latent representation to tackle these challenges effectively.

3. *d*eep Minimum Variance Unbiased (*d*-MVU) Conditions

As outlined in Sec. 1, two primary concerns on latent representation pose significant challenges: 1) *Latent Alignment Bias*: The learned latent space often encodes non-informative or irrelevant features, failing to capture essential aspects of the data. 2) *Structural Inductive Bias*: Even when the latent space is informative, it may still deviate from the true underlying data structure due to imposed constraints or assumptions. These biases collectively lead to suboptimal outcomes, such as low-quality image reconstruction and limited interpolability of the latent space. In this work, we principally derive *d*eep Minimum Variance Unbiased conditions that the latent space distribution should satisfy, ensuring both alignment with the data distribution and structural fidelity.

3.1. *d*-MVU Conditions for Deep Generative Models

Most generative models assume that a real-world dataset $X = \{x_i\}_{i=1}^n$ is sampled from an underlying distribution $p(X)$. A common approach is to use a θ-parameterized probabilistic model $p_\theta(X)$, trained by the maximum likelihood technique:

$$\theta \leftarrow \arg\max_{\theta} \mathbb{E}_{X\sim p(X)} \left[\log p_\theta(X)\right]. \quad (1)$$

In contrast, we adopt a different perspective. We treat the latent representation z itself as the key ”parameter” that describes the data's real distribution, denoted by z^*. Formally, we assume the existence of a hidden factor z^* such that each observed x_i in the dataset is generated via a conditional probability:

$$x_i \sim p(x \mid z^*), \quad i = 1, \ldots, n. \quad (2)$$

Instead of directly estimating a parameter θ for a predefined model, we frame latent representation learning as an estimation problem, where the goal is to construct a function $f(X; z^*)$ that provides an optimal estimator $\hat{z}$ for the true latent variable z^*. And the optimal representation is

$$MVU(X; z^*) \overset{def}{=} \arg\min_{\mathbb{E}[f(X;z^*)]=z^*} Var(f(X; z^*)) \quad (3)$$

To ensure that $\hat{z}$ is an optimal estimator of z^*, we frame the problem within the context of the Minimum Variance Unbiased Estimator (Blackwell & Girshick, 1947; Lee, 1990), inspired by statistical estimation theory. Grounded in this principle, we establish a strategy for latent representation learning informed by the Lehmann–Scheffé Theorem (Lehmann & Scheffé, 1950; 1955), which characterizes the conditions-unbiasedness, sufficiency and completeness-under which a statistic is an minimum variance unbiased estimator.

Under the minimum variance unbiased estimator, the following conditions must be satisfied:

- **Unbiasedness:**

$$\mathbb{E}\left[\hat{z}(x_1, x_2, \ldots, x_n)\right] = z^*. \quad (4)$$

- **Minimum Variance:**

$$\hat{z} = \arg\min_{\mathbb{E}[z]=z^*} \text{Var}\left[\hat{z}(x_1, x_2, \ldots, x_n)\right]. \quad (5)$$

However, verifying that $\hat{z}$ satisfies the conditions of the minimum variance unbiased estimator is often infeasible in practice. Specifically, Ensuring minimum variance in classical estimation requires comparing all unbiased estimators, an approach that becomes infeasible in high-dimensional latent spaces where deep generative models operate. Furthermore, in deep learning, where functions are parameterized by neural networks, verifying sufficiency and completeness is inherently challenging. To address these challenges, we propose $d$eep Minimum Variance Unbiased (d-MVU) conditions that extend classical MVUE principles to deep latent representation learning.

The d-MVU conditions are defined as follows:

- **Unbiased:** The expected value of $\hat{z}$ equals the true latent variable z^*:
$$\mathbb{E}\left[\hat{z}(x_1, x_2, \ldots, x_n)\right] = z^*. \tag{6}$$
- **Sufficiency:** $\hat{z}$ captures all the information in $x_1, \ldots, x_n$ about z^*. This implies:
$$I(X; z^* \mid \hat{z}) = 0, \tag{7}$$
where $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information. This condition ensures that $\hat{z}$ encodes all the relevant information about z^* that is present in X, such that conditioning on $\hat{z}$ leaves no additional information about z^* in X.
- **Completeness:** $\hat{z}$ must not have any unused or redundant dimensions. Formally, if a nontrivial function $f(\hat{z})$ exists such that:
$$\mathbb{E}_{x \sim p(x|z^*)}\left[f(\hat{z}(x))\right] = 0, \tag{8}$$
then $f(\hat{z})$ must be identically zero, i.e., $f \equiv 0$. This ensures that $\hat{z}$ fully utilizes its representational capacity without encoding unnecessary or redundant information, thereby preventing the existence of latent dimensions that do not contribute meaningfully to representing z^*.

Proposition 1: If $\hat{z}$ is a sufficient statistic of z^*, then $I(X; z^* \mid \hat{z}) = 0$.

Proof: According to the Fisher-Neyman Factorization Theorem (Halmos & Savage, 1949), if $\hat{z}$ is sufficient, the joint distribution of X and z^* can be factorized as:

$$p(x_1, \ldots, x_n; z^*) = f_{\hat{z}}(\hat{z}(x_1, \ldots, x_n); z^*) g(x_1, \ldots, x_n). \tag{9}$$

This implies:

$$p(X \mid \hat{z}, z^*) = p(X \mid \hat{z}), \tag{10}$$

and consequently:

$$I(X; z^* \mid \hat{z}) = H(X \mid \hat{z}) - H(X \mid z^*, \hat{z}) = 0. \tag{11}$$

Thus, all information about z^* in X is fully captured by $\hat{z}$.

Proposition 2: If $\hat{z}$ is a complete statistic of z^*, there exists no free dimension or redundant subspace in $\hat{z}$.

Proof (by contradiction): Assume there exists a redundant subspace $\mathcal{U}$ in $\hat{z}$, and let $u \in \mathcal{U}$. If $\mathbb{E}[u] = 0$, we can construct a nontrivial function $g(u)$ with $\mathbb{E}[g(u)] = 0$, violating the definition of completeness. Thus, $\hat{z}$ must have no unused dimensions. The converse direction also holds; a detailed proof is provided in the Appendix. A.1.

By Proposition 1 and Proposition 2, we establish that the d-MVU conditions align with the classical principles of unbiasedness, sufficiency, and completeness in statistical estimation theory. d-MVU conditions provide a practical and theoretically grounded framework for learning optimal latent representations in deep models. This structured approach directly addresses the fundamental challenges in VAE-based models:

Since $\hat{z}$ is a sufficient and complete statistic of z^*, it fully captures the underlying structure of X, preventing mismatches between the learned and true latent distributions, avoiding latent alignment bias. The unbiasedness and completeness conditions ensure that $\hat{z}$ avoids unnecessary constraints or distortions, allowing it to adapt flexibly to the intrinsic data manifold, eliminating structural inductive bias.

Thus, the d-MVU conditions provide a rigorous yet practical criterion for evaluating and optimizing latent representations in deep generative models, enhancing their expressiveness, efficiency, and adaptability in high-dimensional settings.

