

Received 9 July 2025, accepted 21 July 2025, date of publication 23 July 2025, date of current version 29 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3592163

## RESEARCH ARTICLE

# Decoupled Latent Diffusion Model for Enhancing Image Generation

HYUN-TAE CHOI<sup>1</sup>, KENSUKE NAKAMURA<sup>2</sup>, AND BYUNG-WOO HONG<sup>3</sup>

Department of Artificial Intelligence, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Byung-Woo Hong (hong@cau.ac.kr)

This work was supported in part by Korean Government [Ministry of Science and ICT (MSIT)] under Grant IITP-RS-2021-II211341; and in part by the Artificial Intelligence Graduate School, Chung-Ang University under Grant NRF-RS-2023-00251366.

**ABSTRACT** Latent Diffusion Models have emerged as an efficient alternative to conventional diffusion approaches by compressing high-dimensional images into a lower-dimensional latent space using a Variational Autoencoder (VAE) and performing diffusion in that space. In standard Latent Diffusion Model (LDM), the latent code is formed by sampling from a Gaussian distribution (i.e., combining both the mean and the standard deviation), which helps regularize the latent space but appears to contribute little beyond the deterministic component. Motivated by recent empirical observations that the decoder relies primarily on the latent mean, our work reexamines this paradigm and proposes a decoupled latent diffusion model that focuses on a simplified latent representation. Specifically, we compare three configurations: (i) the standard latent code, (ii) a concatenated representation that explicitly preserves both mean and variance, and (iii) a deterministic mean-only representation. Our extensive experiments on multiple benchmark datasets demonstrate that, when compared to the standard approach, the mean-only configuration not only maintains but in many cases improves synthesis quality by producing sharper and more coherent images while reducing unnecessary noise. These findings suggest that a simplified, deterministic latent representation can yield more stable and efficient generative models, challenging the conventional reliance on latent sampling in diffusion-based image synthesis.

**INDEX TERMS** Denoising diffusion model, latent representation, image generation.

## I. INTRODUCTION

Diffusion models have recently emerged as state-of-the-art models across diverse domain such as image synthesis, and audio generation. In particular, Denoising Diffusion Probabilistic Model (DDPM) [1] and their variants have achieved remarkable image quality by learning to reverse a noise-adding process [2], [3], [4], [5]. However, a key challenge with DDPM lies in their high computational cost. Their operation in high-dimensional pixel space, coupled with the requirement for extensive denoising steps to produce samples, results in considerable computational overhead concerning memory and time. This inherent computational load has restricted their broader deployment in applications demanding high-resolution image generation or

real-time synthesis, notwithstanding their excellent generative capacities.

To address this, Latent Diffusion Models (LDM) [6] introduce a two-stage framework in which images are first converted to a lower-dimensional latent space via a Variational Auto-Encoder (VAE) [7] and then refined using a diffusion process within that space. This approach substantially reduces computational cost while maintaining image fidelity.

A key element in this framework is the latent representation produced by the VAE. Typically, the encoder generates a Gaussian distribution by estimating both its mean vector and the corresponding standard deviation, and a latent code is obtained by sampling via the reparameterization trick. Although this stochasticity theoretically enables diverse reconstructions, empirical studies have shown that the standard deviation is often underutilized, with the model

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai<sup>4</sup>.

relying mostly on the mean [8], [9]. This issue is related to the well-known problem of posterior collapse in VAE [8], [10], where the model learns to ignore the latent variables (or parts of them, such as the variance component) and reconstructs outputs based primarily on other information, making the latent space less informative or redundant. This observation raises an important question: Is the stochastic component of the latent code truly necessary for high-quality image synthesis, especially in the context of latent diffusion?

Motivated by these recent empirical observations that the decoder is primarily based on the latent mean, we hypothesize that the latent mean alone is sufficient to capture essential information for high-quality image synthesis. To investigate this phenomenon, our work reexamines the conventional approach, introducing a decoupled latent diffusion model centered on a simplified latent representation. Our study specifically involves comparing three distinct configurations for the latent code: (i) the standard latent representation using both mean and variance, (ii) concatenated representation that preserves the mean and variance as separate channels, and (iii) deterministic mean-only representation which removes the stochastic variance from the standard latent representation. Our analysis spans multiple datasets and includes both quantitative and qualitative comparisons.

Our results demonstrate that using deterministic representation not only preserves image quality but often yields more coherent and stable outputs. In contrast, incorporating the sampled variance can introduce unnecessary noise and, in some cases, even degrade the quality of the generated images. These findings challenge the conventional assumption that latent variance sampling is essential and suggest that a simplified, deterministic representation may reduce variance-induced artifacts without compromising performance.

By focusing on the latent structure itself without modifying the diffusion architecture or training objectives, we provide a new perspective on how generative quality can be retained while simplifying the latent code. Our contributions are as follows:

- We challenge the conventional paradigm in Latent Diffusion Model by demonstrating that the deterministic latent mean alone is sufficient for high-quality image synthesis, thereby eliminating the need for stochastic sampling from the VAE's encoded distribution.
- Through extensive empirical evaluation across multiple datasets, we show that this simplified mean-only approach not only achieves comparable or superior generative performance but also yields more coherent and stable outputs compared to traditional stochastic latent representations.
- Our detailed analyses provide novel insights, revealing that the stochastic variance component of latent codes contributes negligibly to synthesis quality and can, in fact, introduce detrimental noise, advocating for more efficient and robust LDM architectures.

Our study shows the effectiveness of deterministic latent representations in diffusion-based generative modeling and points toward the way for more efficient and robust architectures.

The remainder of the paper is constructed as follows: Section II provides a brief overview of Variational auto-encoder, Diffusion Model and Latent Diffusion Model. Section III details of those generation models with their essential equations. Section IV describes our suggested method. Section V presents and discusses the quantitative and qualitative results. Finally, Section VI concludes the paper and suggests potential future work.

## II. RELATED WORKS

### A. VARIATIONAL AUTO-ENCODER AND LATENT REPRESENTATION

A principled framework is suggested by VAE [7] for learning compact latent representations of high-dimensional data by mapping input to low-dimensional latent, following a Gaussian distribution characterized by mean and variance. In theory, the stochastic component (the variance) captures uncertainty and enhances the representation of sampled output. However, in practice, powerful decoders often ignore this variance, a phenomenon known as posterior collapse [8]. This observation has motivated research into alternative latent representations that do not rely on an explicit variance term. Vector Quantized VAE (VQ-VAE) [11] replaces the continuous Gaussian latent with discrete latent which is called codebook. By removing the need for the encoder to predict a variance, VQ-VAE prevents posterior collapse, effectively forcing the latent space to retain essential information. Building on this, VQ-VAE2 [12] achieves more diverse and high-fidelity images by modeling a multi-level latent code. Concurrently, improved VAE architectures have shown progress in bridging the quality gap with other generative models [13]. Another line of work, hierarchical VAEs such as NVAE [14], stack multiple latent groups and introduce architectural innovations (e.g., residual parameterization of the latent distributions) to improve generative performance. NVAE achieved state-of-the-art likelihoods among VAE-based models by carefully designing the network and employing techniques like KL balancing to prevent entire groups of latents from turning off.

