REGISTERS MATTER FOR PIXEL-SPACE DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Transformers (ViTs) are known to exhibit high-norm patch-token outliers that degrade feature map quality, a problem effectively mitigated by *register tokens*. As diffusion models increasingly adopt transformer architectures and move toward pixel-space training, they become closer in form to ViTs, raising the question of whether register tokens are also useful for Diffusion Transformers (DiTs). In this work, we show that DiTs differ from ViTs in a key respect: *they do not exhibit patch-token outliers but still benefit from registers.* Interestingly, registers are more effective in pixel-space DiTs than in latent-space DiTs. By analyzing intermediate representations, we find that register tokens produce cleaner feature maps at high noise levels, which may contribute to their effectiveness in pixel-space generation. We further observe that recent pixel-space DiT architectures implicitly incorporate register-like mechanisms, which may partially account for their strong empirical performance. Motivated by these observations, we propose *Register Guidance*, a technique that amplifies the contribution of register tokens responsible for improving visual structure and coherence.

1 INTRODUCTION

Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Touvron et al., 2021) have become a dominant architecture for visual representation learning by modeling images as sequences of patch tokens processed via self-attention (Vaswani et al., 2017). Recent advances in self-supervised learning (SSL) (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) demonstrate that ViTs trained on unlabeled data can learn semantically meaningful representations, enabling object- and part-level understanding useful for downstream tasks such as unsupervised segmentation and detection (Siméoni et al., 2021; Hamilton et al., 2022; Amir et al., 2021; Oquab et al., 2023; Wang et al., 2023).

Recent research has focused on understanding the emergence of high-norm tokens in ViTs, which are often associated with artifacts in attention maps (Darcet et al., 2023; Jiang et al., 2025; Lappe & Giese, 2025; Shi et al., 2026; Chen et al., 2025b). As these artifacts lead to less interpretable attention maps and weaker performance on dense prediction tasks, Darcet et al. (2023) proposes using additional *register tokens* to prevent patch tokens from being repurposed for global representations.

In parallel, diffusion models (DMs) (Ho et al., 2020; Song & Ermon, 2019) have widely adopted transformer-based architectures (Peebles & Xie, 2023; Ma et al., 2024), replacing convolutional backbones (Ronneberger et al., 2015; Dhariwal & Nichol, 2021). Recent work has also revisited training directly in pixel space (Li & He, 2025; Yu et al., 2025; Lu et al., 2026), as an alternative to latent diffusion models that rely on pretrained autoencoders (Rombach et al., 2022; Podell et al., 2023). This progress brings Diffusion Transformers (DiTs) closer to ViTs and motivates two questions: *(1) do DiTs inherit high-norm patch-token outliers similar to those observed in ViTs?* and *(2) can register tokens also be effective in these models?*

These questions are also related to broader studies of attention sinks and special-token behavior in transformers (Su et al., 2026), including generative models (Xiao et al., 2023; Gu et al., 2024; Rulli et al., 2025; Liu et al., 2025; Jamal et al., 2026). We defer a detailed discussion of this direction to App. A and focus here on image diffusion transformers, where the presence and role of register-like tokens remain underexplored.

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.

