TAMING OUTLIER TOKENS IN DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study outlier tokens in Diffusion Transformers (DiTs) for image generation. Prior work has shown that Vision Transformers (ViTs) can produce a small number of high-norm tokens that attract disproportionate attention while carrying limited local information, but their role in generative models remains underexplored. We show that this phenomenon appears in both the encoder and denoiser of modern Representation Autoencoder (RAE)-DiT pipelines: pretrained ViT encoders can produce outlier representations, and DiTs themselves can develop internal outlier tokens, especially in intermediate layers. Moreover, simply masking high-norm tokens does not improve performance, indicating that the problem is not only caused by a few extreme values, but is more closely related to corrupted local patch semantics. To address this issue, we introduce *Dual-Stage Registers (DSR)*, a register-based intervention for both components: trained registers when available, recursive test-time registers otherwise, and diffusion registers for the denoiser. Across ImageNet and large-scale text-to-image generation, these interventions consistently reduce outlier artifacts and improve generation quality. Our results highlight outlier-token control as an important ingredient in building stronger DiTs.

1 INTRODUCTION

Vision Transformers (ViTs) Dosovitskiy et al. (2020) have emerged as a foundational architecture in computer vision. Initially, they demonstrated strong performance on visual understanding tasks, including supervised classification (e.g., ViT Dosovitskiy et al. (2020), DeiT Touvron et al. (2021)), self-supervised representation learning (e.g., DINO models Caron et al. (2021); Oquab et al. (2023); Siméoni et al. (2025)), and contrastive language-image pretraining (e.g., CLIP Radford et al. (2021), SigLIP Zhai et al. (2023); Tschannen et al. (2025)). These successes gradually shifted the default backbone for visual recognition models from convolutional networks such as ResNets He et al. (2016) to Vision Transformers Chen et al. (2022); Raghu et al. (2021).

More recently, ViTs have shown remarkable effectiveness in generative modeling, particularly for diffusion models Sohl-Dickstein et al. (2015); Ho et al. (2020); Song & Ermon (2019). Starting from DiT Peebles & Xie (2023), Transformer-based generators have increasingly replaced UNet architectures Ho et al. (2020); Song et al. (2020); Rombach et al. (2022) in both latent-space Ma et al. (2024); Esser et al. (2024) and pixel-space diffusion models Li & He (2025). Beyond the generator itself, recent work such as RAE Zheng et al. (2025) further proposes using pretrained ViT encoders to project images into latents, replacing the UNet-based VAE Rombach et al. (2022), the last remaining convolutional component in state-of-the-art diffusion pipelines.

A commonly cited advantage of Transformers is their ability to model global interactions and their favorable optimization properties, albeit at the cost of weaker inductive biases. Recent studies, however, show that this flexibility can introduce nontrivial artifacts in the feature maps. In particular, Vision Transformers Need Registers Darcet et al. (2023) identifies the ubiquitous presence of outlier tokens in supervised and self-supervised ViTs (e.g., DeiT-III Touvron et al. (2022), OpenCLIP Ilharco et al. (2021), DINOv2 Oquab et al. (2023)). These tokens exhibit abnormally large norms, absorb a disproportionate amount of attention, and yet carry little semantic information, resulting in artifacts in feature maps that hinder downstream utilization, especially for dense visual tasks. This behavior is closely related to the attention sink hypothesis in the context of large language models Xiao et al.

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.

(2024), and can be effectively mitigated by introducing register tokens, i.e., dedicated non-patch tokens that explicitly serve as attention sinks during training.

However, investigations of outlier tokens have thus far focused almost exclusively on recognition models, leaving their role in generative models largely unexplored. This gap is particularly striking given the growing prevalence of Transformers in generative modeling and the heightened importance of locality and spatial fidelity in generation compared to recognition. In this work, we show that outlier token effects are pervasive in both ViT-based autoencoder and diffusion Transformers.

For ViT-based autoencoders, we focus on RAE Zheng et al. (2025), which has pretrained representation encoders. We empirically confirm the existence of outlier tokens in the encoders and demonstrate their negative impact on generation quality. Using the official DINOv2 implementation Oquab et al. (2023); Darcet et al. (2023), we compare DINOv2 encoders with and without register tokens and observe consistent improvements when registers are enabled. For SigLIP 2 Tschannen et al. (2025), which does not include registers by default, we introduce recursive test-time registers Jiang et al. (2025) and show that this modification, too, leads to improved generation.

For Transformer-based generators, identifying outlier tokens is more subtle. In supervised and self-supervised ViTs, inputs and outputs differ in semantic abstraction, and prior work reports that outlier tokens typically emerge in late layers. In contrast, diffusion models operate on inputs and outputs of the same modality—local pixels or latents. Surprisingly, we find that outlier tokens still arise, but predominantly in intermediate layers of the diffusion Transformer. By introducing register tokens into the generator, we achieve consistent gains across a variety of diffusion architectures, including SiT Ma et al. (2024), JiT Li & He (2025), and RAE-based Zheng et al. (2025); Tong et al. (2026) designs. We refer to this unified use of register tokens in both the representation encoder and the diffusion generator as *Dual-Stage Registers* (DSR).

Overall, our results show that DSR provide a simple yet effective way to stabilize Transformer-based diffusion pipelines by addressing outlier tokens in both the tokenizer and the denoiser. This unified intervention consistently improves generation quality across settings. Concretely, for RAE-DiT-XL with SigLIP2-B, it reduces ImageNet-256 FID from 5.89 to 4.58 after 80 epochs, and improves GenEval from 0.426 to 0.466 on a large-scale text-to-image task.

2 Related Work

Outlier Tokens in Transformers. In Vision Transformers, Darcet et al. identify a small fraction of patch tokens with unusually large feature norms, referred to as *outlier tokens* Darcet et al. (2023). These tokens often carry limited local patch information while inducing irregular attention patterns. To mitigate this effect, they introduce extra learnable register tokens during training, which absorb the outlier behavior and lead to cleaner attention over image patches Darcet et al. (2023). Building on this finding, subsequent work traces these outliers to a sparse set of register neurons and proposes a training-free alternative: at inference time, the corresponding activations can be shifted into an additional untrained token, recovering much of the benefit of trained registers in models released without them Jiang et al. (2025).

Related sink-like phenomena have also been studied in language models. In this setting, *attention sinks* attract disproportionate attention and have been linked to large residual-stream activations, motivating neuron-level or activation-level test-time interventions Xiao et al. (2024); Sun et al. (2024); Yona et al. (2025). More recent work explores the connection between sink behavior and outlier-driven dynamics Qiu et al. (2026). Together, these studies suggest that outlier or sink tokens are a recurring phenomenon in Transformer architectures, and register-augmented ViT backbones have become increasingly common in practice Darcet et al. (2023). By contrast, in Diffusion Transformers (DiTs), it remains unclear whether similar outlier tokens emerge during denoising and whether they affect generation quality. Our work addresses this gap.

Diffusion Transformers Across Different Input Spaces. DiTs are originally trained in variational autoencoder (VAE) latent spaces for efficiency, but the tokenizer can bottleneck reconstructed image quality because its latents may not fully preserve semantic information Rombach et al. (2022); Peebles & Xie (2023). Subsequent work improves latent DiT pipelines through representation alignment with visual features from vision foundation models, improving training efficiency and generation quality

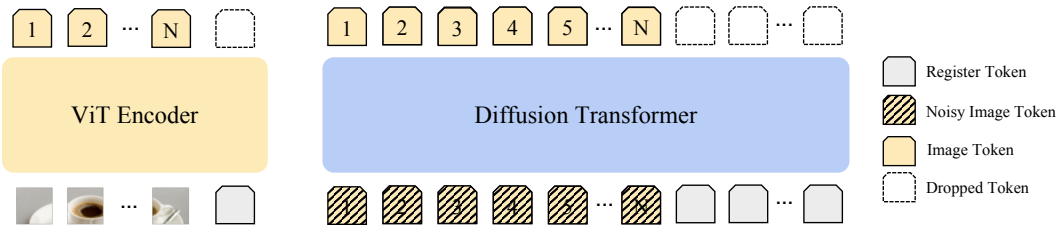

Figure 1: **Framework of DSR.** The encoder uses a *test-time register token* (inference only); the diffusion model uses 36 *trained register tokens* (learned during training). Register-token outputs are discarded before passing features between stages and in the final output.

