VAE-REPA: Variational Autoencoder Representation Alignment for Efficient Diffusion Training

Mengmeng Wang[1]   Dengyang Jiang[2]   Liuzhuozheng Li[2]   Yucheng Lin[1]   Guojiang Shen[1]
Xiangjie Kong[1]   Yong Liu[3]   Guang Dai[2]   Jingdong Wang[4]
[1]Zhejiang University of Technology   [2]SGIT AI Lab, State Grid Corporation of China
[3] Zhejiang University [4]Baidu

Figure 1. **High-Quality Image Generation via VAE-REPA**: Samples from SiT-XL/2+VAE-REPA on ImageNet 256×256. Our method generates images with high fidelity, fine-grained detail, and strong semantic coherence, demonstrating excellent generation quality.

Abstract

*Denoising-based diffusion transformers, despite their strong generation performance, suffer from inefficient training convergence. Existing methods addressing this issue, such as REPA (relying on external representation encoders) or SRA (requiring dual-model setups), inevitably incur heavy computational overhead during training due to external dependencies. To tackle these challenges, this paper proposes **VAE-REPA**, a lightweight intrinsic guidance framework for efficient diffusion training. VAE-REPA leverages off-the-shelf pre-trained Variational Autoencoder (VAE) features: their reconstruction property ensures inherent encoding of visual priors like rich texture details, structural patterns, and basic semantic information. Specifically, VAE-REPA aligns the intermediate latent features of diffusion transformers with VAE features via a lightweight projection layer, supervised by a feature alignment loss. This design accelerates training without extra representation encoders or dual-model maintenance, resulting in a simple yet effective pipeline. Extensive experiments demonstrate that VAE-REPA improves both generation quality and training convergence speed compared to vanilla diffusion transformers, matches or outperforms state-of-the-art acceleration methods, and incurs merely 4% extra GFLOPs with zero additional cost for external guidance models.*
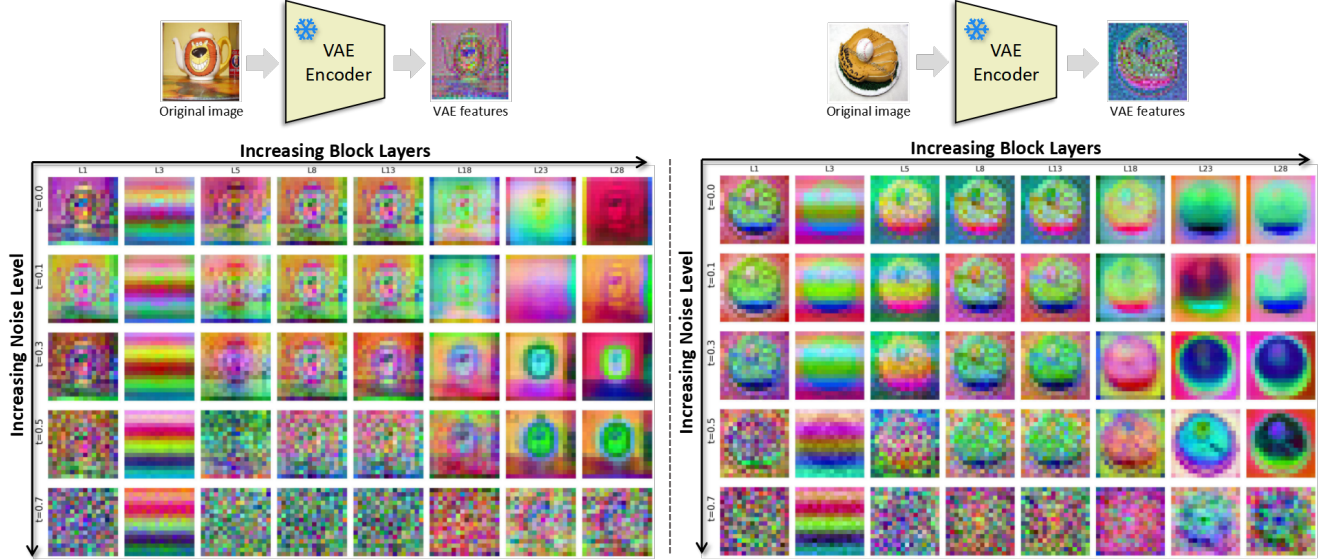
Figure 2. We empirically visualize the feature information richness of SD-VAE [34] and SiT-XL/2 [27] via PCA [1]. **Top**: VAE features, extracted from original images by an SD-VAE encoder. **Bottom**: Latent features of SiT-XL/2 across different block layers and noise levels. We observe that SD-VAE features are significantly superior in delineating visual concepts compared to SiT's latent representations, maintaining clearer details, structural integrity, and stronger semantic coherence. This motivates our use of VAE features for representation alignment.

1. Introduction

Denoising-based generative models, particularly diffusion transformers [27, 31], have attracted significant attention for their exceptional ability to generate diverse, high-fidelity images. They have proven to deliver strong performance across a wide range of domains, including text-to-image synthesis [4, 23], text-to-video generation [39, 43], image editing [10, 36], and 3D asset generation [41, 48], and so on. However, even the most popular Latent Diffusion Model (LDM) [33] architecture still grapples with the critical issue of slow training convergence, often requiring an enormous number of iterations to achieve satisfactory performance.

