GTA: GLOBAL TOKEN ALIGNMENT AS A BRIDGE FROM AUTOENCODING TO DENOISING

Anonymous authors
Paper under double-blind review

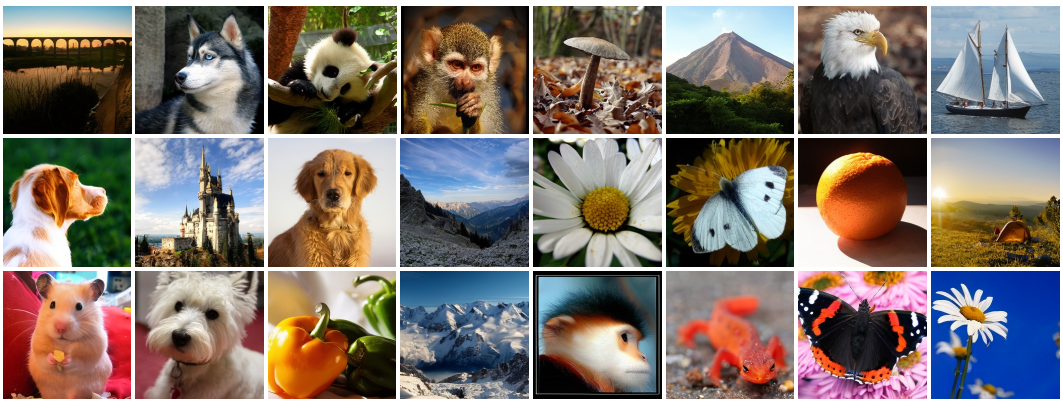

Figure 1: Qualitative Results on Imagenet 256 × 256 using GTA and SiT-XL.

ABSTRACT

Latent generative models are jointly limited by two factors: the quality of the latent space and the modeling capacity of the denoiser. Although recent representation-enhanced methods improve tokenizers or denoisers, the two stages still lack a shared semantic bridge. Global semantics formed during autoencoding may not be effectively inherited by the denoiser. During denoising, the evolving global semantics may also drift from the spatial latent representations. We introduce **GTA**, a **G**lobal **T**oken **A**lignment framework that bridges autoencoding and denoising through a shared global representation. GTA encourages the autoencoder to internalize image-level semantics during latent formation and maintains consistency with spatial latents throughout denoising. Experiments on ImageNet 256×256 show that GTA accelerates training and improves generation quality. With SD-VAE initialization, GTA achieves a gFID of 6.42 after 20 epochs. This outperforms REPA-E, which reaches 7.17 after 40 epochs. With VA-VAE initialization, GTA further achieves a gFID of 1.87 after 200 epochs. GTA also produces a more generation-friendly tokenizer. When GTA-VAE is frozen and new generative models are trained from scratch, it consistently outperforms the corresponding various types of VAEs under SiT, REPA, and REG.

1 INTRODUCTION

Latent generative models have become a dominant paradigm for high-resolution image generation (Esser et al., 2021; Rombach et al., 2022; Song et al., 2020). They decompose generation into two interdependent stages. A tokenizer (e.g., VAE) first compresses an image into a set of latent tokens (Kingma & Welling, 2013; Van Den Oord et al., 2017; Razavi et al., 2019). A denoiser then learns to transform noise into samples within the resulting latent space (Ho et al., 2020; Lipman et al., 2022; Peebles & Xie, 2023; Ma et al., 2024). The decomposition substantially reduces computational cost and enables scalable Transformer-based generators.

This efficiency, however, introduces a coupled representation bottleneck. (Gu et al., 2024) The denoiser can only model what the tokenizer preserves. A tokenizer is therefore judged not only by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

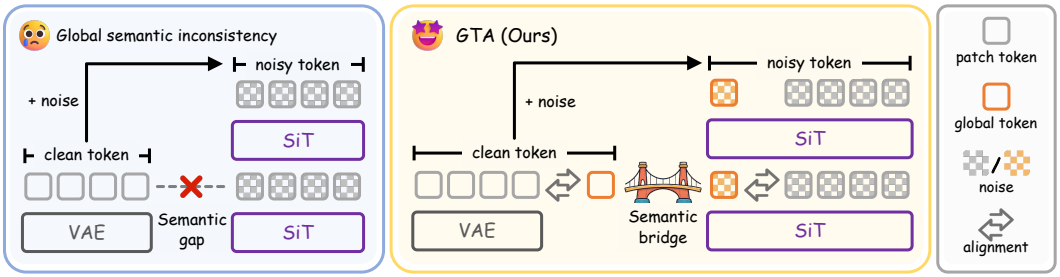

Figure 2: Motivation and overview of GTA. Without global-token alignment, latent formation and denoising may become semantically inconsistent (left). GTA aligns VAE features with a clean global token and synchronizes global and patch tokens during denoising, bridging autoencoding and denoising (right).

reconstruction fidelity, but also by how easily a downstream generative model can learn its latent distribution. (Yu et al., 2022; Chen et al., 2025; Zheng et al., 2025). Consequently, the autoencoding stage is not merely a preprocessing step: it determines what information survives compression and how that information is organized in the space modeled by the denoiser.

Global semantics are particularly important in this process. Conventional autoencoding objectives primarily encourage the preservation of local appearance and reconstruction details, but do not explicitly require latent-forming features to organize image-level factors such as object identity, global layout, or relationships between distant regions (Wen et al., 2025). These factors may be weakly reflected by pixel-level distortion metrics, yet they are essential for generating semantically coherent images (Podell et al., 2024). Meanwhile, large-scale visual encoders such as CLIP, DINO and I-JEPA have demonstrated that global image representations can capture transferable high-level semantics (Radford et al., 2021; Caron et al., 2021; Assran et al., 2023; Oquab et al., 2024). Recent work further shows that supervising diffusion Transformer features with such pretrained representations can accelerate optimization and improve generation quality (Shin et al., 2026; Xu et al., 2026; Singh et al., 2025).