Yu et al. (2024). More recent work further enables end-to-end joint optimization of the VAE tokenizer and diffusion model Leng et al. (2025).

Apart from VAE space, RAE-based pipelines replace VAEs with pretrained representation encoders paired with learned decoders, yielding semantically richer latents and faster convergence in DiT training Zheng et al. (2025). In parallel, another line of work revisits pixel-space diffusion to remove the need of tokenizers, motivated by the desire to avoid lossy encode–decode stages and simplify the modeling stack. These approaches train diffusion transformers directly on images and show competitive performance when paired with appropriate architectures and training recipes Li & He (2025); Baade et al. (2026).

3 Outlier Tokens in RAE-DiT

RAE-DiT consists of two Transformer-based stages: a ViT-based representation encoder that serves as the tokenizer, and a diffusion Transformer that generates in the representation space. We examine outlier tokens in both stages and find distinct outlier patterns in the encoder and the diffusion generator.

3.1 Outlier Tokens in ViT-based Autoencoders

Outlier tokens have been observed in standard ViTs, where they typically appear in the final layers and can be mitigated by register tokens Darcet et al. (2023). In RAE-DiT, such outliers in encoders are especially relevant because the encoder features define the representation space used for diffusion training.

As shown in Fig. 2, using SigLIP2-B as an example, we observe severe high-norm tokens in the last few encoder layers. The penultimate layer exhibits the strongest outliers, while the final output becomes more stable, potentially due to the reconstruction-related training objective in SigLIP2. These observations suggest that outlier tokens already exist in the encoders before diffusion training.

3.2 Outlier Tokens in Transformer-based Generators

Outlier pattern. Beyond prior analyses of vision foundation models (standard ViTs), we identify a consistent outlier-token phenomenon in the RAE-DiT pipeline: a small subset of tokens in the representation encoder exhibits abnormally large norms, and this effect is further amplified when the encoder features are used as conditioning for diffusion training. As shown in Fig. 3, for RAE-DiT with SigLIP2-B as the encoder, outliers concentrate in *intermediate* layers, and their severity decreases as the diffusion noise level increases, suggesting a compounding mechanism in which encoder anomalies are amplified by the denoising objective. This differs from the standard ViT pattern discussed above, where outliers are typically most pronounced in the *final* layers Darcet et al. (2023). A natural question is how these outliers affect generation quality. The most intuitive explanation is that a small number of high-norm tokens produce extreme loss values that dominate optimization. We test this directly below.

Masking the high norms. We apply a simple token-level loss masking strategy in representation-space diffusion training to exclude outlier tokens from the loss. Let $z_0 \in \mathbb{R}^{N\times d}$ denote the clean token representation, and let $z_t = \alpha(t)z_0 + \sigma(t)\epsilon$ be the noisy tokens, where $\epsilon \sim \mathcal{N}(0, I)$ and

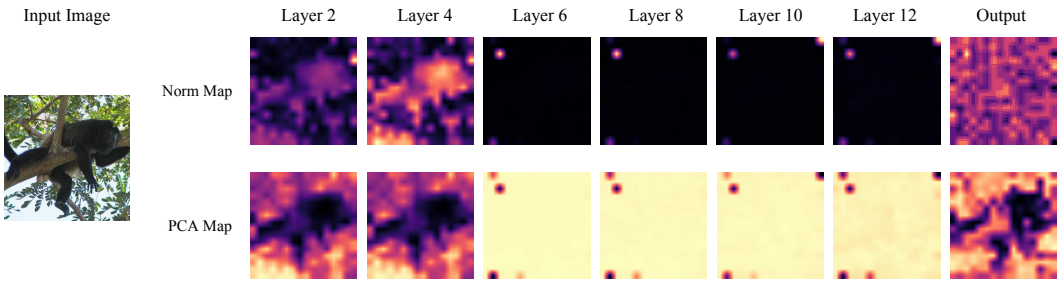

Figure 2: **Outlier tokens in ViT-based autoencoders.** Token-norm maps across SigLIP2-B encoder layers. The penultimate layer shows the strongest outlier pattern; the final output is more stable, potentially due to SigLIP2's reconstruction objective.

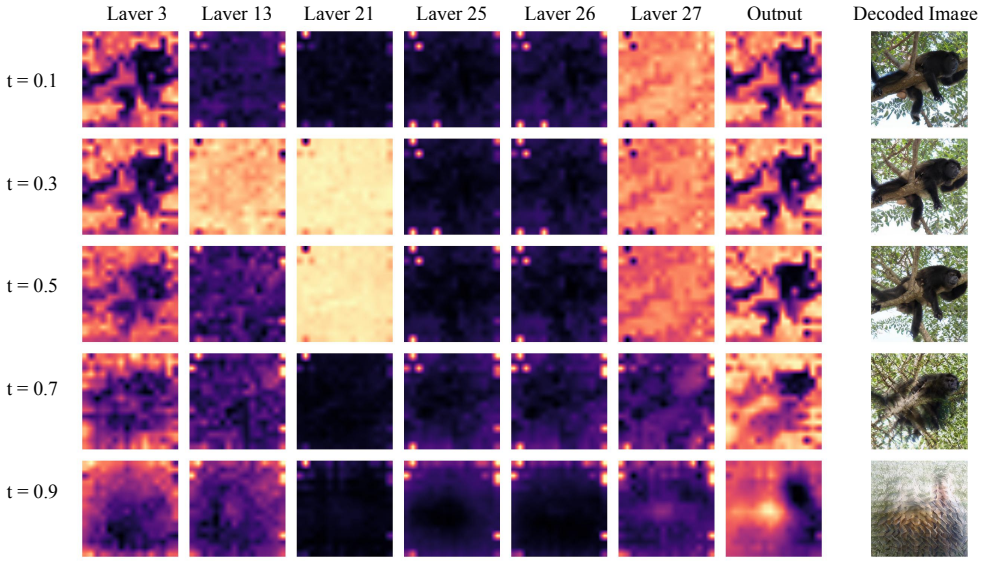

Figure 3: **Outlier tokens in Transformer-based generators.** Token-norm maps of RAE-DiT (SigLIP2-B) across noise scales and layers. High-norm outliers concentrate in *intermediate* layers and diminish at higher noise, differing from standard ViTs where outliers peak in final layers.

$t \in [0, 1]$. Following prior work Li & He (2025), the generator predicts $\hat{z}_0 = x_\theta(z_t, t, c)$, which we convert to a v-prediction as $\hat{v}_\theta = (\hat{z}_0 - z_t)/(1-t)$; the corresponding target is $v = (z_0 - z_t)/(1-t)$.

We then mask tokens according to their representation norms, defining

$$m_i = \mathbf{1}[\|z_{0,i}\|_2 \leq \tau]\,.$$

The masked training objective is

$$\mathcal{L}_v^{\text{mask}}(\theta) = \mathbb{E}_{z_0,\epsilon,t}\left[\frac{1}{\sum_i m_i}\sum_{i=1}^{N} m_i \,\|\hat{v}_\theta(z_t, t, c)_i - v_i\|_2^2\right]. \tag{1}$$

In other words, tokens with unusually large representation norms are excluded from the diffusion loss. If the degradation were driven primarily by a small number of extreme-loss tokens, this masking strategy would be expected to substantially mitigate the problem.

However, as shown in Tab. 1, masking does not improve generation quality. This suggests that the problem lies not in the loss magnitude of individual tokens, but in what these outlier tokens encode in the representation space.

Patch-level semantic degradation. We then hypothesize that outlier tokens reflect corrupted local patch semantics and spatial structure. Prior studies on ViT feature artifacts show that high-norm

Table 1: Effects of loss masking.

