LEARNING WHAT TO ALIGN FOR FASTER DIFFUSION TRANSFORMER TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semantic emergence in Diffusion Transformers motivates alignment with external semantic representations to accelerate training convergence. However, observations suggest that such rigid alignment is often suboptimal, creating an inherent paradox: higher semantic accuracy does not necessarily lead to better generation fidelity. In this paper, we propose allowing the denoising objective itself to ascertain the specific facets of semantic information most beneficial to the generative process. Furthermore, we introduce a semantic internalization mechanism coupled with stage-wise training to estimate and reintegrate those task-relevant signals into the model's intrinsic representation. Our acceleration regime prioritizes generation-critical features over rigid alignment with fixed-target semantic representations. Extensive experiments demonstrate that our approach consistently further accelerates training and improves generation quality across various model scales and resolutions, outperforming existing alignment-based baselines.

1 INTRODUCTION

Several studies (Xiang et al., 2023; Mukhopadhyay et al., 2023) have demonstrated that diffusion transformers implicitly capture object semantics during the generative training process. However, when trained solely under the standard denoising objective, the emergence of these semantics lags significantly behind that of self-supervised learning counterparts. In this context, representation alignment (Yu et al., 2025; Jiang et al., 2025) has recently emerged as a promising direction for accelerating diffusion model training by performing alignment between internal diffusion features and pretrained or self-produced semantic representations. In particular, pioneering approaches such as REPA (Yu et al., 2025) align diffusion features with semantic representations from pre-trained semantic encoders such as DINO (Oquab et al., 2023). As rich semantics are guided to produce, this leads to faster diffusion training convergence and improved generation quality compared to the baseline. However, subsequent works (Singh et al., 2025; Wang et al., 2025b) have shown that the standard alignment objective used in REPA remains suboptimal. It has been discovered that improved semantics—as measured by linear probing accuracy—do not necessarily correlate with enhanced generation performance, revealing a misalignment between semantic accuracy and generation fidelity.

In this work, we revisit representation alignment from a distinct perspective. Rather than enforcing diffusion features to strictly align *fixed-target semantic representations*, we argue that the denoising objective itself should dictate the extent to which external semantic information are more essential for the generative modeling. By allowing the generative task to *select* relevant signals, we can potentially mitigate the previously mentioned misalignment between semantics and generation fidelity. Our empirical findings confirm that tighter alignment does not inherently guarantee superior generation. Instead, the key to performance lies in effectively extracting more task-relevant signals and integrating them into the model's intrinsic representation space.

In this context, we propose a simple yet effective training framework which replaces rigid representation alignment with *denoising-guided semantic selection and internalization*. Our method operates through two key stages. First, parallel to the denoising objective, the model learns to selectively infuse task-relevant information from external semantic representations to a set of learnable latent tokens through a single cross-attention layer. These tokens, now enriched with more task-relevant semantics, are injected back into the diffusion backbone through another cross-attention layer to accelerate training convergence. Second, to further decouple the architecture from external dependen-

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

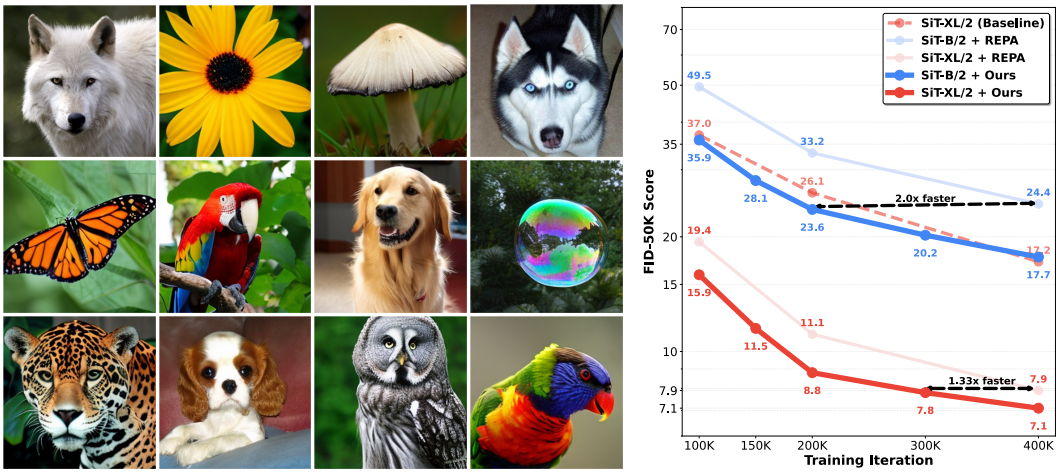

Figure 1: Qualitative examples generated by our method at 400K training iterations and quantitative analysis of FID-50K training efficiency on ImageNet dataset.

cies, we employ these learnable tokens within a cross-attention layer to estimate the infused tokens directly from internal diffusion features. By reintegrating these estimated signals into the model, this mechanism facilitates a progressive internalization of the infused semantic knowledge.

This design allows the diffusion model to incorporate semantic information in a flexible and task-aware manner, rather than being constrained by direct aligning with fixed semantic representations. As a result, the model can effectively utilize semantic guidance during training while maintaining its own expressive capacity for generation. We conduct extensive experiments on ImageNet and demonstrate that our method consistently outperforms REPA and its variants such as iREPA across different model scales and resolutions.

Our contributions are summarized as follows:

- By leveraging learnable latent tokens and a cross-attention layer, our framework enables the denoising objective to autonomously extract specific facets and granularities of semantic information that are more beneficial for generative modeling.
- We introduce a stage-wise training strategy that enables the model to estimate and internalize the selected semantic signals, allowing the diffusion backbone to rely on its internal representations to guide generation.
- We conduct comprehensive experiments across various model scales and resolutions. The results demonstrate that our approach consistently accelerates training and improves generation quality, outperforming the alignment-based baselines including REPA and iREPA.