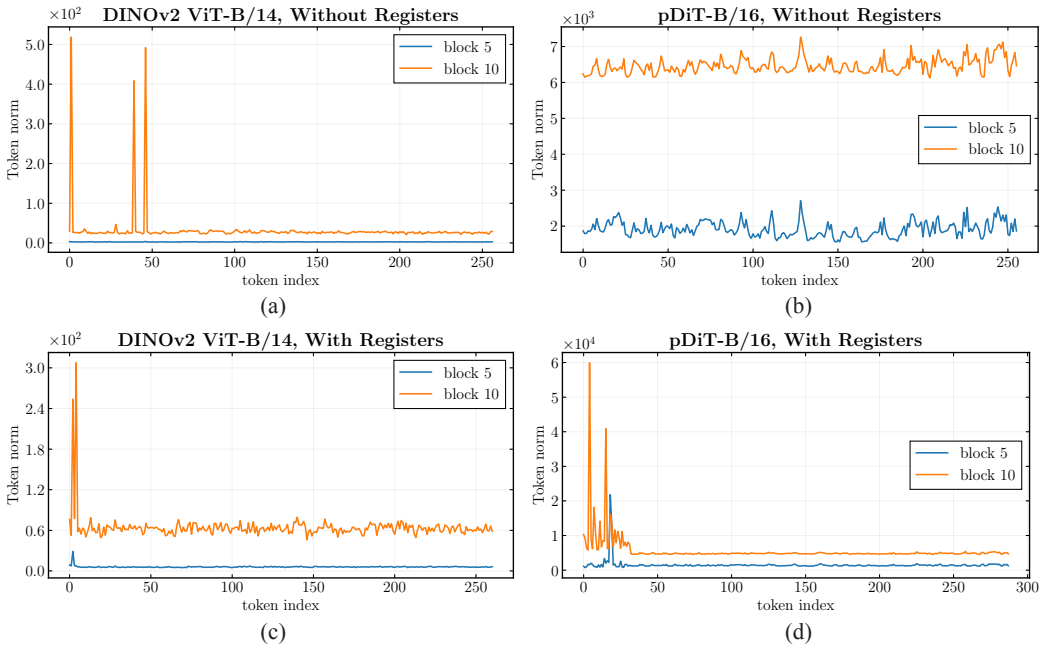


Figure 1: **Token-norm behavior in DINOv2 and pDiTs. Top: without registers; bottom: with registers.** (a) In DINOv2, anomalies are localized to a few image tokens with significantly higher norms than the rest. (b) In contrast, pDiTs show no patch-token norm outliers, suggesting that registers may be unnecessary in this case. (c) As expected, adding register tokens to DINOv2 shifts high-norm outliers into these tokens. (d) Interestingly, pDiTs develop high-norm outliers specifically in register tokens, despite exhibiting no such outliers in their absence.

Contributions. We find that, unlike ViTs, DiTs in both latent and pixel spaces do not exhibit noticeable high-norm outliers among patch tokens. Instead, patch-token norms remain relatively uniform, and attention maps lack the low-information artifacts commonly observed in ViTs. Interestingly, despite the absence of such outliers, adding register tokens to DiTs leads to the emergence of high-norm tokens within the registers themselves.

Importantly, we observe that the impact of register tokens differs across training spaces: pixel-space models benefit the most, whereas latent-space models show only moderate gains or little to no improvement.

Accordingly, we focus our study on pixel-space DiTs and find that registers benefit them through mechanisms distinct from those in ViTs. Our analysis shows that registers consistently reduce patch-token feature norms and produce smoother intermediate feature maps, especially at high noise levels. Moreover, register tokens specialize into distinct roles: some act as norm sinks, while others encode global semantic information.

These findings also provide a rationale for recent pixel-space DiT designs (Li & He, 2025; Lu et al., 2026), which introduce additional in-context class-conditioning and achieve substantial performance gains. Our analysis suggests that these gains may largely arise from register-like behavior rather than from additional class information. In particular, in-context tokens behave similarly to register tokens, with some encoding global semantic information and others acting as norm sinks. We observe a similar phenomenon in the text-to-image setting, where some tokens from the text sequence also become sinks.