REPA aligns intermediate denoiser features with clean visual representations, providing semantic supervision to the generative model (Yu et al., 2024). REPA-E extends representation alignment to end-to-end training, allowing the tokenizer and denoiser to be jointly optimized (Leng et al., 2025). REG introduces a class token as an additional generative variable and denoises it jointly with spatial tokens (Wu et al., 2026). These approaches improve different components of the latent generative pipeline, but two forms of semantic discontinuity remain. On the tokenizer side, end-to-end optimization does not explicitly anchor the latent-forming features to clean image-level semantics before they enter the low-dimensional bottleneck. On the denoiser side, jointly processing global and spatial tokens does not explicitly ensure that the two streams represent consistent global semantics as they evolve across noise levels (Kouzelis et al., 2026). As illustrated on the left of Fig. 2, the semantics encoded during latent formation and those propagated during denoising lack an explicit mechanism for remaining connected.

To address this problem, we propose **Global Token Alignment (GTA)**, a unified framework that establishes semantic continuity from autoencoding to denoising. During autoencoding, GTA globally aggregates the latent-forming features before the VAE output projector and aligns them with the clean global token extracted by a pretrained visual encoder. Applying supervision before dimensional compression allows image-level semantics to directly shape the formation of the latent representation. During denoising, the global token and VAE latent tokens are jointly noised and processed by SiT. GTA further aligns the global token state with the global representation aggregated from the spatial patch-token stream at the same network depth and noise level.

Through this design, the global token is not merely used as an external supervision target or an additional generative variable. Instead, it serves as a shared semantic reference across the two stages of latent generation: VAE GTA writes clean image-level semantics into the latent-forming features, while SiT GTA preserves their consistency as the noisy latent state evolves. As illustrated on the right of Fig. 2, GTA forms an explicit semantic bridge between the latent constructed by the tokenizer and the trajectory modeled by the denoiser.

Experiments on ImageNet 256×256 demonstrate that GTA improves both optimization efficiency and final generation quality. First, GTA further accelerates the convergence of end-to-end generative models. With SD-VAE initialization, GTA trained for only 20 epochs achieves a gFID of 6.42, outperforming REPA-E trained for 40 epochs at 7.17. And it has demonstrated stable performance under different tokenizers and different sizes of SiT models. More importantly, Using GTA for training resulted in a more generation-friendly VAE tokenizer. When the tokenizer is frozen and a new latent generative model is trained from scratch, GTA-VAE consistently yields stronger generation performance across standard SiT, REPA, and REG objectives. In addition, GTA-VAE also transfers effectively to text-to-image generation.

In summary, our main contributions are threefold:

- We identify the lack of explicit global semantic continuity between autoencoding and denoising as an important limitation of existing latent generative models, highlighting two related gaps: semantic anchoring during latent formation and global spatial consistency along the noisy generative trajectory.
- We propose GTA, a better end-to-end training recipe that explicitly connects autoencoders and denoisers via global token alignment.
- We demonstrate that GTA substantially accelerates end-to-end training and improves generation quality across tokenizer initializations and denoiser scales. Standalone LDM and transfer evaluations further show that GTA learns a more generation-friendly tokenizer whose benefits persist beyond the original end-to-end system.

2 Related Work

Semantic Representations and Tokenizers. ViT use a learnable global token to aggregate patch features into an image-level representation Dosovitskiy et al. (2020). Pretrained encoders such as CLIP, DINO and I-JEPA further demonstrate that global representations capture transferable semantics Radford et al. (2021); Caron et al. (2021); Assran et al. (2023); Oquab et al. (2024); Siméoni et al. (2025). Recent tokenizer studies show that reconstruction fidelity alone is insufficient; latent semantic structure and generative learnability also matter Yu et al. (2023); Yao et al. (2025); Zheng et al. (2023). VA-VAE and RAE therefore inject or preserve pretrained semantics during tokenizer construction Yao et al. (2025); Zheng et al. (2025); Bi et al. (2025); Du et al. (2025). Nevertheless, existing approaches typically use pretrained visual representations as external supervision for an individual module, without explicitly modeling how image-level global semantics should propagate from latent formation to latent denoising.

Efficient Training of Diffusion Transformers. Efficient Diffusion Transformer training has also received growing attention, with approaches targeting computational efficiency, optimization, and representation learning. MaskDiT reduces training costs through masked latent modeling, while FasterDiT improves optimization through timestep sampling and velocity-direction supervision Zheng et al. (2023); Yao et al. (2024). REPA aligns intermediate denoiser features from noisy inputs with clean pretrained representations, providing external semantic supervision beyond the denoising objective Yu et al. (2024). REPA-E further extends this alignment to jointly optimize the VAE and denoiser, allowing the latent space to adapt during training Leng et al. (2025). However, these methods do not explicitly connect tokenizer-side global semantics to the semantic representations evolving along the noisy generative trajectory.

Global Semantics in Latent Generation. REG treats a pretrained global token as an additional generative variable and jointly denoises it with spatial latent tokens Wu et al. (2026). Although this introduces image-level semantics into generation, joint processing alone does not enforce consistency between the global-token and patch-token streams across noise levels, nor does it anchor global semantics during VAE latent formation. GTA achieves continuous semantic expression that spans autoencoding and denoising by injecting global semantics into the VAE and synchronizing the semantic consistency between global tokens and patch tokens during the denoising process.

3 METHOD

3.1 OVERVIEW

We propose **Global Token Alignment (GTA)**, which uses an image-level global token as a shared semantic reference to connect latent formation in the VAE with latent denoising in SiT. As illustrated in Fig. 3, GTA comprises two complementary modules: VAE GTA and SiT GTA.

At the autoencoding stage, GTA globally aggregates the high-dimensional latent features before the VAE output projector and aligns the resulting representation with the global token from a pretrained visual encoder. This constraint introduces image-level semantics into latent formation before low-dimensional compression, thereby shaping the global semantic organization of the latent space.

At the denoising stage, we treat the global token as an additional generative variable and jointly model it with the VAE latent tokens. However, joint modeling does not guarantee semantic consistency between the two representations throughout the noisy trajectory. GTA therefore aligns the evolving global-token state with the global representation of the patch-token stream.