2 Related Work

Semantic Emergence in Generation. An increasing number of research has explored and exploited the emergent semantic properties within diffusion models. For instance, Xiang et al. (Xiang et al., 2023) and Mukhopadhyay et al. (Mukhopadhyay et al., 2023) demonstrated that intermediate representations possess strong discriminative capabilities, despite being trained via a purely generative objective. Leveraging these insights, Repfusion (Xingyi Yang, 2023) and related frameworks employ knowledge distillation to repurpose these self-emergent features for downstream discriminative tasks, such as classification and segmentation. These findings underscore the latent semantic potential within diffusion backbones.

Representation Alignment in Generation. While semantics do emerge naturally, another line of study (Yu et al., 2025; Jiang et al., 2025) reveals that this high-level semantic emergence often

[†]Equal Contribution
[*]Corresponding Author

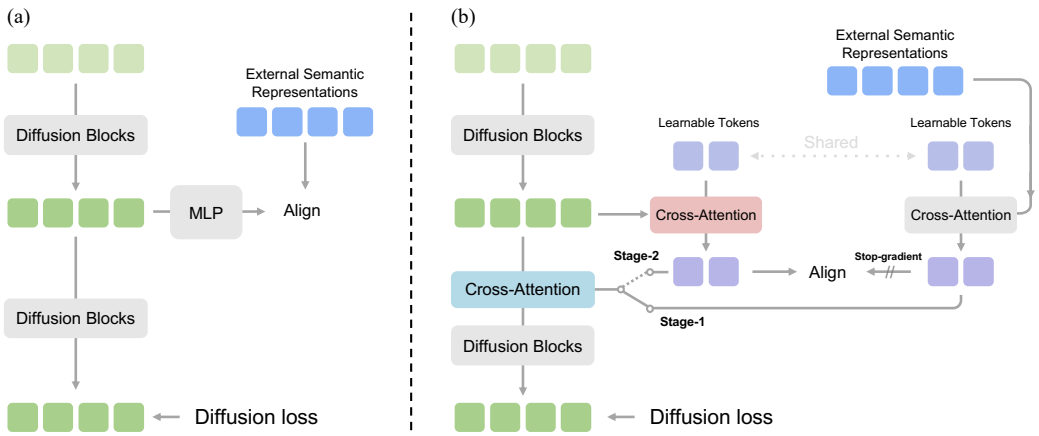

Figure 2: An overview of the standard REPA (a) and our approach (b).

lags significantly behind the denoising progress. To bridge this gap, REPA (Yu et al., 2025) introduces an auxiliary objective to align intermediate diffusion features with pre-trained semantic representations (e.g., DINOv2 (Oquab et al., 2023)). While this alignment accelerates convergence, recent studies (Singh et al., 2025; Wang et al., 2025b) have identified the misalignment phenomenon: higher semantic accuracy, as measured by linear probing, does not consistently yield improvements in generation fidelity. Consequently, these works introduce auxiliary modules or early-stopping heuristics to mitigate this discrepancy.

Building upon REPA, several methods have introduced additional regularization (Wu et al., 2025) or architectural refinements (Wang et al., 2025a). More recently, Self-flow (Chefer et al., 2026) incorporated self-supervised learning directly into the alignment framework, eliminating the need for external encoders. Conversely, RAE (Zheng et al., 2025) takes a more radical approach by employing semantic encoders as the latent encoder itself, thereby performing denoising directly within a semantically rich latent space.

3 Methodology

We propose a simple yet effective training framework to accelerate diffusion training convergence. The key idea is to first allow the model to selectively extract useful semantic information from pre-trained external semantic representations under the denoising objective, and then enable the diffusion model to internally estimate and internalize such information.

3.1 Representation Alignment Preliminaries

Our goal is to improve the convergence efficiency of diffusion models from the perspective of representation alignment. Following prior work (Yu et al., 2025; Singh et al., 2025), we consider latent diffusion transformers such as SiTs (Ma et al., 2024), as well as pixel diffusion models such as JiTs (Li & He, 2025). As a running example, we adopt the standard velocity prediction objective under linear interpolation:

$$\mathcal{L}_{\text{velocity}} = \mathbb{E}_{\mathbf{z},\epsilon,t}\left[\|v_\theta(\mathbf{z}_t, t) - (\epsilon - \mathbf{z})\|^2\right], \tag{1}$$

where $\mathbf{z} = \mathcal{E}(\mathbf{x})$ denotes the image VAE latent $\mathbf{x}$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise, and $\mathbf{z}_t$ is defined as:

$$\mathbf{z}_t = (1-t)\mathbf{z} + t\epsilon, \quad t \in [0, 1]. \tag{2}$$

Beyond standard diffusion training, as shown in Fig. 2 (a), representation alignment employs an external semantic encoder to provide additional semantic supervision. Given a clean image $\mathbf{x}$, let $\mathbf{h}_t \in \mathbb{R}^{N \times D_1}$ denote the intermediate diffusion features corresponding to the noised latent $\mathbf{z}_t$, where N is the number of image latent tokens and D_1 is the feature dimension. Let $\mathbf{y} \in \mathbb{R}^{N \times D_2}$ denote

semantic representations extracted from $\mathbf{x}$ by a pre-trained semantic encoder. The alignment objective is formulated as:

$$\mathcal{L}_{\text{repa}} = -\mathbb{E}_{\mathbf{z},\epsilon,t}\left[\frac{1}{N}\sum_{n=1}^{N}\text{sim}\big(\mathbf{y}^{[n]},\, f_\phi(\mathbf{h}_t)^{[n]}\big)\right], \tag{3}$$

where $f_\phi(\cdot)$ is a projection MLP that maps diffusion features into the semantic space of $\mathbf{y}$, and $\text{sim}(\cdot,\cdot)$ denotes a similarity function such as cosine similarity. The overall objective is formed as:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{velocity}} + \alpha\mathcal{L}_{\text{repa}}, \tag{4}$$

where α balances the denoising and alignment terms.