3.2. d-MVU Conditions for Assessing Latent Representations

In an autoencoder setting, we typically assume that there exists a ”true” latent representation z^* that governs the generation of data points x. Various latent representations in autoencoders can be interpreted as attempts to estimate z^* from a finite sample $\{x_1, \ldots, x_n\}$. Specifically, we can denote the latent representation as:

$$\hat{z}\left(x_i\right) = f_\phi\left(x_i\right), \tag{12}$$

where f_ϕ is the encoder function. From a statistical perspective, $\hat{z}\left(x\right)$ acts as an estimator of the underlying z^*.

We now assess how VAE-based models meet or violate the d-MVU conditions, evaluating their latent representations for unbiasedness, sufficiency, and completeness, and analyzing how structural constraints introduce latent alignment bias and structural inductive bias.

Existing VAEs can be broadly categorized into two categories based on their latent representations:

Latent representations in VAEs with priors do not satisfy d-MVU conditions: VAEs impose a prior $p\left(z\right)$, often $\mathcal{N}\left(0, I\right)$, alongside an encoder $q_\phi\left(z|x\right)$, which approximates $p\left(z|x\right)$, and a decoder $p_\theta\left(x|z\right)$. The Evidence Lower Bound (ELBO) serves as the objective function:

$$\mathcal{L}_{vae} = \mathbb{E}_{z \sim q_\phi(z|x)}\left[\log p_\theta\left(x|z\right)\right] - D\left[q_\phi\left(z|x\right)||p\left(z\right)\right]. \tag{13}$$

The KL term forces $q_\phi\left(z|x\right)$ to be close to $p\left(z\right) = \mathcal{N}\left(0, I\right)$. The estimator of z^* in VAEs is given by:

$$\hat{z}\left(x\right) = \mathbb{E}_{z\sim q_\phi(z|x)}\left[z\right]. \tag{14}$$

The expectation of $\hat{z}$ can be expressed as:

$$\begin{aligned}\mathbb{E}_{x\sim p(x|z^*)}\left[\hat{z}\left(x\right)\right] &= \int \hat{z}\left(x\right) p\left(x|z^*\right) dx \\ &= \int \left(\int z q_\phi\left(z|x\right) dz\right) p\left(x|z^*\right) dx.\end{aligned} \tag{15}$$

Due to the KL term, $q_\phi\left(z|x\right) \rightarrow \mathcal{N}\left(0, I\right)$, so $\int z q_\phi\left(z|x\right) dz \rightarrow 0$. Thus, $\mathbb{E}_{x\sim p(x|z^*)}\left[\hat{z}\left(x\right)\right] \rightarrow 0 \neq z^*$. This demonstrates that the prior "pulls" the expectation of z towards the prior distribution's mean rather than the true z^*. Consequently, the latent representation z in VAEs is neither satisfied with d-MVU conditions of z^* nor an unbiased estimator of z^*. Furthermore, it can be shown that $\hat{z}$ in VAEs is neither sufficient nor complete. Detailed proofs are provided in Appendix A.2.

Thus, VAEs inherently suffer from both latent alignment bias and structural inductive bias. The prior $p(z)$ introduces a structural constraint that misaligns the latent representation with the true data manifold, limiting its expressiveness (structural inductive bias). Simultaneously, the encoder's adherence to the prior causes the latent representation to omit crucial information or encode unnecessary details, further misaligning it with input data (latent alignment bias). These constraints impose an upper bound on VAE, preventing it from fully capturing the true generative factors of X.

Discrete latent representations in VQ-VAE are also not unbiased: In VQ-VAE (Van Den Oord et al., 2017), a codebook $\{e_k\}_{k=1}^K$ is learned; the latent representation is defined as:

$$z_q\left(x\right) = \arg\min_{e_k} \|E\left(x\right) - e_k\|. \tag{16}$$

Let the residual be $r\left(x\right) = E\left(x\right) - z_q\left(x\right)$. Assuming an optimal scenario where the Straight-Through Estimator (STE) trick ensures encoding $E\left(x\right)$ is correctly optimized and unbiased, we have:

$$\mathbb{E}\left[z_q\left(x\right)\right] = \mathbb{E}\left[E\left(x\right) - r\left(x\right)\right] = z^* - \mathbb{E}\left[r\left(x\right)\right]. \tag{17}$$

For $\mathbb{E}\left[z_q\left(x\right)\right] = z^*$, it is necessary that $\mathbb{E}\left[r\left(x\right)\right] = 0$. Unfortunately, there is no guarantee that clustering will result in a zero-mean residual. Therefore, z_q is not an unbiased representation of z^*. However, it can be shown that z_q in VQ-VAE is both sufficient and complete. Detailed proofs are provided in Appendix A.3.

In summary, VQ-VAE primarily suffers from structural inductive bias. The discretization step forces the latent representation to be constrained within a finite set of codebook entries, which may not fully align with the underlying data manifold. This constraint distorts the latent space, limiting its ability to smoothly capture variations in X. Unlike VAEs, VQ-VAE preserves sufficiency and completeness, but its discrete structure introduces an inherent trade-off between expressiveness and reconstruction accuracy.

4. Minimum Variable Unbiased Autoencoder (MVU-AE)

The d-MVU conditions provide a systematic framework for assessing latent representations in deep generative models. However, traditional VAEs and VQ-VAEs do not satisfy these conditions due to the imposed prior and discretization constraints, respectively. To address these limitations, we introduce the Minimum Variance Unbiased Autoencoder (MVU-AE), which directly optimizes a learnable latent distribution to eliminate both latent alignment bias and structural inductive bias while satisfying the three d-MVU conditions.

The full training objective of MVU-AE is given by:

$$\begin{aligned}\mathcal{L}_{\mathrm{MVU-AE}} =& \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{quant}} \\ =& \|x - \hat{x}\|^2 + \|\mathrm{sg}[E(x)] - z_q\|_2^2 \\ &+ \|\mathrm{sg}\left[z_q\right] - E(x)\|_2^2\end{aligned} \tag{18}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation, and z_q is the quantized latent representation.

Unlike traditional approaches that directly replace $E(x)$ with z_q, MVU-AE uses a soft penalty without substituting $\hat{z}$ with e_k.

Lemma 1. Under the constraint of $\mathcal{L}_{\text{quant}}$, any shift in $\hat{z}$ will result in an increase in $\mathcal{L}_{\text{quant}}$.

Proof. Let $\hat{z}_\delta(x) = \hat{z}(x) + \delta$, and for each x, let the best matching centroid be:

$$k^*(x;\delta) = \arg\min_k \|\hat{z}_\delta(x) - e_k\|^2.$$

Define the quantization loss as:

$$\mathcal{L}_{\text{quant}}(\delta) = \lambda \mathbb{E}_x\left[\|\hat{z}_\delta(x) - e_{k^*}(\delta)\|^2\right]. \tag{19}$$

Since $e_{k^*}(\delta)$ represents the centroid of all points assigned to cluster k, we have:

$$\begin{aligned}e_{k^*}(\delta) &= \frac{1}{|S_k(\delta)|}\sum_{x\in S_k(\delta)} \hat{z}_\delta(x), \\ S_k(\delta) &= \{x : k^*(x;\delta) = k\}.\end{aligned} \tag{20}$$

Expanding the residual shift and analyzing the centroids, we find that any deviation in δ increases $\mathcal{L}_{\text{quant}}$. Detailed derivation is provided in Appendix A.4.