In this way, VAE GTA injects global semantics during latent formation, while SiT GTA preserves their consistency throughout denoising, together establishing a continuous semantic pathway from autoencoding to denoising.

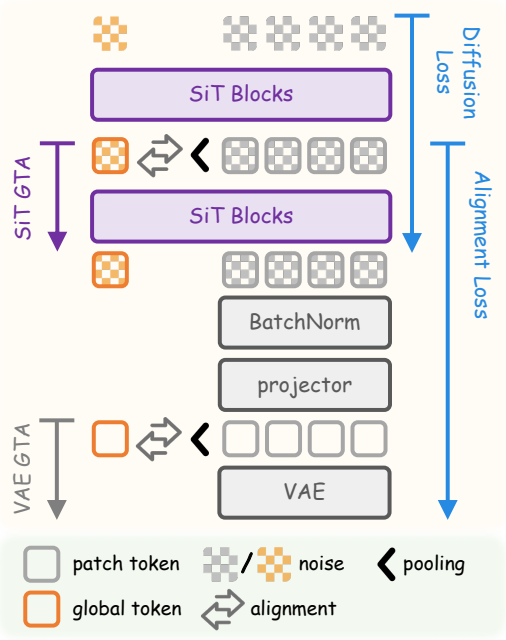

Figure 3: **Overview of Global Token Alignment (GTA).** VAE GTA aligns pooled latent-forming features with a clean global token before dimensional compression, while SiT GTA aligns evolving global-token states with pooled patch-token representations at noise levels, preserving semantic continuity from autoencoding to denoising.

3.2 PRELIMINARIES

Given an input image x and its class condition y, the VAE produces high-dimensional features before its output projector, which are subsequently mapped to and normalized in the latent space:

$$H^{\mathrm{V}} = E_{\phi}^{\mathrm{feat}}(x),\ Z_0 = \mathcal{P}_{\mathrm{VAE}}(H^{\mathrm{V}}),\ \bar{Z}_0 = \mathrm{BN}(Z_0). \tag{1}$$

We refer to H^{V} as the *latent-forming features*, since they directly determine the information retained in the final latent representation. Here, $\mathcal{P}_{\mathrm{VAE}}$ denotes the VAE output projector, while Z_0 and $\bar{Z}_0$ denote the original and dynamically normalized latents, respectively.

A frozen pretrained visual encoder F_ω extracts an image-level global token from the clean image:

$$c_0 = F_\omega(x). \tag{2}$$

Following Wu et al. (2026), at a time step t, we independently add Gaussian noise to $\bar{Z}_0$ and c_0:

$$Z_t = \alpha_t \bar{Z}_0 + \sigma_t \epsilon_z, \qquad \epsilon_z \sim \mathcal{N}(0, I), \tag{3}$$

$$c_t = \alpha_t c_0 + \sigma_t \epsilon_c, \qquad \epsilon_c \sim \mathcal{N}(0, I). \tag{4}$$

The two variables share the same step and noise schedule, while ϵ_z and ϵ_c are sampled independently.

After input embedding and concatenation, the SiT input sequence is

$$H_t^{(0)} = \left[g_t^{(0)}; X_t^{(0)}\right] = \left[\mathcal{E}_c(c_t); \mathcal{E}_z(Z_t)\right], \tag{5}$$

where $\mathcal{E}_c$ and $\mathcal{E}_z$ denote the input embeddings for the global token and VAE latent, respectively. Here, $g_t^{(0)}$ is initialized from the noisy global token c_t, while $X_t^{(0)}$ denotes the spatial latent-token sequence obtained from Z_t.

After the first ℓ SiT blocks, the intermediate state is

$$H_t^{(\ell)} = T_\theta^{1:\ell}\left(H_t^{(0)}, t, y\right) = \left[g_t^{(\ell)}; X_t^{(\ell)}\right], \tag{6}$$

where $T_\theta^{1:\ell}$ denotes the first ℓ SiT blocks, and $g_t^{(\ell)}$ and $X_t^{(\ell)}$ denote the global-token state and spatial latent-token states at layer ℓ, respectively.

3.3 VAE Global Token Alignment

The VAE defines the latent space modeled by the subsequent generative model. However, conventional autoencoding objectives primarily focus on reconstruction fidelity and do not explicitly constrain image-level semantics during latent formation. Since the final latent has already undergone substantial dimensional compression, semantic supervision applied only to the low-dimensional bottleneck may not directly shape the semantic structure of the preceding features.

To address this issue, VAE GTA introduces global semantic supervision before the VAE output projector and operates on the latent-forming features H^{V} defined in Eq. equation 1. Since these features directly determine the information retained in the final latent, they provide a suitable interface for introducing image-level semantics before low-dimensional compression.

Specifically, we first spatially aggregate H^{V} and project the resulting representation into the global-token feature space:

$$\hat{c}^{\mathrm{V}} = \mathcal{P}_{\text{V-GTA}}\left(\mathrm{GAP}\left(H^{\mathrm{V}}\right)\right), \tag{7}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, and $\mathcal{P}_{\text{V-GTA}}$ is a lightweight projection head. We then align this representation with the global token c_0 extracted from the clean image by the pretrained visual encoder:

$$\mathcal{L}_{\text{V-GTA}} = 1 - \cos\left(\hat{c}^{\mathrm{V}}, c_0\right), \tag{8}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity.

VAE GTA thus provides a clean image-level semantic anchor for latent formation. Rather than introducing an additional semantic token into the VAE, it encourages the entire spatial feature field to collectively encode the global semantics of the clean image. Consequently, high-level information such as object identity, global layout, and scene structure can be absorbed by the latent-forming features before low-dimensional compression.

The VAE-side objective is

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{VAE}}^{\mathrm{base}} + \lambda_{\mathrm{V}}\mathcal{L}_{\text{V-GTA}}, \tag{9}$$

where $\mathcal{L}_{\mathrm{VAE}}^{\mathrm{base}}$ denotes the VAE-side objective in the original end-to-end training framework, including the standard VAE objectives and the REPA representation-alignment loss, and λ_{V} controls the weight of VAE GTA. This constraint encourages the learned latent to preserve reconstruction-oriented local information while organizing image-level semantics more explicitly.