3.2 Denoising-Guided Semantic Selection

Instead of directly enforcing rigid alignment to fixed semantic targets, we propose a more flexible mechanism that allows the diffusion model to *select* the semantic information that is more beneficial for generative training. To this end, we introduce a set of M learnable tokens

$$\mathbf{e} = \{e_i\}_{i=1}^{M} \in \mathbb{R}^{M\times D_2}, \quad M \le N, \tag{5}$$

which interact with the external semantic representations $\mathbf{y}$ through a cross-attention module $g_\psi(\cdot,\cdot)$. The semantic infused tokens are obtained as:

$$\tilde{\mathbf{e}} = g_\psi(\mathbf{e}, \mathbf{y}). \tag{6}$$

These semantic infused tokens are then injected into the diffusion model through another cross-attention module $h_\omega(\cdot,\cdot)$:

$$\tilde{\mathbf{h}}_t = h_\omega(\mathbf{h}_t, \tilde{\mathbf{e}}). \tag{7}$$

The resulting features $\tilde{\mathbf{h}}_t$ are passed through the remaining diffusion blocks, yielding the objective:

$$\mathcal{L}^{(1)}_{\text{velocity}} = \mathbb{E}_{\mathbf{z},\epsilon,t}\left[\left\| v_{\theta,\psi,\omega}(\mathbf{z}_t, t; \tilde{\mathbf{h}}_t) - (\epsilon - \mathbf{z})\right\|^2\right]. \tag{8}$$

Unlike rigid alignment, this formulation allows the model to determine both *what* semantic information to extract and *to what extent* such information should be incorporated for denoising. As a result, irrelevant or redundant semantics can be suppressed, leading to improved optimization efficiency and generation quality.

3.3 Stage-wise Semantic Internalization

The semantic selection process described above is only available during training, since the clean target image is unknown at inference time. To bridge this training–inference gap, we introduce a stage-wise training strategy that enables the diffusion model to internalize the selected semantic information and remove the dependency on external semantic encoders at test time.

Specifically, we introduce another cross-attention module $f_\phi(\cdot,\cdot)$, which predicts the semantic infused tokens from the intermediate diffusion features $\mathbf{h}_t$ and the shared learnable queries $\mathbf{e}$:

$$\hat{\mathbf{e}} = f_\phi(\mathbf{e}, \mathbf{h}_t). \tag{9}$$

Note that we use a notation f_ϕ the same as that in REPA introduced in Section 3.1 for simplicity. We then align the self-produced tokens $\hat{\mathbf{e}}$ with the infused semantic tokens $\tilde{\mathbf{e}}$ using

$$\mathcal{L}_{\text{sim}} = -\mathbb{E}_{\mathbf{z},\epsilon,t}\left[\frac{1}{M}\sum_{m=1}^{M}\text{sim}\big(\text{sg}[\tilde{\mathbf{e}}]^{[m]},\, \hat{\mathbf{e}}^{[m]}\big)\right], \tag{10}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator, which stabilizes optimization and prevents degenerate solutions. The predicted tokens are injected back into the diffusion model through the same cross-attention module $h_\omega(\cdot,\cdot)$:

$$\hat{\mathbf{h}}_t = h_\omega(\mathbf{h}_t, \hat{\mathbf{e}}). \tag{11}$$

This gives the second denoising objective:

$$\mathcal{L}^{(2)}_{\text{velocity}} = \mathbb{E}_{\mathbf{z},\epsilon,t}\left[\left\| v_{\theta,\phi,\omega}(\mathbf{z}_t, t; \hat{\mathbf{h}}_t) - (\epsilon - \mathbf{z}) \right\|^2\right]. \tag{12}$$

As shown in Fig. 2 (b), we divide training into two stages:

$$\mathcal{L}_{\text{overall}} = \begin{cases} \mathcal{L}^{(1)}_{\text{velocity}} + \alpha\mathcal{L}_{\text{sim}}, & s < r, \\ \mathcal{L}^{(2)}_{\text{velocity}} + \alpha\mathcal{L}^{*}_{\text{sim}}, & s \geq r, \end{cases} \tag{13}$$

where s denotes the current training step and r is the stage transition step. In the second stage, $\mathcal{L}^{*}_{\text{sim}}$ indicates that the parameters of g_ψ and the learnable tokens $\mathbf{e}$ are frozen, such that the selected semantic target remains fixed.

In the first stage, the model learns to extract semantic information from the external encoder that is useful for denoising, while simultaneously learning to predict such selected information from its own internal features. In the second stage, the model no longer relies on externally selected semantics for feature injection, and instead uses its self-predicted semantic tokens to guide denoising. This progressively encourages the diffusion model to internalize the selected semantics into its own representation space.

During inference, the external selection module $g_\psi(\cdot,\cdot)$ is removed entirely. The model instead relies on $f_\phi(\cdot,\cdot)$ to estimate semantic tokens from its internal features and inject them back through $h_\omega(\cdot,\cdot)$. In this way, our method eliminates the dependency on external semantic encoders at test time while preserving the training benefits of semantic guidance.