To explore the efficient training of large transformer-based diffusion models, self-supervised learning techniques such as masked modeling [15] have been adopted [49, 50], which help reduce training costs and accelerate convergence to a certain extent, but they come at the cost of network architecture adjustments such as the need for an additional diffusion decoder. Recently, several methods [38, 40, 45, 47] represented by REPA [45] have started guiding the optimization of diffusion models by incorporating external large-scale pre-trained representative encoders like DINOv2 [30]. However, despite their promising performance, we observe that integrating these extra representation encoders not only increases training computational overhead but also creates dependencies on external large-scale pre-trained models. In practice, suitable pre-trained models are not available across all domains, such as the video do-

main or certain specialized downstream tasks, where there is a lack of encoders with strong generalization capabilities. This will significantly limit the applicability of such methods. Therefore, attempts have also been made to accelerate training by leveraging the diffusion transformer's inherent discriminative information [19, 28], such as that from different layers or time steps. Nevertheless, their additional cost lies in the need to maintain an extra teacher diffusion model to provide self-alignment guidance during training. Thus, we wonder: ***Does there exist a simpler and more lightweight guidance approach that can avoid external representative encoders or dual-model maintenance?***

In this paper, we focus our solution on finding more suitable supervisory features, and in this process, we turn to a readily available resource: pre-trained Variational Autoencoders (VAEs). This may be a promising solution to addressing the aforementioned pain point. In the two-stage LDM, the pre-trained VAE in the first stage, having been trained on large-scale natural image datasets, possesses inherent feature encoding capabilities. It can be used for high-quality image reconstruction, thus ensuring that its encoded features encapsulate the image's texture details, low-level structural patterns, and basic semantic information, as illustrated in Fig. 2. Importantly, these VAE features are usually pre-extracted offline for training the diffusion models in the second stage; thus, they can be conveniently reused directly as our target built-in guidance at no extra cost. If we can leverage the prior information in VAE features to pro-
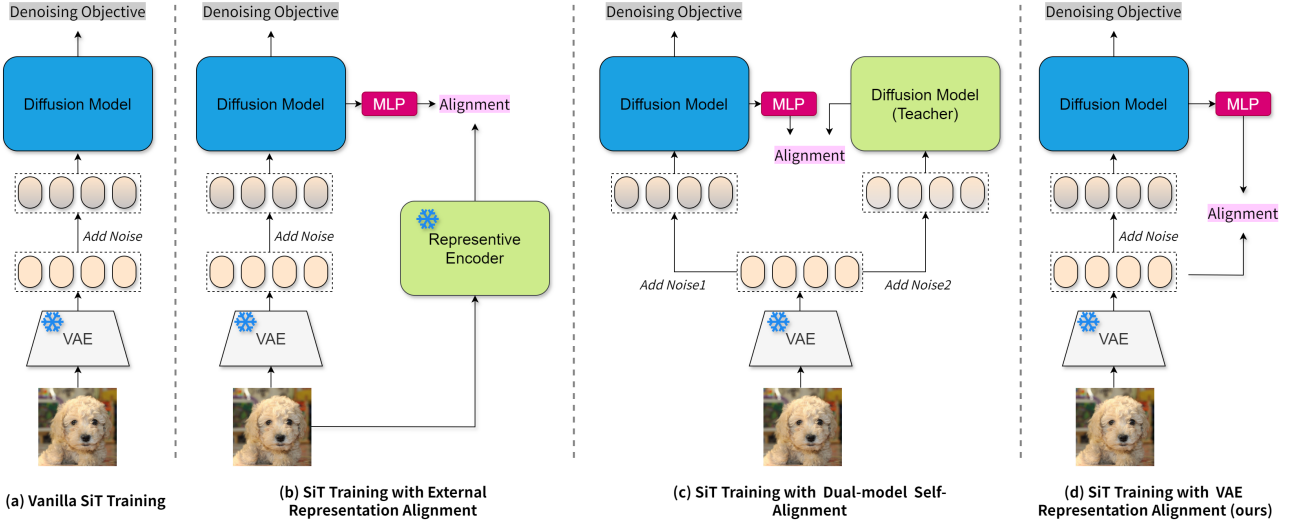
Figure 3. **Comparison of typical SiT training paradigms.** (a) Vanilla SiT Training: Images are encoded by a VAE, added with noise, and processed by the diffusion model for denoising. (b) SiT Training with External Representation Alignment (e.g., REPA [45]): SiT training augmented with an external representative encoder and an MLP for alignment. (c) SiT Training with Dual-model Self-Alignment (e.g., SRA [19]): SiT training leveraging a dual-model setup with an MLP for self-alignment, guided by a teacher diffusion model. (d) SiT Training with VAE Representation Alignment (ours): SiT utilizes VAE features as representation guidance and an MLP for alignment, efficiently combining VAE's semantic richness with SiT's denoising capability without introducing additional heavy models.

vide diffusion models with feature-rich, noise-free targets, we can thereby fundamentally help the model improve both optimization efficiency and generation quality.

To address this, we propose introducing <u>VAE</u> features for <u>REP</u>resentation <u>A</u>lignment (**VAE-REPA**), a lightweight intrinsic guidance framework for training diffusion transformer models. Specifically, it leverages off-the-shelf pretrained VAE features to guide the intermediate layer representations of diffusion models. The intermediate diffusion latent feature is first passed through a projection layer to perform the nonlinear and dimension transformation of the feature space, and then aligned with the target VAE features. To achieve effective alignment, VAE-REPA incorporates a simple yet efficient feature alignment loss, which minimizes the discrepancy between the diffusion model's intermediate features and the VAE's representations. With this simple design, VAE-REPA keeps the overall training framework highly concise and lightweight as shown in Fig. 3(d), eliminating the need for extra representative encoders (as in Fig. 3(b)) or dual-model maintenance (as in Fig. 3(c)) by directly reusing pre-extracted VAE features. Finally, we conduct extensive experiments to demonstrate the effectiveness of our VAE-REPA by applying it to the recent diffusion transformer SiT [27]. It achieves notable improvements in both generation quality and training convergence speed, while incurring zero additional guidance feature extraction cost and with only 4% extra GFLOPs for feature alignment during the training process.