3.4 SiT Global Token Alignment

To extend image-level semantics from latent formation to latent denoising, we treat the global token extracted by the pretrained visual encoder as an additional generative variable and jointly noise and denoise it with the VAE latent tokens.

However, joint modeling does not guarantee that these representations remain semantically consistent throughout generation. The global token provides a compact representation of image-level semantics, whereas the latent patch tokens encode the evolving generative state in a spatially distributed manner. Although the two token streams interact through self-attention, they may evolve along semantically mismatched trajectories without an explicit consistency constraint.

SiT GTA addresses this issue by synchronizing the global semantics represented by the global-token stream and the latent patch-token stream at each sampled noise state. After the first ℓ SiT blocks,

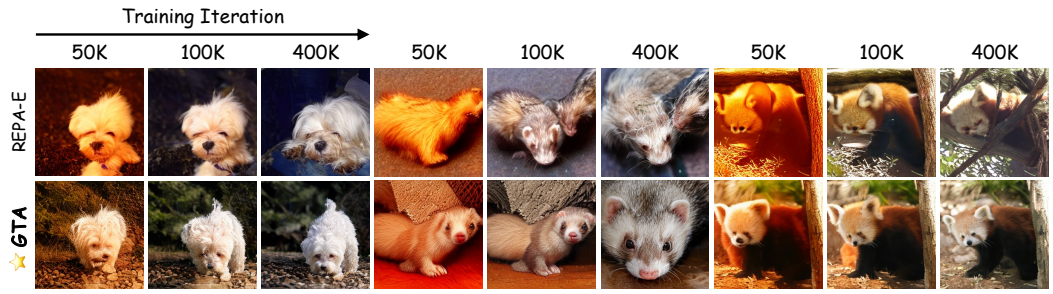

Figure 4: Qualitative comparison along training. Samples generated by REPA-E and GTA at 50k, 100k, and 400k iterations under identical class conditions, initial noise, sampler, and sampling steps.

the joint representation is

$$H_t^{(\ell)} = \left[g_t^{(\ell)}; X_t^{(\ell)} \right], \tag{10}$$

where $g_t^{(\ell)}$ denotes the global-token state and $X_t^{(\ell)}$ denotes the spatial patch-token states.

Image-level semantics are implicitly distributed across the spatial patch-token stream. We therefore apply global average pooling to $X_t^{(\ell)}$ and align its global representation with the evolving global-token state $g_t^{(\ell)}$:

$$\mathcal{L}_{\text{S-GTA}} = 1 - \cos\left(\text{GAP}\left(X_t^{(\ell)} \right), g_t^{(\ell)} \right), \tag{11}$$

where $\text{GAP}(\cdot)$ denotes global average pooling and $\cos(\cdot, \cdot)$ denotes cosine similarity.

Since different noise levels are sampled during training, SiT GTA does not merely align two intermediate representations at a single state. Instead, it imposes a unified semantic-consistency constraint over the entire noisy trajectory. At high noise levels, the two token streams are encouraged to agree on coarse and uncertain global structures. As the noise level decreases, they progressively synchronize toward clearer and more complete image-level semantics.

The SiT-side objective is

$$\mathcal{L}_{\text{SiT}} = \mathcal{L}_{\text{SiT}}^{\text{base}} + \lambda_{\text{S}} \mathcal{L}_{\text{S-GTA}}, \tag{12}$$

where $\mathcal{L}_{\text{SiT}}^{\text{base}}$ denotes the SiT-side objective in the original end-to-end training framework, including the latent-token denoising, global-token denoising, and representation-alignment objectives, and λ_{S} controls the weight of SiT GTA.

Through this trajectory-wise global semantic synchronization, SiT GTA maintains semantic consistency between the explicit global-token state and the spatial latent-token stream throughout denoising, allowing image-level semantics to persist along the entire latent generative trajectory.

Overall, VAE GTA and SiT GTA operate at two critical stages of latent generation. VAE GTA injects image-level semantics from pretrained visual representations into the latent-forming features before low-dimensional compression, while SiT GTA synchronizes the global semantics expressed by the dynamic global-token state and the spatial patch-token stream across different noise levels. Together, they establish a continuous global semantic pathway from latent formation to latent denoising, ensuring that image-level semantics are not only encoded into the latent space but also preserved throughout the entire generative trajectory.

4 Experiments

We conduct extensive experiments to evaluate GTA and test our central hypothesis. Specifically, we address the following three questions:

1. Can GTA improve generation quality and accelerate convergence in end-to-end training?
2. Does GTA produce a more generation-friendly VAE that benefits independently trained downstream generative models?
3. How do the individual GTA components and the choice of pretrained visual encoder affect generation performance?

4.1 Experimental Setup

Dataset, models, and protocols. We conduct experiments on ImageNet-1K at a resolution of 256×256 and adopt Scalable Interpolant Transformer (SiT) Ma et al. (2024) as the latent generative model. Unless otherwise specified, we initialize the tokenizer from SD-VAE, use DINOv2-B as the pretrained visual encoder and SiT-L as the denoiser, and train for 100k iterations following REPA-E Leng et al. (2025). We follow the optimization, sampling, and evaluation protocols of SiT, REPA, and REPA-E. Results are reported without classifier-free guidance unless noted. We evaluate generation quality using generation FID (gFID), structural FID (sFID), and Inception Score (IS), and additionally report precision and recall for the model-scale and visual-encoder studies. More detailed experimental parameters are presented in the appendix to ensure that our results are reproducible.

Table 1: Accelerated generation performance. * indicates that the VAE is updated during end-to-end training. All results are reported with SiT-XL unless otherwise specified.

