

Scaling Text-to-Image Diffusion Transformers with Representation Autoencoders

Shengbang Tong*, Boyang Zheng*, Ziteng Wang*, Bingda Tang, Nanye Ma,
 Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, Saining Xie
 New York University

Website

Code

Models

Data

Abstract

Representation Autoencoders (RAEs) have shown distinct advantages in diffusion modeling on ImageNet by training in high-dimensional semantic latent spaces. In this work, we investigate whether this framework can scale to large-scale, freeform text-to-image (T2I) generation. We first scale RAE decoders on the frozen representation encoder (SigLIP-2) beyond ImageNet by training on web, synthetic, and text-rendering data, finding that while scale improves general fidelity, targeted data composition is essential for specific domains like text. We then rigorously stress-test the RAE design choices originally proposed for ImageNet. Our analysis reveals that scaling simplifies the framework: while dimension-dependent noise scheduling remains critical, architectural complexities such as wide diffusion heads and noise-augmented decoding offer negligible benefits at scale. Building on this simplified framework, we conduct a controlled comparison of RAE against the state-of-the-art FLUX VAE across diffusion transformer scales from 0.5B to 9.8B parameters. RAEs consistently outperform VAEs during pretraining across all model scales. Further, during finetuning on high-quality datasets, VAE-based models catastrophically overfit after 64 epochs, while RAE models remain stable through 256 epochs and achieve consistently better performance. Across all experiments, RAE-based diffusion models demonstrate faster convergence and better generation quality, establishing RAEs as a simpler and stronger foundation than VAEs for large-scale T2I generation. Additionally, because both visual understanding and generation can operate in a shared representation space, the multimodal model can directly reason over generated latents, opening new possibilities for unified models.

1. Introduction

Diffusion-based generative modeling [19, 38, 47] has made rapid progress, giving rise to state-of-the-art systems across

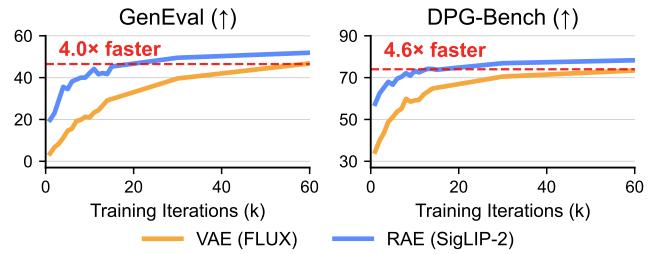


Figure 1. **RAE converges faster than VAE in text-to-image pretraining.** We train Qwen-2.5 1.5B + DiT 2.4B models from scratch on both RAE (SigLIP-2) and VAE (FLUX) latent spaces for up to 60k iterations. RAE converges significantly faster than VAE on both GenEval (4.0×) and DPG-Bench (4.6×).

visual generative domains such as text-to-image generation [1, 46, 88]. A key factor in this success is the adoption of *latent diffusion* [65], where generation occurs in a compact latent space encoded by a variational autoencoder (VAE) [43], rather than directly in pixel space.

In parallel with advances in generative modeling, visual representation learning has progressed through self-supervised learning (SSL) [6, 12, 34, 35], language supervision [62, 98], and their combinations [54, 83]. These models produce semantically structured, high-dimensional representations that generalize well across visual understanding tasks. Unlike VAE encoders, which compress images into *low-dimensional* latents, the representation encoders operate on *high-dimensional* latents that can capture much more semantically rich features.

Such high-dimensional latents were previously considered too “abstract” for effective generative modeling [72, 92], or outright intractable [13, 48]. However, a recent approach, Representation Autoencoder (RAE) [100], has paved a path forward by training decoders on frozen representation encoders. RAE pairs a powerful frozen representation encoder with a lightweight trained decoder to reconstruct pixels from high-dimensional embeddings, enabling diffusion directly in this semantic latent space. In the

*Core contributor.



Figure 2. RAE decoders trained on more data (web, synthetic & text) generalize across domains. Decoders trained only on ImageNet reconstruct natural images well but struggle with text-rendering scenes (see second column). Adding web and text data greatly improves text reconstruction while maintaining natural-image quality. We also observe that both the language-supervised model and the SSL model learn representations suitable for reconstructing diverse images, including natural languages. Compared to proprietary VAEs, our RAE models achieve competitive overall fidelity.

highly controlled class-conditional ImageNet [18] setting, RAE demonstrates that diffusion in such frozen representation spaces can achieve more efficient and effective training than conventional VAE-based diffusion.

However, ImageNet represents a best-case scenario: fixed resolution, curated content, and class-conditional generation. A critical question remains unanswered: *can RAE truly scale to the complexities of freeform text-to-image generation?* This setting involves broader visual diversity, open-ended compositions, and substantially larger models and compute—challenges for which high-dimensional latent diffusion remains unproven.

In this work, we investigate whether RAEs can succeed at scale by training diffusion models for large-scale text-to-image (T2I) generation. We adopt SigLIP-2 [84] as the frozen representation encoder and use the MetaQuery framework [56] to train a unified T2I model, leveraging a powerful pretrained large language model (LLM) [61].

As a first step, we study decoder training beyond ImageNet supervision (Sec. 2). Expanding from ImageNet to web-scale and synthetic aesthetic data yields only *small* gains on ImageNet itself, but provides *moderate* improvements on more diverse natural images such as YFCC [80], showing that broader distributions enhance generalization.

However, we find that text reconstruction requires targeted supervision: without text-specific data, the decoder fails to reproduce fine glyph details. Adding text-rendering data leads to substantial improvements, highlighting that data *composition*, not just scale, is crucial.

Next, we analyze design choices in the RAE framework [100] and evaluate their importance under large-scale T2I training (Sec. 3). We find that scale acts as a amplifier. Dimension-aware noise scheduling remains essential: removing the shift leads to substantially worse performance. The *wide DDT head* (DiT^{DH}) provides clear benefits for smaller backbones, but its advantage fades as Diffusion Transformers (DiT) scale to the billion-parameters. Finally, the effect of *noise-augmented decoding* is modest at T2I scale, with gains saturating quickly.

We then systematically compare RAEs with *SOTA* VAEs under matched training conditions (Sec. 4). We train DiTs from **scratch** following the conventional two-stage T2I setup [15, 60]: (i) large-scale pretraining with randomly initialized DiTs, and (ii) finetuning on smaller high-quality datasets. During pretraining, RAE-based models converge significantly faster and achieve higher performance on both GenEval and DPG-Bench. As shown in Fig. 1, training a 1.5B LLM + 2.4B DiT with RAE (SigLIP-2) achieves

Table 1. Data matters for RAE’s reconstruction fidelity. We train RAE (SigLIP-2) on different data sources. Compared with ImageNet-only training, using web-scale images consistently improves reconstruction quality across all domains.

Data Sources	#Data	ImageNet ↓	YFCC ↓	Text ↓
ImageNet	1.28M	0.462	0.970	2.640
Web	39.3M	0.529	0.629	2.325
Web + Synthetic	64.0M	0.437	0.683	2.406
Web + Synthetic + Text	73.0M	0.435	0.702	1.621

a $4.0\times$ speedup on GenEval and a $4.6\times$ speedup on DPG-Bench compared to its VAE counterpart. This advantage is consistent across both language backbones (Qwen-2.5 [2] 1.5B–7B) and diffusion scales (DiT 0.5B–9.8B). In finetuning, RAE models continue to outperform their VAE counterparts and are less prone to overfitting.