The main contributions can be summarized as:

- We discover that the features of the pre-trained VAE, by virtue of their reconstruction property, inherently encode rich visual priors, which can serve as a readily available guidance source for diffusion transformer training.
- We propose **VAE-REPA**, a simple and lightweight built-in guidance framework that leverages off-the-shelf pretrained VAE features to align diffusion transformer's intermediate representations, avoiding external model dependencies.
- On the ImageNet 256×256 benchmark, our proposed VAE-REPA achieves notable improvements over the vanilla SiT and matches or surpasses the performance of methods with external model dependencies, while incurring zero additional cost for guidance feature extraction.

2. Related Works

Diffusion Transformers for image generation. Diffusion models have emerged as a powerful paradigm for high-fidelity image generation, evolving from early pixel-space approaches [7, 17, 18, 20] to latent diffusion frameworks (LDM) [33]. With the integration of transformer architectures [8, 37], diffusion transformers [4, 9, 23, 27, 31, 32] have further advanced this paradigm: they leverage transformer-based attention mechanisms to model complex semantic dependencies, enabling scalable high-resolution generation and more sophisticated task adaptation. Foundational works like DiT [31] first demonstrated that transformers can model diffusion's denoising dynamics in latent

spaces with minimal structural overhead, while SiT [27] advanced this paradigm via linear flow diffusion to balance scalability and performance. Yet a critical limitation remains: their strong performance relies on massive training iterations, slowing convergence. Our work addresses this inefficiency with VAE-REPA, a compact method that accelerates training while preserving modern diffusion transformers' architectural elegance.

Diffusion Training Guided by External Dependency. Diffusion training guided by external dependencies has provided an effective pathway for improving convergence efficiency and generative quality, with three distinct paradigms: masking modeling [11, 12, 49, 50], pre-trained representation guidance [25, 38, 40, 45], and self-alignment [19, 28, 46]. Masking modeling methods exemplified by methods like SD-DiT [50] and MaskDiT [49] introduce structured noise through partial masking of input tokens or latent features, creating an external diffusion decoder dependency on reconstructing missing information. Pre-trained representation guidance methods leverage extra features from external foundation models [14, 30] as a dependency, explicitly aligning the diffusion model's intermediate representations with these robust, pre-learned representation priors. For instance, RCG [25] uses a pretrained self-supervised encoder [14] to map image distributions to aligned representation distributions, while REPA [45] and its successors like REG [40] and DDT [38] improve semantic representation quality via feature alignment between early diffusion layers and pretrained vision features of DINOv2 [30]. Self-alignment approaches like SRA [19] and SSD [28] employ a dual-model framework with an online teacher model that supplies higher quality features characterized by more robust semantics or lower noise. While these methods do accelerate training, they often involve external dependencies that are not always accessible and incur additional training computational overhead. In contrast, our work eschews such external dependencies, aiming for a lightweight intrinsic solution that preserves the simplicity of vanilla diffusion frameworks while achieving comparable or superior acceleration.

Leveraging VAEs for Diffusion Acceleration. Stable Diffusion [34] popularized the latent diffusion paradigm, employing a VAE [21], termed SD-VAE, to encode images into a compact latent space and decode latent tokens back into images. This design enables efficient training and scaling of diffusion models, making it a dominant choice for visual generation. We focus on related methods that optimize VAEs for accelerating diffusion models, which can be categorized into two types. The first type is latent compression methods [3, 5], which emphasize deeper compression and structured latent subspaces to reduce computation. The second type involves representation alignment for VAE refinement [24, 42, 44], which aligns VAE features

with pre-trained representations to speed up convergence via semantic priors. For example, REPA-E [24] attempts to train latent diffusion models together with the VAE tokenizer in an end-to-end manner using representation alignment loss. VAVAE [44] mitigates the optimization dilemma between VAE reconstruction and diffusion generation by aligning VAE representations with external semantic priors. In contrast, our work differs by adopting a sample alignment strategy compatible with SD-VAE, directly leveraging pre-extracted VAE features to enhance the diffusion model's convergence speed without extra VAE training or compression.

3. Method

3.1. Preliminaries

We apply our VAE-REPA upon Scalable Interpolant Transformers (SiT) [27], a unified framework that bridges flow and diffusion models using interpolating processes. We focus on its core mechanisms relevant to our method, and briefly elaborate on its foundational principles for clarity.

Here we denote $\boldsymbol{y}_t \in \mathbb{R}^n$ as a sample at time t, and SiT models the probability density flowing from a data sample $\boldsymbol{z} \sim p_{data}(\boldsymbol{z})$ to Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{I})$ through a time-dependent interpolation:

$$\boldsymbol{y}_t = a_t \boldsymbol{z} + b_t \boldsymbol{\epsilon}, \tag{1}$$

where $t \in [0, T]$ is the time variable, a_t decreases from 1 to 0, and b_t increases from 0 to 1 (generalizing both flow models, which use finite intervals, and diffusion models, which extend to $T \to \infty$).