Although those VAE frameworks continue to utilize the full latent code, these approaches demonstrate that removing or constraining the stochastic element in the latent space (e.g. using deterministic encoders or discrete latents) can still yield powerful generative models. This insight lays the groundwork for our investigation into using only the mean of the latent code in a diffusion-model framework.

### B. LATENT DIFFUSION MODELS FOR IMAGE SYNTHESIS

DDPM has emerged as a powerful image generation model, achieving state-of-the-art image synthesis quality [1], [3], [15]. The diffusion model defines a forward process

with Markov chain [16] that progressively corrupts an image by incrementally adding noise and defines a reverse process that remove the noise to recover the original data distribution. Notably, DDPM demonstrated that iterative denoising can produce high-fidelity images, even surpassing GAN without mode collapse [3]. Further improvements introduced techniques like classifier guidance to trade diversity for realism [3] and optimized noise schedules for faster sampling [2]. Crucially, the original DDPM fixed the variance and trained the network to predict only mean at each denoising step, a design that simplified training without degrading sample quality in practice, since the estimation of variance is not important to the final result.

Despite their success, diffusion models operating in the original image space suffer very high computational costs because of the large dimensionality and the need for many iterative steps for diversity. This motivated the development of LDM [6], which performs the diffusion process in a learned lower-dimensional latent space. In LDM, a high-resolution image is first compressed into a latent code by a pre-trained auto-encoder, and diffusion process is applied in this compact low-dimensional latent domain. LDM shows that applying diffusion in a suitably chosen latent space can retain image detail while dramatically reducing computational requirements. The auto-encoder is trained with perceptual and adversarial losses to ensure the latent representations preserve important visual features. To avoid unbounded latent distributions, the auto-encoder latent is lightly regularized (e.g., a small KL penalty toward  $\mathcal{N}(0, \mathbf{I})$  or vector quantization) during training. This strategy yields a “near-optimal point between complexity reduction and detail preservation”, enabling high-resolution image generation with far less computational cost than image-space diffusion.

This careful regularization suggests that the stochastic variance may play a limited role in the final synthesis, motivating our investigation into deterministic representation. By focusing solely on the deterministic mean, we aim to simplify the latent diffusion process further while maintaining, or even improving, reconstruction quality.

### C. EMPIRICAL STUDIES ON DETERMINISTIC LATENT REPRESENTATIONS

A growing body of empirical work directly evaluates the extreme case of using only the posterior mean (i.e., without noise) in latent representations. Early evidence came from deterministic variants of VAE, which showed that removing latent sampling has little adverse effect on performance. Reference [17] report that VAE with deterministic encoder achieves log-likelihoods on par with a stochastic Gaussian encoder, indicating that the random latent variance was largely unnecessary. Reference [9] provide a comprehensive study demonstrating that regularized auto-encoder can produce generated samples comparable to or better than those from standard VAE. In their experiments, these deterministic models learned smooth, meaningful latent spaces and even

improved sample quality in some settings. Reference [18] show that training an auto-encoder deterministically (by embedding it in an invertible flow) produces substantially sharper and higher-fidelity images than a variational counterpart, especially when the latent dimensionality is limited. Furthermore, generative models with quantized of fixed latents support that stochastic noise is not required for rich generation. VQ-VAE-2 [12] attains GAN-level image quality on ImageNet [19] using only discrete latent codes (without continuous noise), underlining that deterministic code can support diversity and high-fidelity.

Even during training, certain techniques deliberately operate in a mean-only regime: [20] introduce a capacity annealing schedule in  $\beta$ -VAE that initially constrains the KL term to zero, effectively training without stochastic information (i.e. using only the deterministic reconstruction loss) and gradually increases effect of KL term. Such practices further confirm that the decoder can rely almost entirely on the deterministic component without collapse.

In summary, these regularized auto-encoders, quantized latent models, and training heuristics studies all converge to the finding that generative models can perform remarkably well with deterministic representations. The diversity and quality of outputs remain high [9], [12], validating our hypothesis that injecting stochastic noise into latent is often not strictly necessary to achieve high-performance generative results. Each of these empirical results reinforces the intuition that the encoded mean carries the essential information needed for generation, while the omitted noise tends to have only minor influence on the overall output distribution. Deep learning’s influence extends far beyond generative modeling, driving significant advancements across various image processing applications. Recent comprehensive surveys demonstrate this broad impact, covering tasks from image restoration, including denoising [21], to intricate analyses like image matching [22], [23]. This widespread progress emphasizes deep learning’s versatility and power in extracting insights and enhancing visual data across diverse domains.

## III. PRELIMINARIES

In this section, we outline the fundamental concepts underlying VAE, DDPM and LDM. These models form the backbone of our work, and a clear understanding of their standing formulations is essential for appreciating our investigation into simplifying the latent representation by relying solely on the mean.

### A. VARIATIONAL AUTO-ENCODER

VAE [7] provide a probabilistic framework for learning compact latent representations of high-dimensional data. Given an input  $x$  drawn from the data distribution  $q(x)$ , the encoder  $E(x)$  maps  $x$  into the parameters of a Gaussian latent distribution characterized with mean  $\mu$  and standard

deviation  $\sigma$ :

$$q(z|x) = \mathcal{N}(z; \mu, \sigma^2 \mathbf{I}), \quad (1)$$

where  $z$  is latent code obtained using the reparameterization trick [7], [24]

$$z = \mu + \sigma \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (2)$$

where  $\odot$  is the element-wise product of two vectors, also known as the Hadamard product. The decoder  $D(z)$  reconstructs the input as  $\hat{x} = D(z)$ . The model is trained by minimizing a loss function ( $L_{\text{VAE}}$ ) that combines a reconstruction term (e.g., the  $L_2$  norm between  $x$  and  $\hat{x}$ ) with a Kullback-Leibler divergence term that regularizes the approximate posterior  $q(z|x)$  to be close to a standard Gaussian:

$$L_{\text{VAE}} = \mathbb{E}_{q(z|x)} \|x - D(z)\|^2 + \beta \cdot \mathbf{KL}(q(z|x) \| \mathcal{N}(0, \mathbf{I})) \quad (3)$$

In theory, the stochastic component  $\sigma \odot \epsilon$  in Eq. (2) is designed to capture uncertainty and variability in the data. However, in practice, powerful decoders can lead to under-utilization of the variance—a phenomenon known as posterior collapse [8]. This observation has motivated subsequent research into simplified latent representations that rely primarily on the deterministic mean  $\mu$ , while still preserving the essential information for reconstruction. Such insights are critical for the present work, which investigates the efficacy of using only the latent mean in a diffusion-model framework.