Finally, we examine unified models in which RAE enables understanding and generation to operate in the same high-dimensional semantic space (Sec. 5). We find that adding generative training does *not* degrade understanding performance, and the choice of RAE vs. VAE in the generative path has little effect because both rely on the same frozen understanding encoder. Moreover, the shared latent space allows the LLM to process generated latents directly, without decoding back to pixels. We take a first exploratory step toward leveraging this property through latent-space test-time scaling, which proves both feasible and effective.

Ultimately, we aim to convey one primary message: **Representation Autoencoders provide a simpler and stronger foundation than VAEs for training large-scale text-to-image diffusion models.** They offer a simple yet effective path to scaling generation within semantic representation spaces. We will release *all* code, data, and model checkpoints related to this work to foster open and reproducible research in multimodal generation.

2. Scaling Decoder Training Beyond ImageNet

To adapt the RAE framework for open-world T2I generation, we first train a RAE decoder on a larger and more diverse dataset than ImageNet [18]. Throughout this section, we choose SigLIP-2 So400M (patch size 14) [84] as the frozen encoder, and train a ViT-based [21] decoder to reconstruct images from these tokens at 224×224 resolution. We present the architectural details in Sec. A. Given an input image $x \in \mathbb{R}^{3 \times 224 \times 224}$, the encoder produces $N = 16 \times 16$ tokens with channel dimension $d = 1152$.

Training objective. Following RAE, we adopt ℓ_1 , LPIPS [99], and adversarial losses [33, 68]. Additionally, we integrate Gram Loss [29], which is found beneficial for reconstruction [52]. The training objective is set as

Table 2. Comparison of reconstruction performance. After expanding the training data, RAE outperforms SDXL-VAE across all three domains, though it still falls short of FLUX-VAE. Within RAE variants, WebSSL reconstructs better than SigLIP-2.

Family	Model	ImageNet ↓	YFCC ↓	Text ↓
VAE	SDXL	0.930	1.168	2.057
	FLUX	0.288	0.410	0.638
RAE	WebSSL ViT-L	0.388	0.558	1.372
	SigLIP-2 ViT-So	0.435	0.702	1.621

$L(x, \hat{x}) = \ell_1(x, \hat{x}) + \omega_L \text{LPIPS}(x, \hat{x}) + \omega_G \text{Gram}(x, \hat{x}) + \omega_A \text{Adv}(x, \hat{x}), \hat{x} = \text{RAE}(x)$. We include the weights and training details in Sec. A.

Training data. We use a dataset combining roughly 73M data from three data sources: web image sources from FuseDiT [77], synthetic images generated by FLUX.1-schnell [46], and RenderedText [87], which focuses on text-rendering scenes. Details are provided in Sec. 4.

Evaluation. We evaluate rFID-50k [36] of reconstructed images in three representative domains: (i) ImageNet-1k [67] for classic object-centric evaluation, (ii) YFCC [80] for diverse web-scale imagery, and (iii) RenderedText [87] held-out set for text-rendering and typography-specific evaluation. We evaluate rFID on 50k samples from each data source and present our results in Tabs. 1 and 2.

Web-scale training of RAE decoders. As shown in Tab. 1, expanding decoder training beyond ImageNet to include web-scale and synthetic data yields only marginal gains on ImageNet itself, but provides moderate improvements on more diverse images (YFCC). This indicates that exposure to a broader distribution enhances the decoder’s generalizability. However, generic web data is insufficient for text reconstruction. Training on Web + Synthetic data yields little improvement over ImageNet-only training. In contrast, performance improves substantially once text-specific data is included, highlighting that reconstruction quality is very sensitive to the composition of the training data. As shown in Fig. 2, training the RAE decoder with additional text data is essential for accurate text reconstruction. Overall, RAE reconstruction improves with scale, but the composition of data—not just its size—matters: each domain benefits most from domain-matched coverage.

Different encoders. We also experiment training RAE using different pretrained encoders. In particular, we replace SigLIP-2 with WebSSL-DINO [26], a large-scale self-supervised model. As shown in Tab. 2, WebSSL-DINO achieves stronger reconstruction performance than SigLIP-2 across all domains, including text reconstruction. Both SigLIP-2 and WebSSL-L consistently outperform SDXL VAE [60], though they still fall short of FLUX VAE [46].

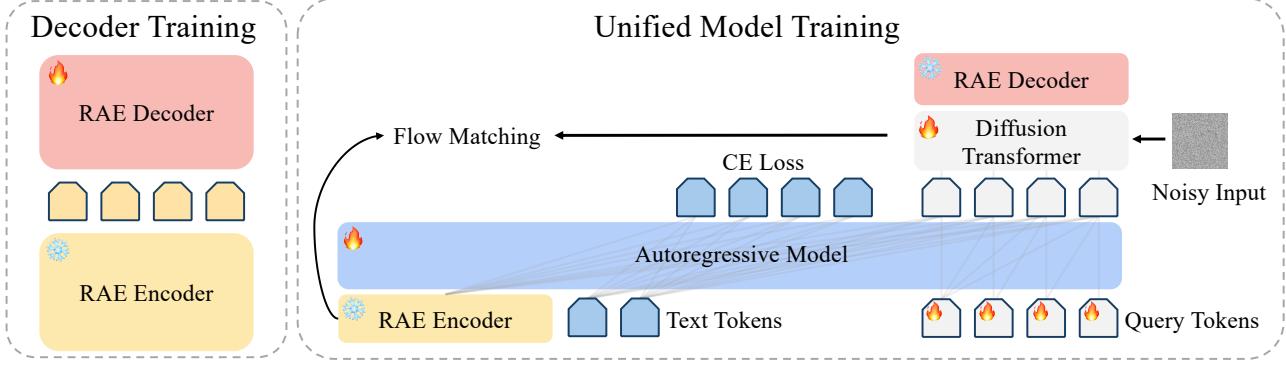


Figure 3. **Overview of training pipeline.** *Left:* RAE decoder training stage. We train a decoder on the representations (yellow tokens) produced by the frozen RAE encoder. *Right:* End-to-end unified training of the autoregressive model, diffusion transformer, and learnable query tokens (gray tokens) using cross-entropy (CE) loss for text prediction and a flow-matching objective for image prediction.

3. RAE is Simpler in T2I

The original RAE framework [100] introduced a suite of specialized design choices—including dimension-dependent noise scheduling, noise-augmented decoding, and a modified backbone (DiT^{DH})—to enable diffusion on high-dimensional latents. While these modifications proved effective for class-conditional ImageNet generation, it remains unclear which are fundamental requirements for high-dimensional diffusion and which are adaptations for lower-capacity regimes.

In this section, we systematically stress-test these components under large-scale T2I settings. We systematically evaluate these components to determine their necessity in large-scale T2I generation. Our analysis reveals that adapting the noise schedule to the latent dimension is critical for convergence, whereas the architectural modifications proposed in the original work—such as wide diffusion heads and noise augmentation—become redundant at scale.