Method	Tokenizer	Epochs	gFID↓	sFID↓	IS↑
Without End-to-End Tuning					
MaskDiT (2023)	SD-VAE	1600	5.69	10.34	177.9
DiT (2023)		1400	9.62	6.85	121.5
SiT (2024)		1400	8.61	6.32	131.7
FasterDiT (2024)		400	7.91	5.45	131.3
REPA (2024)	SD-VAE	20	19.40	6.06	67.4
		40	11.10	5.05	100.4
		80	7.90	5.06	122.6
		800	5.90	5.73	157.8
With End-to-End Tuning					
REPA-E (2025)	SD-VAE*	20	12.83	5.04	88.8
		40	7.17	4.39	123.7
		80	4.07	4.60	161.8
	VA-VAE*	20	8.92	5.34	100.4
		40	5.41	4.38	132.3
		80	3.94	4.48	152.0
GTA (Ours)	SD-VAE*	20	6.42	4.94	146.0
		40	3.52	4.59	179.6
		80	2.51	4.67	200.3
	VA-VAE*	20	5.62	4.81	136.3
		40	3.55	5.11	168.0
		80	2.47	4.44	189.1
		200	1.87	4.03	208.6

4.2 End-to-End Training Efficiency and Generation Performance

Quantitative comparison. Table 1 shows that GTA provides a substantial convergence advantage from early training onward. With SD-VAE initialization, GTA achieves gFID scores of 6.42, 3.52, and 2.51 at 20, 40, and 80 epochs, substantially outperforming REPA-E at 12.83, 7.17, and 4.07 under the same budgets. Notably, GTA at 20 epochs surpasses REPA-E at 40 epochs (6.42 vs. 7.17), and GTA at 40 epochs surpasses REPA-E at 80 epochs (3.52 vs. 4.07), indicating stronger performance with approximately half of the training budget. Compared with a fixed tokenizer, GTA at 20 epochs also outperforms REPA at 80 epochs (6.42 vs. 7.90). Its IS further increases from 146.0 to 200.3 between 20 and 80 epochs, showing that GTA accelerates early convergence while continuing to improve at later stages.

The same trend holds with VA-VAE initialization: GTA reaches gFID scores of 5.62, 3.55, and 2.47 at 20, 40, and 80 epochs and consistently outperforms the corresponding REPA-E results. Extending training to 200 epochs further reduces gFID to 1.87. Thus, GTA is not merely an auxiliary regularizer for early optimization. It continues to benefit end-to-end latent generation under longer training schedules and with a stronger tokenizer initialization.

Scalability with model size. Table 2 further evaluates SiT-B, SiT-L, and SiT-XL. GTA reduces gFID from 49.5 to 21.8 for SiT-B, from 24.1 to 8.1 for SiT-L, and from 19.4 to 6.4 for SiT-XL. These improvements correspond to relative reductions of approximately 56.0%, 66.6%, and 66.9%, respectively. Meanwhile, IS increases from 27.5 to 67.1, from 55.7 to 128.4, and from 67.4 to 146.0 across the three model scales. GTA achieves greater gains on larger denoisers, and this trend highlights the scalability of GTA.

Qualitative comparison along training. We provide a qualitative comparison between GTA and REPA-E in Fig. 4. We use checkpoints at 50K, 100K, and 400K training iterations to generate images from the same initial noise and class labels. As shown in Fig. 4, GTA produces higher-quality samples than REPA-E and generates more structurally meaningful images, particularly during the early stages of training.

Table 2: Scalability across SiT model sizes.

Diff. Model	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
SiT-B (130M)	49.5	7.00	27.5	0.46	0.59
+ GTA (Ours)	21.8	6.28	67.1	0.65	0.57
SiT-L (458M)	24.1	6.25	55.7	0.62	0.60
+ GTA (Ours)	8.1	4.79	128.4	0.76	0.56
SiT-XL (675M)	19.4	6.06	67.4	0.64	0.61
+ GTA (Ours)	6.4	4.94	146.0	0.78	0.55

Table 3: Standalone gFID with frozen tokenizers and different downstream objectives.

VAE	rFID	Alignment Method		
		None	REPA	REG
SD-VAE	0.74	46.81	22.74	11.72
E2E-SDVAE	0.53	25.18	14.76	9.02
GTA-SDVAE (Ours)	0.47	22.78	13.21	7.56
VA-VAE	0.28	15.57	12.54	9.76
E2E-VAVAE	0.28	14.30	9.84	7.62
GTA-VAVAE (Ours)	0.27	13.26	8.90	7.14

Table 4: Disentangling system-level gains from tokenizer-level improvements.

Method	VAE Source	gFID↓	sFID↓	IS↑
End-to-End Tuning				
REPA-E	SD-VAE	16.3	5.69	75.0
+ Global Modeling		9.2	4.68	120.2
+ GTA		8.1	4.79	128.4
Standalone LDM Training				
SiT-L	REPA-E	25.2	4.93	48.6
	+ Global Modeling	26.1	4.53	47.6
	+ GTA	22.8	5.32	53.4

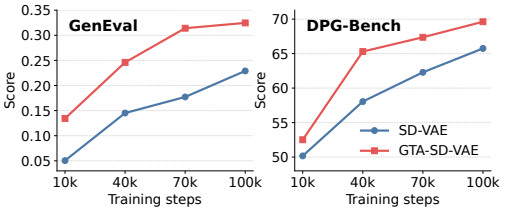

Figure 5: **Text-to-image transfer.** GTA-SD-VAE consistently outperforms SD-VAE on GenEval and DPG-Bench.

4.3 Generation-Friendliness of GTA-VAE

End-to-end generation is jointly determined by the tokenizer and denoiser, so stronger system-level performance does not necessarily imply a more generation-friendly VAE. We freeze the VAEs obtained from end-to-end training and independently retrain downstream latent generative models, directly testing whether GTA improves the tokenizer itself rather than only facilitating optimization of the original denoiser.

Standalone generation and transfer across alignment methods. For a fair comparison across different VAEs, Table 3 further evaluates each frozen tokenizer using the same downstream alignment objectives and training budget. GTA-VAE consistently achieves the best gFID under standard SiT, REPA, and REG for both the SD-VAE and VA-VAE families. This consistent advantage shows that the improvements do not arise from using a stronger downstream objective; instead, GTA improves the underlying latent space itself. Overall, GTA-VAE can serve as a drop-in tokenizer that accelerates downstream generation and retains its advantage under different training budgets and generative objectives.