### B. DENOISING DIFFUSION PROBABILISTIC MODEL

DDPMs are generative models that define a gradual noising process followed by a corresponding denoising process. Starting from a data  $x_0$  drawn from the true data distribution  $q(x_0)$ , the forward process progressively corrupts the sample by adding Gaussian noise over  $T$  time steps. This process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I}), \quad (4)$$

where  $\beta_t$  is a variance schedule that increase with  $t$ , ensuring that the data become increasingly noisy, until at  $t = T$ . At time  $T$ , the distribution is close to a standard Gaussian. We can re-define  $\alpha_t$  as  $\alpha_t = 1 - \beta_t$  and define the cumulative product  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ , which allows to relate  $x_t$  directly to  $x_0$  as follows:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (5)$$

The reverse (denoising) process is modeled as a Markov chain that starts from a noise  $x_T$  which is sampled from  $\mathcal{N}(0, \mathbf{I})$  and iteratively reconstruct  $x_0$ . The joint distribution is given by:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t), \quad (6)$$

where  $p(x_T) = \mathcal{N}(x_T; 0, \mathbf{I})$ ,

and each reverse transition is parameterized as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)), \quad (7)$$

where  $\mu_\theta$  and  $\Sigma_\theta$  are functions learned by a neural network. The overall training objective is to minimize the discrepancy between the true forward process and the learned reverse process, typically leading to a simplified objective:

$$L = \mathbb{E} \|\epsilon - \epsilon_\theta(x_t, t)\|^2, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (8)$$

with  $x_t$  expressed in closed form as a function of  $x_0$  and noise  $\epsilon$ :

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad (9)$$

which allows to directly sample noisy data  $x_t$  at time step  $t$  directly from the clean image  $x_0$ .

Due to the iterative noise addition and removal in image space, DDPM incur high computational costs. To mitigate this, the framework has been extended to LDM [6], which employ a VAE to compress images into a lower-dimensional latent space for more efficient training and inference.

### C. LATENT DIFFUSION MODEL

While DDPM operates directly in image space, LDM reduces computational overhead by performing diffusion in a lower-dimensional latent space. A pretrained VAE is employed to compress the high-dimensional image  $x$  into a latent representation  $z$ :  $z = E(x)$ , where the encoder  $E$  outputs parameters of a Gaussian distribution, typically represented as:

$$z = \mu + \sigma \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (10)$$

with  $\mu$  representing the deterministic component and  $\sigma \odot \epsilon$  the stochastic noise. The diffusion process is then applied in this latent space. Specifically, a forward noising process is defined on  $z$  similar to the pixel-space process:

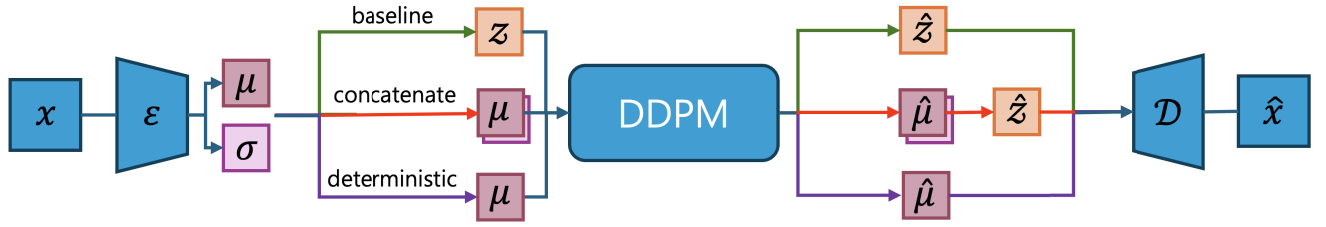
$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (11)$$

using the same cumulative product  $\bar{\alpha}_t$  as defined above. The reverse process is parameterized as:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)). \quad (12)$$

Once the reverse diffusion yields the denoised latent  $z_0$ , the VAE decoder  $D$  reconstructs the high-resolution image  $\hat{x} = D(z_0)$ .

Notably, while the standard LDM use the full latent code, recent empirical observations indicate that the VAE decoder is predominantly conditioned on the deterministic mean  $\mu$ . This finding has motivated our investigation into whether deterministic representation is sufficient for high-quality synthesis, potentially simplifying the latent space without sacrificing reconstruction performance.



**FIGURE 1.** Pipeline of our proposed method. Baseline corresponds to the original LDM latent code; Concatenated denotes the channel-wise concatenation of the mean and standard deviation; and Deterministic represents the use of only the VAE encoder’s mean. Each configuration follows the same overall pipeline from latent input to reconstructed output. In the pipeline,  $x$  is the input image, and  $\epsilon$  and  $\mathcal{D}$  denote the encoder and decoder of the VAE, respectively. The encoder outputs the mean  $\mu$  and standard deviation  $\sigma$ . Each colored arrow represents the flow of a specific configuration.

#### IV. METHOD

In this work, we extend the standard LDM framework by re-examining the latent representation generated by the VAE. In the original formulation, the latent code is sampled from a Gaussian distribution, combining both the mean and the standard deviation using the reparameterization trick. However, empirical observations indicate that the stochastic noise from the standard deviation is underutilized, and the decoder appears to rely predominantly on the latent mean. Motivated by this insight, we propose alternative latent configurations that either decouple or entirely remove the stochastic component. By keeping the diffusion process itself unchanged, our goal is to rigorously evaluate how different latent representations affect synthesis quality, while isolating the effect of latent stochasticity. The overall pipeline is represented in Fig. 1.

##### A. PROPOSED DECOUPLE LATENT DIFFUSION

Recall that in the original LDM framework [6], the latent code  $z$  is constructed with the mean  $\mu$  and standard deviation  $\sigma$  of a Gaussian distribution using Eq. (2):

$$z = \mu + \sigma \odot \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(0, \mathbf{I}), \quad (13)$$

which is then used as input of the diffusion process for image synthesis. Based on our observations, we hypothesize that the final generated output is predominantly determined by the deterministic component mean  $\mu$ . To investigate this hypothesis, we explore three latent representation configurations: (1) **Full Latent Code** ( $z$ ), (2) **Concatenated Representation**, and (3) **Deterministic Representation**. The first configuration, Full Latent Code ( $z$ ), is the baseline and uses the standard formulation in Eq. (13), which is the original latent representation in LDM. In the Concat Representation configuration, explicitly preserves both the mean and the standard deviation by setting:

$$z = \text{Concat}(\mu, \sigma), \quad (14)$$

where Concat represents concatenating the mean and the standard deviation channel-wise. This configuration effectively doubles the latent dimensionality by providing separate channels for the deterministic and stochastic components. For instance, if mean and standard deviation each

have dimensions [Batch size, Channels, Height, Width], the resulting concatenated latent code will have dimensions [Batch size,  $2 \times$  Channels, Height, Width]. In the final configuration Deterministic Representation configuration, the stochastic component is completely removed by setting

$$z = \mu, \quad (15)$$

which is equivalent to  $z = \mu + 0 \cdot \sigma \odot \epsilon$ . This variant retains the same dimensionality as the full latent code but omits the injected noise.