The key insight is that sampling can be driven by a *velocity function* $\boldsymbol{v}(\boldsymbol{y}_t, t)$, governing the probability flow ordinary differential equation (ODE):

$$\frac{d\boldsymbol{y}_t}{dt} = \boldsymbol{v}(\boldsymbol{y}_t, t). \tag{2}$$

This velocity function captures the expected trajectory of $\boldsymbol{y}_t$ over time, derived as:

$$\boldsymbol{v}(\boldsymbol{y}_t, t) = \dot{a}_t \cdot \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{y}_t] + \dot{b}_t \cdot \mathbb{E}[\boldsymbol{\epsilon} \mid \boldsymbol{y}_t], \tag{3}$$

where $\dot{a}_t$ and $\dot{b}_t$ denote time derivatives of a_t and b_t, respectively.

To learn $\boldsymbol{v}(\boldsymbol{y}_t, t)$, SiT trains a parameterized model $\boldsymbol{v}_\phi(\boldsymbol{y}_t, t)$ to minimize the squared error between predicted and true velocities:

$$\mathcal{L}_\phi = \mathbb{E}_{t, \boldsymbol{z}, \boldsymbol{\epsilon}} \left[\left\| \boldsymbol{v}_\phi(\boldsymbol{y}_t, t) - \left(\dot{a}_t \boldsymbol{z} + \dot{b}_t \boldsymbol{\epsilon} \right) \right\|^2 \right] dt. \tag{4}$$

For generation, one integrates the reverse of Eq. (2) starting from pure noise ($\boldsymbol{y}_T = \boldsymbol{\epsilon}$) to recover $\boldsymbol{z}$. Notably,

SiT connects velocity to the score function $s(y_t, t)$ (used in stochastic differential equation (SDE) sampling) via:

$$s(y_t, t) = -\frac{1}{b_t}\mathbb{E}[\epsilon \mid y_t], \qquad (5)$$

which can be rewritten using the velocity function:

$$s(y_t, t) = \frac{\dot{a}_t y_t - a_t v(y_t, t)}{b_t(a_t \dot{b}_t - \dot{a}_t b_t)} \qquad (6)$$

Vanilla SiT is trained with this denoising objective $\mathcal{L}_\phi$, which enables learning a unified velocity function and provides a strong foundation for generative modeling as our baseline.

3.2. VAE Feature Alignment

Vanilla SiT's velocity prediction relies solely on the denoising signal without leveraging visual priors guidance, a strategy that has been validated as highly effective in recent works [19, 40, 45]. To identify a readily available visual feature source rich in informative cues but without introducing additional pre-trained feature encoders or online extra teacher diffusion models, we turn to the first-stage output of the LDM framework itself: the off-the-shelf features of the pre-trained VAE, whose reconstruction property ensures encoded texture details, structural patterns, and basic semantic information as shown in Fig. 2. This motivates us to propose a novel VAE feature representation alignment framework to accelerate the training and enhance the generative fidelity of the vanilla diffusion models. As shown in Fig. 3(d), our VAE-REPA retains SiT's core framework while introducing an alignment component to bridge VAE features and SiT's latent space, requiring only the addition of a lightweight MLP.

3.2.1. VAE Feature Extraction

Since SiT is a latent generative model, it follows the two-stage LDM [33] training pipeline: the first stage trains a VAE to map raw images into a compact latent space, and the second stage trains a generative model (e.g., diffusion model) in this latent space. We follow SiT's setting and use the same pre-trained VAE from Stable Diffusion [34] (SD-VAE), leveraging its capability in learning meaningful, reconstruction-capable visual representations.

For an input image x, the SD-VAE encoder maps it to a compact latent embedding space as a VAE feature tensor $f_{\text{VAE}} \in \mathbb{R}^{C \times H \times W}$ (with a shape of 4×32×32 for 3×256×256 input images), where C denotes the feature channel dimension, and H, W are the spatial dimensions of the feature maps. Notably, in the training process of the second-stage diffusion models in LDM, f_{VAE} is typically pre-extracted and stored for later use to eliminate the on-the-fly feature extraction during training [27, 31, 40, 45]. Thus, we refer to these as off-the-shelf features.

3.2.2. Latent Representation Alignment

The core of our VAE-REPA is to align intermediate features of SiT with the VAE's prior-rich features during training. Let y_t denote SiT's latent state at time step t, generated by the interpolating process $y_t = a_t z + b_t \epsilon$. We extract an intermediate feature tensor h_{SiT} from a certain hidden layer of SiT, then project this intermediate feature into the same feature space as f_{VAE} using a lightweight MLP $\mathcal{P}(\cdot)$, yielding the aligned SiT feature $f_{\text{SiT}} = \mathcal{P}(h_{\text{SiT}})$, as shown in Fig. 3(d).

To enforce alignment, we adopt a *smooth $\mathcal{L}_1$ loss* as the alignment objective $\mathcal{L}_{\text{align}}$ on the element-wised feature difference $\Delta f = f_{\text{SiT}} - f_{\text{VAE}}$ as:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{z,\epsilon,t}\left[\sum_{i=1}^{N}\begin{cases}\frac{1}{2\beta}(\Delta f_i)^2 & \text{if } |\Delta f_i| \leq \beta, \\ \frac{|\Delta f_i|}{\beta} - \frac{1}{2} & \text{otherwise}\end{cases}\right], \quad (7)$$

where $N = C \times W \times H$ and β controls the threshold between the quadratic and linear regions and is set to 0.05 in all our experiments, the sum is taken over all elements of Δf, and the expectation is computed over the input z, Gaussian noise ϵ, and time step t. This loss encourages SiT's intermediate features to capture similar fine-grained details, structural patterns and semantic information as the VAE's feature maps, infusing valuable visual priors into the diffusion learning process.