3.1. Experiment Setup

Model architecture. We adopt the MetaQuery architecture [56] for text-to-image (T2I) generation and unified modeling. The model initializes with a pretrained language model (LLM) and prepends a sequence of learnable query tokens to the text prompt. The number of query tokens is set to 256, matching the number of visual tokens (16×16) produced by the representation encoder. The LLM jointly processes the text and queries, producing query-token representations that serve as the conditioning signal. A 2-layer MLP connector then projects these representations from the LLM’s hidden space into the DiT model [58].

For this DiT model, we adopt the design based on LightningDiT [92] and train it using the flow matching objective [47]. Critically, our model does not operate in a compressed VAE space. Instead, the DiT learns to model the distribution of high-dimensional, semantic representations

generated by the frozen representation encoder. During inference, the DiT generates a set of features conditioned on the query tokens, which are then passed to our trained RAE decoder for rendering into pixel space.

We also train visual instruction tuning [49, 50] for image understanding. For this, we use a separate 2-layer MLP projector that maps visual tokens into the LLM’s embedding space. Importantly, these visual tokens come from the **same** frozen representation encoder whose features the diffusion model is trained to generate.

Unless otherwise specified, we use SigLIP-2 So400M (patch size 14) [84] as our representation encoder and Qwen-2.5 1.5B [61] as the LLM in our experiments. We fix the number of visual tokens to 256, resulting in 224-resolution images for RAE and 256-resolution for VAE.

Flow matching. Following standard practice, we adopt the flow matching objective [47, 51] with linear interpolation $\mathbf{x}_t = (1 - t)\mathbf{x} + t\varepsilon$, where $\mathbf{x} \sim p(\mathbf{x})$ and $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$, and train the model to predict the velocity $v(\mathbf{x}_t, t)$. Unless otherwise noted, we employ a 50-step Euler sampler for generation, consistent with RAE [100].

Evaluation. We evaluate using two widely adopted metrics: the GenEval score [32] and the DPG-Bench score [39].

3.2. Noise scheduling remains crucial for T2I

The RAE work [100] argues that conventional noise schedules become suboptimal when applied to high-dimensional latent spaces. The paper proposes a *dimension-dependent noise schedule shift* [25] that rescales the diffusion timestep according to the effective data dimension $m = N \times d$ (number of tokens \times token dimension). Formally, given a base schedule $t_n \in [0, 1]$ defined for a reference dimension n , the shifted timestep is computed as

$$t_m = \frac{\alpha t_n}{1 + (\alpha - 1)t_n}, \quad \text{where } \alpha = \sqrt{\frac{m}{n}}.$$

We follow the RAE setting and use $n=4096$ as the base dimension for computing the scaling factor α . We experiment with and without applying the dimension-dependent shift when training text-to-image diffusion models on RAE latents, as shown below.

Setting	GenEval \uparrow	DPG-Bench \uparrow
w/o shift	23.6	54.8
w/ shift	49.6	76.8

Consistent with Zheng et al. [100], applying the noise shift dramatically improves both GenEval and DPG-Bench scores, demonstrating that adjusting the schedule to the effective latent dimension is critical for T2I.

3.3. Design Choices that Saturate at Scale

While dimension-aware noise scheduling proves essential, we find that other design choices in RAE, which was originally developed for smaller-scale ImageNet models, provide diminishing returns at T2I scale.

Noise-augmented decoding. RAE originally proposed training decoders on perturbed latents to bridge the gap between training and inference distributions. Formally, it trains the RAE decoder on smoothed inputs $z' = z + n$, where $n \sim \mathcal{N}(0, \sigma^2 I)$ and σ is sampled from $|\mathcal{N}(0, \tau^2)|$. We set $\tau = 0.2$ as we find a too high τ makes decoder training hard to converge.

We visualize the effect of noise-augmented decoding at different training stages in Fig. 4a. The gains are noticeable early in training (before ~ 15 k steps), when the model is still far from convergence, but become negligible at later stages. This suggests that noise-augmented decoding acts as a form of regularization that matters most when the model has not yet learned a robust latent manifold.

Wide DDT head. The DiT^{DH} architecture augments a standard DiT with a shallow but wide DDT head, increasing denoising width without widening the entire backbone. In standard ImageNet-scale DiTs, the backbone width ($d \approx 1024$) is often narrower than the high-dimensional RAE latent targets ($d = 1152$). DiT^{DH} circumvents this by appending a wide, shallow denoising head ($d = 2688$) without incurring the cost of widening the full backbone.

However, T2I setting operates in a different regime. Modern large-scale T2I DiTs [46, 88] (≥ 2 B parameters) possess hidden dimensions ($d \geq 2048$) that inherently exceed the latent dimension. We hypothesize that this natural width eliminates the bottleneck DiT^{DH} was designed to fix.

To verify this, we train DiT variants across three scales—0.5B, 2.4B, and 3.1B—comparing standard architectures against counterparts augmented with the +0.28B parameter DiT^{DH} head. As shown in Fig. 4b, the results confirm our hypothesis: at 0.5B, where the backbone is narrow, DiT^{DH}

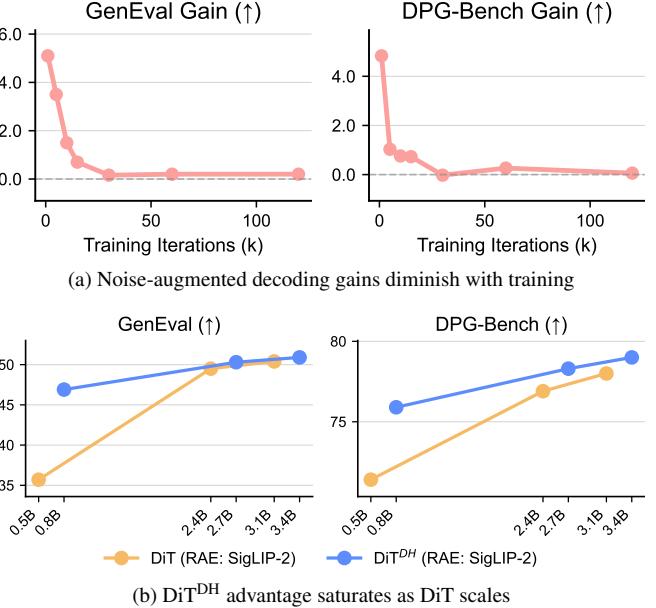


Figure 4. **Design choices that saturate at T2I scale.** *Left:* Noise-augmented decoding provides substantial gains early in training but becomes negligible by 120k steps. *Right:* DiT^{DH} yields large gains at 0.5B (+11.2 GenEval), but the advantage diminishes at >2.4B, where backbone capacity dominates.

provides a critical +11.2 GenEval boost. Yet as the model scales to 2.4B and beyond, this advantage saturates greatly.

This finding clarifies that DiT^{DH} is a patch for capacity-constrained models, not a fundamental requirement for RAE. For scalable T2I training, standard DiT architectures are already sufficient.

Summary. Our experiments reveal clear design principles for scaling RAE-based diffusion models: dimension-aware noise scheduling remains non-negotiable, as it directly addresses the mathematical properties of high-dimensional latent spaces. In contrast, architectural refinements (DiT^{DH}) and training augmentations (noise-augmented) that help at small scales provide diminishing returns as models grow—backbone capacity increasingly dominates performance. From here on, we adopt standard DiT architectures with proper noise scheduling and no noise-augmented decoding.