In this article, the term “decoupled” refers to our approach of separating or independently analyzing the distinct components of the VAE’s latent representation that are fed into the LDM. Conventionally, the latent code  $z$  for an LDM is derived from the VAE encoder as a sample from a Gaussian distribution, typically formulated as  $z = \mu + \sigma \odot \epsilon$ . This standard approach inherently couples the generation process to both the deterministic information (mean) and the stochastic information (derived from standard deviation through sampling). Our “decoupled” approach investigates what happens when we break or modify this traditional coupling. Specifically, we decouple the latent representation in the following ways: (1) **Challenging the Stochastic Coupling**: Our primary focus is on decoupling the generation process from the inherent stochasticity introduced by sampling from the variance component. We hypothesize that the deterministic latent  $\mu$  alone might contain sufficient information for high-quality image synthesis. By introducing a deterministic mean-only representation, we essentially “decouple” the diffusion model’s input from the sampled stochastic noise originating from  $\sigma$ . This allows us to directly test if the mean alone is a sufficient condition for effective generation, without the potential complications or redundancy introduced by the variance. (2) **Separate Component Analysis**: Even in our “concatenated representation” configuration, where both  $\mu$  and  $\sigma$  are present (as  $\mu$  and  $\sigma$  concatenated channel-wise), we are still “decoupling” them from the sampling process that creates a single  $z$ . Instead of being combined through sampling, they are presented as distinct, separate information channels to the diffusion model. In essence, “decoupled” signifies our move away from a fixed, combined latent representation towards an exploration

of isolated or independently presented latent components to determine their individual contributions and necessity for high-quality image generation in Latent Diffusion Models. This enables a more precise understanding of the role of mean and variance in the generative process. Our goal is to assess whether removing the stochastic component standard deviation adversely affects synthesis quality or if the mean alone is sufficient for high-quality image generation.

For each configuration, the corresponding latent code is fed into the same diffusion model architecture with identical training hyper-parameters (e.g., learning rate, scheduler, etc.), except that the concatenated configuration requires twice the number of input channels. We deliberately do not modify the diffusion process; our focus is solely on the impact of the latent representation on the generative quality. In Fig. 1, the first configuration (baseline) is shown in the top branch, the second configuration (concatenated) in the middle branch, and the third configuration (deterministic) in the bottom branch. Although all branches follow the same overall pipeline from latent input to reconstructed output, they differ solely in the manner in which the latent code is formed prior to diffusion.

## B. TRAINING AND INFERENCE

We first describe the motivation and overall pipeline in Fig. 1, then mathematically formalize the different latent representations, and finally detail the training and sampling algorithms in Alg. 1 and Alg. 2. The key idea is to isolate the effect of the latent representation by keeping all other components of the diffusion process unchanged.

During training, given an input image  $x$ , the VAE encoder  $E(x)$  produces the parameters  $\mu$  and  $\sigma$ . Depending on the chosen configuration  $c$  (baseline, concatenated, or deterministic), the latent code  $z_0$  is computed as follows:

$$z_0 = \begin{cases} \mu + \sigma \odot \epsilon, & \text{if } c \text{ is baseline,} \\ \text{Concat}(\mu, \sigma), & \text{if } c \text{ is concatenated,} \\ \mu, & \text{if } c \text{ is deterministic.} \end{cases}$$

This  $z_0$  is then processed by the diffusion model, which performs the forward and backward diffusion processes in the latent space, ultimately yielding a refined latent  $z'_0$ . The VAE decoder  $D(z'_0)$  then reconstructs the high-resolution image  $\hat{x}$ . The training algorithm is outlined in Algorithm 1, and the corresponding sampling procedure is provided in Algorithm 2. These algorithms ensure that all three configurations use the same diffusion process and training schedule, so that any differences in performance are attributable solely to the latent representation.

This formulation encapsulates the concept and methodology of our proposed decoupled latent diffusion model. By keeping the diffusion process unchanged across configurations, our design ensures that any observed differences in generative performance are attributable solely to the choice of latent representation.

## Algorithm 1 Training With Three Latent Configurations $c$

---

```

1: Input: Image  $x$  and
   configuration  $c \in \{\text{baseline, concat, deterministic}\}$ 
2: Encode the image:  $[\mu, \sigma] \leftarrow E(x)$ 
3: if  $c = \text{baseline}$  then
4:   Sample  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
5:   Compute  $z_0 \leftarrow \mu + \sigma \odot \epsilon$ 
6: else if  $c = \text{concat}$  then
7:   Compute  $z_0 \leftarrow \text{Concat}(\mu, \sigma)$ 
8: else
9:   Compute  $z_0 \leftarrow \mu$ 
10: end if
11: repeat
12:   Sample timestep  $t \sim \text{Uniform}(\{1, \dots, T\})$ 
13:   Compute  $z_t \leftarrow \text{AddNoise}(z_0, t)$   $\triangleright$  Apply forward
      diffusion
14:   Predict noise:  $\hat{n} \leftarrow \text{DiffusionModel}_\theta(z_t, t)$ 
15:   Compute loss:  $\mathcal{L} = \|\text{TrueNoise}(z_0, t) - \hat{n}\|^2$ 
16:   Update model parameters via gradient descent
17: until convergence

```

---

## Algorithm 2 Sampling With Latent Configuration $c$

---

```

1: Input:
   configuration  $c \in \{\text{baseline, concat, deterministic}\}$ 
2: Initialize  $z_T \sim \mathcal{N}(0, \mathbf{I})$ 
3: for  $t = T, T-1, \dots, 1$  do
4:   Predict noise:  $\hat{n} \leftarrow \text{DiffusionModel}_\theta(z_t, t, z_{\text{cond}})$ 
5:   Compute  $z_{t-1} \leftarrow \text{RemoveNoise}(z_t, \hat{n}, t)$ 
6: end for
7: if  $c = \text{concat}$  then
8:   Split  $z_0$  into  $\mu$  and  $\sigma$ 
9:   Sample  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
10:  Reconstruct  $z \leftarrow \mu + \sigma \odot \epsilon$ 
11: else
12:  Set  $z \leftarrow z_0$ 
13: end if
14: Decode image:  $\hat{x} \leftarrow D(z)$ 
15: return Generated image  $\hat{x}$ 

```

---

## V. EXPERIMENTAL RESULTS

Through our experiments, we demonstrate that our proposed method, which relies solely on the latent mean, effectively enhances image synthesis quality and stability compared to conventional approaches that utilize the full latent representation. Our experiments are designed to isolate the effect of different latent representations within the LDM framework. To ensure that any observed performance differences stem exclusively from the chosen latent configuration, we fix the training settings and use the same diffusion model architecture and hyper-parameters across all experiments.

### A. EXPERIMENTAL SETUP

Our implementation builds upon the official Latent Diffusion Model (LDM) framework, leveraging its established

architecture and training methodologies as detailed in LDM [6].