3.2.3. Overall Training Objective

The overall training objective of our VAE-REPA framework is a weighted combination of vanilla SiT's denoising loss and our proposed alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_\phi + \lambda \cdot \mathcal{L}_{\text{align}}, \qquad (8)$$

where λ is a hyperparameter that balances the contribution of the two losses. By optimizing $\mathcal{L}_{\text{total}}$, the model retains SiT's advantage while leveraging the VAE's representations to refine latent space learning, ultimately accelerating the training process and improving the generation fidelity.

4. Experiments

4.1. Experimental Setup

Implementation details. All experiments are performed on the ImageNet dataset [6], where images are preprocessed to 256×256 resolution through center cropping and resizing. To ensure comparability, our training settings strictly follow the configurations specified in SiT [27] and REPA [45]. This encompasses the AdamW optimizer [26] with a constant learning rate of 1e-4, no weight decay, a fixed batch size of 256, and the use of SD-VAE [34] for latent VAE feature extraction. Regarding model architectures, we employ the B/2, L/2, and XL/2 designs from SiT, which process inputs with a patch size of 2. Sampling follows the SDE

Table 1. **Ablation studies** on ImageNet 256×256 without classifier-free guidance (CFG), which employs SiT-B/2 architectures trained for 400K iterations (with a batch size of 256). ↓ and ↑ indicate whether lower or higher values are better, respectively.

Align Depth	Timesteps	Align Objective	λ	MLP Setting	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑
Vanilla SiT-B/2 [27]	-	-	-	-	33.02	6.46	43.71	0.53	0.63
2	[0, 1]	smooth-ℓ_1	1.0	5-layer	28.89	6.20	51.64	0.56	0.63
3	[0, 1]	smooth-ℓ_1	1.0	5-layer	29.11	6.07	50.93	0.56	0.64
4	[0, 1]	smooth-ℓ_1	1.0	5-layer	29.51	6.15	50.50	0.56	0.65
6	[0, 1]	smooth-ℓ_1	1.0	5-layer	32.44	6.24	45.65	0.54	0.64
8	[0, 1]	smooth-ℓ_1	1.0	5-layer	36.20	6.50	40.88	0.52	0.64
2	[0, 1]	smooth-ℓ_1	1.0	5-layer	28.89	6.20	51.64	0.56	0.63
2	[0, 0.5]	smooth-ℓ_1	1.0	5-layer	30.04	6.24	49.80	0.56	0.64
2	[0.5, 1]	smooth-ℓ_1	1.0	5-layer	29.59	6.18	49.71	0.56	0.64
2	[0, 1]	smooth-ℓ_1	1.0	5-layer	28.89	6.20	51.64	0.56	0.63
2	[0, 1]	cosine	1.0	5-layer	29.30	6.20	50.51	0.56	0.63
2	[0, 1]	ℓ_1	1.0	5-layer	29.50	6.20	50.14	0.56	0.64
2	[0, 1]	ℓ_2	1.0	5-layer	29.40	6.16	50.46	0.56	0.64
2	[0, 1]	smooth-ℓ_1	1	5-layer	28.89	6.20	51.64	0.56	0.63
2	[0, 1]	smooth-ℓ_1	0.1	5-layer	30.10	6.21	50.38	0.55	0.64
2	[0, 1]	smooth-ℓ_1	0.5	5-layer	29.50	6.19	50.50	0.56	0.64
2	[0, 1]	smooth-ℓ_1	1.0	5-layer	28.89	6.20	51.64	0.56	0.63
2	[0, 1]	smooth-ℓ_1	1.0	2-layer	31.32	6.51	48.15	0.55	0.64

Euler–Maruyama solver with 250 steps, as in REPA [45]. Additional implementation specifics are available in the Appendix.

Evaluation protocol. To comprehensively assess image generation quality across multiple dimensions, we employ a rigorous set of quantitative metrics, including Fréchet Inception Distance (FID) [16] for realism, structural FID (sFID) [29] for spatial coherence, Inception Score (IS) [35] for class-conditional diversity, precision (Pre.) and recall (Rec.) [22] for sample fidelity and target distribution coverage, respectively. All metrics are computed on a standardized set of 50K generated samples to ensure statistical reliability [19, 40, 44, 45].

4.2. Ablation Studies and Discussions

Alignment Depth. We investigate the effects of applying our VAE-REPA at different network depths in Tab. 1. Our analysis reveals that applying the VAE-REPA alignment in earlier layers yields superior results, which is also consistent with previous works' [19, 40, 45] findings. Performance gradually degrades as the alignment shifts to deeper layers; we hypothesize this is because deeper layers focus on precise fine-grained details or semantic information, which go beyond the ability of VAE features, and imposing constraints here disrupts their natural refinement of such details. Notably, our method demonstrates the best results in layer 2, achieving an FID of 28.89, a reduction of 4.13 compared to the baseline. Based on this observation, we strategically apply VAE-REPA alignment at layers 2, 8, and 8 for the B, L, and XL architectures, respectively.

Timesteps. We also examine VAE-REPA's effect across timestep ranges in Tab. 1. Our analysis shows applying alignment over the full range $t \in [0, 1]$ yields the best performance, outperforming both $t \in [0, 0.5]$ and $t \in [0.5, 1]$. This is due to complementary alignment across noise levels: in low-noise stages, VAE-REPA refines coherent representations using the VAE's texture details and structural patterns; in high-noise stages, it helps the model against degradation via the VAE's visual properties. Covering $[0, 1]$ ensures the VAE's rich features are leveraged consistently throughout the entire diffusion process. Thus, we adopt the full timestep range $t \in [0, 1]$ for VAE-REPA to fully utilize the guidance of VAE features.

Alignment Objective. Next, we evaluate the impact of different alignment objectives in Tab. 1, including ℓ_1, ℓ_2, and smooth-ℓ_1. The results show that the smooth-ℓ_1 loss achieves the best overall performance, while the others also perform reasonably well. Consequently, this objective was adopted as the default for all subsequent experiments.