4. Training Diffusion Model with RAE vs. VAE

In this section, we compare text-to-image diffusion training using the RAE (SigLIP-2) encoder versus a standard VAE (FLUX-VAE). For the VAE baseline, we adopt the state-of-the-art model from FLUX [46]. All experiments follow the same setup described in Sec. 3.1, with identical training configurations; the only difference lies in whether diffusion is performed in the RAE or VAE latent space. We defer implementation details to Sec. A.

Table 3. Data composition matters more than scale. Synthetic data substantially outperforms web data, and their combination (49.5 GenEval) surpasses even doubled synthetic data (48.0), demonstrating synergistic benefits from complementary data sources rather than volume alone.

Training Data	GenEval \uparrow	DPG-Bench \uparrow
Synthetic	45.1	73.8
Synthetic $\times 2$	48.0	75.2
Web	25.9	69.5
Web $\times 2$	26.3	70.6
Synthetic + Web	49.5	76.9

Experimental Protocol. We organize our comparison into two stages: *pretraining* and *finetuning*. We train the Diffusion Transformer *from scratch* in both settings to ensure a fair comparison of convergence speed and data efficiency. We ensure apples-to-apples comparison. The *only* component that differs is the latent space and its decoder (SigLIP-2 RAE vs. FLUX VAE). For the VAE baseline, we employ FLUX VAE for generation while retaining the SigLIP encoder for understanding, as VAE latents are insufficient for perception [101]. This design choice effectively forms a two-tower architecture, mirroring the design of recent unified models like Bagel [17] and UniFluid [27].

Pretrain Data. We follow the data mixture developed in FuseDiT [77] and adopt the recaptioned texts and remixing ratios released by BLIP-3o [10]. The mixture combines mostly webdata like CC12M [7], SA-1B [44], and JourneyDB [73], totaling approximately 39.3M images. In addition, we use FLUX.1-schnell [46] to generate 24.7M synthetic images. We also train on Cambrian-7M [81] to develop the model’s visual understanding capabilities.

We experiment with a Qwen-2.5 1.5B LLM and a 2.4B DiT to study how different pretraining corpora influence text-to-image performance. We train three variants: (i) Web-39M + Cambrian-7M, (ii) FLUX-generated synthetic data + Cambrian-7M, and (iii) their union. As shown in Tab. 3, the mixed dataset yields the best performance.

To ensure the gains are not simply due to more data, we also double the size of each individual source (Web $\times 2$, Synthetic $\times 2$). These runs yield much smaller improvements, indicating that the benefits arise from the complementary nature of the two data types rather than data volume alone.

We also find that synthetic data results in lower training loss and faster convergence, suggesting that FLUX images provide more stylistically consistent signals. Web-scale data, by contrast, is harder to fit but provides more diverse signals. When combined, the model inherits visual style from synthetic data and rich semantics from web data, leading to clear and robust improvements in generation quality.

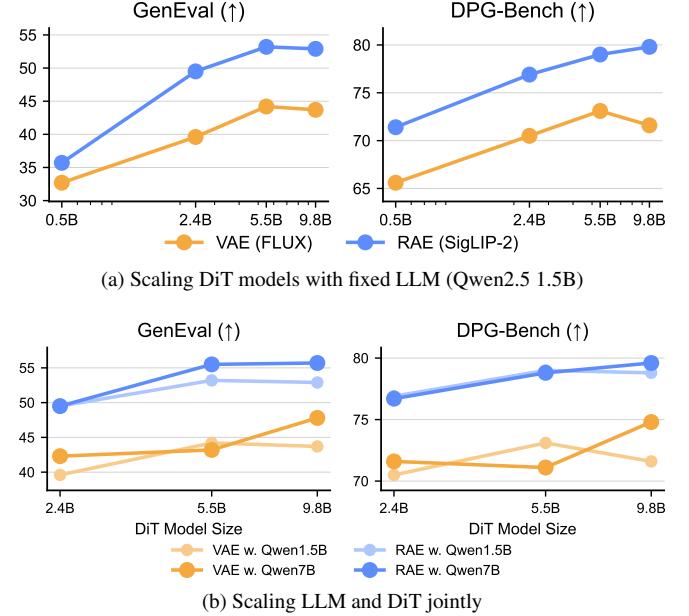


Figure 5. RAE outperforms VAE across LLM and DiT scales. *Top:* With a 1.5B LLM, RAE-based models outperform VAE-based ones at all DiT sizes (0.5B, 2.4B, 5.5B, 9.8B). *Bottom:* Using a larger 7B LLM, RAE continues to maintain its advantage.

4.1. Pretraining

Convergence. We first compare the convergence behavior. We train a Qwen2.5-1.5B LLM with a 2.4B DiT backbone. As shown in Fig. 1, the RAE-based model converges significantly faster than its VAE counterpart, achieving a 4.0 \times speedup on GenEval and a 4.6 \times speedup on DPG-Bench.

Scaling DiT models. We use Qwen-2.5 1.5B as the language backbone, and train DiT variants of 0.5B, 2.4B, 5.5B, and 9.8B parameters. The architectures of these DiT variants are designed following recent advances in large-scale vision models [24, 26, 85, 88], and detailed model specifications are provided in Sec. B. In this experiment, we train all the models for 30k iterations with a batch size of 2048.

In Fig. 5a, we find that RAE-based models consistently outperform their VAE counterparts at all scales. Even for the smallest 0.5B DiT, where the network width only slightly exceeds the RAE latent dimension, the RAE-based model still shows clear advantages over the VAE baseline.

We also observe diminishing returns when scaling DiT models beyond 6B parameters. The performance trend appears to plateau, suggesting that simply increasing model size without proportionally improving data quality and diversity may lead to underutilized capacity. This observation aligns with discussions in large-scale visual SSL literature [26], which highlight the need for high-quality data scaling to fully exploit model capacity.

Scaling LLM backbones. We study how scaling the LLM backbone influences text-to-image performance when

Table 4. **SSL encoders are effective RAE backbones for T2I.**
A WebSSL-based RAE performs slightly worse than SigLIP-2 but remains stronger than FLUX VAE in T2I.

Model Variant	GenEval \uparrow	DPG-Bench \uparrow
VAE-based models		
FLUX VAE	39.6	70.5
RAE-based models		
WebSSL ViT-L	46.0	72.8
SigLIP-2 ViT-So	49.5	76.9

paired with diffusion models of different sizes. To this end, we train both RAE- and VAE-based models using LLMs of {1.5B, 7B} parameters combined with DiTs of {2.4B, 5.5B, 9.8B}, and present the results in Fig. 5b.

We observe performance gains from scaling the LLM to 7B, particularly when paired with RAE. We note that prior studies, such as MetaQuery [56], reported limited benefits from LLM scaling. Our results diverge from this conclusion, likely due to two key factors: (1) we evaluate on significantly larger diffusion backbones (up to 9.8B) which can better exploit rich text representations, and (2) we finetune the LLM backbone, allowing it to adapt its latent space for generative tasks more effectively than frozen approaches.

Generalizing to other vision encoders. We also experiment with training RAE with WebSSL ViT-L [26]. Under the same 1.5B LLM and 2.4B DiT setup, the WebSSL RAE performs slightly below the SigLIP-2 version but still exceeds the FLUX VAE baseline (Tab. 4). This finding is notable because WebSSL is not explicitly aligned with text; it suggests that the RAE framework in T2I training is robust to the choice of encoder.