Table 5 compares the system-level performance of GTA-VAVAE with other tokenizers under conventional LDM training on ImageNet 256×256. GTA training substantially improves the downstream generation performance of the VAE. At 80 epochs, GTA-VAVAE with REG achieves gFID scores of 2.36 without CFG and 1.48 with CFG, outperforming E2E-VAVAE with REPA at 3.46 and 1.67, respectively. At 800 epochs, GTA-VAVAE further reaches gFID scores of 1.47 and 1.14 without and with CFG, respectively.

From system-level gains to tokenizer-level improvements. Table 4 reveals an important distinction between improving the jointly trained system and improving the tokenizer itself. Adding global-token modeling to REPA-E substantially improves end-to-end generation, reducing gFID from 16.3 to 9.2. However, this gain does not transfer to the tokenizer: after freezing the resulting VAE and training a new LDM from scratch, standalone gFID slightly degrades from 25.2 to 26.1.

GTA exhibits a different behavior. It further improves end-to-end gFID to 8.1, while simultaneously reducing standalone gFID to 22.8. Therefore, the main role of GTA is not merely to introduce an additional global generative variable, but to make the semantic information carried by this variable beneficial to latent formation itself. This distinction is important: global-token modeling primarily benefits the jointly optimized generation system, whereas GTA additionally produces a latent space whose advantage persists after the original denoiser is removed.

Table 5: System-Level Performance on ImageNet 256 × 256

Tokenizer	Method	Training Epochs	#params	rFID	Generation w/o CFG					Generation w/ CFG				
					gFID	sFID	IS	Prec.	Rec.	gFID	sFID	IS	Prec.	Rec.
AutoRegressive (AR)														
VQGAN (2021)	LlamaGen (2024)	300	3.1B	0.59	9.38	8.24	112.9	0.69	0.67	2.18	5.97	263.3	0.81	0.58
VQVAE (2017)	VAR (2024)	350	2.0B	-	-	-	-	-	-	1.80	-	365.4	0.83	0.57
Latent Diffusion Models (LDM)														
SD-VAE (2022)	SiT (2024)	1400	675M	0.61	8.61	6.32	131.7	0.68	0.67	2.06	4.50	270.3	0.82	0.59
	FasterDiT (2024)	400	675M		7.91	5.45	131.3	0.67	0.69	2.03	4.63	264.0	0.81	0.60
Representation Alignment Methods														
VA-VAE (2025)	LightningDiT (2025)	80	675M	0.28	4.29	-	-	-	-	-	-	-	-	-
		800	675M		2.05	4.37	207.7	0.77	0.66	1.35	4.15	295.3	0.79	0.65
E2E-VAVAE (2025)	REPA (2024)	80	675M	0.28	3.46	4.17	159.8	0.77	0.63	1.67	4.12	266.3	0.80	0.63
		800	675M		1.83	4.22	217.3	0.77	0.66	1.26	4.11	314.9	0.79	0.66
GTA-VAVAE (Ours)	REG (2026)	80	677M	0.27	2.36	4.49	190.7	0.79	0.61	1.48	4.45	276.1	0.81	0.61
		800	677M		1.47	4.38	230.7	0.79	0.64	1.14	4.33	309.0	0.79	0.65

Table 6: Ablation of the GTA components.

Method	Model	gFID↓	sFID↓	IS↑
End-to-End Tuning				
REPA-E	SiT-L	16.3	5.69	75.0
GTA (w/o VAE Align)	SiT-L	9.18	5.18	119.6
GTA (w/o SiT Align)		8.43	4.85	125.8
GTA		8.05	4.79	128.4

Table 7: Variation in representation encoder.

Target Repr.	gFID↓	sFID↓	IS↑	Prec.↑	Rec.↑
I-JEPA-H (2023)	15.88	5.38	77.72	0.69	0.57
CLIP-L (2021)	17.53	5.24	73.20	0.68	0.58
DINOv2-B (2024)	8.05	4.79	128.39	0.76	0.56
DINOv2-L (2024)	11.43	4.91	106.74	0.73	0.56

Transfer to text-to-image generation. Following Fuse-DiT Tang et al. (2025), we further evaluate GTA-VAE for text-to-image generation using DiT-2.5B, a Gemma-2B Team et al. (2024) text encoder, CC-12M training, and 25-step sampling with a guidance scale of 6. As shown in Fig. 5, GTA-SD-VAE consistently outperforms SD-VAE on both GenEval and DPG-Bench from 10k to 100k training steps, demonstrating effective transfer to text-conditioned generation.

4.4 Ablation Studies

Contributions of VAE GTA and SiT GTA. Table 6 shows that VAE GTA and SiT GTA independently reduce the REPA-E baseline gFID from 16.3 to 8.43 and 9.18, respectively. VAE GTA introduces a clean image-level semantic boundary condition during latent formation. SiT GTA further preserves and leverages the benefits of this semantic injection by synchronizing the noisy global-token and patch-token streams across different noise states. Combining both gives the best overall result 8.05 gFID, confirming that the objectives are complementary rather than redundant.

Choice of pretrained visual encoder. Table 7 shows that GTA is effective across diverse pretrained encoders. GTA also follows a trend similar to prior work across different representation sources Wu et al. (2026); Yu et al. (2024). DINOv2-B performs best overall, achieving 8.05 gFID, and is therefore used as the default encoder.

5 Conclusion

We propose **Global Token Alignment (GTA)**, which establishes a continuous semantic pathway between VAE latent formation and SiT latent denoising through explicit global semantic alignment. GTA injects image-level global semantics from a pretrained visual encoder into the latent-forming features of the VAE, allowing global information to be explicitly preserved before entering the low-dimensional latent space. During denoising, GTA further aligns the dynamic global-token state with the implicit global representation encoded by the spatial patch-token stream, thereby maintaining semantic consistency throughout the denoising trajectory. Experiments demonstrate that GTA substantially accelerates the convergence of end-to-end latent generative models and consistently improves performance across different VAE initializations, SiT model scales, and pretrained visual encoders. Moreover, GTA produces a more generation-friendly VAE tokenizer, referred to as GTA-VAE. Overall, GTA highlights the importance of explicitly modeling global semantic continuity between autoencoding and denoising, and provides a simple yet effective approach for learning semantically organized and generation-friendly latent representations.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Tianci Bi, Xiaoyi Zhang, Yan Lu, and Nanning Zheng. Vision foundation models can be good tokenizers for latent diffusion models. *arXiv e-prints*, pp. arXiv–2510, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2025.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Sinan Du, Jiahao Guo, Bo Li, Shuhao Cui, Zhengzhuo Xu, Yifu Luo, Yongxian Wei, Kun Gai, Xinggang Wang, Kai Wu, et al. Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction. *arXiv preprint arXiv:2511.23386*, 2025.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12868–12878. IEEE, 2021.