4 Experiments

4.1 Experimental Setup

To ensure a fair evaluation, our experimental setup strictly adheres to the configurations established by SiT (Ma et al., 2024) and REPA (Yu et al., 2025). We conduct our experiments on the ImageNet dataset (Deng et al., 2009). All images are resized and center-cropped to a 256×256 or 512×512 resolution, following the data preprocessing protocols defined in ADM (Dhariwal & Nichol, 2021). We employ the pre-trained Stable Diffusion VAE (Rombach et al., 2021) to encode each image into a compressed latent vector $\mathbf{z} \in \mathbb{R}^{32\times32\times4}$. We evaluate our method across three model scales—B/2, L/2, and XL/2—as introduced in the SiT frameworks. All configurations utilize a patch size of 2. For optimization, we maintain a consistent batch size of 256 across all training stages to facilitate a direct and fair comparison with baseline SiT and REPA models. Unless otherwise specified, all hyperparameters and optimization schedules remain unchanged from the reference implementations. Following SiT (Ma et al., 2024) and REPA (Yu et al., 2025), we consistently employ the SDE Euler-Maruyama sampler. By default, we set the number of function evaluations (NFE) to 250 for both 256×256 and 512×512 resolution. For experiments involving JiT, we follow the original work and utilize 50 sampling steps.

Evaluation metrics. We report the Fréchet Inception Distance (FID; (Heusel et al., 2017)), sFID (Nash et al., 2021), Inception Score (IS; (Salimans et al., 2016)), and Precision (Pre.) and Recall (Rec.) (Kynkäänniemi et al., 2019) using 50,000 samples.

Baselines. We compare against three groups of baselines: i) vanilla diffusion transformers without representation alignment, ii) representation alignment methods, including REPA and iREPA where available, and iii) stronger variants that augment REPA with additional cross-attention or MLP fusion modules. These baselines are particularly important since they help disentangle whether the gains come merely from adding extra parameters, rather than from our module and training designs.

4.2 Quantitative Results

Results on ImageNet 256×256 benchmark. Table 1 summarizes the primary results on the ImageNet 256×256 benchmark. Our proposed method consistently outperforms both the vanilla SiT baseline and alignment-based counterparts across all model configurations.

For the SiT-B/2 scale, vanilla training achieves an FID of 33.0. While REPA and iREPA lower this to 24.4 and 21.4, respectively, our method further reduces the FID to 17.7. This represents a substantial improvement over REPA by 6.7 points and iREPA by 3.7 points. This performance advantage persists as we scale up the architecture. On SiT-L/2, our approach achieves an FID of 8.2, surpassing REPA (9.7) and iREPA (9.4). Similarly, on SiT-XL/2, our method reaches an FID of 7.1, outperforming both REPA (7.9) and iREPA (7.5).

These results demonstrate that the efficacy of our approach is agnostic to model scale; the proposed training mechanism provides a robust performance boost for diffusion transformers ranging from small to large. Most notably, our SiT-B/2 model achieves an FID of 17.7, which remarkably approaches the performance of the vanilla SiT-XL/2 at the same 400K training milestone (17.7 vs. 17.2). This indicates that our proposed method significantly compensates for limited model capacity, enabling a base-scale architecture to rival the early-stage performance of a significantly larger model.

Table 1: FID comparisons with vanilla SiTs on ImageNet 256×256. We do not use classifier-free guidance(CFG). ↓ denotes that lower values are better. Iter. indicates the training iteration.

Model	#Params	Iter.	FID↓	sFID ↓	IS ↑	Pre ↑	Rec. ↑
SiT-B/2	130M	400K	33.0	6.6	36.6	0.51	0.63
+ REPA	130M	400K	24.4	6.6	65.3	0.59	0.64
+ iREPA	130M	400K	21.4	-	-	-	-
+ Ours	135M	400K	**17.7**	6.5	**79.7**	**0.63**	0.64
+ REG	132M	400K	15.2	6.7	97.2	0.63	0.62
+ REG + Ours	137M	400K	**13.8**	**6.6**	**101.5**	**0.65**	**0.64**
SiT-L/2	458M	400K	18.8	5.4	20.3	0.62	0.62
+ REPA	458M	400K	9.7	5.3	113.0	0.69	0.65
+ iREPA	458M	400K	9.4	-	-	-	-
+ Ours	464M	400K	**8.2**	5.3	**122.5**	**0.70**	0.65
+ REG	677M	400K	4.6	5.1	170.5	0.74	0.64
+ REG + Ours	684M	400K	**4.2**	5.1	**170.6**	0.73	**0.64**
SiT-XL/2	675M	400K	17.2	5.2	76.8	0.64	0.63
SiT-XL/2	675M	7M	8.3	-	-	-	-
+ REPA	675M	400K	7.9	5.1	125.7	0.70	0.65
+ iREPA	675M	400K	7.5	-	-	-	-
+ Ours	682M	400K	**7.1**	**5.0**	**131.4**	**0.71**	0.65
+ REG	677M	400K	3.4	4.7	**186.2**	0.75	0.64
+ REG + Ours	684M	400K	**3.2**	4.7	185.6	0.75	**0.65**
RAE	676M	400K	4.3	-	-	-	-

Results on ImageNet 512×512 **benchmark.** We further investigate whether the observed performance gains generalize to higher-resolution synthesis. As summarized in Table 2, our method consistently outperforms the REPA baseline on the ImageNet 512×512 benchmark across all scales. Specifically, for the SiT-B/2 configuration, our approach reduces the FID from 31.7 to 26.1. Similar improvements are observed in larger models: the FID drops from 14.1 to 12.8 for SiT-L/2, and from 11.6 to 10.5 for SiT-XL/2. These results demonstrate that the efficacy of our proposed method is not confined to a specific resolution; rather, it remains highly robust as the generation task becomes increasingly complex.