We employ a standard LDM architecture based on a UNet [25] backbone. The model is trained with the following fixed configuration: a spatial downsampling factor 4 resulting in a latent shape of  $64 \times 64 \times 3$  when the original input shape is  $256 \times 256 \times 3$ . We used 1000 diffusion steps with a linear noise schedule (ranging from 0.0015 to 0.0155). Training is performed using a batch size 16 over 120k iterations on an NVIDIA GeForce RTX 4090, with a constant base learning rate of  $1.28 \times 10^{-4}$  using AdamW optimization [26].

For the first stage, we use a pre-trained Variational Auto-Encoder (VAE) with a latent resolution of  $64 \times 64 \times 3$  to compress the high-dimensional image  $x$  into its latent representation  $z$ . This VAE architecture, including its specific components and parameters, is consistent with the one provided in the official LDM framework, which is fully detailed in our primary LDM reference [6]. Unlike many implementations that leverage VQ-VAE for higher performance, we adopt a standard VAE with reduced latent dimensionality to specifically study the impact of the latent variance. Importantly, in original LDM setup, the VAE decoder is solely used to reconstruct images from the encoded latent and it cannot generate images from random noise.

Our evaluation covers four benchmark dataset: LSUN-Churches, LSUN-Bedrooms [27], AFHQv2 [28] and Celeba-HQ [29].

By keeping the training configuration and diffusion process identical across experiments, we ensure that any differences in performance can be imputed solely to the three latent representations used. This unified set-up provides a controlled environment to rigorously evaluate the influence of latent stochasticity on generative quality.

## B. ROLE OF STANDARD DEVIATION IN IMAGE RECONSTRUCTION

VAE encodes input images into a latent space characterized by mean and standard deviation, enabling stochastic sampling for reconstruction and generation. Traditional approaches utilize both mean and standard deviation in latent representations to capture variations in data. However, the necessity of the standard deviation component in image generation remains an open question. To systematically examine its role, we conduct experiments focusing on reconstruction performance when only the mean is used versus when both mean and standard deviation are incorporated. This section explores how these latent representations influence the quality, consistency, and interpretability of generated images. To analyze the impact of the standard deviation in reconstruction, we utilize a pre-trained VAE and manipulate the latent space by adjusting the weight of the standard deviation component. Specifically, we reconstruct images by feeding different latent representations into the decoder of the VAE, formulated as

$$z = \mu + \lambda \cdot \sigma \odot \epsilon, \quad (16)$$

where  $\lambda$  controls the contribution of the standard deviation component. This allows us to observe how increasing or decreasing the standard deviation component affects the reconstructed images. As shown in Fig. 2, each column represents the input images, the mean and standard deviation extracted from the encoder, and the corresponding reconstruction results in order. Notably, when  $\lambda = 0$ , the reconstruction is solely based on the mean and when  $\lambda = 1$ , it is totally same as traditional VAE. From this perspective, we change the value of  $\lambda$  from 0 to 1.

Our findings reveal that varying  $\lambda$  has minimal effect on the quality of the reconstructed images. Regardless of the value of  $\lambda$ , the reconstructed images remain visually indistinguishable, even the mean square errors between the each reconstructed result are nearly zero, suggesting that the standard deviation does not play a crucial role in preserving the details of the image. Even when  $\lambda$  is a high value, the reconstructions do not show significant variations compared to those generated using only the mean. This indicates that the standard deviation does not provide additional useful information for reconstruction.

Additionally, when visualizing the mean and standard deviation outputs of the VAE encoder, we observe a stark contrast in the type of information they capture. The mean representation effectively preserves the input style, form, and fine details, making it a reliable descriptor of the original content. However, the standard deviation contains very little information. Since the value of standard deviation merely 0, we calculate log scale and apply min-max normalization for visualization.

To produce these mean and standard deviation visualizations, we calculate the mean along the channel dimension, revealing that the standard deviation lacks clear structural patterns and contributes no meaningful high-level features.

These results suggest that the mean component alone is sufficient for high-quality image reconstruction, and the inclusion of the standard deviation does not improve the perceptual quality of the images. This finding challenges the conventional assumption that the standard deviation is necessary for generating diverse outputs and indicates that a simpler latent representation using only the mean could be a more efficient approach in generative modeling.

Building upon the findings in this section, we offer a mathematical analysis to explain why the standard deviation component of the latent code have a negligible effect on reconstruction fidelity. We can assume the decoder of VAE  $D : \mathbb{R}^d \rightarrow \mathbb{R}^n$  is  $L$ -Lipschitz [30], [31], [32], [33]:

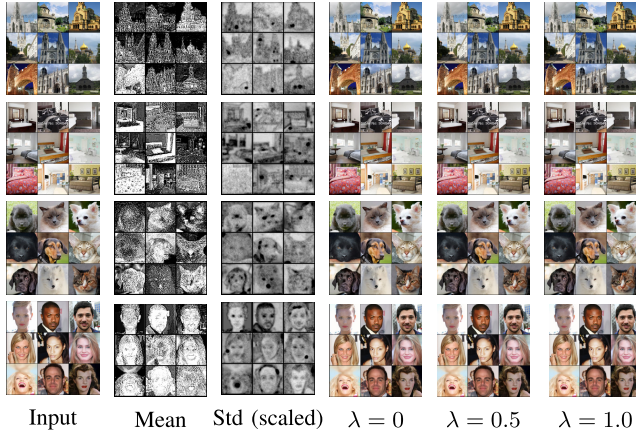
$$\|D(z_1) - D(z_2)\|_2 \leq L \|z_1 - z_2\|_2, \quad \text{for all } z_1 \text{ and } z_2. \quad (17)$$

When we set  $z_1 = \mu + \sigma \odot \epsilon$  and  $z_2 = \mu$ , Lipschitz continuity gives

$$\|D(\mu + \sigma \odot \epsilon) - D(\mu)\|_2 \leq L \|\sigma \odot \epsilon\|_2. \quad (18)$$

Squaring and taking expectation yields

$$\mathbb{E}_\epsilon[\|D(\mu + \sigma \odot \epsilon) - D(\mu)\|_2^2] \leq L^2 \mathbb{E}_\epsilon[\|\sigma \odot \epsilon\|_2^2]. \quad (19)$$



**FIGURE 2.** Reconstruction results on LSUN-Churches, LSUN-Bedrooms, AFHQv2, and CelebA-HQ (top to bottom). Columns show: input, mean, std (scaled), and reconstructed outputs using  $\lambda \in \{0, 0.5, 1.0\}$ . Labels shown only for bottom row; alignment is preserved using spacing.