Yuchao Gu, Xintao Wang, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Rethinking the objectives of vector-quantized tokenizers for image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7631–7640, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Theodoros Kouzelis, Efstathios Karypidis, Ioannis Kakogeorgiou, Spyridon Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. *Advances in Neural Information Processing Systems*, 38:16685–16714, 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18262–18272, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The eleventh international conference on learning representations*, 2022.

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*, pp. 23–40. Springer, 2024.

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. *Transactions on Machine Learning Research Journal*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 10684–10695, 2022.

Jaeyo Shin, Jiwook Kim, and Hyunjung Shim. Representation alignment for just image transformers is not easier than you think. *arXiv preprint arXiv:2603.14366*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Bingda Tang, Boyang Zheng, Xichen Pan, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. *arXiv preprint arXiv:2505.10046*, 2025.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

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

Xin Wen, Bingchen Zhao, Ismail Elezi, Jiankang Deng, and Xiaojuan Qi. " principal components" enable a new language of images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16641–16651, 2025.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Ming-Ming Cheng, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *Advances in Neural Information Processing Systems*, 38:7714–7743, 2026.

Shaodong Xu, Zhendong Wang, Litong Gong, Zexian Li, Wengang Zhou, Tiezheng Ge, and Houqiang Li. Beyond point-wise matching: Structural representation alignment for accelerating diffusion transformers. *arXiv preprint arXiv:2605.16949*, 2026.

Jingfeng Yao, Wang Cheng, Wenyu Liu, and Xinggang Wang. Fasterdit: Towards faster diffusion transformers training without architecture modification. *arXiv preprint arXiv:2410.10356*, 2024.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved VQ-GAN. In *International Conference on Learning Representations*, 2022.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

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. *arXiv preprint arXiv:2410.06940*, 2024.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

A ETHICAL STATEMENT

This work studies Global Token Alignment (GTA) for improving latent generative models using existing datasets, including ImageNet-1K and CC-12M, and pretrained models. The resulting models may inherit data biases or be misused to generate deceptive or harmful content. Our evaluations address generation quality and training convergence, not fairness, privacy, or safety. Model use and release should respect applicable dataset and model terms and intellectual-property rights. We recommend application-specific risk assessment, content screening, and transparent identification of synthetic images before deployment.

B REPRODUCIBILITY STATEMENT

We will provide comprehensive information to facilitate the control of randomness during training and validation, including but not limited to hardware specifications, software environment configurations, and model hyperparameters. Furthermore, we will release the complete source code alongside detailed training logs. All relevant details will be exhaustively documented in the supplementary materials.

C DISCLOSURE OF LLM USAGE

In the research and preparation of this manuscript, large language models (LLMs) were utilized for code implementation and language polishing. All code generated by an LLM underwent rigorous review, debugging, and comprehensive testing by the authors. The core algorithms and experimental design were conceived and executed directly by the authors. The LLM served solely as an assistive tool, with its outputs consistently under the authors' supervision and control. The authors assume full responsibility for all content presented in this work.

D SUPPLEMENTARY MATERIALS OVERVIEW

This supplementary material provides additional experimental details and results for the main paper. The first section presents a detailed analysis of the reconstruction performance of GTA-VAE. The second section reports the implementation details and hyperparameters of GTA, including the training configurations, loss weights, and classifier-free guidance settings. The third section provides additional qualitative generation results produced by GTA.

E RECONSTRUCTION QUALITY VERSUS GENERATIVE ADAPTABILITY

To further examine whether GTA improves generative adaptability at the cost of reconstruction quality, we evaluate both reconstruction-oriented metrics and standalone generation performance for different frozen tokenizers. For standalone generation, we train the same downstream latent generative model from scratch without representation-alignment objectives, such that the resulting gFID primarily reflects the learnability of the corresponding latent space.

As shown in Table 8, GTA improves both distribution-level reconstruction quality and standalone generation performance while largely preserving conventional reconstruction fidelity. For the SD-VAE family, GTA-SDVAE reduces rFID from 0.74 for the original SD-VAE and 0.53 for E2E-SDVAE to 0.47. This improvement is accompanied by a substantial reduction in standalone gFID, from 46.81 and 25.18 to 22.78, respectively. For the VA-VAE family, whose original reconstruction quality is already strong, GTA-VAVAE further reduces rFID from 0.28 to 0.27 and improves standalone gFID from 15.57 for VA-VAE and 14.30 for E2E-VAVAE to 13.26.

The remaining reconstruction metrics exhibit only modest numerical variations. Compared with the original pretrained VAEs, GTA results in small decreases in PSNR and SSIM, while the changes in LPIPS also remain limited. In particular, GTA-SDVAE achieves an LPIPS of 0.132, which is close to the 0.130 of SD-VAE and better than the 0.139 of E2E-SDVAE. Similarly, GTA-VAVAE improves LPIPS from 0.110 for E2E-VAVAE to 0.105, while remaining close to the 0.096 of the original

Tokenizer	rFID↓	PSNR↑	LPIPS↓	SSIM↑	gFID↓
SD-VAE	0.74	26.91	0.130	0.74	46.81
E2E-SDVAE	0.53	26.10	0.139	0.72	25.18
GTA-SDVAE	0.47	25.70	0.132	0.71	22.78
VA-VAE	0.28	27.96	0.096	0.79	15.57
E2E-VAVAE	0.28	27.63	0.110	0.77	14.30
GTA-VAVAE	0.27	27.20	0.105	0.76	13.26

Table 8: Reconstruction quality and standalone generation performance. All tokenizers are frozen when training the downstream generative model. The gFID results are obtained using the same standalone training protocol without downstream representation alignment. Lower rFID, LPIPS, and gFID are better, while higher PSNR and SSIM are better.