From a practical standpoint, we exploit the observation that registers improve object structure and coherence while preserving the main content. This motivates *Register Guidance*, which treats the prediction of a model without registers as a weak model (Karras et al., 2024) and uses it as the negative direction. Combined with CFG (Ho & Salimans, 2022), RG consistently improves the performance of recent pixel-space DiTs.

	pDiT-B/16			pDiT-L/16			pDiT-H/16			PixelDiT-XL/16			
Epoch	w/o regs.	w/ regs.	IC	w/o regs.	w/ regs.	IC	w/o regs.	w/ regs.	IC	Epoch	w/o regs.	w/ regs.	IC
200	7.39	**5.30**	4.71	4.13	**3.17**	2.95	3.52	**2.69**	2.38	80	2.74	**2.24**	2.28
600	4.80	**3.80**	3.71	2.80	**2.47**	2.47	2.35	**2.02**	1.90	120	2.25	**2.12**	2.12

Table 1: **Generation quality (FID ↓) of pDiTs and PixelDiT (Yu et al., 2025) with and without register tokens.** We compare generation quality across model sizes and training epochs. Here, "w/o regs." and "w/ regs." denote training without and with register tokens, respectively. The shaded IC column reports performance with in-context class conditioning (Section 2.6).

2 Register Tokens for Image Diffusion Transformers

In this section, we analyze the role of register tokens in image DiTs and highlight key differences from their use in ViTs. As a representative ViT-based model, we use DINOv2 (Oquab et al., 2023).

For DMs, we primarily focus on pixel-space DiTs based on the standard architecture (Peebles & Xie, 2023) with widely used transformer improvements (Yao et al., 2025; Li & He, 2025); we refer to these models as pDiTs. We train pDiTs using flow matching (Lipman et al., 2023) on ImageNet (Deng et al., 2009) at resolution 256×256 with patch size 16. We also consider the more advanced pixel-space architecture PixelDiT (Yu et al., 2025), following the original setup.

We also analyze latent-space architectures, SiT (Ma et al., 2024) and RAE (Zheng et al., 2026), using their original backbones and training pipelines. For RAE, we use a DINOv2-based model (Oquab et al., 2023) with registers in the encoder. Quality is evaluated using FID (Heusel et al., 2017).

We study models of varying sizes, with and without register tokens. Registers are implemented as additional learnable tokens appended to the patch-token sequence following Darcet et al. (2023) and are not used in the training objective. Further details are provided in App. B.

2.1 Registers Benefit Diffusion Transformers Despite the Absence of Outliers

The motivation for introducing register tokens in ViTs is to mitigate outliers in feature maps. These outliers manifest as high-norm tokens, often localized in low-information regions, e.g., background.

We first investigate whether such outliers arise in DiTs without registers across different spaces by comparing token-wise feature norms across layers. As shown in Figure 1(top), pDiTs exhibit a nearly uniform distribution of patch-token norms, in contrast to DINOv2, where a few tokens attain significantly larger norms. As shown in Figures 12, 13, and 14, this behavior consistently holds across larger model variants and latent-space architectures.

We further support this observation by comparing the corresponding attention maps to those of ViTs. In particular, DiT attention maps remain free of anomalies in low-information regions, which in ViTs are typically associated with unusually high-norm tokens, as shown in Figure 7.

Based on these observations, DiTs would not be expected to benefit from register tokens, as the feature map artifacts that originally motivated their use are absent. However, contrary to this expectation, we observe the opposite.

First, in Figure 1(bottom), we compare token-wise feature norms for DINOv2 and pDiTs with register tokens. As expected, in DINOv2, registers primarily absorb pre-existing outliers from patch tokens. In contrast, pDiTs develop high-norm tokens within the registers, despite the absence of such outliers in models without registers. Figures 11 and 12 show that this effect consistently holds across different timesteps and model sizes. It also extends to latent-space models (Figures 13 and 14) and few-step models (Figures 15 and 16). In App. C.1, we show that outliers mainly form after the MLP layer.

Second, as shown in Table 1, introducing register tokens in pDiTs consistently improves generation quality across model sizes. We observe the same effect for PixelDiT and consider a higher resolution in Table 9. However, Table 2 shows that registers provide significantly smaller gains in latent-space models, a phenomenon we discuss in Section 2.4.