Because the components of  $\epsilon$  are independent with unit variance,

$$\mathbb{E}_{\epsilon}[\|\sigma \odot \epsilon\|_2^2] = \sum_{i=1}^d \sigma_i^2 \mathbb{E}[\epsilon_i^2] = \sum_{i=1}^d \sigma_i^2. \quad (20)$$

With eq. (20), eq. (19) can be written as

$$\mathbb{E}_{\epsilon}[\|D(\mu + \sigma \odot \epsilon) - D(\mu)\|_2^2] \leq L^2 \sum_{i=1}^d \sigma_i^2. \quad (21)$$

We check the magnitude of Eq. (20), using the pre-trained VAE which is used in original LDM. For each dataset we forward a batch which has  $N$  images, through the encoder of VAE and get the mean value of Eq. (20).

$$\frac{1}{N} \sum \sigma^2 = \begin{cases} 2.52 \times 10^{-3} & \text{AFHQ-v2,} \\ 2.78 \times 10^{-3} & \text{CelebA-HQ,} \\ 5.65 \times 10^{-3} & \text{LSUN-Bedrooms,} \\ 6.07 \times 10^{-3} & \text{LSUN-Churches.} \end{cases} \quad (22)$$

Treating the decoder as unit-Lipschitz—a worst-case but always valid assumption because every activation is non-expansive [34]—turns Eq. (21) into an explicit numeric bound. For LSUN-Churches we obtained the worst-case numeric bound:

$$\mathbb{E}_{\epsilon}[\|D(\mu + \sigma \odot \epsilon) - D(\mu)\|_2^2] \leq 6.07 \times 10^{-3},$$

which implies an root-mean-square (RMS) error of

$$\text{RMS} = \sqrt{6.07 \times 10^{-3}} \approx 0.078.$$

The pixel values in all our experiments are normalized to the interval  $[-1, 1]$ , so the dynamic range is  $\text{DR} = 1 - (-1) = 2$ . Expressed relative to this full scale,

$$\frac{\text{RMS}}{\text{DR}} = \frac{0.078}{2} = 0.039 \quad (3.9\%).$$

In other words, the average value of the pixels moves by less than 4% of its total range. Across all four datasets, setting  $\sigma = 0$  changes the reconstructions by less than 0.1 RMS—too small to see. This shows our deterministic-mean approach works well even without extra regularization.

### C. ROLE OF STANDARD DEVIATION IN IMAGE GENERATION

Our results demonstrate that excluding standard deviation in latent representations does not degrade image quality but instead leads to more stable and visually coherent results. As shown in Fig. 3, the first column shows images generated using the original model, while the second column displays images generated using the deterministic model (which excludes standard deviation). The third column highlights cases where the original model generates excessive details, resulting in over-textured or visually corrupted images, whereas the fourth column shows corresponding results from the deterministic model, which successfully mitigates such artifacts and produces more structured outputs. These observations suggest that while the original model introduces high-frequency variations and unnecessary textures, the deterministic model produces cleaner, more interpretable, and structurally coherent images without sacrificing realism. Notably, the excessive details produced by the original model often lead to ambiguous object boundaries, causing textures to blend unnaturally and reducing the overall clarity of the generated images. In contrast, the mean-only model preserves well-defined object structures, ensuring clear separations between different elements in the scene. This indicates that the standard deviation component amplifies irrelevant noise, leading to artificial artifacts rather than meaningful enhancements. By removing standard deviation, we achieve greater stability and consistency in image generation, demonstrating that the mean alone is sufficient for high-quality synthesis. These findings challenge the conventional assumption that standard deviation is necessary in latent representations and support a simplified modeling approach that maintains generation quality while reducing computational complexity and eliminating unwanted noise artifacts.

### D. QUANTITATIVE AND QUALITATIVE COMPARISON TO LDM

Based on our previous experiments, we observed that the standard deviation has no discernible effect on image reconstruction and may even impair image quality when the latent is constructed with the standard deviation. Motivated by these findings, we further evaluated image synthesis performance by training a DDPM within the LDM framework using three latent configurations: the baseline (full latent code), the concatenated representation, and the deterministic representation. Our primary evaluation metrics is the Fréchet Inception Distance (FID), which provides an indication of both the quality and diversity of the generated images on LSUN-Churches [27], LSUN-Bedrooms [27], AFHQv2 [28] and CelebA-HQ [29]. The results of image generation are



**FIGURE 3.** The first column represents images generated using the original method, while the second column shows images generated using only the mean of the latent representation. The third column displays images where the original approach produces excessive detail, leading to visual artifacts or unrealistic textures. The fourth column shows the corresponding images generated using only the mean, which mitigate such excessive details and produce more visually coherent results.

shown in Fig. 4 and evaluation metrics are shown in Tab. 1. We calculated the FID using 15,000 images for AFHQv2, CelebA-HQ, and 50,000 images for LSUN-Bedrooms and LSUN Churches.

For datasets such as LSUN-Bedrooms and AFHQv2, the mean-only configuration outperforms the standard model. On LSUN-Bedrooms, the FID improves from 27.3467 to 22.8598, and on AFHQv2, a substantial reduction is observed—from 9.5429 to 4.4371. These improvements suggest that, in these cases, the additional noise introduced by sampling from the variance can be detrimental. By using only the deterministic mean, the model produces a more stable and consistent latent code, leading to better alignment with the real data distribution.

On LSUN-Churches, the standard configuration yields the lowest FID (15.8321), while the mean-only model has a moderately higher FID (22.2465). Similarly, on CelebA-HQ, the standard model achieves an FID of 5.6427 versus 6.2689 for the mean-only variant. These results indicate that while the mean-only representation may introduce slight degradation in some cases, the differences are relatively modest. This suggests that, for certain datasets, the stochastic noise may play a small beneficial role, but overall the contribution of the variance remains minimal.

Across all datasets, the concatenated representation produces significantly higher FID scores. This consistent degradation demonstrates that merely preserving the variance as a separate channel does not enhance generative performance; on the contrary, it appears to amplify noise, leading to artifacts that increase the divergence from real image distributions.

In addition to the quantitative FID differences, qualitative analysis (as shown in Fig. 4) reveals that images generated with the concatenate configuration are noticeably blurrier and lack sharp details compared to those from the standard

and deterministic models. The extra stochastic component appears to introduce excessive detail in an uncontrolled manner, which results in a loss of fine structure and overall coherence in the generated images.

Taken together, these results strongly suggest that within the LDM framework, the latent mean captures the essential information for high-quality image synthesis. The extra noise contributed by the latent variance not only fails to improve performance but can, in fact, degrade both quantitative metrics and perceptual quality. Therefore, using a deterministic latent representation offers a promising direction: it simplifies the generative process, and, in many cases, yields sharper, more coherent images. These findings motivate further investigation into optimizing latent diffusion models by omitting the stochastic component of the latent representation.