Lemma 2. A shift in $\hat{z}$ does not increase the reconstruction loss $\mathcal{L}_{\text{rec}}$.

Proof. The reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_x \left[\|x - D_\theta(\hat{z}(x))\|^2\right]. \tag{21}$$

Under a shift $\hat{z}_\delta(x) = \hat{z}(x) + \delta$, the decoder parameters D_θ can be updated to accommodate the shift, ensuring that $\mathcal{L}_{\text{rec}}$ remains unchanged. Detailed proof is in Appendix A.5.

The d-MVU conditions provide a practical method for evaluating whether a latent representation is a good estimator. Unlike traditional approaches in classical statistics, which are infeasible in high-dimensional or implicit-function settings, d-MVU conditions simplify the verification process by breaking it into three manageable criteria: unbiased representation, sufficient representation, and complete representation.

Proposition 3. The latent representation of MVU-AE satisfies the d-MVU conditions.

Proof. Suppose:

$$\delta = \mathbb{E}_{x \sim p(x|z^*)}\left[\hat{z}(x)\right] - z^* \neq 0. \tag{22}$$

Lemma 1 establishes that the quantization loss penalizes deviation δ from the optimal latent representation. Lemma 2 ensures that reconstruction quality remains stable despite shifts in $\hat{z}$, implying that learning dynamics will not introduce bias. Since any deviation from the true z^* increases the total loss, the optimal solution must satisfy $\mathbb{E}[\hat{z}(x)] = z^*$, proving unbiasedness.

We further prove $\hat{z}$ satisfies sufficiency:

$$\begin{aligned} I\left(z^*, x|\hat{z}\right) &= H\left(x|\hat{z}\right) - H\left(x|z^*, \hat{z}\right) \\ &= H\left(x|\hat{z}\right) \leq H\left(x|z_q\right) \to 0. \end{aligned} \tag{23}$$

Additionally, $\hat{z}$ satisfies completeness, as completeness in the d-MVU framework ensures that every latent dimension meaningfully contributes to the representation. Any unused subspace in $\hat{z}$ increases the quantization loss, as the residual variance in these dimensions creates additional reconstruction error. As training progresses, this redundancy is penalized, causing the optimization process to collapse these subspaces to zero. Thus, MVU-AE inherently removes free dimensions, ensuring a complete representation.

The optimization of $\mathcal{L}_{\text{quant}}$ encourages all latent dimensions to align with the nearest cluster centroid. Suppose there exists a leftover subspace $\mathcal{U}$, with some dimension $u \in \mathcal{U}$ contributing noise or redundancy. During training, the quantization process penalizes this redundancy, as the distance between $\hat{z}$ and the assigned centroid e_k increases in the presence of unused dimensions. This penalty forces the optimization to reduce the variance in u, effectively collapsing the subspace $\mathcal{U}$ and ensuring that all latent dimensions contribute to meaningful representation.

Moreover, the iterative adjustment of centroids based on the assigned points (Eq. 20) ensures that the latent clusters are refined to minimize redundancy. Any residual variance in unused subspaces directly contributes to the quantization loss, driving the model to eliminate such residuals over successive updates. As a result, the latent representation $\hat{z}$ becomes complete, with no leftover dimensions.

Formally, we prove it by contradiction. Based on Proposition 2 in 3.1, it's sufficient to prove there is no leftover space in $\hat{z}$. Assume there exists a leftover subspace $\mathcal{U}$ in $\hat{z}$, with u being one of its dimensions. This implies that $\hat{z}$ is uniformly random with a mean of 0 across all clusters on dimension u:

$$\sum_{k=1}^{K} \sum_{x \in S_k} \left(\hat{z}_u(x) - \left(e_k\right)_u\right)^2. \tag{24}$$

Since $\hat{z_u}$ has zero-average in k-th cluster, that is,

$$\left(e_k\right)_u = \frac{1}{|S_k|} \sum_{x \in S_k} \hat{z}_u(x) = 0, \tag{25}$$

and this term contributes directly to $\mathcal{L}_{\text{quant}}$ and will be minimized during optimization. Consequently, $\hat{z}_u(x) \approx 0$ for all x, as the optimization progresses.

Therefore, there is no leftover subspace in $\hat{z}$, and the latent representation satisfies completeness.

Verifying the completeness of $\hat{z}$ under the minimum variance unbiased estimator concept directly is abstract and inherently complex. The proof requires explicitly constructing functions related to $\hat{z}$, which is particularly challenging in general cases due to the implicit nature of $\hat{z}$ in deep learning. This difficulty highlights the significant advantage of the d-MVU conditions, which offers a more practical and systematic approach to evaluating completeness.

In summary, MVU-AE is the first deep generative model that explicitly optimizes a latent representation to satisfy the d-MVU conditions. MVU-AE learns an adaptive latent distribution that remains unbiased, sufficient, and complete. By ensuring d-MVU conditions, MVU-AE not only eliminates structural inductive bias and latent alignment bias but also improves expressiveness and generative quality. Empirical results validate its superiority in image reconstruction and downstream tasks, demonstrating its potential as a new paradigm for deep generative modeling.

5. Experiments

5.1. Experimental Settings

Datasets. For comparison with exisiting methods, we conduct the mdoel training on ADE20K (Zhou et al., 2017), CelebA-HQ (Liu et al., 2015) and FFHQ (Karras et al., 2019) datasets respectively, for image reconstruction. To further demonstrate the generalization capabilities of our method, we extend our evaluation to cross-domain datasets, including MS-COCO (Lin et al., 2014), LSDIR (Yu et al., 2021), and DIV2K (Timofte et al., 2017). Additionally, we undertake unconditional generation task on the CelebA-HQ dataset integrating to diffusion model. We also evaluate interpolativity by Perceptual Path Length (Karras et al., 2019).

Implementation details. In our implementation, all the models compress the input 256×256 images into the 64×64 latent representations, corresponding to a downsampling factor of $f = 4$. For the reconstruction task, we set the batch size to 40 and the model is trained for 80 epochs. For the generation task, the batch size is increased to 120, and the model is trained for 200 epochs. All the experiments are conducted on 8 NVIDIA RTX 3090 GPUs with 24GB of memory. For the optimization process, we utilize the AdamW optimizer (Loshchilov, 2017) with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.95$, and a base learning rate of 2×10^{-6}.

Metrics. On image reconstruction and cross-domain evaluation tasks, we use PSNR, SSIM, LPIPS to evaluate the model's capabilities from various aspects. For unconditional generation tasks, we use FID (Heusel et al., 2017), IS, Precision and Recall for comparison.