Effect of λ. We further explore the impact of the alignment loss weight λ in Tab. 1. Among the tested values, $\lambda = 1.0$ achieves the best overall performance across the major evaluation metrics. Therefore, we set $\lambda = 1.0$ as the default parameter in our experiments.

MLP Setting. We finally demonstrate the impact of different MLP layers in Tab. 1. The results show that using a 2-layer MLP (1M) yields suboptimal performance, while a 5-layer MLP (8M) achieves promising results. We believe that there exists a significant feature space discrepancy between SiT features and VAE features, as illustrated in Fig. 2, and a deeper MLP can better perform the feature transformations needed to align with the VAE's features, thus enabling more
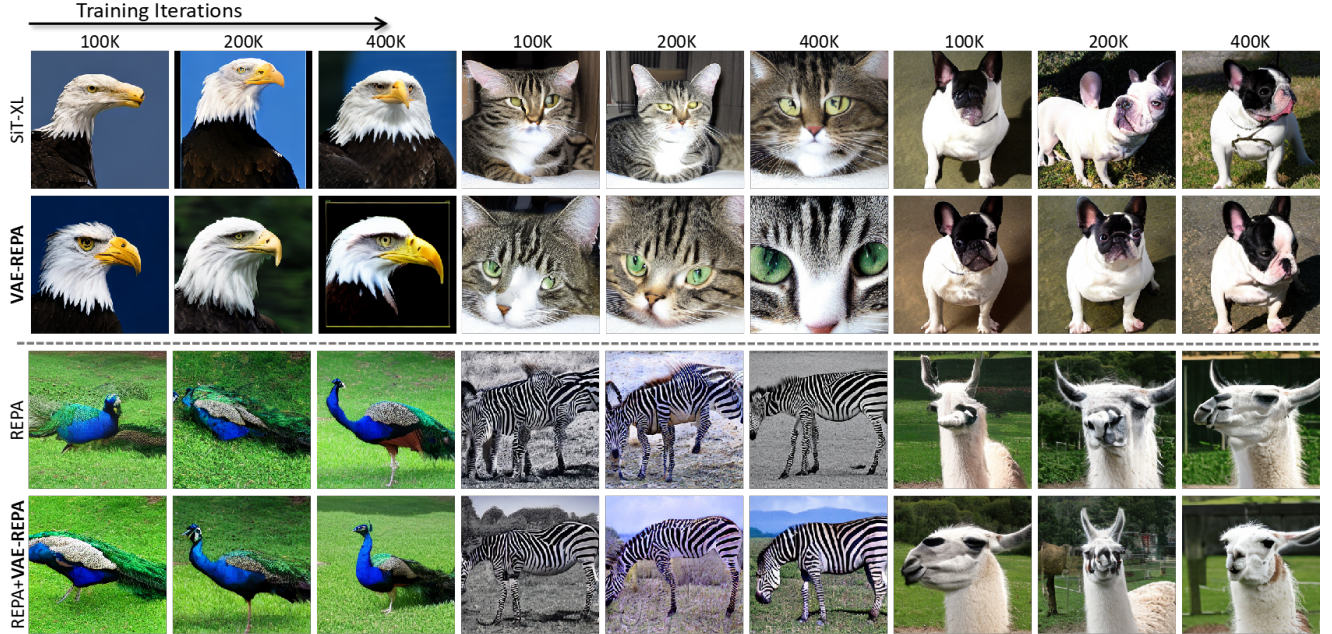
Figure 4. **VAE-REPA Improves Visual Scaling. Top**: Vanilla SiT-XL/2 and SiT-XL/2+VAE-REPA. **Bottom**: Vanilla REPA and REPA+VAE-REPA. Our method is verified to produce images with higher structural fidelity, finer details, and stronger semantic coherence at the same training steps compared with both vanilla SiT and vanilla REPA. Results for all methods are sampled using the same seed, noise, and class label, with a classifier-free guidance scale of 4.0 employed during sampling.

effective feature integration and refinement throughout the diffusion process. We apply a 5-layer MLP as the default setting for subsequent experiments.

4.3. System-level Comparisons

Accelerating training convergence. Tab. 2 provides a detailed comparison of training convergence across different model scales and iterations on ImageNet 256×256 without classifier-free guidance (CFG). The proposed VAE-REPA method consistently accelerates training convergence while reducing FID scores. For SiT-B/2, our VAE-REPA achieves an FID of 28.9 at 400K iterations, outperforming the baseline SiT-B/2 by 4.1 points. For SiT-L/2, VAE-REPA reaches an FID of 14.3 at 400K iterations, surpassing SiT-L/2's 18.8 at 400K and even SiT-XL/2's 14.6 at 600K. For SiT-XL/2, VAE-REPA gets an FID of 8.2 at 1M iterations, better than SiT-XL/2's 8.3 at 7M, representing a 7× training acceleration for better performance. Moreover, VAE-REPA continues to improve, reaching 6.6 at 4M iterations. This demonstrates that our VAE-REPA effectively accelerates training convergence across different model scales. Additionally, we present the generation quality comparison at different training steps in Fig. 4, where VAE-REPA stably enhances generation quality, further validating its effectiveness.