VA-VAE. These relatively small differences indicate that GTA largely retains the reconstruction capability of the pretrained tokenizers rather than introducing a substantial reconstruction penalty.

In contrast to these minor variations, the improvement in rFID is more consistent, particularly for the SD-VAE family, suggesting that GTA enhances the perceptual and distributional quality of reconstructed images. More importantly, the improved reconstruction distribution is accompanied by consistently lower standalone gFID across both tokenizer families. Taken together, these results demonstrate that GTA produces a more generation-friendly latent representation while maintaining comparable pixel- and perceptual-level reconstruction quality. This supports the effectiveness of global token alignment in organizing image-level semantic information without materially compromising the reconstruction capability of the tokenizer.

F Implementation Details

Setup. Unless otherwise specified, we follow the same experimental settings as REPA-E. All ImageNet experiments are conducted on the ImageNet-1K training set with an input resolution of 256×256. Training is performed on 8 NVIDIA H200 GPUs with a global batch size of 256. We use the AdamW optimizer with a constant learning rate of 1×10^{-4}. All models are trained using FP16 mixed precision with TF32 enabled.

We employ SiT-B, SiT-L, and SiT-XL as the generative models. For SD-VAE, which has a downsampling factor of 8 and 4 latent channels, the latent spatial resolution is 32×32. SiT uses a patch size of 2, resulting in 256 spatial tokens. For VA-VAE, which has a downsampling factor of 16 and 32 latent channels, the latent spatial resolution is 16×16. SiT uses a patch size of 1, likewise producing 256 spatial tokens. When the global token is enabled, an additional image-level semantic token is appended to the input sequence.

VAE training objective. The VAE training loss is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{LPIPS}} \\ + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} \\ + \lambda_{\mathrm{E}}\mathcal{L}_{\text{REPA-E}} + \lambda_{\mathrm{V}}\mathcal{L}_{\text{V-GTA}}.\end{aligned} \tag{13}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ denotes the L_1 reconstruction loss, $\mathcal{L}_{\mathrm{LPIPS}}$ is the perceptual reconstruction loss, and $\mathcal{L}_{\text{REPA-E}}$ provides end-to-end representation-alignment supervision through the intermediate SiT features. The proposed $\mathcal{L}_{\text{V-GTA}}$ aligns the latent-forming features before the VAE output projector with the clean global representation extracted by the frozen visual encoder. We set

$$\lambda_{\mathrm{KL}} = 10^{-6},\ \lambda_{\mathrm{adv}} = 0.1,\ \lambda_{\mathrm{E}} = 2.0,\ \lambda_{\mathrm{V}} = 0.5. \tag{14}$$

SiT training objective. The SiT training loss is given by

$$\mathcal{L}_{\mathrm{SiT}} = \mathcal{L}_{\mathrm{denoise}} + \lambda_{\mathrm{G}}\mathcal{L}_{\text{global-denoise}} + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{REPA}} + \lambda_{\mathrm{S}}\mathcal{L}_{\text{S-GTA}}. \tag{15}$$

Here, $\mathcal{L}_{\mathrm{denoise}}$ is the velocity-prediction objective for the spatial latent stream, while $\mathcal{L}_{\text{global-denoise}}$ provides denoising supervision for the global token treated as an additional generative variable. The REPA loss $\mathcal{L}_{\mathrm{REPA}}$ aligns the intermediate SiT representation with the pretrained visual representation. The proposed $\mathcal{L}_{\text{S-GTA}}$ aligns the global-token state with the global representation aggregated from the spatial patch-token stream at the same network depth and noise level. We set

$$\lambda_{\mathrm{G}} = 0.03,\ \lambda_{\mathrm{R}} = 0.5,\ \lambda_{\mathrm{S}} = 0.05. \tag{16}$$

Classifier-free guidance analysis. We further analyze the effect of classifier-free guidance (CFG) on the standalone latent generative model trained with the frozen GTA-VAVAE tokenizer and the REG objective for 4M iterations. As shown in Table 9, GTA-VAVAE exhibits stable performance over CFG scales ranging from 2.3 to 2.7 when the upper guidance boundary is fixed at $g_{\mathrm{high}} = 0.85$. Increasing the CFG scale from 2.3 to 2.7 slightly improves gFID from 1.15 to 1.14. Meanwhile, IS increases consistently from 295.04 to 309.04, indicating improved conditional fidelity without a noticeable loss of structural quality.

G More visualization results

We present additional qualitative visualization results of SiT-XL/1 + GTA with CFG in Figs. 6 - 15.

Model	VAE	Iter.	CFG scale	g_{high}	gFID↓	sFID↓	IS↑
SiT-XL	GTA-VAVAE	4M	1.0	1.00	1.47	4.38	230.74
SiT-XL	GTA-VAVAE	4M	2.3	0.75	1.23	4.37	260.81
SiT-XL	GTA-VAVAE	4M	2.3	0.80	1.18	4.35	273.09
SiT-XL	GTA-VAVAE	4M	2.3	0.85	1.15	4.33	295.04
SiT-XL	GTA-VAVAE	4M	2.5	0.85	1.14	4.33	302.40
SiT-XL	GTA-VAVAE	4M	2.7	0.85	1.14	4.33	309.04

Table 9: Classifier-free guidance analysis for standalone GTA-VAVAE generation. The latent generative model is trained with the REG objective for 4M iterations. Here, g_{high} denotes the upper boundary of the time interval over which guidance is applied.



Figure 6: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 104.



Figure 7: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 207.



Figure 8: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 232.



Figure 9: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 258.



Figure 10: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 358.



Figure 11: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 387.



Figure 12: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 388.



Figure 13: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 649.



Figure 14: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 970.



Figure 15: The visualization results of SiT-XL/1 + GTA use CFG with $w = 2.7$, and the class label is 980.