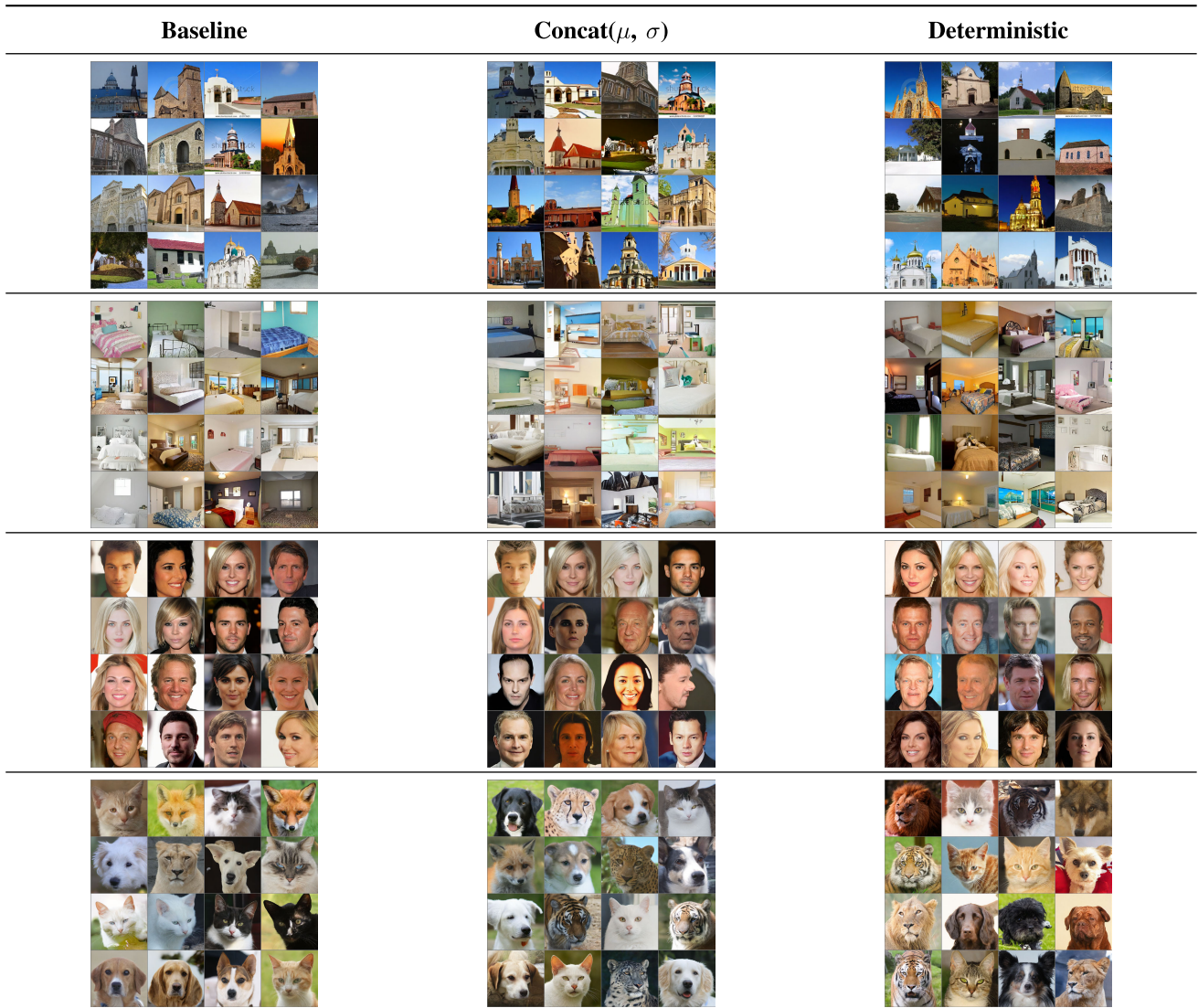
## E. DISCUSSION OF RESULTS

Our comprehensive evaluation across the three distinct latent representation configurations: (1) Full Latent Code (Baseline), (2) Concatenated Representation, and (3) Deterministic Representation yielded clear insights into the role of stochasticity in Latent Diffusion Models. As quantitatively summarized in Table 1, the Deterministic Representation consistently demonstrated strong performance, notably outperforming the Baseline on datasets like LSUN-Bedrooms and AFHQv2, while showing comparable results on others. In contrast, the Concatenated Representation consistently resulted in significantly higher FID scores across all datasets, indicating a detrimental effect on generative quality.

The superior performance of the Deterministic method primarily stems from its ability to provide a cleaner and more focused latent input to the diffusion model by eliminating the potentially noisy stochastic component derived from the VAE's standard deviation. Our experiments on image reconstruction (Section V-B) indicated that the latent mean alone is largely sufficient for high-quality image reconstruction, suggesting that additional stochasticity from variance might be worthless.

Building upon observations in general VAEs regarding latent space utilization, we propose a detailed theoretical hypothesis for why variance is underutilized in practice within the LDM framework. As far as we know, our work is the first to systematically investigate and demonstrate the efficacy of exclusively using the deterministic latent mean for image generation in LDM. We identify two primary theoretical considerations for this phenomenon:

Firstly, posterior collapse may occur even in well-trained VAEs used within LDM. While the KL divergence regularization aims for a rich latent distribution, it can be minimally satisfied, leading the encoder to produce a very narrow or near-deterministic posterior. Consequently, the decoder learns to predominantly rely on the more stable and informative mean component for reconstruction, effectively marginalizing any sampled noise from the variance term.



**FIGURE 4.** Image generation results on LSUN-Churches, LSUN-Bedrooms, CelebA-HQ, and AFHQv2 (top to bottom). Each image is  $256 \times 256$ , sampled in batches of 16 without cherry-picking. Columns: (1) Baseline ( $\mu + \sigma \odot \epsilon$ ), (2) Concatenated ( $(\mu, \sigma)$ ), and (3) Deterministic ( $\mu$ ).

**TABLE 1.** FID scores for each latent configuration across datasets (lower is better).

| Method        | LSUN-Churches | LSUN-Bedrooms | AFHQv2  | CelebA-HQ |
|---------------|---------------|---------------|---------|-----------|
| Baseline      | 15.8321       | 27.3467       | 9.5429  | 5.6427    |
| Concatenate   | 34.3693       | 55.5091       | 13.0355 | 12.8054   |
| Deterministic | 22.2464       | 22.8598       | 4.4371  | 6.2689    |

This creates an optimization shortcut, where the model finds it more efficient to ignore the stochasticity.

Secondly, the inherent stochasticity and robust denoising capabilities of the diffusion process itself may render the additional, less structured noise derived from the VAE's latent variance redundant or even counterproductive. Diffusion models are explicitly designed to learn intricate denoising functions across a predefined noise schedule. We hypothesize that the noise sampled from the VAE's latent variance might

not align optimally with these learned noise patterns. As a result, the diffusion model, optimized for generation from its own controlled noise, could implicitly learn to 'filter out' or 'discount' this extraneous VAE-induced noise, perceiving it as interference rather than a beneficial contributor to high-fidelity generation. This suggests that while stochasticity is crucial for diffusion, the specific stochasticity from the VAE's variance might not add value and can, in fact, hinder the generative process.

These theoretical considerations are strongly corroborated by our empirical findings, particularly the superior generative performance of the Deterministic Representation as shown in Table 1 and the qualitative observations in Fig. 4.

By relying solely on the precise semantic and structural information within the latent mean, the Deterministic method facilitates a more stable and consistent generation process. This results in sharper, more coherent images that better align

with the real data distribution, underscoring the efficiency of a simplified latent representation.