4.2. Finetuning

Following standard practice in T2I training [8, 15, 60], models are finetuned on a smaller high-quality dataset after large-scale pretraining. We evaluate this finetuning stage for both RAE- and VAE-based models under identical settings. Unless otherwise noted, we use the BLIP-3o 60k dataset [10] and start from the 1.5B LLM + 2.4B DiT checkpoint trained for 30k steps in Sec. 4.1. We update both the LLM and the DiT; additional details are provided in Sec. A.

RAE-based models consistently outperform VAE-based models. We finetune both family of models for {4, 16, 64, 128, 256} epochs and compare the performance on GenEval and DPG-Bench in Fig. 6. We observe that across all iterations, the RAE-based model shows an advantage on both GenEval and DPG-Bench across all settings.

RAE-based models are less prone to overfitting. As shown in Fig. 6, VAE-based models degrade significantly after 64 epochs. Training loss analysis (Appendix Fig. 9) reveals that the VAE loss collapses rapidly to near-zero, suggesting the model is memorizing individual training sam-

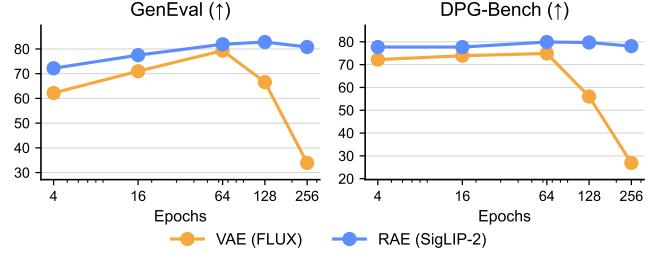


Figure 6. **RAE-based models outperform VAE-based models and are less prone to overfitting.** We train both models for 256 epochs and observe that (1) RAE-based models consistently achieve higher performance, and (2) VAE-based models begin to overfit rapidly after 64 epochs.

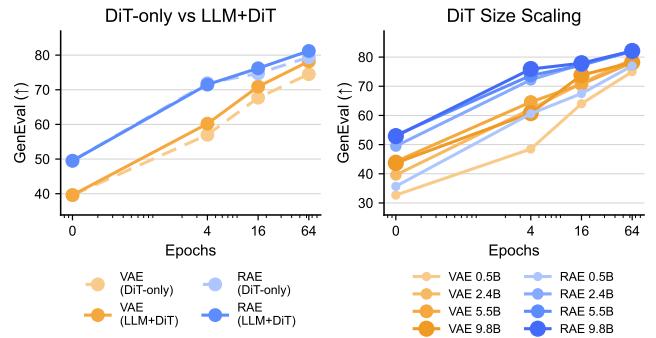


Figure 7. **RAE-based models outperform VAEs across different settings.** Left: When fine-tuning only the DiT versus the full LLM+DiT system, RAE models consistently achieve higher GenEval scores. Right: RAE models maintain their advantage over VAE across all DiT model scales (0.5B–9.8B parameters), with the performance gap widening as model size increases.

ples rather than learning the underlying distribution. In contrast, RAE-based models remain stable and show only a mild decline. We hypothesize that the higher-dimensional and semantically structured latent space of the RAE¹ may provide an implicit regularization effect, helping mitigate overfitting during finetuning.

RAE’s advantage generalizes across settings. To verify whether RAE’s advantage over VAE extends beyond our main setup, we conduct two additional experiments: 1) finetuning only the DiT while freezing the LLM (following recent works [10, 56]), and 2) scaling to different sizes DiT models (0.5B–9.8B parameters). Figure 7 shows that RAE consistently outperforms VAE in both settings. The left panel shows that both selective fine-tuning (DiT-only) and joint fine-tuning (LLM+DiT) favor RAE over VAE; notably, the top-performing VAE configuration reaches 78.2, while the weakest RAE approach achieves 79.4. The right panel shows continued RAE gains across the scaling range, with larger models exhibiting greater improvements.

¹SigLIP-2 produces 1152-dim. tokens vs. <100 in typical VAEs

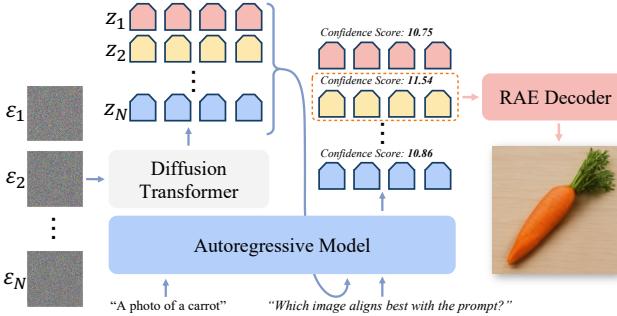


Figure 8. **Test-time scaling in latent space.** Our framework allows the LLM to directly evaluate and select generation results within the latent space, bypassing the decode-re-encode process.

5. Implications for Unified Models

A key advantage of the RAE framework is that it unifies visual understanding and generation within a *single, shared, high-dimensional latent space*. This contrasts with the conventional two-tower design (used in our Section 4 baseline). In two-tower models, the generation head operates in a latent space alien to the LLM’s understanding encoder. This effectively makes the unified model ‘blind’ to its own output distribution without a VAE decoder. In contrast, RAE forces generation to occur in the *same representation space* of the visual encoder. This means the model generates the exact same high-dimensional features it uses to see.

Test-time scaling in latent space. A direct benefit of this shared representation is that the LLM can interpret the latents produced by the diffusion model without needing to decode them into pixels and re-encode them, leaving the representation and pixel spaces fully decoupled. We leverage this property to implement **Latent Test-Time Scaling (TTS)**, where the LLM acts as a verifier for its own generations directly and only in the feature space (Fig. 8).

We explore two verifier metrics that leverage the LLM’s understanding capabilities to score generated latents: (1) **Prompt Confidence**: We re-inject the generated latents and the original text prompt back into the LLM and measure the aggregate token-level confidence of the prompt, following Kang et al. [42]. (2) **Answer Logits**: We explicitly query the LLM with: “*Does this generated image ⟨image⟩ align with the ⟨prompt⟩?*” and use the logit probability of the “Yes” token as the score.

With the verifier defined, we adopt the standard test-time scaling protocol [53, 90] using a best-of- N selection strategy. As shown in Tab. 5, both verification metrics yield consistent improvements on GenEval, demonstrating that latent-space TTS is not only feasible but also an effective way to enhance generation quality. Crucially, this improvement is achieved entirely within the semantic latent space, demonstrating that the model can verify the quality of its own generations without ever needing to render pixels.

Table 5. **TTS results across LLM-DiT configurations.** Substantial performance improvements observed with both verifier metrics on GenEval. 4/8 refers to “selecting best 4 out of 8”.

Best-of- N	Prompt Confidence	Answer Logits
1.5B LLM + 5.5B DiT	(GenEval = 53.2)	
4/8	56.7	59.6
4/16	57.5	62.5
4/32	60.0	64.3
7.0B LLM + 5.5B DiT	(GenEval = 55.5)	
4/8	58.3	62.5
4/16	59.6	65.8
4/32	60.1	67.8

Table 6. **Generative training leaves understanding intact; RAE and VAE perform similarly.** Across VL benchmarks, both latent choices produce comparable understanding performance.