Training Strategies	% of tokens filtered	FID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (SigLIP2-B)	0%	5.89	156.54	0.686	0.562
+ loss masking ($\tau = 100$)	0.1%	6.06	152.72	0.686	0.562

tokens degrade dense prediction tasks such as segmentation, and that register tokens can absorb these artifacts and restore patch-level representations Darcet et al. (2023). We verify this in our setting by training linear probes on frozen intermediate DiT features for ImageNet classification and VOC segmentation (Appendix Sec. C): register interventions consistently improve dense-prediction mIoU, confirming that they restore patch-level information rather than merely suppressing extreme values. Analyses in iREPA further suggest that patchwise structure, rather than global semantics alone, is closely tied to generative performance Singh et al. (2025); since the diffusion model is trained to denoise directly in this representation space, restoring patch-level semantics may directly improve generation quality.

4 Improving Generation with Dual-Stage Registers (DSR)

Motivated by the analysis above, we introduce *Dual-Stage Registers* (DSR), a lightweight method that applies register tokens to both stages of the RAE-DiT pipeline to absorb token-level artifacts and stabilize patch representations. As shown in Fig. 1, DSR patches the vision encoder and the diffusion transformer, addressing each source of outliers in turn.

4.1 Registers in Vision Encoders

We begin by validating the effect of *trained* registers in a vision encoder where such registers are available. Specifically, we compare DINOv2 Oquab et al. (2023) trained with and without register tokens[1], measuring outlier severity and downstream generation quality when a DiT is trained on the resulting representation space. As shown in Fig. 4 and Tab. 2, trained registers reduce high-norm token artifacts and improve the quality of patch-level representations, leading to stronger generative performance.

Table 2: RAE-DiT (DINOv2-B) on ImageNet, w/ and w/o encoder registers.

Training Strategies	FID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (DINOV2-B-no-reg)	4.16	203.46	0.787	0.525
RAE-DiT-XL (DINOV2-B-with-reg)	3.95	216.84	0.737	0.547

We next study SigLIP2 Tschannen et al. (2025), a widely used vision-language encoder in recent multimodal models, to test whether our approach remains effective in a more practically important setting. In our experiments, the outlier phenomenon is substantially more severe for SigLIP2, while the off-the-shelf checkpoints do not provide trained registers. We therefore adopt *test-time register* tokens (TTR), following prior work Jiang et al. (2025), by appending an additional token to the input sequence at inference time without further encoder training. Empirically, we find that the SigLIP2-So400 pipeline exhibits *two* distinct sources of outliers, as discussed in Appendix Sec. A. To address this, we apply TTR *recursively*: we first use TTR to stabilize the encoder output, and then apply it again to the resulting representation when a second-stage outlier pattern is detected. As shown in Tab. 4, this recursive TTR scheme consistently reduces outlier severity and improves generation quality.

4.2 Registers in Diffusion Transformers

While encoder-side registers substantially reduce representation outliers, we still observe persistent outlier tokens inside the diffusion transformer itself, especially in intermediate layers, as shown in Fig. 5. This suggests that the generator can also develop a small set of high-norm tokens, potentially

[1]For this comparison, we use the released DINOv2 checkpoints with and without register tokens.

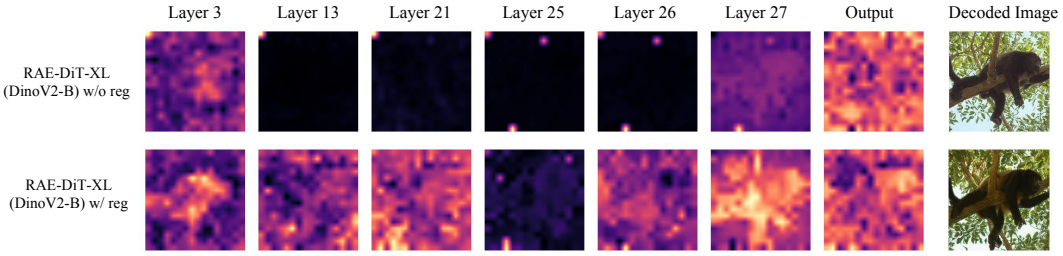

Figure 4: **Norm map comparison for trained encoder registers.** RAE-DiT (DINOv2-B) at t=0.5: trained registers suppress high-norm outliers and improve patch-level representations.

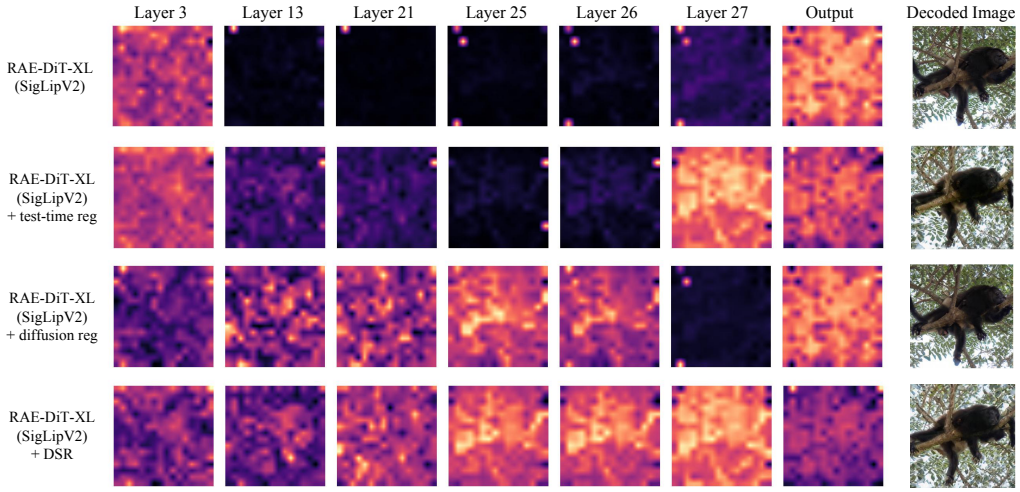

Figure 5: **Norm map comparison across variants.** Outliers are suppressed only when registers are applied to *both* the encoder and the diffusion model.

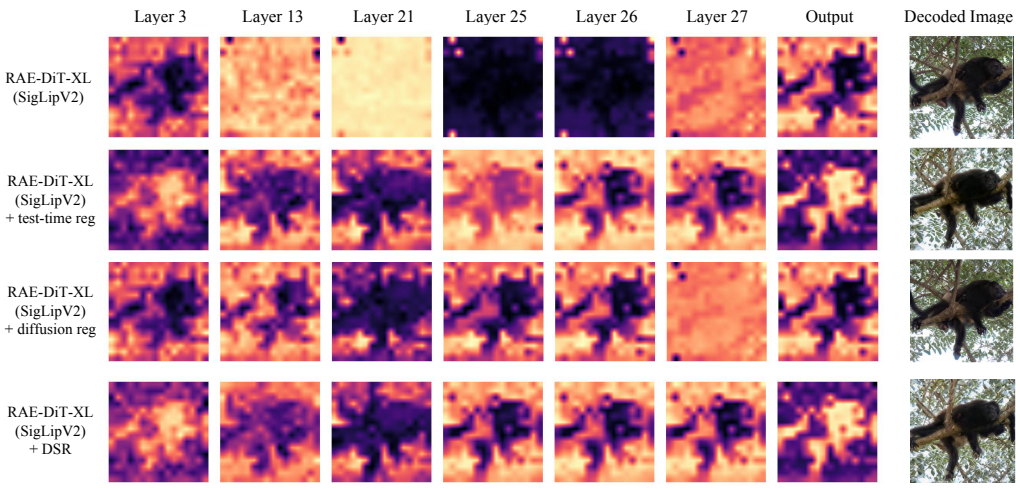

Figure 6: **PCA map comparison across variants.** We observe that adding test-time registers in the encoder yields a strong and visible improvement in the PCA map. Further adding trained registers in the diffusion model brings some additional improvements.

encouraged by the global aggregation behavior of self-attention and the need to maintain global context throughout the denoising process Xiao et al. (2024). Motivated by this observation, we introduce a small number of *trainable diffusion registers* into the diffusion transformer. These

Table 3: **Effect of diffusion registers across different DiT on ImageNet.** Indented rows add diffusion registers on top of the baseline.