Compatibility with Other Methods. Tab. 2 also shows the compatibility of VAE-REPA with other methods. When combined with SiT-XL/2+REPA [45], using VAE-REPA achieves consistent improvements across 100K, 200K, and 400K iterations, with FID score reductions of 3.1, 1.9, and 1.1, respectively, outperforming SiT-XL/2+REPA at each corresponding iteration. Additionally, when integrated with VAVAE [44], VAE-REPA further reduces the FID to 4.4 at 400K iterations from 4.9, indicating that integrating our method with other advanced methods is compatible and can bring additional performance gains. Qualitatively, we present the generation quality of vanilla REPA and REPA+VAE-REPA at 100K, 200K, and 400K iterations in Fig. 4. On top of REPA, VAE-REPA generates more realistic images with more plausible details, further validating its compatibility.

Comparison with SOTA methods. Tab. 3 presents a comprehensive comparison against recent SOTA methods on ImageNet 256×256 with CFG, focusing on differences in external dependencies. Our proposed VAE-REPA method achieves competitive performance while demonstrating efficient training convergence **without any external dependencies**. Against methods with external diffusion decoders (e.g., MaskDiT [49], SD-DiT [50]), VAE-REPA surpasses MaskDiT's performance at 1300 epochs with an FID of 1.98 at only 200 epochs, and outperforms SD-DiT's result at 480 epochs using just 100 epochs. Against methods with external pretrained encoders (e.g., REPA [45], REG [40]), VAE-REPA achieves a comparable FID (1.52 vs. REPA's 1.42 at 800 epochs) and superior IS (316.2 vs. REPA's 311.4). Against methods with external teacher diffusion

Table 2. **FID comparison across training iterations for accelerated alignment methods.** All experiments are conducted on ImageNet (256×256) with a batch size of 256 and without CFG.

Method	#Params	Iter.	FID↓
SiT-B/2	130M	400K	33.0
+ VAE-REPA(ours)	130M	**400K**	**28.9**
SiT-L/2	458M	200K	28.2
SiT-L/2	458M	400K	18.8
+ VAE-REPA(ours)	458M	**200K**	**22.5**
+ VAE-REPA(ours)	458M	**400K**	**14.3**
SiT-XL/2	675M	400K	17.2
SiT-XL/2	675M	600K	14.6
SiT-XL/2	675M	7M	8.3
+ VAE-REPA(ours)	675M	**400K**	**11.7**
+ VAE-REPA(ours)	675M	**1M**	**8.2**
+ VAE-REPA(ours)	675M	**2M**	**7.0**
+ VAE-REPA(ours)	675M	**4M**	**6.6**
SiT-XL/2 + REPA [45]	675M	100K	19.4
SiT-XL/2 + REPA [45]	675M	200K	11.1
SiT-XL/2 + REPA [45]	675M	400K	7.9
+ VAE-REPA(ours)	675M	**100K**	**16.3**
+ VAE-REPA(ours)	675M	**200K**	**9.2**
+ VAE-REPA(ours)	675M	**400K**	**6.8**
VAVAE [44]	675M	400K	4.9
+ VAE-REPA(ours)	675M	**400K**	**4.4**

Table 3. **Comparison of the performance of different methods on ImageNet 256×256 with CFG.** Performance metrics are annotated with ↑ (higher is better) and ↓ (lower is better). "External" indicates whether they rely on external dependencies (e.g., extra encoders, diffusion teachers, or diffusion decoders).

Method	Epochs	FID↓	sFID↓	IS↑	Pre.↑	Rec.↑	External
Pixel diffusion							
ADM-U [7]	400	3.94	6.14	186.7	0.82	0.52	✗
VDM++ [20]	560	2.40	-	225.3	-	-	✗
Simple diffusion [18]	800	2.77	-	211.8	-	-	✗
CDM [17]	2160	4.88	-	158.7	-	-	
Latent diffusion, U-Net							
LDM-4 [33]	200	3.60	-	247.7	0.87	0.48	✗
Latent diffusion, Transformer + U-Net hybrid							
U-ViT-H/2 [2]	240	2.29	5.68	263.9	0.82	0.57	✗
DiffiT [13]	-	1.73	-	276.5	0.80	0.62	✗
Latent diffusion, Transformer							
DiT-XL/2 [31]	1400	2.27	4.60	278.2	**0.83**	0.57	✗
+ iSSD [28]	1400+0.8	2.02	4.22	250.0	0.81	0.60	✓
MaskDiT [49]	1600	2.28	5.67	276.6	0.80	0.61	✓
SD-DiT [50]	480	3.23	-	-	-	-	✓
SiT-XL/2 [27]	1400	2.06	4.50	270.3	0.82	0.59	✗
+ SRA [19]	800	1.58	4.65	305.7	0.80	0.65	✓
+ REPA [45]	800	1.42	4.70	311.4	0.80	0.63	✓
+ REG [40]	800	**1.36**	4.25	299.4	0.77	**0.66**	✓
+ VAE-REPA(ours)	100	2.39	4.54	252.5	**0.83**	0.56	✗
+ VAE-REPA(ours)	200	1.98	4.53	284.5	**0.83**	0.59	✗
+ VAE-REPA(ours)	400	1.84	4.50	306.5	0.82	0.61	✗
+ VAE-REPA(ours)	800	1.52	4.63	**316.2**	0.82	0.62	✗

Table 4. **Training computational computational cost comparison.** This table compares REPA, SRA, and VAE-REPA on ImageNet 256×256, detailing external forward parameters (EFP, formatted as external model parameters + MLP head parameters), training speed per batch (size 256) (TS), GFLOPs, and forward latency. Values in red parentheses indicate changes relative to the SiT-XL/2 baseline. These results were tested on H100 GPUs.