5.2. Quantitive Evaluation

Reconstruction experiments are conducted on the ADE20K, CelebA-HQ and FFHQ datasets. The comparative results are detailed in Tab 1. It is noteworthy that our model outperforms VAE (Kingma, 2013; Kingma et al., 2014), VQGAN (Esser et al., 2021), WAE (Tolstikhin et al., 2017), RV-VAE (Nicodemou et al., 2023) by 1.86, 1.49, 1.31, 0.65 dB in PSNR, 0.07, 0.65, 0.04, 0.02 in SSIM and 0.031, 0.032, 0.018, 0.012 in LPIPS on ADE20K. On CelebA-HQ, compared to VAE, VQGAN, WAE and RV-VAE, there is a 3.11, 1.54, 2.15, 1.77 dB improvement in PSNR, 0.09, 0.04, 0.05 and 0.04 performance gains in SSIM and 0.032, 0.011, 0.014 improvement in LPIPS. Additionally, our model surpasses VAE, VQGAN, WAE, RV-VAE by 0.75, 2.71, 0.63, 1.07 dB in PSNR, 0.02, 0.07, 0.02, 0.02 in SSIM, and 0.016, 0.09, 0.01, 0.009 in LPIPS on FFHQ dataset. Moreover, it is important to highlight that our model has lower variance of latent representation, shown in Tab. 4. These observations demonstrate that our model can reconstruct well on different datasets. Previous methods face challenges in reconstructing images without unbiased or sufficient or complete latent representation. In contrast, our model effectively reconstructs image details and reduces artifacts by applying reasonable regularization to ensure the latent representation is an unbiased representation with lower variance. We further conduct unconditional image generation on CelebA-HQ dataset and the results are shown in Tab. 2. Our model still shows improvement in the FID and IS on different kinds of datasets. It shows that an accurate, informative, and complete latent representation is beneficial in producing unconditional images for diffusion model.

Table 1. Quantitative comparison of image reconstruction with VQGAN (Esser et al., 2021), VAE (Kingma, 2013), RV-VAE (Nicodemou et al., 2023), WAE (Tolstikhin et al., 2017) on the ADE20K, CelebA-HQ and FFHQ datasets.

Method	ADE20K			CelebA-HQ			FFHQ		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VAE	25.56	0.76	0.098	28.40	0.80	0.081	28.95	0.84	0.058
VQGAN	25.93	0.78	0.099	29.97	0.85	0.060	26.99	0.79	0.132
WAE	26.11	0.79	0.085	29.36	0.84	0.063	29.07	0.84	0.052
RV-VAE	26.77	0.81	0.079	29.74	0.85	0.058	28.63	0.84	0.051
Ours	**27.42**	**0.83**	**0.067**	**31.51**	**0.89**	**0.049**	**29.70**	**0.86**	**0.042**

Table 2. Unconditional genereation results on CelebA-HQ of different models.

Method	FID↓	IS↑	Prec.↑	Recall↑
VAE	31.73	3.18	0.41	0.50
VQGAN	81.01	3.20	0.25	0.29
RV-VAE	33.82	3.19	0.31	0.40
MVU-AE (Ours)	**26.82**	**3.23**	**0.52**	**0.53**

5.3. Qualitative Evaluation

We present the reconstructed images generated by our model on the ADE20K, CelebA-HQ, and FFHQ datasets, as illustrated in Appendix B.1. The results demonstrate high-quality reconstruction with realistic details and natural, vivid colors. The reconstructed images effectively preserve fine-grained features, showcasing the capability of our model to faithfully represent both texture and structure in complex scenes. Fig. 3 shows the results generated by each model unconditionally. Our model produces images with reasonable details in human faces.

5.4. Cross-domain Evaluation

The generalization performance of VAE, VQGAN, WAE, RV-VAE and our model on MS-COCO, LSDIR, and DIV2K is examined. These models are trained on ADE20K with a resolution of 256×256 and tested on the aforementioned datasets. As shown in Fig. 4, our model produces impressive visual results. This demonstrates the accuracy of latent representations and the ability to generalize to cross-domain images while still achieving accurate reconstruction compared to other models. Additional visualization results can be found in the Appendix B.3.

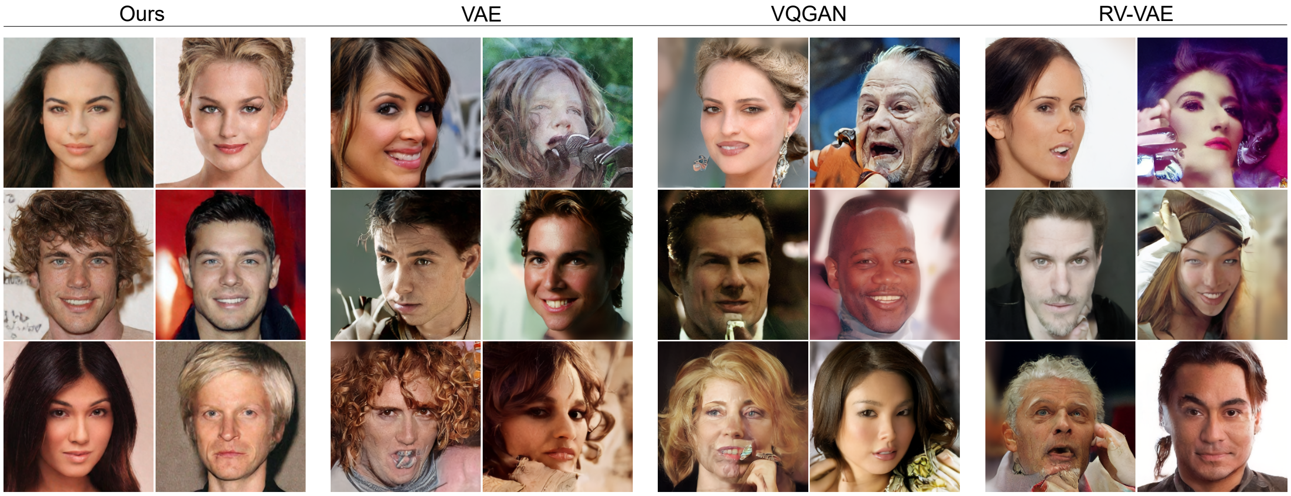

Figure 3. Unconditional generation results on CelebA-HQ from different models.

Table 3. Quantitative cross-domain evaluation. All the models are trained on the ADE20K dataset and tested on MS-COCO, LSDIR and DIV2K datasets.