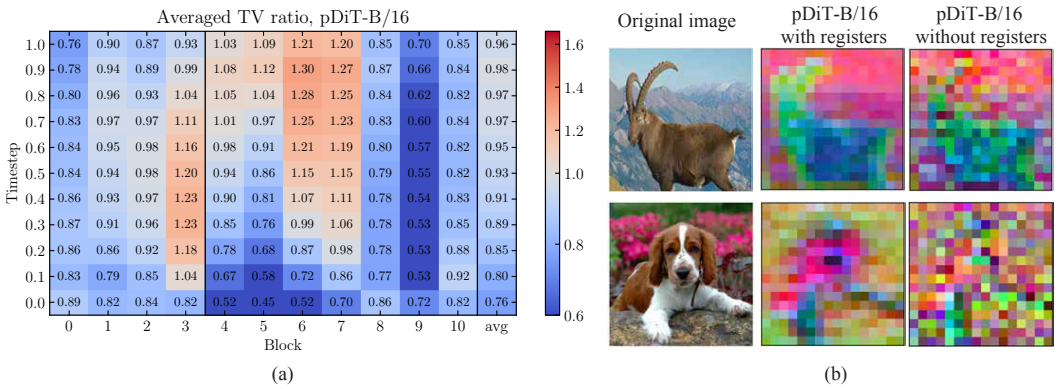


Figure 2: **Registers improve intermediate representations**. **(a)** Total Variation of intermediate features for models with and without register tokens. We report the ratio (with / without registers); lower values indicate smoother features. Registers improve feature smoothness most at high noise levels ($t \in [0, 0.2]$). **(b)** PCA visualizations of feature maps at t=0.1 qualitatively confirm this effect.

Based on these observations, pDiTs appear to benefit from outliers but lack a mechanism to accommodate them in the absence of special tokens. Introducing registers provides convenient slots for high-norm outliers.

In App. C.2, we find that patch tokens can occasionally become outliers in large text-to-image DiTs, but this behavior is inconsistent and vanishes in the later layers. This further suggests the need for dedicated outlier slots, since patches are constrained by the loss and cannot sustain high norms.

2.2 Register Tokens Lead to Cleaner Internal Feature Maps

The previous analysis shows that register tokens significantly improve pDiT performance, but their functional role remains unclear. In particular, their effect differs from that in ViTs, where registers primarily absorb pre-existing outliers. We therefore investigate how registers influence the internal representations of DMs, focusing on pixel-space models where their impact is strongest.

First, we find that register tokens influence all image tokens by consistently reducing their feature norms (Figure 8). Interestingly, DINOv2 register tokens do not exhibit this behavior for non-outlier patch tokens (Figure 23).

A possible explanation is that larger feature norms may reflect high local variability in hidden representations. In DiTs, such variability may arise from the requirement to carefully predict high-dimensional targets, causing all information, including global semantics and low-level signals, to propagate through patch tokens. Register tokens may absorb part of this information, reducing patch-token norms and allowing patch tokens to form smoother, more spatially structured representations.

To examine this, we consider Total Variation (TV) (Rudin et al., 1992; Aly & Dubois, 2005), which measures spatial smoothness by quantifying intensity differences between adjacent pixels. In our case, we use TV to quantify the spatial smoothness of intermediate transformer features. Specifically, we extract feature maps after each transformer block at different diffusion timesteps and compute their TV averaged over 1K images. We then evaluate TV value ratios (with registers / without registers).

We present the results in Figure 2(a) and observe that the ratios remain below 1 at lower timesteps (noisier images) and gradually approach 1 at higher timesteps (less noisy images). This suggests that register tokens improve feature smoothness primarily at high-noise levels ($t \in [0, 0.2]$). Figure 2(b) provides qualitative support for this observation: PCA visualizations of intermediate features at t=0.1 show that models with registers produce smoother and more coherent feature representations.