Model	MME _P	TVQA	AI2D	Seed	MMMU	MMMU _P
Und.-only	1374.8	44.7	63.9	67.1	40.2	20.5
RAE-based	1468.7	39.6	66.7	69.8	41.1	19.8
VAE-based	1481.7	39.3	66.7	69.7	37.2	18.7

Visual understanding. Finally, we conduct a comparative study to study how the choice of visual generation backbone—VAE versus RAE—affects multimodal understanding performance. We evaluate the trained models on standard benchmarks: MME [28], TextVQA [71], AI2D [37], SeedBench [30], MMMU [95], and MMMU-Pro [96]. We emphasize that the goal of this work is not to build a SOTA VQA model; achieving that would require additional components such as any-resolution inputs, multimodal continual pretraining, and very high-quality data.

Similar to prior findings [17, 27, 82], we observe in Tab. 6 that adding generative modeling does not degrade visual understanding performance. The choice of RAE vs. VAE in the generative path has little impact, likely because both variants share the same frozen understanding encoder.

6. Related Work

VAE, representation and representation autoencoder. A common line of work uses VAEs [43] as autoencoders to compress images into low-dimensional latent spaces, typically with channel dimensions below 64 [25, 46]. Many studies [9, 93] have pursued aggressive compression, while others [55, 63] reduce dimensionality further by quantizing continuous latents into discrete codes. However, both directions unavoidably result in information loss.

Representation Autoencoders (RAE) [100] take a different route: use a frozen, pretrained representation encoder and train only the decoder to reconstruct images from high-dimensional semantic features. In ImageNet experiments, training diffusion transformers [58] in this latent

space yields faster convergence and better performance than VAEs. In this work, we extend RAE to text-to-image generation and show that its reconstruction and generative advantages transfer naturally to the multimodal setting.

Recently, several works have explored leveraging representation encoders for reconstruction. SVG [69, 70] employs a residual encoder to refine visual details during reconstruction, while VTP [91] incorporates a reconstruction loss into the pretraining of representation encoders. VQRAE [23] further applies quantization on top of representation encoders to construct discrete representations for generation. In a related direction, another line of work [45, 57, 59] integrates representation encoders with VAEs to improve generation fidelity.

VAE in Text-to-image models. VAE has also been widely used in text-to-image models. Stable Diffusion [64] uses an off-the-shelf VAE and a text-conditioned U-Net [66] for T2I training. Subsequent work [25, 46, 60] improves VAE through higher-quality and larger-scale training data.

Recently, Stable Diffusion 3 [25] shows that widening the VAE channels boosts reconstruction fidelity and enhances the scalability of the downstream diffusion model, while Hunyuan-Image-3 [5] further incorporates representation alignment [94] into VAE training.

This work takes the representation route a step further: instead of modifying VAEs, we train T2I models directly on high-dimensional representation spaces with RAE. This approach yields clear advantages over VAE in both convergence speed and final generation quality.

Unified Multimodal Models. Recently, many works focus on unifying multimodal understanding and generation into one modeling paradigm. One stream of work discretizes visual input and trains next token prediction modeling [14, 41, 79, 86, 89]. Another stream of research incorporates diffusion model into LLMs [10, 15, 17, 20, 31, 56, 75, 82, 101]. However, it has been viewed that understanding and generation require *different* visual representations—high-dimensional CLIP features for understanding and low-dimensional VAE latents for generation—leading most unified models to adopt a two-tower design.

An emerging direction in unified multimodal modeling is to unify *understanding* and *generation* into a shared latent space. To work around this mismatch, recent approaches [13, 40, 52, 78, 97] adopt continuous representation spaces but introduce substantial downsampling for generation. For example, Chen et al. [13] uses high-dimensional, uncompressed features for understanding but falls back to compressed, lower-dimensional latents for generation. Jiao et al. [41] and Yue et al. [97] employ compressed embeddings for both understanding and generation, limiting the model’s perception ability. BLIP-3o [10] experiments with using a Qwen2.5-VL encoder [3] for understanding and EVA-CLIP [74, 76] for generation; However,

because the model does not apply noise-schedule shifting and its DiT width is smaller than the EVA-CLIP embedding dimension, it relies on a strong diffusion decoder [60] to map these features back to pixels.

Our work takes a step forward by using a *single high-dimensional encoder* for both understanding and generation. Leveraging RAE designs, the model enjoys a simpler architecture that understands and generates directly in this semantic space, surpassing VAE-based designs in T2I.

7. Conclusion

In this work, we investigate scaling Representation Autoencoders (RAEs) to text-to-image generation. Our study begins by scaling the decoder, where we find that while larger data scales improve general fidelity, specific domains such as text require targeted data composition. We then examine the RAE framework itself, revealing that scaling simplifies the design: dimension-dependent noise scheduling remains essential, but architectural modifications like DiT^{DH} yield diminishing returns as model capacity increases. Building on this streamlined recipe, we show that RAE-based diffusion models consistently outperform state-of-the-art VAE baselines in convergence speed and generation quality, while being less prone to overfitting during finetuning. Collectively, these results establish RAE as a simple and effective foundation for large-scale generation. Moreover, by enabling understanding and generation to operate in a shared representation space, RAEs open new possibilities for unified models, such as the latent-space test-time scaling demonstrated in this work. We believe RAE serve as a strong foundation for future research in both scalable generation and unified multimodal modeling.

8. Acknowledgements

The authors would like to thank Xichen Pan, Shusheng Yang, David Fan, John Nguyen for insightful discussions and feedback on the manuscript. This work was mainly supported by the Google TPU Research Cloud (TRC) program and the Open Path AI Foundation. ST is supported by Meta AI Mentorship Program. SX also acknowledges support from the MSIT IITP grant (RS-2024-00457882) and the NSF award IIS-2443404.

References

- [1] Stability AI. Stable diffusion 3.5, 2024. 1
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingen Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3
- [3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 9
- [4] Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Daniel Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, et al. Pathways: Asynchronous distributed dataflow for ml. *Proceedings of Machine Learning and Systems*, 4:430–449, 2022. 14
- [5] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuaniimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 9
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1
- [7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricu. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 6
- [8] 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 *ICLR*, 2024. 7
- [9] 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. In *ICLR*, 2025. 8
- [10] Juhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025. 6, 7, 9
- [11] Juhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025. 14
- [12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 1
- [13] Xiangyi Chen, Théophane Vallaeyns, Maha Elbayad, John Nguyen, and Jakob Verbeek. Vugen: Visual understanding priors for generation, 2025. 1, 9
- [14] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025. 9
- [15] Xueyan Dai and et al. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023. 2, 7, 9
- [16] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 14
- [17] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 6, 8, 9
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 3
- [19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021. 1
- [20] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamilm: Synergistic multimodal comprehension and creation. In *ICLR*, 2024. 9
- [21] 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. In *ICLR*, 2021. 3
- [22] 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. In *ICLR*, 2021. 14
- [23] Sinan Du, Jiahao Guo, Bo Li, Shuhao Cui, Zhengzhuo Xu, Yifu Luo, Yongxian Wei, Kun Gai, Xinggang Wang, Kai Wu, and Chun Yuan. Vqrae: Representation quantization autoencoders for multimodal understanding, generation and reconstruction, 2025. 9
- [24] 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 *ICML*, 2024. 6
- [25] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yanik Marek, and Robin Rombach. Scaling rectified flow