Method	FID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (DINOv2-B)	4.16	203.46	0.787	0.525
+ diffusion reg	**3.92**	226.92	0.773	0.542
RAE-DiT-XL (SigLIP2-B)	5.89	156.54	0.686	0.562
+ diffusion reg	**5.33**	166.2	0.702	0.556
RAE-DiT-XL (SigLIP2-B, w/ test-time reg)	4.63	177.2	0.748	0.542
+ diffusion reg	**4.58**	165.99	0.725	0.56
VAE-SiT-XL	16.05	70.11	0.550	0.647
+ diffusion reg	**14.47**	78.50	0.554	0.651
JIT-H	30.34	22.34	0.424	0.621
+ diffusion reg	**23.14**	26.36	0.475	0.611

registers are learned jointly with the generator and removed at inference time, following common practice in prior work on registers Darcet et al. (2023); Jiang et al. (2025). With diffusion registers, the internal outlier pattern is largely eliminated, as shown in Fig. 5 and Fig. 9. Intermediate-layer semantics are also improved, as illustrated by the PCA visualization in Fig. 6, and generation quality improves consistently, as reported in Tab. 3.

More precisely, we find that diffusion registers yield consistent improvements across a wide range of settings. In particular, they improve performance for every input-space variant we evaluate, including pixel space, VAE latents, and multiple representation encoders. This suggests that the gains are not tied to any particular encoder or representation family. Instead, the broad effectiveness of diffusion registers indicates that they address token-level artifacts arising *within* the diffusion transformer itself, rather than serving as a modification tailored to a specific representation.

We also observe a close relationship between diffusion registers and in-context conditioning Li & He (2025). Both mechanisms fill the same auxiliary token slots with different content: registers use freely learned embeddings, while in-context conditioning fills these slots with repeated class-conditioning embeddings. In-context conditioning can therefore be viewed as combining a register-like function with additional class information. As the input-space semantics grow richer, the marginal value of injecting class information decreases and content-free registers suffice; in semantically poorer spaces such as raw pixels, the additional conditioning remains beneficial. A detailed comparison is provided in Appendix Sec. D.

5 Experiments

We evaluate DSR on ImageNet-1K class-conditional generation at 256×256 and text-to-image generation. For ImageNet-1K, we follow the RAE setting (Zheng et al., 2025) and keep the same training epochs, learning rate, and model architecture for fair comparison; our main configuration uses SigLIP2-B as the encoder. While our encoder is slightly different due to the test-time register modification, we use the same decoder as in RAE; we find that retraining the decoder has limited impact (Appendix Sec. E). Additionally, we use data prediction with velocity loss to train our model and RAE baseline, which has been proven effective in JiT (Li & He, 2025).

Our focus on SigLIP2 as the primary encoder is motivated by the scaling setting of RAE (Zheng et al., 2025). While DINOv2 remains the strongest setting for ImageNet class-conditional generation, recent large-scale text-to-image and unified multimodal RAE systems, such as Scale-RAE (Tong et al., 2026), adopt SigLIP2 as the visual encoder. We therefore consider SigLIP2 particularly relevant to larger-scale, real-world applications; Tab. 5 evaluates this setting. DSR substantially reduces gFID for RAE (SigLIP2-B). Combining DSR with the DDT head yields further gains and achieves competitive performance. Fig. 7 plots the per-epoch trajectory and shows that DSR reaches comparable quality with $4\times$ fewer epochs.

We also apply DSR on RAE's strongest ImageNet configuration (DINOv2-B + DiT^{DH}-XL, Tab. 6), where diffusion registers further improve gFID from 1.51 to 1.49; compared with DINOv2-B without

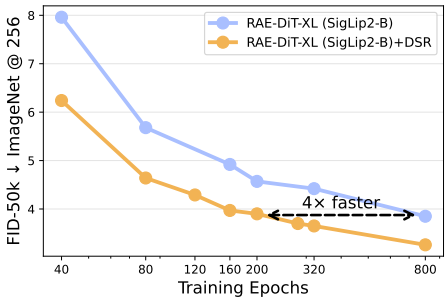

Figure 7: FID vs. epochs on IN-1K 256^2.

Table 4: Test-time registers (SigLIP2-B / So400, 80 epochs).

Training Strategies	FID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (SigLIP2-B)	5.89	156.54	0.686	0.562
+ test-time register	4.63	177.2	0.748	0.542
RAE-DiT-XL (SigLIP2-So400)	7.04	167.01	0.682	0.515
+ test-time register	6.66	166.88	0.687	0.527
+ test-time register (recursive)	6.48	163.35	0.684	0.531

Table 5: RAE-DiT with SigLIP2-B on ImageNet 256×256.

Method	Epochs	#Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (SigLiP2-B)	80	676M	5.89	156.54	0.686	0.562	–	–	–	–
	800		3.85	179.82	0.692	0.613	3.58	194.67	0.691	0.619
RAE-DiT-XL (SigLiP2-B) + DSR	80	676M	4.58	165.99	0.725	0.560	–	–	–	–
	800		3.26	185.54	0.704	0.62	2.97	203.95	0.709	0.621
RAE-DiT^{DH} (SigLiP2-B)	80	839M	3.74	179.25	0.707	0.600	–	–	–	–
	800		2.91	204.73	0.686	0.642	2.77	221.76	0.685	0.649
RAE-DiT^{DH}-XL (SigLiP2-B) + DSR	80	839M	3.57	181.08	0.722	0.590	–	–	–	–
	800		2.72	207.76	0.697	0.646	2.62	223.85	0.697	0.651

Table 6: Class-conditional performance on ImageNet 256×256.

Method	Epochs	#Params	Generation@256 w/o guidance				Generation@256 w/ guidance			
			gFID↓	IS↑	Prec.↑	Rec.↑	gFID↓	IS↑	Prec.↑	Rec.↑
Pixel Diffusion										
ADM (Dhariwal & Nichol, 2021)	400	554M	10.94	101.0	0.69	0.63	3.94	215.8	0.83	0.53
JiT-H/16 (Li & He, 2025)	600	953M	–	–	–	–	1.86	303.4	–	–
Latent Diffusion										
SiT-XL (Ma et al., 2024)	1400	675M	8.61	131.7	0.68	0.67	2.06	270.3	0.82	0.59
REPA (Yu et al., 2024)	800	675M	5.78	158.3	0.70	0.68	1.29	306.3	0.79	0.64
REPA-E (Leng et al., 2025)	800	675M	1.70	217.3	0.77	0.66	1.15	304.0	0.79	0.66
RAE-DiT^{DH}-XL (DINOv2-B w/o reg)	800	839M	1.83	226.0	0.79	0.64	1.52	228.2	0.77	0.66
RAE-DiT^{DH}-XL (DINOv2-B)	800	839M	1.51	242.9	0.79	0.63	1.23	262.6	0.78	0.67
+ DSR	800	839M	**1.49**	237.8	0.79	0.64	**1.17**	256.6	0.78	0.67

registers, the overall improvement is more substantial, reducing unguided gFID from 1.83 to 1.49 and guided gFID from 1.52 to 1.17.

5.1 Scalability and Ablation

A key motivation for decoupling the Transformer design from the task is to better leverage scaling (Li & He, 2025; Peebles & Xie, 2023). We reports ImageNet 256 × 256 results across model sizes (DiT-B/L/XL). As shown in Tab. 7, DSR consistently improves gFID across all parameter scales, while introducing only a minor increase in GFLOPs.

We conduct an ablation study of diffusion registers on RAE-DiT-XL with SigLIP2-B on ImageNet 256×256. The baseline model uses no diffusion registers. We study two factors: the insertion depth, defined as the transformer block from which registers are introduced, and the number of registers. In the first set of experiments, we fix the number of registers to 36 and vary the insertion depth. In the second, we fix the insertion depth to block 8 and vary the number of registers.