Table 2: FID on ImageNet 512×512.

Model	#Params	Iter.	FID↓
SiT-B/2	130M	400K	42.6
+ REPA	130M	400K	25.2
+ Ours	135M	400K	**19.3**
SiT-L/2	458M	400K	22.4
+ REPA	458M	400K	9.6
+ Ours	464M	400K	**8.2**
SiT-XL/2	675M	400K	19.9
+ REPA	675M	400K	7.6
+ Ours	682M	400K	**7.0**

Generalization to JiTs. To evaluate the architectural generality of our approach, we extend its application to pixel diffusion models, JiTs, at a 256×256 resolution. As shown in Table 3, our method consistently outperforms both the vanilla JiT and the REPA-enhanced version. Specifically, for JiT-B/16, our method achieves an FID of 33.4, marking a significant improvement over REPA (37.2) and a substantial leap from the vanilla JiT baseline (46.1). On the larger JiT-L/16 configuration, our approach reaches an FID of 15.7, further surpassing REPA's 16.8. These results confirm that the performance gains are not confined to the SiT framework.

Table 3: FID comparisons with vanilla JiTs on ImageNet 256×256.

Model	#Params	Iter.	FID↓
JiT-B/16	131M	400K	46.1
+ REPA	131M	400K	37.2
+ Ours	136M	400K	**33.4**
JiT-L/16	459M	400K	26.4
+ REPA	459M	400K	16.8
+ Ours	465M	400K	**16.1**

Effect of Different Pretrained Feature Encoders. Figure 3 illustrates the comparative performance between REPA and our method across a diverse suite of pretrained feature encoders, including

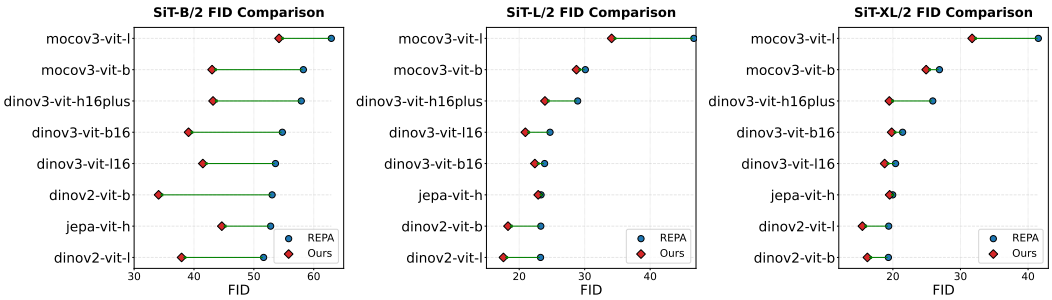

Figure 3: FID comparison between REPA and our method using different pretrained visual encoders on ImageNet 256×256. Results are reported for SiT-B/2, SiT-L/2, and SiT-XL/2. Across a diverse set of pretrained representations, our method consistently achieves lower FID, demonstrating that the proposed infusion mechanism is robust to the choice of semantic encoder.

variants of DINOv2, DINOv3, JEPA, and MoCo. Across all tested architectures—SiT-B/2, SiT-L/2, and SiT-XL/2—our method consistently achieves superior FID scores compared to REPA for nearly every encoder. These findings demonstrate that our proposed framework is not coupled to any specific semantic encoder; rather, it functions as a versatile mechanism for effectively integrating external visual representations into the diffusion training process. While the absolute performance fluctuates depending on the choice of encoder—indicating that the inherent quality of the teacher representation remains a decisive factor—our method significantly mitigates the sensitivity of the student model to the injection process. In essence, while the choice of pretrained teacher still matters, our approach provides a more robust and efficient pathway for the student to inherit and leverage those teacher-derived semantics.

Effect of Alignment and Fusion Module, and Stage-wise Training. A natural question is whether our gains simply come from introducing extra parameters or a more expressive fusion module. To answer this, we compare with several stronger baseline variants on ImageNet 256×256, including MLP-based alignment and fusion, and Cross-Attention (CA) based alignment and fusion.

Table 4: Module design ablation on ImageNet 256×256. We do not use classifier-free guidance(CFG). ↓ denotes that lower values are better. #Params. indicates the additional number of parameters required for inference. As $g(\cdot)$ is discarded during inference, we do not take into account its parameters. Iter. indicates the training iteration.

$f(\cdot)$	$g(\cdot)$	$h(\cdot)$	#Params.	Stage-wise	FID↓
MLP	-	-	-	×	23.0
MLP	Linear	-	8M	×	20.8
MLP	MLP	-	16M	×	18.7
CA	-	-	2M	×	19.8
CA	CA	-	5M	×	19.9
MLP	MLP	MLP	16M	✓	20.0
CA	CA	CA	5M	✓	17.7

As shown in Table 4, the first row using MLP as $f(\cdot)$ is identical to REPA. Based on this, we add stronger fusion modules using a linear layer and MLP as $g(\cdot)$. For example, REPA with linear fusion reduces FID from 23.0 to 20.8, and deeper MLP fusion further reaches 18.7. Instead of a linear layer or MLP, we use CA to serve as $f(\cdot)$ and $g(\cdot)$, but find that this setting performs worse than linear or MLP based settings. However, by integrating our core idea to let the model learn what matters for generation in stage-wise training with CA as $h(\cdot)$, our method achieves the best result of 17.7, despite a relatively modest parameter increase. This significantly demonstrates the effectiveness of our approach.