Method	MS-COCO			LSDIR			DIV2K		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VAE	26.22	0.79	0.082	21.71	0.71	0.110	21.15	0.70	0.109
VQGAN	23.48	0.68	0.221	20.43	0.61	0.282	19.88	0.58	0.292
WAE	25.61	0.78	0.091	21.47	0.69	0.121	20.99	0.68	0.120
RV-VAE	26.20	0.79	0.083	21.55	0.70	0.113	21.08	0.69	0.111
Ours	**26.91**	**0.82**	**0.069**	**21.96**	**0.73**	**0.090**	**21.44**	**0.73**	**0.089**

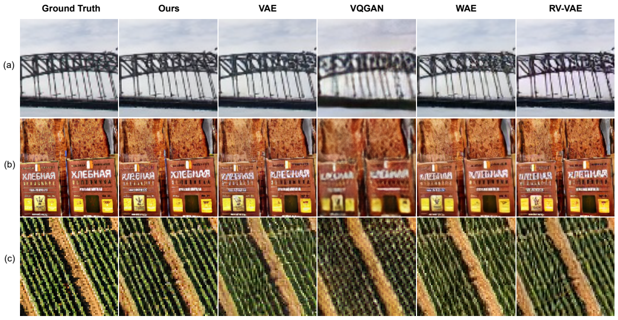

Figure 4. Visualization for cross-domain evaluation. All the models are trained on the ADE20K dataset and tested on (a) MS-COCO, (b) LSDIR, and (c) DIV2K datasets.

Table 4. Interpolability and Variance results on ADE20K, CelebA-HQ and FFHQ from different models.

Method	ADE20K		CelebA-HQ		FFHQ	
	Var.↓	PPL↓	Var.↓	PPL↓	Var.↓	PPL↓
VAE	57.39	12.93	81.81	10.12	95.06	10.05
VQGAN	11.01	18031.42	9.75	7442.83	9.96	7870.31
WAE	103.36	12.25	56.32	10.24	215.29	11.62
RV-VAE	17.76	16.01	26.73	21.55	25.14	11.87
Ours	**0.95**	**10.82**	**0.73**	**9.16**	**0.76**	**8.61**

5.5. Interpolability and Variance Evaluation

To further evaluate the interpolability of the latent space, we compute the Perceptual Path Length (PPL) (Karras, 2019) for our model and compare it with other state-of-the-art models. The results are shown in Tab. 4. Our MVU-AE model achieves a lower PPL score, indicating smoother transitions in the latent space.

The experimental results in Tab. 4 present the variance in the latent representation z across various models. Our proposed MVU-AE model achieves lower variance, significantly surpassing other models. This reduced variance in MVU-AE signifies a more stable and consistent latent representation, which is essential for dependable downstream tasks such as reconstruction and generation. These results confirm that our MVU-AE not only generates high-quality images but also maintains a smooth and interpolative latent space, which is crucial for various downstream tasks such as image interpolation and manipulation.

6. Conclusion

This study addresses key challenges in Variational Autoencoders (VAEs), including latent alignment bias, structural inductive bias, and reconstruction quality. Leveraging classical statistical principles, we propose the $d$eep Minimum Variance Unbiased (d-MVU) Conditions, which optimize latent space by reducing variance while preserving unbiasedness. Our MVU-AE model outperforms existing approaches, demonstrating lower latent variance, improved stability, and enhanced reconstruction and generative capabilities. By integrating deterministic modeling, alternative divergence metrics, and advanced statistical techniques, our work bridges traditional statistics and modern deep learning. This research lays the foundation for future exploration of unbiased estimation in generative modeling, fostering more efficient and interpretable models.

Impact Statement

Our work advances Machine Learning by improving the effectiveness and interpretability of Variational Autoencoders (VAEs) through the Deep Minimum Variance Unbiased conditions, enhancing stability and reliability in generative modeling. Applications include medical imaging, scientific simulations, and content creation, benefiting from improved representation learning and data compression. Overall, our study contributes to the advancement of generative modeling with no immediate ethical concerns beyond standard best practices.

References

Bai, J., Kong, S., and Gomes, C. P. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *Proceedings of International Conference on Machine Learning*, pp. 1383–1398. PMLR, 2022.

Bendekgey, H. C., Hope, G., and Sudderth, E. Unbiased learning of deep generative models with structured discrete representations. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69849–69886. Curran Associates, Inc., 2023.

Blackwell, D. and Girshick, M. A. A lower bound for the variance of some unbiased sequential estimates. *The Annals of Mathematical Statistics*, 18(2):277–280, 1947. ISSN 00034851.

Caterini, A. L., Doucet, A., and Sejdinovic, D. Hamiltonian variational auto-encoder. *Advances in Neural Information Processing Systems*, 31, 2018.

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

Chien, J.-T. Deep bayesian natural language processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 25–30, 2019.

Dai, B., Wang, Z., and Wipf, D. The usual suspects? reassessing blame for vae posterior collapse. In *Proceedings of International Conference on Machine Learning*, pp. 2313–2322. PMLR, 2020.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021.

Halmos, P. R. and Savage, L. J. Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics. *The Annals of Mathematical Statistics*, 20(2):225 – 241, 1949. doi: 10.1214/aoms/1177730032.

Hao, X. and Shafto, P. Coupled variational autoencoder. *arXiv preprint arXiv:2306.02565*, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 3, 2017.

Hu, Y., Luo, C., and Chen, Z. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18219–18228, 2022.

Karras, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Mohamed, S., Jimenez Rezende, D., and Welling, M. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 27, 2014.

Lee, A. U-statistics: theory and practice. statistics: textbooks and monographs, new york, 1990.

Lee, D. B., Min, D., Lee, S., and Hwang, S. J. Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning. In *International Conference on Learning Representations*, 2020.

Lehmann, E. L. and Scheffé, H. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 10(4):305–340, 1950. ISSN 00364452.

Lehmann, E. L. and Scheffé, H. Completeness, similar regions, and unbiased estimation: Part ii. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(3):219–236, 1955. ISSN 00364452.

Li, J., Kang, D., Pei, W., Zhe, X., Zhang, Y., He, Z., and Bao, L. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11293–11302, 2021.

Li, Y., Swersky, K., and Zemel, R. Learning unbiased features. *arXiv preprint arXiv:1412.5244*, 2014.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3730–3738, 2015.

Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lucas, J., Tucker, G., Grosse, R., and Norouzi, M. Understanding posterior collapse in generative latent variable models, 2019.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Mao, Y., Zhang, J., Xiang, M., Zhong, Y., and Dai, Y. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 954–965, 2023.

Nasiri, A. and Bepler, T. Unsupervised object representation learning using translation and rotation group equivariant vae. *Advances in Neural Information Processing Systems*, 35:15255–15267, 2022.

Nicodemou, V. C., Oikonomidis, I., and Argyros, A. Rv-vae: Integrating random variable algebra into variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 196–205, 2023.

Razavi, A., Van den Oord, A., and Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32, 2019.

Salimans, T., Kingma, D., and Welling, M. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of International Conference on Machine Learning*, pp. 1218–1226. PMLR, 2015.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016.

Timofte, R., Agustsson, E., Van Gool, L., Yang, M.-H., and Zhang, L. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 114–125, 2017.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

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

Wolf, C., Karl, M., and van der Smagt, P. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.

Yang, Z.-L., Zhang, S.-Y., Hu, Y.-T., Hu, Z.-W., and Huang, Y.-F. Vae-stega: linguistic steganography based on variational auto-encoder. *IEEE Transactions on Information Forensics and Security*, 16:880–895, 2020.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 633–641, 2017.