Method	EFP(M)	TS (Iter/s)↑	GFLOPs↓	Latency (ms)↓
SiT-XL/2	0 + 0	8.1	114.46	18.05
+REPA [45]	303 + 8	5.4(-33%)	194.50 (+70%)	29.03 (+61%)
+SRA [19]	481 + 2	5.1(-37%)	197.57 (+73%)	30.90 (+71%)
+VAE-REPA	0 + 18	7.2(-11%)	118.55 (+4%)	19.20 (+6%)

models (e.g., SRA [19]), VAE-REPA maintains better performance on both FID and IS at the same epochs. Additionally, VAE-REPA shows consistent improvement with increasing training epochs, validating its robustness and scalability. These results highlight that our method can compete with SOTA approaches across different external dependency paradigms while maintaining training efficiency and architectural simplicity. We provide selected qualitative results of SiT-XL/2+VAE-REPAin Fig. 1 and additional examples in the Appendix, all of which exhibit excellent image quality.

Training Computational Cost Comparison. Tab. 4 compares the training computational cost of typical extra-encoder-required REPA [45], dual-model SRA [19], and our VAE-REPA on ImageNet 256×256. Among the three methods, our VAE-REPA demonstrates the most significant advantages in computational efficiency: it requires 0 external model parameters, in stark contrast to REPA (303M) and SRA (481M). In terms of training speed, VAE-REPA maintains performance close to the baseline SiT-XL/2 (only 11% slower per batch), outperforming REPA (-33%) and SRA (-37%), which suffer larger slowdowns. For GFLOPs, VAE-REPA adds merely 4% compared to the baseline, far superior to the substantial increments of REPA (+70%) and SRA (+73%). Regarding forward latency, VAE-REPA introduces only a 6% increase relative to the baseline (attributed to the lightweight MLP alignment head), while REPA and SRA show much larger overheads (+61% and +71%, respectively). These results highlight VAE-REPA's remarkable training efficiency, with no external encoder parameters and minimal increments in training speed, GFLOPs and latency, making it significantly more lightweight than existing methods.

5. Conclusion

This work addresses the training acceleration of diffusion transformers. We propose VAE-REPA, a lightweight framework that leverages off-the-shelf pre-trained SD-VAE features for representation alignment. These features are inherently rich in texture details, structural patterns, and basic semantic information. Unlike existing methods that rely on external encoders that are not available in all domains or dual-model setups, VAE-REPA reuses pre-extracted

VAE features and only introduces a lightweight projection layer and an alignment loss. Experiments confirm that VAE-REPA accelerates training, enhances generation quality, and complements other methods, all with minimal additional computational overhead. This work demonstrates that pre-trained VAE visual priors are a powerful low-cost resource for efficient diffusion training. It offers a practical path to balancing efficiency and generation quality.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 2

[2] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023. 8

[3] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 4

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024. 2, 3

[5] Junyu Chen, Dongyun Zou, Wenkun He, Junsong Chen, Enze Xie, Song Han, and Han Cai. Dc-ae 1.5: Accelerating diffusion model convergence with structured latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19628–19637, 2025. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[10] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 2

[11] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Masked diffusion transformer is a strong image synthesizer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23164–23173, 2023. 4

[12] Shanghua Gao, Pan Zhou, Ming-Ming Cheng, and Shuicheng Yan. Mdtv2: Masked diffusion transformer is a strong image synthesizer. *arXiv preprint arXiv:2303.14389*, 2023. 4

[13] Ali Hatamizadeh, Jiaming Song, Guilin Liu, Jan Kautz, and Arash Vahdat. Diffit: Diffusion vision transformers for image generation. In *European Conference on Computer Vision*, pages 37–55. Springer, 2024. 8

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 3, 8

[18] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023. 3, 8

[19] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 2, 3, 4, 5, 6, 8

[20] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36: 65484–65516, 2023. 3, 8

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[22] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019. 6

[23] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 3

[24] Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025. 4

[25] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024. 4

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[27] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024. 2, 3, 4, 5, 6, 8

[28] Xinyin Ma, Runpeng Yu, Songhua Liu, Gongfan Fang, and Xinchao Wang. Diffusion model is effectively its own teacher. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12901–12911, 2025. 2, 4, 8

[29] Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021. 6

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3, 5, 8

[32] Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, et al. Lumina-image 2.0: A unified and efficient image generative framework. *arXiv preprint arXiv:2503.21758*, 2025. 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 5, 8

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 4, 5

[35] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[36] Wensong Song, Hong Jiang, Zongxing Yang, Ruijie Quan, and Yi Yang. Insert anything: Image insertion via in-context editing in dit. *arXiv preprint arXiv:2504.15009*, 2025. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[38] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 2, 4

[39] WanTeam. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[40] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467*, 2025. 2, 4, 5, 6, 7, 8

[41] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *Advances in Neural Information Processing Systems*, 37:121859–121881, 2024. 2

[42] Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025. 4

[43] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[44] Jingfeng Yao and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *arXiv preprint arXiv:2501.01423*, 2025. 4, 6, 7, 8

[45] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025. 2, 3, 4, 5, 6, 7, 8

[46] Xiaoyu Yue, Zidong Wang, Zeyu Lu, Shuyang Sun, Meng Wei, Wanli Ouyang, Lei Bai, and Luping Zhou. Diffusion models need visual priors for image generation. *arXiv preprint arXiv:2410.08531*, 2024. 4

[47] Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *arXiv preprint arXiv:2505.23656*, 2025. 2

[48] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 2

[49] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 2, 4, 7, 8

[50] Rui Zhu, Yingwei Pan, Yehao Li, Ting Yao, Zhenglong Sun, Tao Mei, and Chang Wen Chen. Sd-dit: Unleashing the power of self-supervised discrimination in diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8435–8445, 2024. 2, 4, 7, 8