Ablation studies on r, α, and M. As shown in Table 5, we observe that: (i) Increasing r to 75K consistently improves FID, suggesting that a longer initial phase stabilizes representations. (ii) A higher weight ($\alpha = 1.75$) significantly boosts performance, confirming the importance of semantic alignment. (iii) $M = 32$ is the optimal sweet spot, whereas $M = 64$ leads to a slight performance drop due to increased complexity.

Table 5: Ablation studies on r, α, and M.

r	FID	α	FID	M	FID
55K	18.4	0.75	19.0	16	18.1
65K	18.1	1.25	18.4	32	17.7
75K	17.7	1.75	17.7	64	18.3

4.3 Analyses

Does better semantic accuracy lead to better generation quality? To better understand the learned representations, we perform linear probing on aligned diffusion features and report top-1 accuracy on the ImageNet-1K validation set in Table 6. Interestingly, improved generation quality does not consistently correlate with stronger linear separability. For SiT-B/2, our method achieves better FID while yielding lower linear probing accuracy than REPA (59.83% vs. 61.86%). For SiT-L/2, the two methods are comparable, with our method slightly underperforming REPA (69.01% vs. 69.39%). For SiT-XL/2, our method slightly surpasses REPA (70.75% vs. 70.06%).

Table 6: Linear probing accuracy (%) of aligned diffusion features on ImageNet-1K validation set. All results are based on models trained with 400K iterations.

	SiT-B/2	SiT-L/2	SiT-XL/2
REPA	61.86%	69.39%	70.06%
Ours	59.83%	69.01%	70.75%

These results suggest that better generation does not necessarily require diffusion features to be more directly aligned with discriminative semantics. This finding supports our central hypothesis: enforcing strict imitation of pretrained representations may improve alignment metrics, but can also impose constraints that are suboptimal for generative modeling. In contrast, our approach enables the model to retain only the information that is most beneficial for the denoising process.

What really drives the improved generation quality? We further evaluate the Local-vs-Distance Score (LDS), which measures the gap between local feature similarity and long-range similarity to quantify spatial coherence. As shown in Fig. 4, our method consistently achieves higher LDS scores than both SiT and REPA across model scales (L and XL), and even surpasses DINO features at intermediate timesteps. Since higher LDS generally correlates with better structural coherence, these results suggest that our method learns representations that are more beneficial for generation.

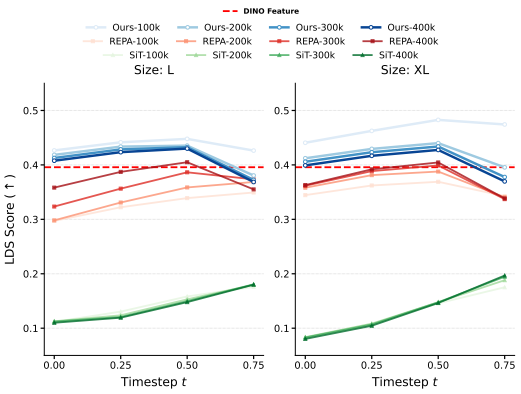

Figure 4: LDS across diffusion timesteps and training iterations. Our method consistently achieves higher LDS than SiT and REPA, indicating stronger local semantic coherence. While REPA shows increasing LDS (closer alignment to DINO), our method exhibits a decreasing trend, suggesting selective deviation from pretrained features. The dashed line denotes DINO.

Notably, our method exhibits distinct training dynamics. At early stages, LDS increases and exceeds DINO, indicating that the model selectively extracts and amplifies the most useful structural information from pretrained features. This aligns with our core idea of *aligning what matters more for generation*. As training progresses, however, LDS gradually decreases, in contrast to REPA which continues to increase and closely match DINO. This suggests while structural coherence is important, excessive alignment to pretrained semantics may become restrictive.

We hypothesize that in later training stages, the model intentionally suppresses overly strong semantic alignment to recover fine-grained details necessary for high-quality generation. In this sense, LDS should not be maximized indefinitely: an optimal representation balances structural consistency and detailed generation. Overall, our method achieves this balance by first leveraging and then selectively deviating from pretrained priors.

4.4 Qualitative Results

Visual comparison across training iterations (without CFG). Fig. 5 shows generation results at different training stages (100K, 200K, and 400K iterations) without classifier-free guidance. At early stages, both methods capture coarse structures but lack fine details. As training progresses, REPA gradually improves global appearance but remains relatively smooth. In contrast, our method produces progressively sharper structures and richer details, with clearer object boundaries and more realistic textures at later iterations.

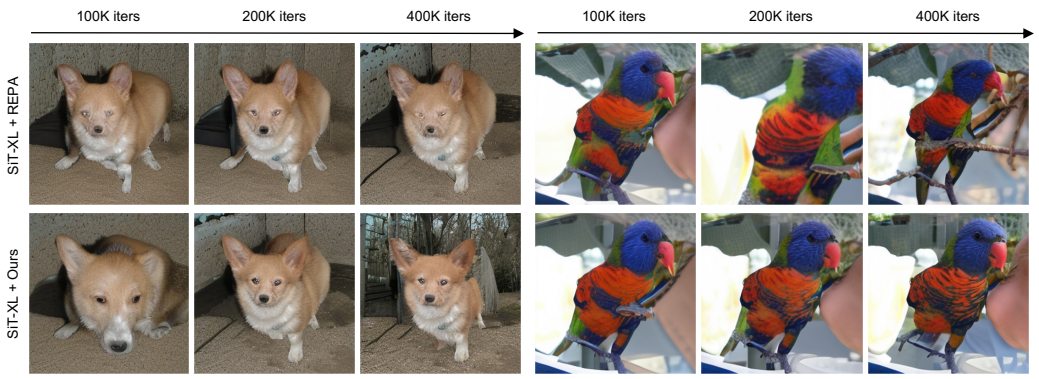

Figure 5: Generation results across training iterations (100K, 200K, 400K) without classifier-free guidance. Our method progressively improves structural sharpness and detail compared to REPA.