Conversely, the consistent degradation observed with the Concatenated Representation, as seen in Fig. 4, highlights that merely preserving the variance as a separate information channel does not enhance generative performance. This counter-intuitive result underscores a critical point: simply providing the stochastic variance alongside the mean, even in a distinct channel, does not inherently make it useful for the diffusion process. Instead, it appears that the additional variance information, despite being explicitly separated, introduces noise or complexities that the diffusion model struggles to effectively utilize. As hypothesized earlier, the diffusion model, being optimized for its own noise schedule, may perceive this auxiliary VAE-derived variance as extraneous or even conflicting information. When this superfluous information is explicitly concatenated, it potentially adds a burden to the model's learning and generation process, leading to a loss of fine structure, reduced coherence, and overall degradation in perceptual quality. This suggests that the diffusion model might not only ignore but actively struggle with processing such auxiliary stochastic inputs when they don't contribute constructively to the denoising task.

These findings collectively suggest that the generative power within the LDM framework predominantly resides in the deterministic mean of the latent space. The stochastic component, as traditionally incorporated or even separately provided, does not consistently improve performance and can, in fact, introduce detrimental noise. This motivates a promising direction for optimizing latent diffusion models by focusing on more streamlined, deterministic latent representations. This work, therefore, not only challenges conventional paradigms but also opens avenues for designing more efficient and perceptually robust generative models by focusing on the most informative components of the latent space.

## VI. CONCLUSION

In this work, we reexamine the conventional LDM framework by challenging the necessity of incorporating a stochastic latent component. We investigated three latent representation configurations: the standard full latent code (as used in the original LDM), a concatenated representation that explicitly preserves both the mean and the standard deviation, and a deterministic representation that only using deterministic component  $\mu$ . Our experiments show that the deterministic configuration frequently yields competitive—and in some cases improved—generative quality as measured by FID. Conversely, the concatenated configuration consistently produces significantly higher FID scores, indicating that simply preserving the variance as an independent channel is not effective without further optimization.

All configurations were implemented using the same network architecture and hyper-parameters, ensuring that the observed performance differences can be solely attributed to

the latent representation. While our primary focus was on the synthesis quality of generated images, the use of a deterministic latent representation may also offer computational advantages by eliminating the need for stochastic sampling during inference. Future work will explore the trade-offs between generation quality and computational efficiency, as well as adaptive strategies to leverage latent variance when it is beneficial.

Overall, our findings suggest that the latent mean encapsulates the essential information needed for high-quality image synthesis within the LDM framework, whereas the stochastic variance component can introduce unnecessary noise. Simplifying the latent representation to rely solely on the mean thus offers a promising direction for developing more efficient and robust generative models.

## REFERENCES

- [1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
- [2] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.
- [3] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 8780–8794, 2021.
- [4] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. 38th Int. Conf. Mach. Learn.*, Jul. 2021, pp. 8162–8171.
- [5] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11451–11461.
- [6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
- [7] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
- [8] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2015, *arXiv:1511.06349*.
- [9] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black, and B. Schölkopf, "From variational to deterministic autoencoders," 2019, *arXiv:1903.12436*.
- [10] J. Lucas, G. Tucker, R. Grosse, and M. Norouzi, "Understanding posterior collapse in generative latent variable models," in *Proc. ICLR*, 2019, pp. 1–19.
- [11] A. Van Den Oord and O. Vinyals, "Neural discrete representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6309–6318.
- [12] A. Razavi, A. Van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 14837–14847.
- [13] R. Child, "Very deep VAEs generalize autoregressive models and can outperform them on images," 2020, *arXiv:2011.10650*.
- [14] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 19667–19679.
- [15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
- [16] P. Brémaud, *Markov Chains: Gibbs Fields Monte Carlo Simulation, and Queues*, vol. 31, Berlin, Germany: Springer, 2001.
- [17] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, *arXiv:1511.05644*.
- [18] G. Silvestri, D. Roos, and L. Ambrogioni, "Deterministic training of generative autoencoders using invertible layers," 2022, *arXiv:2205.09546*.
- [19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

- [20] V. Prokhorov, E. Shareghi, Y. Li, M. Taher Pilehvar, and N. Collier, "On the importance of the Kullback–Leibler divergence term in variational autoencoders for text generation," 2019, *arXiv:1909.13668*.
- [21] S. Rezvani, F. Soleymani Siahrkar, Y. Rezvani, A. Alavi Gharahbagh, and V. Abolghasemi, "Single image denoising via a new lightweight learning-based model," *IEEE Access*, vol. 12, pp. 121077–121092, 2024.
- [22] M. Trigka and E. Dritsas, "A comprehensive survey of deep learning approaches in image processing," *Sensors*, vol. 25, no. 2, p. 531, Jan. 2025.
- [23] J. Valente, J. António, C. Mora, and S. Jardim, "Developments in image processing using deep learning and reinforcement learning," *J. Imag.*, vol. 9, no. 10, p. 207, Sep. 2023.
- [24] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 2575–2583.
- [25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Munich, Germany: Springer, 2015, pp. 234–241.
- [26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
- [27] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.
- [28] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8188–8197.
- [29] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.
- [30] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6240–6249.
- [31] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*.
- [32] T. Avant and K. A. Morgansen, "Analytical bounds on the local Lipschitz constants of ReLU networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 1–12, Oct. 2024.
- [33] Z. Yang, R. Feng, H. Zhang, Y. Shen, K. Zhu, L. Huang, Y. Zhang, Y. Liu, D. Zhao, and J. Zhou, "Lipschitz singularities in diffusion models," in *Proc. 12th Int. Conf. Learn. Represent.*, 2023.
- [34] A. Virmaux and K. Scaman, "Lipschitz regularity of deep neural networks: Analysis and efficient estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 3839–3848.



**HYUN-TAE CHOI** received the B.S. and M.S. degrees in computer science from CAU. He is currently pursuing the Ph.D. degree in artificial intelligence. His current research interests include computer vision, machine learning, image segmentation, image translation, and image generation.



**KENSUKE NAKAMURA** received the M.Sc. degree in engineering and the D.Phil. degree in engineering from the Institute of Technology, Kyoto Institute of Technology, Japan, in 2004 and 2012, respectively. He is currently a Research Associate with Chung-Ang University. He has participated for studies on models and applications of the 2-D and 3-D human body shapes. His current research interests include computer vision and a support systems for fashion design.



**BYUNG-WOO HONG** received the M.Sc. degree in computer vision from the Weizmann Institute of Science with Prof. Shimon Ullman, in 2001, and the D.Phil. degree in computer vision from the University of Oxford with Prof. Michael Brady, in 2005. He joined as a Faculty Member with the Computer Science Department, CAU, in 2008, after the postdoctoral research at the Computer Science Department, University of California Los Angeles, with Prof. Stefano Soatto.

His research interests include image processing, computer vision, machine learning, and medical image analysis.

• • •