As shown in Tab. 8, diffusion registers provide clear but non-monotonic gains, with a distinct sweet spot in both insertion depth and register count. When fixing the register count at 36, introducing registers too early or too late is less effective, while starting from block 8 gives the best performance.

Table 7: Scalability of DSR across DiT model sizes (SigLIP2-B, 100k iterations).

Model	FID↓	IS↑	Prec.↑	Rec.↑	Gflops.↓
RAE-DiT-S (SigLIP2-B)	28.03	66.15	0.277	0.302	12.4
+ DSR	23.93	63.86	0.498	0.470	$13.7_{+10.4\%}$
RAE-DiT-B (SigLIP2-B)	20.36	78.76	0.539	0.493	46.6
+ DSR	9.81	110.23	0.637	0.543	$51.16_{+9.9\%}$
RAE-DiT-XL (SigLIP2-B)	5.89	156.54	0.686	0.562	238.1
+ DSR	4.58	165.99	0.725	0.560	$262.9_{+10.4\%}$

Table 8: Ablation on diffusion registers (SigLIP2-B, ImageNet 256×256). We vary insertion depth (fixed 36 registers) and register count (fixed block 8).

Starting block	#Regs	FID↓	IS↑	Pre.↑	Rec.↑	Gflops ↓
-	0	5.89	156.54	0.686	0.562	238.1
0	36	5.54	163.45	0.701	0.552	$272.8_{+14.6\%}$
8	36	**5.33**	166.20	0.702	0.556	$262.9_{+10.4\%}$
16	36	5.49	168.01	0.700	0.549	$253.0_{+6.3\%}$
24	36	5.68	159.04	0.685	0.560	$243.1_{+2.1\%}$
8	1	6.16	153.95	0.685	0.555	$238.9_{+0.3\%}$
8	4	5.47	165.55	0.700	0.553	$240.8_{+1.1\%}$
8	36	**5.33**	166.20	0.702	0.556	$262.9_{+10.4\%}$
8	100	5.58	165.61	0.701	0.546	$307.5_{+29.1\%}$

This suggests that registers are most useful when introduced in the early-to-middle part of the generator, where they can influence a substantial portion of the computation without interfering too strongly with the earliest layers. When fixing the starting block to 8, varying the number of registers shows a similar trend: very small numbers do not help, moderate numbers improve performance, the best results are obtained with 36 registers, and performance degrades at 100 registers. Overall, these results suggest that diffusion registers are most effective at an appropriate depth and capacity.

Table 9: Text-to-image perf. on Scale-RAE.

Models	GenEval↑	DPG-Bench↑
Baseline	42.6	74.3
Ours	46.6	75.4

We also demonstrate the scalability of DSR on Scale-RAE (Tong et al., 2026), which leverages the MetaQuery architecture (Pan et al., 2025) for text-to-image (T2I) generation and unified modeling. We use SigLIP2-B (Tschannen et al., 2025) as the encoder and train the model on the Scale-RAE dataset (Tong et al., 2026), which contains 24.7M synthetic images generated by FLUX.1-schnell (Black Forest Labs, 2024).

The training process takes 12 hours on a google cloud v5p-128 TPU for 10k training steps with a batch size of 2048. We then sample images from the trained model using text prompts and decode the generated latents with SigLIP-B pretrained decoder provided by RAE (Zheng et al., 2025). We evaluate the sampled images on GenEval (Ghosh et al., 2023) and DPG-bench (Hu et al., 2024). As shown in Tab. 9, DSR achieves better performance than the original RAE baseline on both benchmarks. More details can be found in Appendix Sec. G.

6 Conclusion

We study outlier tokens in Diffusion Transformers and show that they arise not only in pretrained vision encoders, but also within the diffusion model itself. Our analysis suggests that these outliers reflect degraded patch-level semantics rather than merely a few abnormally large token norms. Motivated by this insight, we introduce simple register-based interventions for both the encoder and the diffusion transformer. These interventions consistently improve training stability and generation quality in both ImageNet generation and large-scale text-to-image generation, and generalize across diverse input spaces and model variants. Overall, our results identify outlier-token control as an important ingredient for building stronger and more robust Transformer-based diffusion pipelines.

Reproducibility Statement

We provide detailed experimental configurations including model architectures, training hyperparameters, and evaluation protocols in the main text and appendix. Our method builds on publicly available codebases and pretrained models. Code is available via `https://anonymous.4open.science/r/TAMING-OUTLIER-EB01`

AI Use Statement

In this work, we used generative AI tools to assist with research execution and paper writing. Specifically, AI coding assistants were used to write and edit experiment code and to launch and manage training and evaluation runs; all AI-assisted code was reviewed by the authors, and all reported results were verified against the raw evaluation outputs. We also used generative AI tools to edit the paper for readability (grammar, spelling, and word choice). We have not used generative AI tools to develop the core research ideas, theoretical models, or conceptual frameworks, or to formulate the claims of this paper. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Alan Baade, Eric Ryan Chan, Kyle Sargent, Changan Chen, Justin Johnson, Ehsan Adeli, and Li Fei-Fei. Latent forcing: Reordering the diffusion trajectory for pixel-space image generation. *arXiv preprint arXiv:2602.11401*, 2026.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. GitHub repository.

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.

Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representation*, 2022.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

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

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.

Nicolas Dufour, Alexei A. Efros, and Patrick Pérez. The FID lottery: Quantifying hidden randomness in generative-model evaluation. *arXiv preprint arXiv:2606.20536*, 2026.

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

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025.

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.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *CVPR*, 2022.

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.

Zihan Qiu, Zeyu Huang, Kaiyue Wen, Peng Jin, Bo Zheng, Yuxin Zhou, Haofeng Huang, Zekun Wang, Xiao Li, Huaqing Zhang, et al. A unified view of attention and residual sinks: Outlier-driven rescaling is essential for transformer training. *arXiv preprint arXiv:2601.22966*, 2026.

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.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

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.

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

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.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fréchet loss for visual generation. *arXiv preprint arXiv:2604.28190*, 2026.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712. IEEE, 2025.

Itay Yona, Ilia Shumailov, Jamie Hayes, Federico Barbero, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*, 2025.

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.

Zhengrong Yue, Taihang Hu, Mengting Chen, Haiyu Zhang, Zihao Pan, Tao Liu, Zikang Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, et al. What matters for diffusion-friendly latent manifold? prior-aligned autoencoders for latent diffusion. *arXiv preprint arXiv:2605.07915*, 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

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

A Outlier Visualization on SigLIP2-So400

We further analyze the norm distribution of SigLIP2-So400 and observe that its outliers do not form a single homogeneous group. Instead, they appear as two clearly separated clusters, indicating that a single-pass filtering strategy is insufficient. As shown in Fig. 8, filtering once removes only the more salient cluster, while the other remains visible. To address this, we adopt a recursive procedure that re-localizes outlier neurons after the first filtering pass and then jointly filters all detected outliers, allowing us to suppress both groups more effectively.

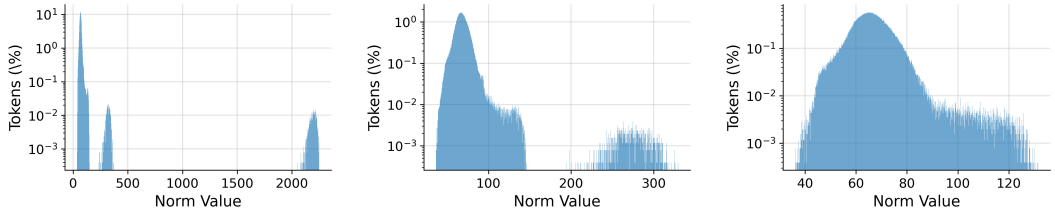

Figure 8: **Two outlier sources in the norm distribution of SigLIP2-So400.** We compute the ℓ_2 norm of SigLIP2-So400 output features on 10k randomly selected images from the ImageNet-1K validation set. *Left:* the original norm distribution shows two separated outlier groups. *Middle:* applying our filtering once removes only one group, leaving the other largely intact. *Right:* we therefore use a recursive procedure that re-localizes outlier neurons after the first pass and filters all detected outliers jointly, which effectively suppresses both groups.