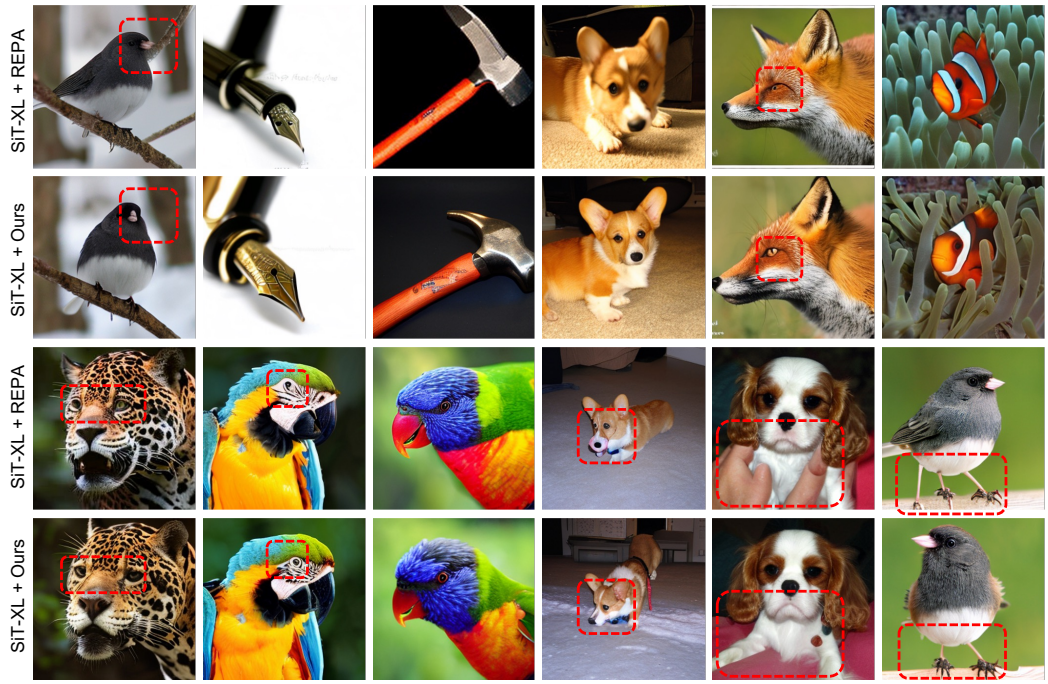

Figure 6: Visual comparison at the final checkpoint with classifier-free guidance. Our method produces sharper and more detailed structures than REPA, as highlighted in the zoomed-in regions.

Visual comparison at final checkpoint (with CFG). Fig. 6 presents results at the final checkpoint with classifier-free guidance. Both methods generate visually plausible images, but clear differences remain in fine-grained details. REPA tends to produce smoother and less precise structures in localized regions, while our method yields sharper patterns and more accurate geometry, particularly in challenging areas such as eyes, edges, and small object parts. These differences are highlighted in the zoomed-in regions.

5 Conclusion

We presented a simple yet effective training framework to enhance diffusion model convergence through denoising-guided semantic selection and internalization. By replacing rigid representation alignment with a flexible selection mechanism, we enable the model to prioritize task-relevant semantic signals that directly benefit generative modeling. The introduced stage-wise strategy further ensures that the model can bridge the training-inference gap by internalizing external knowledge. Empirical results on ImageNet confirm that our method achieves superior generation performance and faster convergence compared to state-of-the-art alignment methods. This work underscores the importance of task-awareness in representation alignment and offers a robust, scalable solution for efficient large-scale diffusion training.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting, language revision, and presentation-oriented code, including figure formatting. They were not used to generate datasets or experimental results. All AI-assisted content was critically reviewed and independently verified by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript.

ETHICS STATEMENT

This work adhered to the relevant Code of Ethics. The study involved no human participants or private or sensitive personal data. All evaluated models and datasets were publicly available, and their sources and evaluation settings are described in Section 4. We considered potential dataset biases and limitations when interpreting the results and conducted all analyses solely for academic research.

REPRODUCIBILITY STATEMENT

Section 4 provides detailed descriptions of the model architecture, training objectives, and evaluation protocols. Additional implementation details and hyperparameter settings are included in AppendixÃ, while the supplementary materials contain further results.

REFERENCES

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

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

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

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

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

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

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.

Soumik Mukhopadhyay, Matthew Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. Diffusion models beat gans on image classification, 2023.

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

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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

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.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025a.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025b.

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

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Xinchao Wang Xingyi Yang. Diffusion model as representation learner. *International Conference on Computer Vision (ICCV)*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders, 2025.

A Appendix

A.1 More Implementation Details

Learnable Tokens. These tokens are initialized from a zero-centered normal distribution with a standard deviation of 0.02. To provide the model with sequential information, we augment these tokens with fixed 1D sinusoidal positional embeddings. Specifically, the positional embeddings are pre-computed using the sine and cosine functions of different frequencies and are added to the infusion tokens. These embeddings are kept frozen during the entire training process.