A. Proof Details

A.1. Proof: If there is no leftover subspace in $\hat{z}$, then $\hat{z}$ is complete.

Proof.(By contradiction) Suppose $\hat{z}$ is not complete, that is, there exists a non-trivial function of $\hat{z}$ with zero-average. Formally,

$$\mathbb{E}[f(\hat{z})] = 0 \quad \text{and} \quad f(\hat{z}) \not\equiv 0. \tag{26}$$

Then there exists z_0 such that $f(z_0) \neq 0$. Let $f^+ = \{f(\hat{z}) : f(\hat{z}) > 0\}$ and $f^- = \{f(\hat{z}) : f(\hat{z}) < 0\}$. For each points $(\hat{z}_0, f(\hat{z}_0))$, $\hat{z}$ can roam in the hyperplane, akin the set $S_{f(\hat{z}_0)} = \{\hat{z} : f(\hat{z}) = f(\hat{z}_0)\}$ without changing $\mathbb{E}[f(\hat{z})]$. In order to keep $\mathbb{E}[f(\hat{z})] = 0$, we could let $\hat{z}$ move across the hyperplanes and only to adjust its corresponding points in f^+ or f^-. For instance, if $f(\hat{z}) \in f^+$, change some points in f^- correspondingly could maintain $\mathbb{E}[f(\hat{z})] = 0$. Therefore, $\hat{z}$ can move freely in the direction that those hyperplanes form and determine, that is, $\hat{z}$ has free degrees which form a leftover subspace.

A.2. Proof: $\hat{z}$ in VAE is not sufficient nor complete in d-MVU conditions sense.

Proof. Since $q_\phi\left(z|x\right) \to \mathcal{N}\left(0, I\right)$, which means some information in X is lost, thus $I\left(x|z\right) \neq 0$.Therefore,

$$I\left(z^*; x|z\right) = H\left(x|z\right) - H\left(x|z, z^*\right) = H\left(x|z\right) > 0 \tag{27}$$

violating the definition of sufficiency. As for completeness, consider the KL term of VAE loss:

$$\frac{1}{2}\sum_j \left(\mu_j\left(x\right)^2 + \sigma_j\left(x\right)^2 - 2\log\sigma_j\left(x\right) - 1\right), \tag{28}$$

which encourages μ_j and σ_j to be 0 and 1 respectively. Since $\hat{z}\left(x\right) = \mu\left(x\right)$ and $\mathbb{E}_{x\sim p(x|z^*)}\left[\hat{z}\left(x\right)\right] = 0$, define a subspace $\mathcal{U} = \{j : z_j \approx 0\}$, then we can construct a function

$$f\left(\hat{z}\right) = \sum_{j\in\mathcal{U}} z_j. \tag{29}$$

Every z_j is sampled from $\mathcal{N}\left(0, 1\right)$, thus

$$\mathbb{E}\left[f\left(\hat{z}\right)\right] = 0 \tag{30}$$

but $f\left(\hat{z}\right) \neq 0$. This proves that there is a leftover subspace $\mathcal{U}$ of $\hat{z}$ that being ignored during training. Therefore, the latent representation $\hat{z}$ is not sufficient nor complete.

A.3. Proof: $\hat{z}$ in VQ-VAE is both sufficient and complete in d-MVU conditions sense.

Proof. In VQVAE, a codebook $\{e_k\}_{k=1}^K$ is learned and

$$z_q\left(x\right) = \arg\min_{e_k} ||E\left(x\right) - e_k||. \tag{31}$$

We define the residual $r\left(x\right) = E\left(x\right) - z_q\left(x\right)$, consider the optimal situation that the STE trick works so $E\left(x\right)$ is correctly optimized and thus become unbiased. Then

$$\mathbb{E}_x\left[z_q\left(x\right)\right] = \mathbb{E}_x\left[E\left(x\right) - r\left(x\right)\right] = z^* - \mathbb{E}_x\left[r\left(x\right)\right]. \tag{32}$$

$\mathbb{E}\left[z_q\left(x\right)\right] = z^*$ if and only if $\mathbb{E}\left[r\left(x\right)\right] = 0$. Unfortunately, nothing ensures this will always happen because clustering cannot ensure zero-expectation residual. Therefore, z_q is not an unbiased estimator of z^*. As for sufficiency,

$$I\left(z^*; x|z_q\right) = H\left(x|z_q\right) - H\left(x|z_q, z^*\right) = H\left(x|z_q\right) \approx H\left(x|E\left(X\right)\right) \to 0 \tag{33}$$

The approximation means if the codebook is well trained, it can store the same information as $E\left(X\right)$ does. And if encoder is well-trained, all information of dataset X has been stored in $E\left(X\right)$. Thus z_q is a sufficient latent representation. As for completeness, we claim that there is no free degrees in z_q. This is because, if there exists a free degree, which means z_q can be any value in that dimension with zero average. However, this violates the quantization loss because each degree in z_q determines the exact position of z_q. Moving freely has possibility to affect the assignments of z_q thus increasing $\mathcal{L}_{\text{quant}}$ as we have proved in Sec. 3. Hence the quantization loss prevents leftover subspace in z_q and ensures the completeness.

A.4. Lemma 1. Proof Details

Lemma 1. Under the constraint of $\mathcal{L}_{\text{quant}}$, any shift in $\hat{z}$ will result in an increase in $\mathcal{L}_{\text{quant}}$.

Proof. Let $\hat{z}_\delta(x) = \hat{z}(x) + \delta$, and for each x, let the best matching centroid be

$$k^*(x;\delta) = \arg\min_k \|\hat{z}_\delta(x) - e_k\|^2.$$

Define the quantization loss as:

$$\mathcal{L}_{\text{quant}}(\delta) = \lambda \mathbb{E}_x \left[\|\hat{z}_\delta(x) - e_{k^*}(\delta)\|^2 \right]. \tag{34}$$

Since $e_{k^*}(\delta)$ is the centroid of all points assigned to cluster k, it follows that:

$$\begin{aligned} e_{k^*}(\delta) &= \frac{1}{|S_k(\delta)|} \sum_{x \in S_k(\delta)} \hat{z}_\delta(x), \\ \text{where} \quad S_k(\delta) &= \{x : k^*(x;\delta) = k\}. \end{aligned} \tag{35}$$

Let $c_k(\delta) = \frac{1}{|S_k(\delta)|} \sum_{x \in S_k(\delta)} \hat{z}_\delta(x)$ and $c_k(0) = \frac{1}{|S_k(\delta)|} \sum_{x \in S_k(\delta)} \hat{z}_0(x)$. Here, $c_k(\delta)$ represents the centroid of points in $\hat{z}_\delta$ assigned to cluster k, under the decision boundary determined by $\hat{z}_\delta$. Similarly, $c_k(0)$ represents the centroid of points in $\hat{z}_0$ assigned to cluster k, under the same decision boundary. Then:

$$\begin{aligned} c_k(\delta) &= \frac{1}{|S_k(\delta)|} \sum_{x \in S_k(\delta)} \hat{z}_\delta(x) \\ &= \frac{1}{|S_k(\delta)|} \sum_{x \in S_k(\delta)} \big(\hat{z}(x) + \delta(x)\big) \\ &= c_k(0) + \delta(x). \end{aligned} \tag{36}$$