- transformers for high-resolution image synthesis. In *ICML*, 2024. 4, 8, 9
- [26] David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, et al. Scaling language-free visual representation learning. In *ICCV*, 2025. 3, 6, 7, 14
- [27] Lijie Fan, Luming Tang, Siyang Qin, Tianhong Li, Xuan Yang, Siyuan Qiao, Andreas Steiner, Chen Sun, Yuanzhen Li, Tao Zhu, et al. Unified autoregressive visual generation and understanding with continuous tokens. *arXiv preprint arXiv:2503.13436*, 2025. 6, 8
- [28] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. *corr abs/2306.13394* (2023), 2023. 8
- [29] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 3
- [30] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv preprint arXiv:2307.08041*, 2023. 8
- [31] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 9
- [32] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 4
- [33] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 3
- [34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arxiv e-prints, art.* *arXiv preprint arXiv:1911.05722*, 2, 2019. 1
- [35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2021. 1
- [36] 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. In *NeurIPS*, 2017. 3
- [37] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kallikoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55:661–688, 2021. 8
- [38] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 1
- [39] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 4
- [40] Ziyuan Huang, DanDan Zheng, Cheng Zou, Rui Liu, Xiaolong Wang, Kaixiang Ji, Weilong Chai, Jianxin Sun, Libin Wang, Yongjie Lv, et al. Ming-univision: Joint image understanding and generation with a unified continuous tokenizer. *arXiv preprint arXiv:2510.06590*, 2025. 9
- [41] Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *CVPR*, 2025. 9
- [42] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025. 8
- [43] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1, 8
- [44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 6
- [45] Theodoros Kouzelis, Efstrathios Karaypidis, Ioannis Kakogiorgiou, Spyros Gidaris, and Nikos Komodakis. Boosting generative image modeling via joint image-feature synthesis. In *NeurIPS*, 2025. 9
- [46] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 1, 3, 5, 6, 8, 9, 14
- [47] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 1, 4
- [48] Bingchen Liu, Ehsan Akhgari, Alexander Vishератин, Aleks Kamko, Linmiao Xu, Shivam Shrira, Chase Lambert, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models, 2024. 1
- [49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4, 14
- [50] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 4
- [51] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 4
- [52] Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision, 2025. 3, 9
- [53] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 8
- [54] Norman Mu, Alexander Kirillov, David A Wagner, and Saining Xie. Slip: self-supervision meets language-image pre-training. *corr abs/2112.12750* (2021). *arXiv preprint arXiv:2112.12750*, 3, 2021. 1
- [55] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 8

- [56] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jihui Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries, 2025. 2, 4, 7, 9
- [57] Yueming Pan, Ruoyu Feng, Qi Dai, Yuqi Wang, Wenfeng Lin, Mingyu Guo, Chong Luo, and Nanning Zheng. Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion, 2025. 9
- [58] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 4, 8
- [59] Giorgos Petsangourakis, Christos Sgouropoulos, Bill Psomas, Theodoros Giannakopoulos, Giorgos Sfikas, and Ioannis Kakogeorgiou. Reglue your latents with global and local semantics for entangled diffusion, 2025. 9
- [60] David Podell and et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 3, 7, 9
- [61] A Yang Qwen, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen-2.5 technical report. *arXiv preprint*, 2024. 2, 4, 14
- [62] 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 *ICML*, 2021. 1
- [63] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *NeurIPS*, 2019. 8
- [64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 9
- [65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1
- [66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 9
- [67] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3
- [68] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *SIGGRAPH*, 2022. 3
- [69] Minglei Shi, Haolin Wang, Borui Zhang, Wenzhao Zheng, Bohan Zeng, Ziyang Yuan, Xiaoshi Wu, Yuanxing Zhang, Huan Yang, Xintao Wang, Pengfei Wan, Kun Gai, Jie Zhou, and Jiwen Lu. Svg-t2i: Scaling up text-to-image latent diffusion model without variational autoencoder, 2025. 9
- [70] Minglei Shi, Haolin Wang, Wenzhao Zheng, Ziyang Yuan, Xiaoshi Wu, Xintao Wang, Pengfei Wan, Jie Zhou, and Jiwen Lu. Latent diffusion model without variational autoencoder, 2025. 9
- [71] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 8
- [72] Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *ICML*, 2025. 1
- [73] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding, 2023. 6
- [74] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 9
- [75] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 9
- [76] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 9
- [77] Bingda Tang, Boyang Zheng, Sayak Paul, and Saining Xie. Exploring the deep fusion of large language models and diffusion transformers for text-to-image synthesis. In *CVPR*, 2025. 3, 6
- [78] Hao Tang, Chenwei Xie, Xiaoyi Bao, Tingyu Weng, Pandeng Li, Yun Zheng, and Liwei Wang. Unilip: Adapting clip for unified multimodal understanding, generation and editing, 2025. 9
- [79] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 9
- [80] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2, 3
- [81] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024. 6, 14
- [82] Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *ICCV*, 2025. 8, 9
- [83] Michael Tschanne, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1

- [84] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 3, 4
- [85] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6, 14
- [86] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiyi Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 9
- [87] C. Wendler. Renderedtext. <https://huggingface.co/datasets/wendlerc/RenderedText>, 2024. 3
- [88] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1, 5, 6
- [89] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 9
- [90] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Chengyue Wu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, et al. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*, 2025. 8
- [91] Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation, 2025. 9
- [92] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, 2025. 1, 4, 14
- [93] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. In *NeurIPS*, 2024. 8
- [94] 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 *ICLR*, 2025. 9
- [95] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 8
- [96] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. 8
- [97] Zhengrong Yue, Haiyu Zhang, Xiangyu Zeng, Boyu Chen, Chenting Wang, Shaobin Zhuang, Lu Dong, KunPeng Du, Yi Wang, Limin Wang, and Yali Wang. Uniflow: A unified pixel flow tokenizer for visual understanding and generation, 2025. 9
- [98] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1
- [99] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3
- [100] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 1, 2, 4, 5, 8, 14
- [101] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR*, 2025. 6, 9

A. Implementation

Our experiments are conducted on TPU v4, v5p, and v6e with TorchXLA.

Decoder training. We largely follow RAE for decoder architecture and adopt ViT-XL [22] as the default decoder. The decoder contains 28 blocks with a hidden size of 1152 and 16 heads. Decoder training details are included in Table 7. We find the GAN training recipe provided in [100] is not stable on web-scale images. To tackle the issue, we tune the recipe to as Tab. 7. On web-scale images, we find using DINO-S/16 already suffices as a strong discriminator, and using DINO-S/8 as in [100] makes it hard to converge. Therefore, we use DINO-S/16 as the default discriminator. All input is interpolated to 224×224 resolution before feeding into the discriminator. We use an epoch-based training scheme and set the sample of each virtual epoch to be the same as ImageNet (1.28M). For loss coefficients, we set $\omega_G = 100.0$, $\omega_L = 1.0$, $\omega_A = 10.0$.

T2I & unified model pretraining. For pretraining experiments in Sec. 4.1, we primarily train on TPU-v5p-128 and TPU-v6e-64. Detailed training configurations are provided in Tab. 8. We find that finetuning a pretrained LLM while training the DiT from scratch benefits from using separate optimizers, and properly decoupling their optimizer settings substantially improves training stability. We use SPMD sharding [4] together with TorchXLA to train the LLM, adapters, and DiT models.