B More Visualization

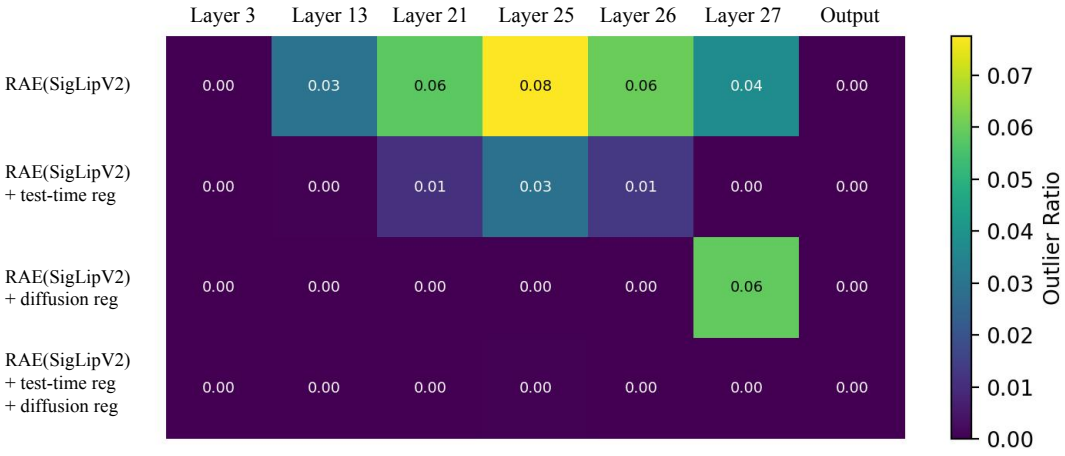

Figure 9: **Quantitative measurements of outliers.** We report the fraction of outlier tokens across layers under different setups. An outlier is defined as a token whose ℓ_2 norm exceeds $2\times$ the median token norm.

To complement the qualitative visualization, we also provide a quantitative view of outlier behavior across layers. Figure 9 reports the fraction of outlier tokens under different setups, where we define an outlier as a token whose ℓ_2 norm is greater than $2\times$ the median token norm. These results show that outlier tokens are not uniformly distributed across depth, but instead exhibit clear layer-dependent patterns that vary across settings.

C Linear Probing of Patch Features

We hypothesize that outlier tokens corrupt patch-level semantics, and that register tokens restore them. To test this, we freeze intermediate DiT features (t=0.75, Block 15) and train linear probes

on the patch tokens for two tasks: ImageNet classification (top-1) and PASCAL VOC segmentation (mIoU). Segmentation depends on each patch encoding its local region correctly, making it a direct proxy for patch-level feature quality.

Table 10: **Linear probing of intermediate DiT patch features.** We freeze DiT features at t=0.75, Block 15 and train linear probes for ImageNet classification and VOC segmentation.

Backbone	Encoder registers	Diffusion registers	ImageNet top-1↑	VOC mIoU↑
SigLIP2	–	–	81.641±0.046	74.579±0.068
SigLIP2	Test-time	–	81.108±0.028	74.842±0.019
SigLIP2	Test-time	✓	81.180±0.024	**74.858±0.061**
DINOv2	–	–	79.822±0.030	79.224±0.068
DINOv2	Trained	–	**80.625±0.029**	79.465±0.026
DINOv2	Trained	✓	80.611±0.034	**79.615±0.023**

As shown in Tab. 10, register interventions consistently improve VOC mIoU across both encoder backbones, confirming that the intermediate DiT features carry better patch-level spatial information after the register intervention. This supports the hypothesis in Sec. 3 that the benefit of registers stems from restoring patch-level semantics in the representation space.

D Diffusion Registers vs. In-Context Conditioning

We observe a close relationship between diffusion registers and in-context conditioning Li & He (2025). Both augment the input with extra tokens that can carry global context: in-context conditioning uses input-dependent tokens, while registers provide dedicated learnable slots. As shown in Tab. 11, in-context conditioning is most competitive when the input space contains less semantic or structured token representations, while diffusion registers remain beneficial as representations become more semantic. These results suggest that registers provide a simpler and more robust mechanism for supplying global capacity, without relying on the semantics or formatting of in-context tokens.

Table 11: **Register tokens vs. in-context conditioning across DiTs and input spaces.** In-context conditioning yields smaller gains as the input representation becomes more semantic, whereas register tokens provide consistent improvements.

Method	FID↓	IS↑	Prec.↑	Rec.↑
RAE-DiT-XL (SigLIP2-B)	5.89	156.54	0.686	0.562
+ diffusion reg	**5.33**	166.2	0.702	0.556
+ diffusion in-context condition	5.79	164.99	0.696	0.546
VAE-SiT-XL	16.05	70.11	0.550	0.647
+ diffusion reg	14.47	78.50	0.554	0.651
+ diffusion in-context condition	**14.31**	77.04	0.559	0.646
JIT-H	30.34	22.34	0.424	0.621
+ diffusion reg	23.14	26.36	0.475	0.611
+ diffusion in-context condition	**15.51**	33.05	0.519	0.618

E Retrained Decoder

Applying test-time registers alters the encoder feature distribution. In principle, this distribution shift may require retraining the Stage-2 decoder. Interestingly, we find that the shift is mild: retraining the decoder yields no visible improvement over using the pretrained decoder.

For SigLIP2-B, we search over augmented noise scales σ, a standard knob for balancing reconstruction quality and gFID in prior work Zheng et al. (2025). As shown in Tab. 12, We find no setting that

consistently outperforms the baseline, retraining under different σ yields results comparable to the pretrained decoder. It suggests that test-time registers only mildly shift the encoder feature distribution seen by the decoder.

Table 12: **Retraining the decoder.** Retraining the Stage-2 decoder after applying test-time registers brings negligible gains, indicating that the encoder distribution shift is mild.

Training Strategy	Decoder	rFID↓	gFID↓	IS↑
DiT-XL (SigLIP2-B)	Pretrained	0.82	5.89	156.54
DiT-XL (SigLIP2-B) + DSR	Pretrained	0.82	4.58	165.99
DiT-XL (SigLIP2-B) + DSR	Retrained (σ=0.8)	0.58	6.00	164.20
DiT-XL (SigLIP2-B) + DSR	Retrained (σ=1.5)	0.67	5.35	157.59
DiT-XL (SigLIP2-B) + DSR	Retrained (σ=2.0)	0.75	5.28	157.98

F Remarks on FID Evaluation

F.1 Sensitivity to the Reference Set

FID is not solely a property of the generator; it depends on the full evaluation pipeline, including the choice of reference statistics Parmar et al. (2022). We present two controlled checks that quantify how much the reference alone can move the score. First, Tab. 13 scores the *same* generated samples from publicly released models against three reference statistics: the 50k validation set, a random 50k subset of training images, and the ADM pre-computed reference Dhariwal & Nichol (2021). The shift is model-dependent in both sign and magnitude, since a generator whose outputs sit closer to the training distribution scores relatively better against training-based references. It can even flip rankings: VA-VAE Yao et al. (2025) appears slightly better than RAE Zheng et al. (2025) under the validation reference but worse under the training-based references. Second, Tab. 14 computes the FID of real ImageNet training images against each reference over 100 independent random 50k draws: even with a fixed reference, the floor fluctuates by a few hundredths. Combined with the observation that retraining the same recipe under different random seeds moves FID by $\pm$0.4 on ImageNet-256 Dufour et al. (2026), small FID gaps can easily come from the evaluation pipeline rather than the generator.

For these reasons, we compute all FID numbers in this paper against the ImageNet validation statistics (FID_{VAL}), which avoids overlap between the reference distribution and the training data and yields a more faithful estimate of generalization quality. The exceptions are the main comparison in Tab. 6 and the Inception column of Tab. 15, where we adopt the ADM pre-computed reference to stay consistent with the evaluation setting used by the baselines. To further highlight the advantage of our method beyond single-space Inception FID, we additionally report FDr-6 in the next subsection.