Since $c_k(\delta) = e_{k^*}(\delta)$, as both denote the centroid of cluster k under the decision boundary of $\hat{z}_\delta$, the quantization loss becomes:

$$\begin{aligned} \mathcal{L}_{\text{quant}}(\delta) &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(\delta)} \|\hat{z}_\delta(x) - e_{k^*}(\delta)\|^2 \\ &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(\delta)} \|\hat{z}(x) + \delta - \big(c_k(0) + \delta\big)\|^2 \\ &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(\delta)} \|\hat{z}(x) - c_k(0)\|^2. \end{aligned} \tag{37}$$

If the decision boundary of the clusters does not change after the shift δ, we have:

$$\begin{aligned} \mathcal{L}_{\text{quant}}(\delta) &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(\delta)} \|\hat{z}(x) - c_k(0)\|^2 \\ &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(0)} \|\hat{z}(x) - c_k(0)\|^2 \\ &= \mathcal{L}_{\text{quant}}(0). \end{aligned} \tag{38}$$

However, if the decision boundary of the clusters changes, we have:

$$\begin{aligned}\mathcal{L}_{\text{quant}}(\delta) &= \lambda \sum_{k=1}^{K} \sum_{x \in S_k(\delta)} \|\hat{z}(x) - c_k(0)\|^2 \\ &\geq \lambda \sum_{k=1}^{K} \sum_{x \in S_k(0)} \|\hat{z}(x) - c_k(0)\|^2 \\ &= \mathcal{L}_{\text{quant}}(0).\end{aligned} \tag{39}$$

Therefore, we have proven that:

$$\mathcal{L}_{\text{quant}}(\delta) \geq \mathcal{L}_{\text{quant}}(0). \tag{40}$$

A.5. Lemma 2. Proof Details

Lemma 2. A shift in $\hat{z}$ will not increase the reconstruction loss $\mathcal{L}_{\text{rec}}$.

Proof. Consider the reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_x \left[\|x - D_\theta(\hat{z}(x))\|^2 \right]. \tag{41}$$

Define the shifted reconstruction loss as:

$$\begin{aligned}\mathcal{L}_{\text{rec}}(\delta) &= \mathbb{E}_x \left[\|x - D_\theta(\hat{z}_\delta(x))\|^2 \right] \\ &= \mathbb{E}_x \left[\|x - D_\theta(\hat{z}(x) + \delta)\|^2 \right] \\ &= \mathbb{E}_x \left[\|x - D_{\theta'}(\hat{z}(x))\|^2 \right],\end{aligned} \tag{42}$$

where θ' represents the updated decoder parameters corresponding to the shifted latent representation.

It is worth noting that if $D_{\theta'}(z) = D_\theta(z + \delta)$, the reconstruction loss $\mathcal{L}_{\text{rec}}$ will remain unchanged. This adjustment can be achieved by updating the decoder parameters during training. Therefore, a shift δ applied to $\hat{z}$ does not affect $\mathcal{L}_{\text{rec}}$.

B. More Visualization Results

B.1. More reconsturction results

We provide reconstruction results on the ADE20K, CelebA-HQ and FFHQ datasets in Fig. 6. The results indicate that our approach consistently achieves high-fidelity reconstructions, even for complex and diverse image distributions.

B.2. More generation results

We provide additional unconditional generation results on the CelebA-HQ datasets in Fig. 5, showcasing the qualitative performance of our proposed MVU-AE model, diverse, and realistic images. Compared to existing methods, MVU-AE reduces artifacts, preserves structural consistency, and captures fine details like facial features, textures, and lighting. Its stable latent space ensures consistency while maintaining diversity, avoiding mode collapse. Future work may explore scaling to larger datasets or integrating diffusion-based methods for further improvement.

B.3. More cross-domain results

We provide additional cross-domain reconstruction results on MS-COCO, LSDIR, and DIV2K datasets, as shown in Fig. 7 respectively. In comparison of different datasets, our model outperforms other models in reconstructing various elements such as animals, buildings, text, and landscapes. For instance, our model reproduces the shape of the railing more accurately, and the fox's eyes are more faithful. These results highlight the potential of the d-MVU conditions in cross-domain. The flexible and informative latent representation enhance the capability to capture more details of image.