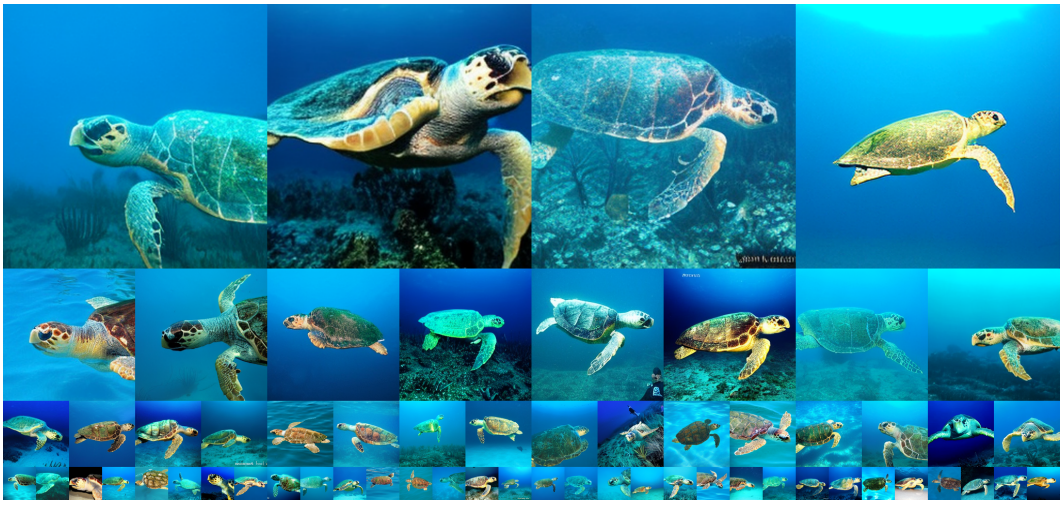

Figure 7: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “loggerhead sea turtle” (33).

A.2 More Visualization Results

In this section, we provide additional qualitative results to further demonstrate the generative capabilities of our proposed method. Figures 7 to 13 display uncurated samples across various ImageNet classes, including animals, plants, and man-made objects. All samples are generated using SiT-XL/2 integrated with our approach, employing classifier-free guidance with a scale of $w = 4.0$. These results highlight the consistency and high fidelity of our model across diverse semantic categories.



Figure 8: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “macaw” (88).



Figure 9: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “golden retriever” (207).



Figure 10: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “boathouse” (449).



Figure 11: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “mushroom” (947).

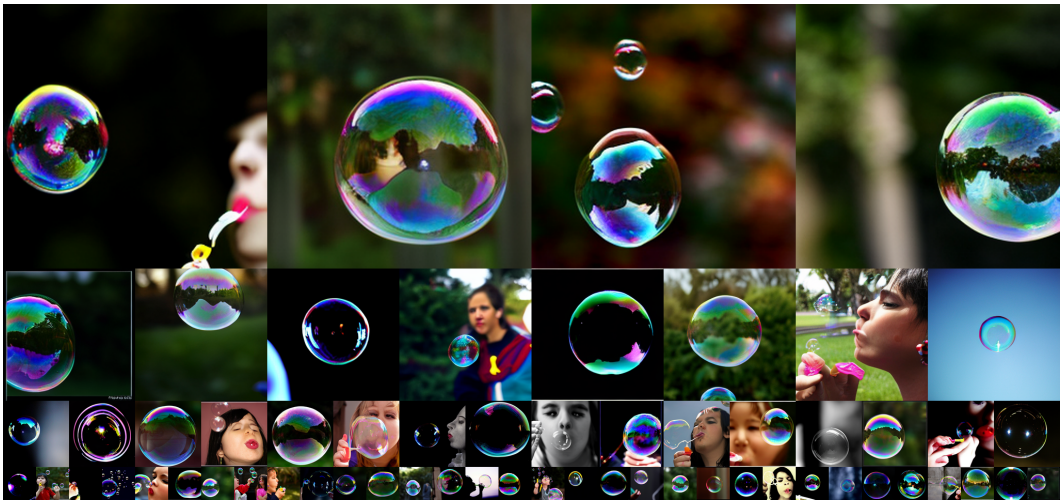

Figure 12: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “bubble” (971).

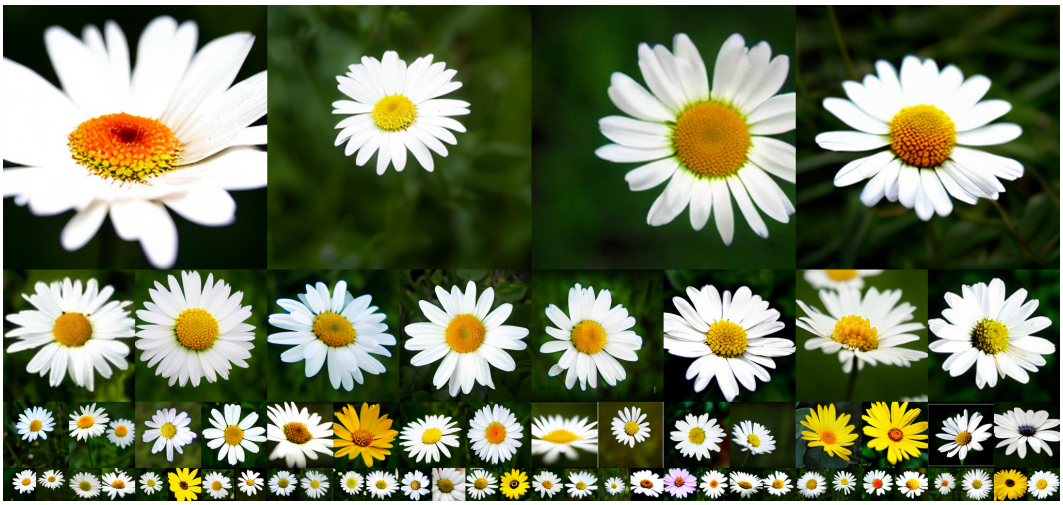

Figure 13: Uncurated generation results of SiT-XL/2 + Ours. We use classifier-free guidance with w = 4.0. Class label = “daisy” (985).