Table 13: **FID of the same generated samples under three reference statistics.** The shift is model-dependent in both sign and magnitude, and can flip rankings between models.

Input space	Val-50k	Train-50k	ADM ref
RAE Zheng et al. (2025)	4.295	4.008	3.535
PAE Yue et al. (2026)	2.976	3.539	2.889
VA-VAE Yao et al. (2025)	4.204	5.524	4.859

F.2 Per-Space FDr-6 Breakdown

When models are already strong, Inception FID saturates and may no longer resolve the remaining gap between a method and its baseline. Following the Representation Fréchet Loss Yang et al. (2026), we therefore additionally report FDr-6, which normalizes the Fréchet distance in six representation spaces (Inception-V3, ConvNeXt-V2-B, DINOv2-L, MAE-L, SigLIP2-so400m, and CLIP-L) and averages them. As shown in Tab. 15, FDr-6 clearly separates our method from the baselines: on

Table 14: **FID floor of real ImageNet training images against each reference**, over 100 independent random 50k draws. Even a fixed reference carries sampling variance.

Reference	n	Mean	Std	Variance	Range
vs Val	100	2.223	0.029	0.00084	[2.137, 2.276]
vs VIRTUAL (ADM)	100	0.857	0.011	0.00013	[0.831, 0.880]
vs JiT	100	0.754	0.011	0.00012	[0.726, 0.777]

RAE's strongest configuration (DINOv2-B + DiT^{DH}-XL), DSR improve every one of the six spaces, both with and without guidance.

Table 15: **Per-space raw Fréchet distances underlying FDr-6.** DSR (applying both encoder registers and diffusion registers) improves every one of the six spaces, both with and without guidance. Guided rows use AutoGuidance with the same guiding model and guidance strength (1.35). The Inception column uses ADM pre-computed reference statistics for consistency with Tab. 6.

Configuration	**FDr-6↓**	Inception	ConvNeXt	DINOv2	MAE	SigLIP	CLIP
RAE-DiT^{DH}-XL (DINOv2-B), 800 epochs							
w/o encoder reg	4.42	1.83	137.88	40.39	0.22	3.33	51.14
w/ encoder reg	4.05	1.51	122.00	39.82	0.18	3.12	49.01
+ diffusion reg	**3.81**	**1.49**	**115.72**	**37.18**	**0.17**	**2.88**	**46.82**
RAE-DiT^{DH}-XL (DINOv2-B), 800 epochs, w/ AutoGuidance							
w/o encoder reg	4.05	1.52	133.15	37.17	0.19	2.97	48.63
w/ encoder reg	3.92	1.23	124.64	39.50	0.17	2.98	48.29
+ diffusion reg	**3.25**	**1.17**	**105.59**	**29.80**	**0.14**	**2.22**	**42.80**

G More Results on Text-to-Image Experiments

Training configuration. Tab. 16 summarizes the computational resources required and the training hyperparameter configuration for the scaling-up experiment. Following Scale-RAE Tong et al. (2026), We use SPMD sharding together with TorchXLA to train the LLM and DiT models. Due to limited computational resources, we use only the Scale-RAE dataset, which is one quarter the size of the data used in the original Scale-RAE Tong et al. (2026), and train the model for one epoch.

Step-to-step comparison. We also perform a step-to-step comparison between DSR and Scale-RAE baseline. The detailed comparison is shown in Fig. 11. We observe consistent improvements of DSR over the Scale-RAE Tong et al. (2026) baseline on GenEval Ghosh et al. (2023). On DPG-Bench Hu et al. (2024), DSR shows a steadier improvement. We further compare the training losses of the baseline and DSR. As shown in Fig. 10, the green curve corresponding to DSR is consistently lower and more stable overall, and the same trend is also visible in the diffusion loss. In contrast, the language-model loss of DSR exhibits more spikes during training, which may help explain why it sometimes underperforms the baseline on DPG-Bench, a benchmark that emphasizes following longer and more complex instructions. This suggests that additional improvements on the language side, such as mitigating the effect of outlier tokens, may further enhance performance. We leave this direction for future work.

Extended samples. We provide extended samples from Scale-RAE Tong et al. (2026) and DSR on DPG-Bench Hu et al. (2024) and Geneval Ghosh et al. (2023) (Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16, Fig 17, Fig 18,Fig 19). The prompts for generated images are:

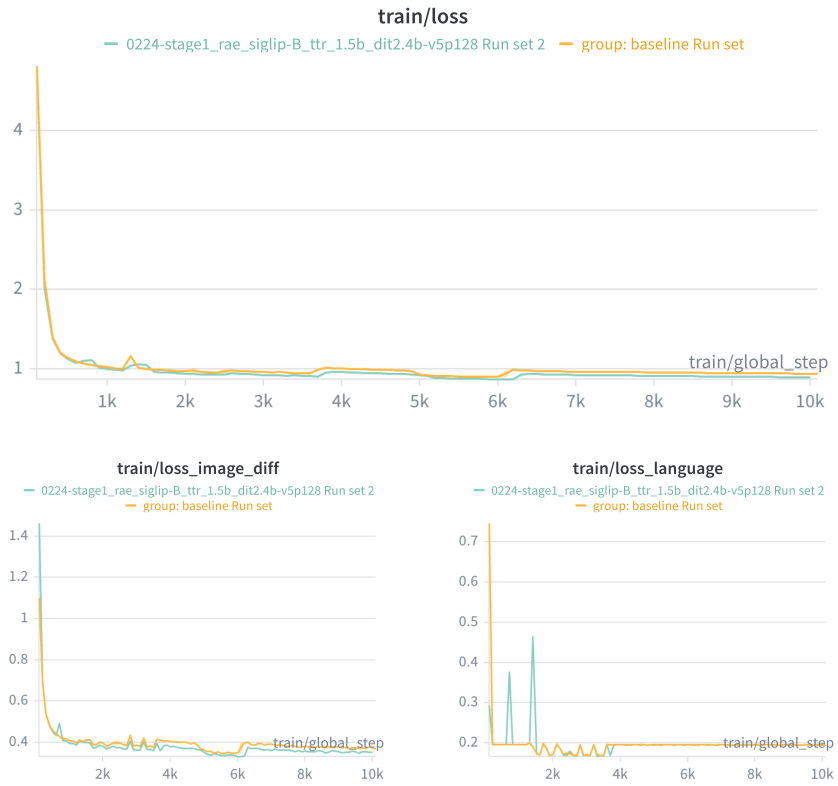

Figure 10: Training loss comparison between Scale-RAE and DSR. Top: total loss. Bottom left: image diffusion loss. Bottom right: language loss.

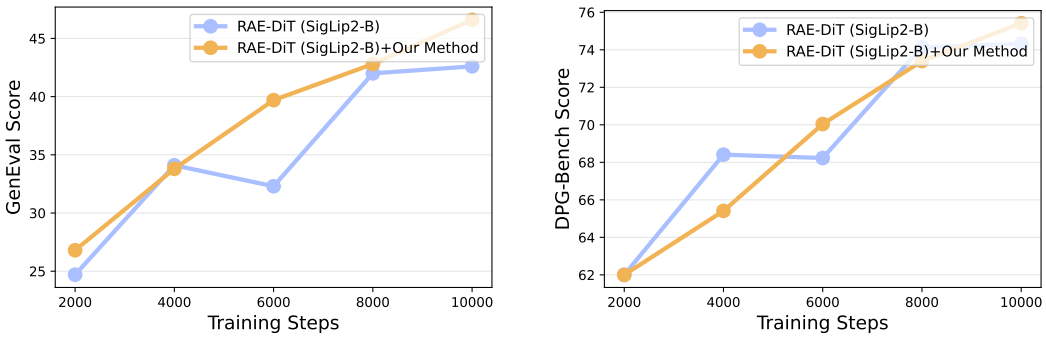

Figure 11: Step-to-step comparison between Scale-RAE and DSR on Geneval and DPG-Bench.