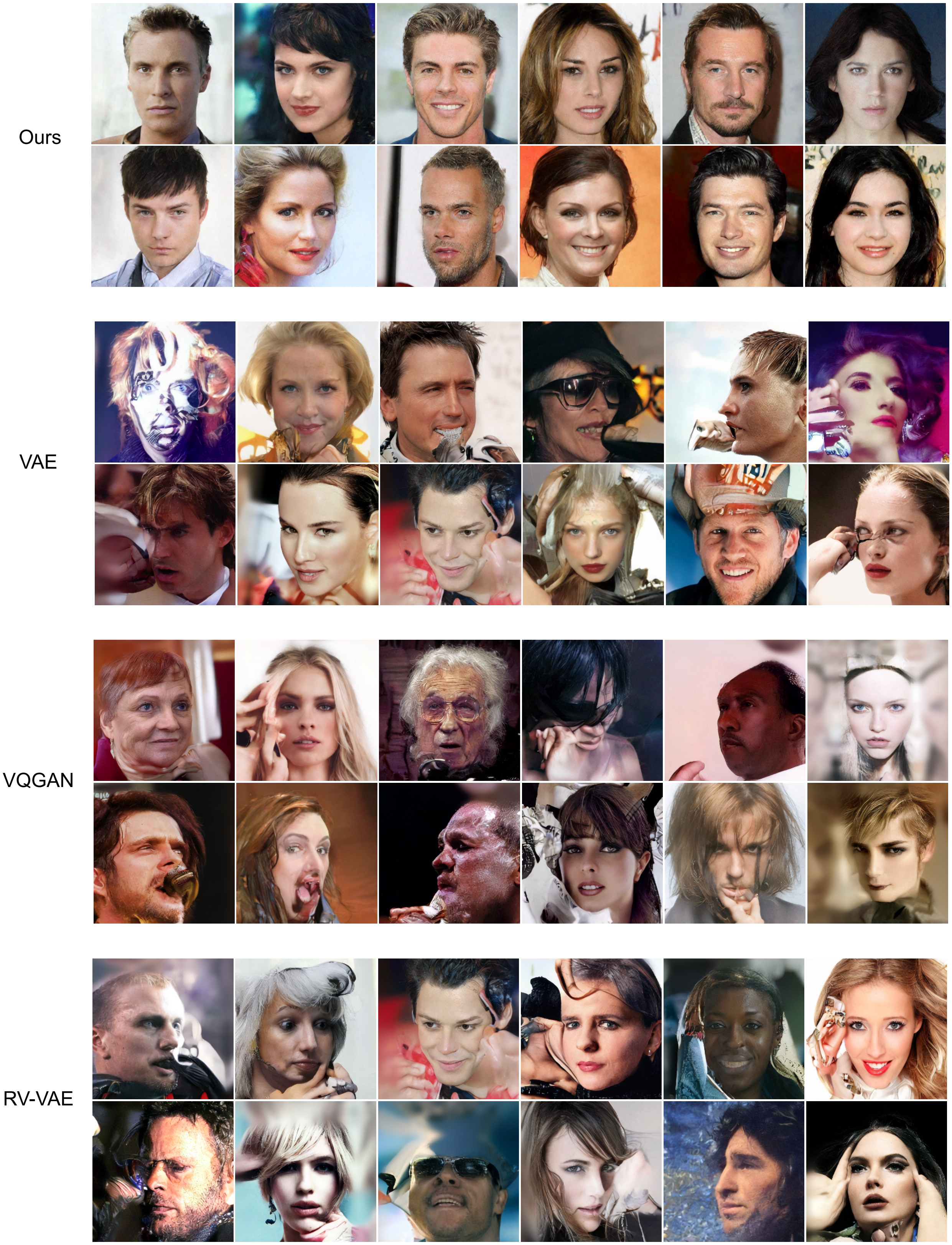

Figure 5. Additional unconditional generation results on CelebA-HQ dataset

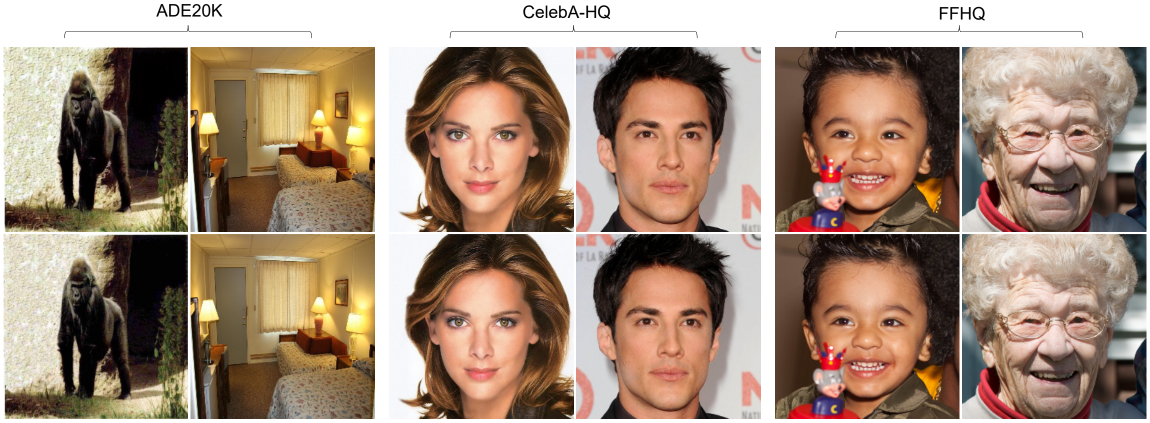

Figure 6. Top: original 256 × 256 images in ADE20K, CelebA-HQ and FFHQ datasets. Bottom: reconstructed images from our MVU-AE.

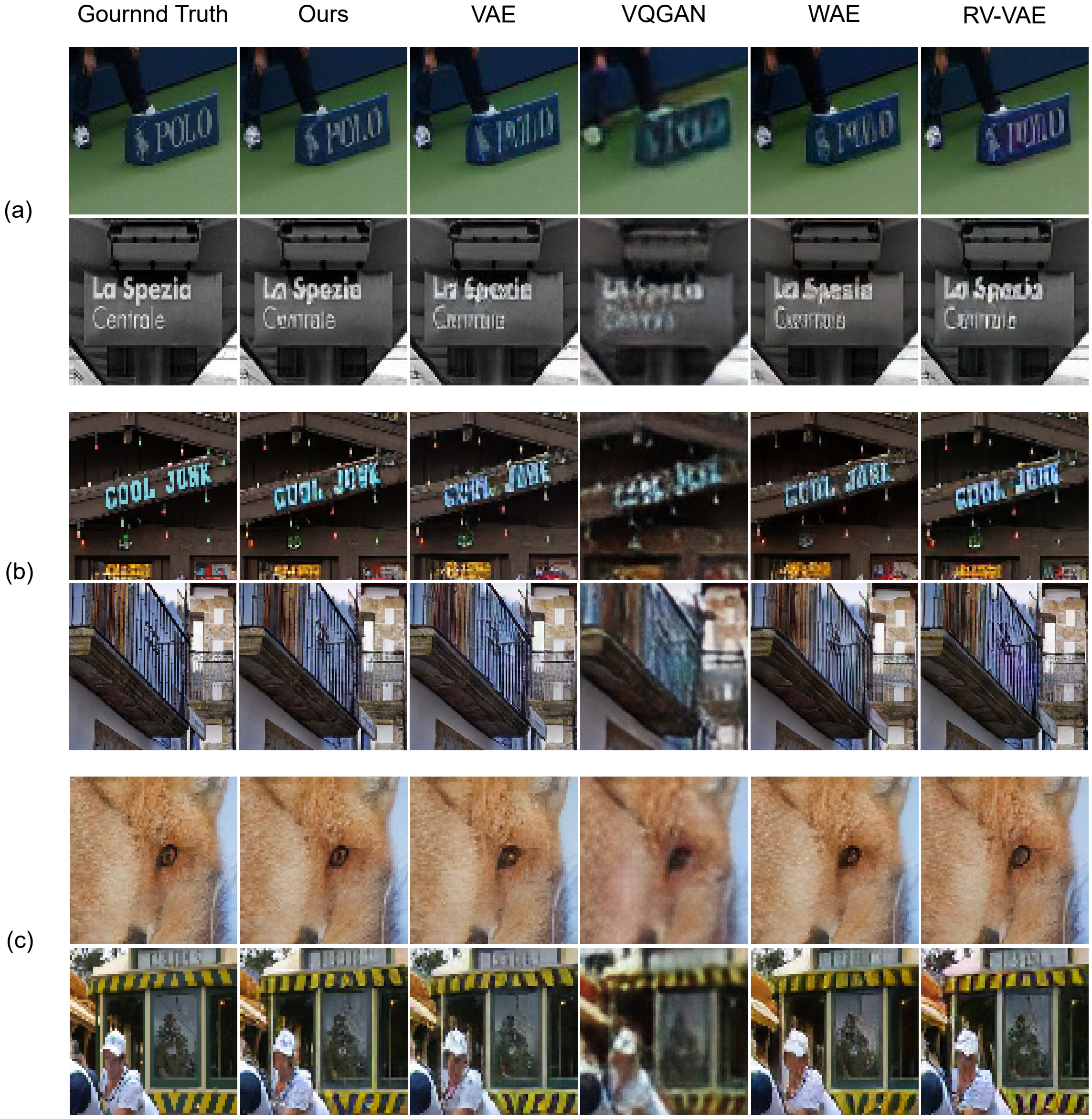

Figure 7. Additional cross-domain reconstruction results on (a) MS-COCO, (b) LSDIR, (c) DIV2K datasets