T2I & unified model finetuning. We finetune pre-trained models on the BLIP3o-60k dataset [11] using TPU-v4-128 and TPU-v5p-64. To ensure a fair comparison, we apply identical training configurations to both RAE and VAE models across 4, 16, 64, and 256 epochs. We utilize the same codebase and training infrastructure as the pretraining stage. We use a global batch size of 1024 and the optimization settings detailed in Tab. 9.

Synthetic data generation. For synthetic image generation, we compile prompts from publicly available prompt datasets². Using these prompts, we generate 24.7M synthetic images with FLUX.1-schnell [46], which form part of our decoder training and T2I training corpus. We perform large-scale generation using TPU-v6e and will open-source the inference pipeline to facilitate future research.

²<https://huggingface.co/datasets/Geonmo/midjourney-prompts-only>, <https://huggingface.co/datasets/FredZhang7/stable-diffusion-prompts-2.47M>, <https://huggingface.co/datasets/neuralworm/stable-diffusion-discord-prompts>, <https://huggingface.co/datasets/isidentical/random-stable-diffusion-prompts>, <https://huggingface.co/datasets/CaptionEmporium/coyo-hd-11m-llavanext>

B. Models

LLM model and unified model configs. We use pre-trained Qwen2.5 [61] language models at the 1.5B and 7B scales in our experiments. Following prior work [49, 81], we use a 2-layer MLP to project visual features from the representation encoder into the LLM embedding space, and a separate linear layer to map the LLM’s query-token outputs into the input space of the diffusion model.

DiT Model configs. We design our diffusion architecture following LightningDiT [92]. Motivated by recent findings in scaling vision backbones [16, 26, 85], we prioritize increasing model *width* rather than depth when scaling DiT models. Consistent with insights from the RAE paper [100], we also ensure that the DiT hidden dimension remains strictly larger than the target latent dimension (e.g. 1152 for the SigLIP2 ViT-So model), including at small scales such as DiT-0.5B. The detailed model specifications are provided in Tab. 10.

C. Additional Results

Training losses. To complement the results in Sec. 4.2, we additionally compare the training loss curves of RAE and VAE models during finetuning in Fig. 9. We observe that the VAE model’s loss decreases rapidly to a very low value, which correlates with the performance degradation observed in Fig. 6, a clear sign of overfitting. In contrast, the RAE model’s loss decreases more gradually and stabilizes at a higher value, maintaining robust generation performance throughout the training process. This suggests that the high-dimensional semantic space of RAE provides a form of implicit regularization that prevents the model from memorizing the small finetuning dataset.

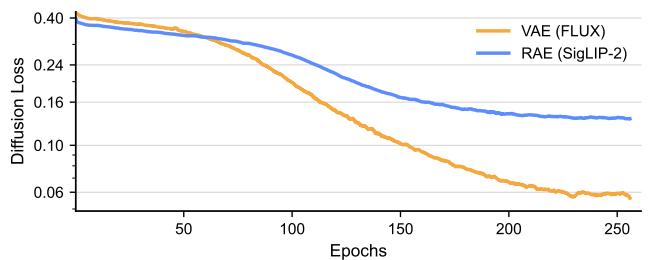


Figure 9. **Diffusion loss during finetuning (256 epochs).** RAE overfits less and later than VAE: the VAE loss plunges early to very low values, while the RAE loss decreases more gradually and plateaus at higher values, indicating reduced overfitting.

Extending finetuning to 512 epochs. We extend the finetuning experiment from Sec. 4.2 to 512 epochs. As shown in Fig. 6, VAE-based models already suffer substantial performance drops by 256 epochs, so we do not continue training them further. In contrast, the RAE-based model remains stable: even after 512 epochs (Fig. 10), it shows only a small

Component	Decoder	Discriminator	Component	Decoder	Discriminator
optimizer	AdamW	AdamW	optimizer	AdamW	AdamW
max learning rate	4×10^{-4}	5×10^{-5}	max learning rate	2×10^{-4}	2×10^{-5}
min learning rate	4×10^{-5}	5×10^{-6}	min learning rate	2×10^{-5}	2×10^{-6}
learning rate schedule	cosine decay	cosine decay	learning rate schedule	cosine decay	cosine decay
optimizer betas	(0.9, 0.95)	(0.9, 0.95)	optimizer betas	(0.9, 0.95)	(0.9, 0.95)
weight decay	0.0	0.0	weight decay	0.0	0.0
batch size	512	512	batch size	512	512
warmup	2 epoch	1 epoch	warmup	2 epoch	1 epoch
loss	$\ell_1 + \text{LPIPS} + \text{GAN} + \text{Gram}$	adv.	loss	$\ell_1 + \text{LPIPS} + \text{GAN} + \text{Gram}$	adv.
Model	ViT-XL	Dino-S/16 (frozen)	Model	ViT-XL	Dino-S/16 (frozen)
LPIPS start epoch	0	–	LPIPS start epoch	0	–
disc. start epoch	–	7	disc. start epoch	–	10
adv. loss start epoch	8	–	adv. loss start epoch	11	–
Training epochs	80	73	Training epochs	80	70

Table 7. **Training configuration for decoder and discriminator.** *Left:* Configuration used for SigLIP2-So. *Right:* Configuration used for WebSSL ViT-L. Different encoders require slightly different training recipes for achieving strong decoder performance.

Component	LLM	DiT
optimizer	AdamW	
learning rate schedule	cosine w/ warmup ratio 0.0134	
global batch size	2048	
max learning rate	5×10^{-5}	5×10^{-4}
optimizer betas	(0.9, 0.999)	(0.9, 0.95)
loss	autoregressive LM	diffusion loss
model	Qwen2.5 (1.5B / 7B)	DiT (0.5B–9.8B)

Table 8. **Optimization hyperparameters for the LLM backbone and the DiT diffusion head in the unified T2I model.**

Model	Hidden size	Heads	Depth
DiT-0.5B	1280	32	16
DiT-2.4B	2048	32	32
DiT-3.3B	2304	32	32
DiT-5.5B	3072	32	32
DiT-9.8B	4096	32	32

Table 10. **Architectural specifications of DiT variants.**

Component	LLM	DiT
optimizer	AdamW	
learning rate schedule	cosine w/ warmup ratio 0.03	
global batch size	1024	
training epochs	4, 16, 64, 256	
max learning rate	5.66×10^{-5}	5.66×10^{-4}
optimizer betas	(0.9, 0.999)	(0.9, 0.95)
loss	autoregressive LM	diffusion loss
model	Qwen2.5 (1.5B / 7B)	DiT (0.5B–9.8B)

Table 9. **Finetuning hyperparameters.**

decline in performance. This further supports the robustness of RAE-based methods under long-horizon finetuning.

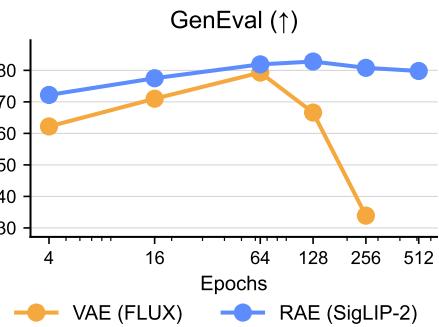


Figure 10. **Extended finetuning to 512 epochs.** RAE maintains robust performance even with 512 epochs of training, while VAE suffers catastrophic overfitting after 64 epochs.