Method	Scale-RAE		Scale-RAE+DSR	
component	LLM	DiT	LLM	DiT
optimizer	AdamW			
learning rate schedule	cosine w/ warmup ratio 0.0134			
global batch size	2048			
Training epoch	1			
Dataset	Scale-RAE dataset (24.7M)			
max learning rate	5e-5	5e-4	5e-5	5e-4
optimizer betas	(0.9, 0.999)	(0.9, 0.95)	(0.9, 0.999)	(0.9, 0.95)
loss	autoregressive loss	diffusion loss	autoregressive loss	diffusion loss
model	Qwen2.5 1.5B	DiT 2.4B	Qwen2.5 1.5B	DiT 2.4B
Register num	-	0	-	36
Register adding block	-	0	-	8
Vision encoder	google/siglip2-base-patch16-256			
Vision decoder	nyu-visionx/RAE-siglip2-base-p16-i256-ViTXL-n08			
Encoder test time register	-	No	-	Yes
Training device	v5p-128			
TPU memory usage per chip	22.34GiB		23.68GiB	
Training speed	4.00s/iter		4.09s/iter	

Table 16: Training configuration comparison between Scale-RAE and DSR.

- During the twilight hour, an individual can be seen extending an arm towards the sky, pointing at a trio of wild birds gliding through the rich deep blue of the early evening sky. The birds' silhouettes contrast distinctly against the fading light, their wings spread wide as they soar. The person is silhouetted against the dusky sky, creating a peaceful scene of human connection with nature.
- An expansive palace constructed from iridescent materials that shimmer with hues reminiscent of a vivid, Slime-like substance, majestically stands at the heart of a fantastical realm. Its towers twist skyward, defying conventional architecture with their organic, flowing forms. In the foreground, a field of exotic flowers blooms, each petal displaying an array of otherworldly colors that could have been plucked from a Lovecraftian spectrum, while overhead, a radiant sun bathes the surreal landscape in brilliant light.
- a highly intricate and vibrant cityscape that reflects a fusion of Moebius's imaginative design and Makoto Shinkai's detailed animation style. The streets are aglow with neon signs in a kaleidoscope of colors, casting reflections on the glossy, rain-slicked pavements. Towering skyscrapers with glowing windows rise towards a starless night sky, as the artwork garners significant attention and praise on ArtStation.
- Adjacent to each other in a room, a large rectangular bed draped in a navy-blue comforter sits parallel to a square-shaped nightstand with a matte finish. The nightstand holds an angular lamp and a small stack of hardcover books. The two pieces of furniture are positioned on a plush beige carpet that covers the majority of the floor space.
- a photo of a cow.
- a photo of a purple potted plant.
- a photo of an elephant below a surfboard.
- a photo of a white toilet and a red apple.

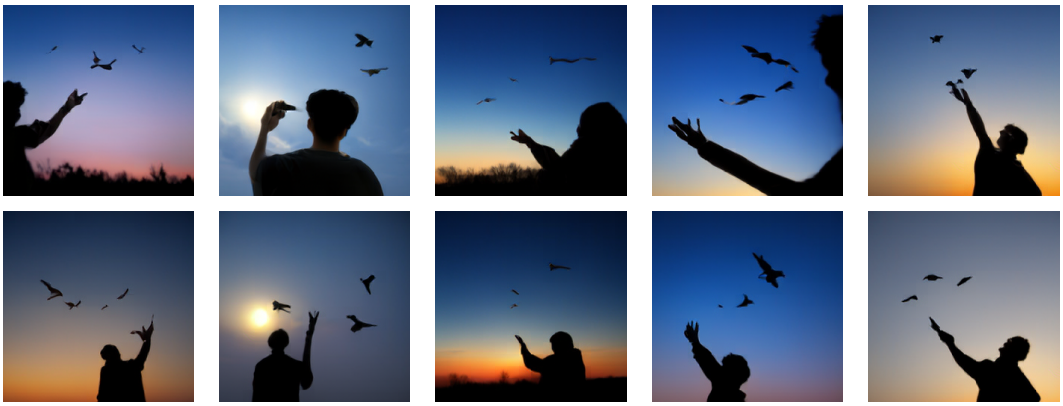

Figure 12: Scale-RAE baseline (top) vs. DSR (bottom) on DPG-Bench Hu et al. (2024).



Figure 13: Scale-RAE baseline (top) vs. DSR (bottom) on DPG-Bench Hu et al. (2024).

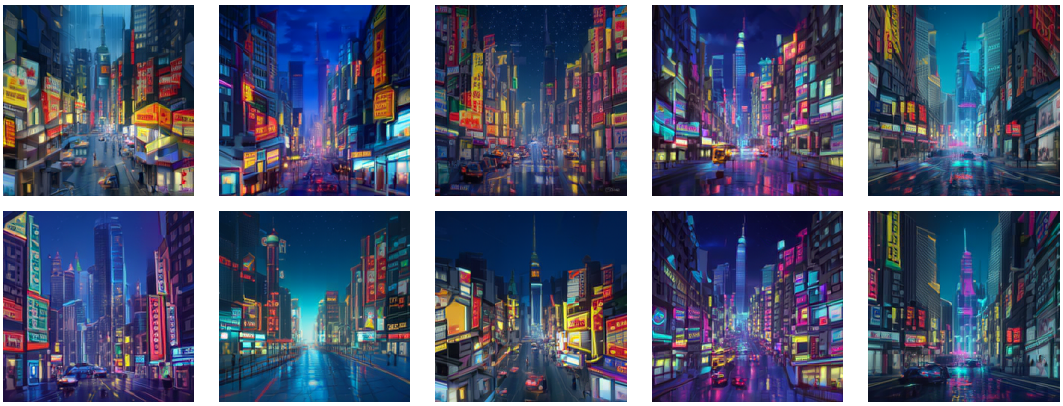

Figure 14: Scale-RAE baseline (top) vs. DSR (bottom) on DPG-Bench Hu et al. (2024).



Figure 15: Scale-RAE baseline (top) vs. DSR (bottom) on DPG-Bench Hu et al. (2024).



Figure 16: Scale-RAE baseline (top) vs. DSR (bottom) on GenEval Ghosh et al. (2023).



Figure 17: Scale-RAE baseline (top) vs. DSR (bottom) on GenEval Ghosh et al. (2023).

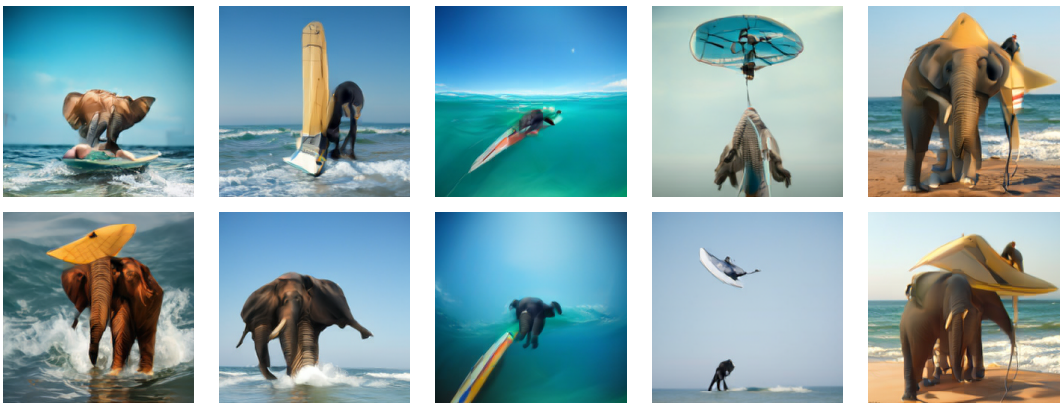

Figure 18: Scale-RAE baseline (top) vs. DSR (bottom) on GenEval Ghosh et al. (2023).



Figure 19: Scale-RAE baseline (top) vs. DSR (bottom) on GenEval Ghosh et al. (2023).