Note that high-noise levels are particularly important for flow-matching models in high-dimensional spaces (Esser et al., 2024; Black Forest Labs, 2025; Yun et al., 2025), as they shape the main image content. Therefore, the improved representations induced by registers at these stages provide a plausible explanation for the observed quality gains.

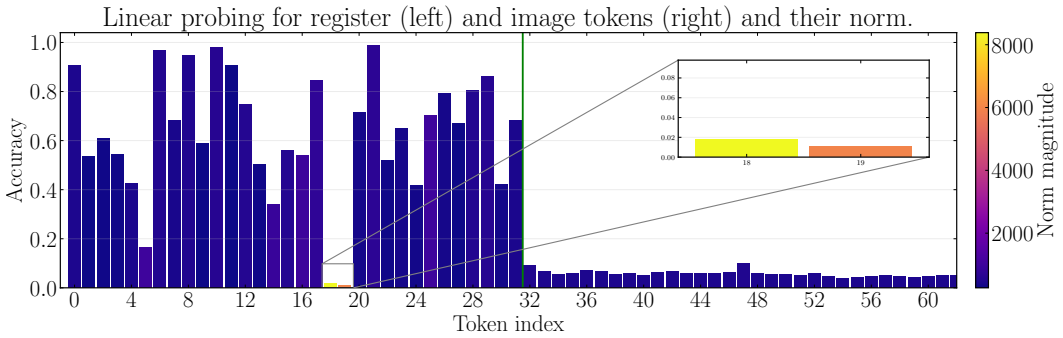


Figure 3: **Register tokens act as both global information carriers and norm sinks**. Linear probing reveals that low-norm register tokens encode meaningful global semantics and achieve strong classification accuracy, whereas low-accuracy registers exhibit much larger norms, suggesting that they primarily function as norm sinks that absorb magnitude from patch tokens.

2.3 Registers Do Both: Encode Global Information and Act as Norm Sinks

Next, we find that, beyond acting as norm sinks for patch tokens, register tokens can encode diverse global information about the input image. Specifically, we perform linear probing using register-token features extracted from an intermediate transformer block (the 5th out of 12 blocks).

Figure 3 shows that classification accuracy is highly diverse: some tokens achieve high accuracy (≈ 0.9), others moderate (≈ 0.4), and some very low (≈ 0.02). These results suggest the following: (a) tokens with the highest norms act as norm sinks, yielding the lowest accuracy; (b) tokens with moderate norms encode diverse information about the image beyond class-specific features.

To validate that non-sink tokens encode diverse information, we visualize attention maps of different registers (Figure 9a). We observe that registers attend to distinct semantic regions of the image. In the example, some tokens focus on background elements (e.g., jungle), while others attend to object parts (e.g., bird, branch, beak). In contrast, norm sinks lack meaningful semantic information (Figure 9b).

Discussion. The insights from Sections 2.2 and 2.3 may relate to recent representation-alignment methods (Yu et al., 2024; Singh et al., 2025), which improve DM convergence by aligning diffusion internal representations with vision encoders, e.g., DINOv2. Notably, iREPA (Singh et al., 2025) shows that generation quality benefits more from aligning spatial structure than global semantics. This aligns with our analysis: register tokens absorb global semantic information while improving the spatial coherence of patch tokens, suggesting that registers may play a regularizing role similar to REPA. We therefore explore whether register tokens complement REPA-like objectives in App. D.3.

2.4 Registers Are More Effective in Pixel Space

In Table 2, we compare the models operating in different spaces with and without registers. We find that register tokens show the largest improvements in pixel space, provide smaller gains in VAE space (Ma et al., 2024), and yield no improvement for RAE-based models (Zheng et al., 2026).

To isolate the effect of different diffusion backbones, we apply the pDiT backbone to RAE and VAE spaces as well. The results presented in Table 10 show the same trend. This indicates that the effect of register tokens is not related to corresponding architectural differences.

To better understand this effect, we analyze the smoothness of intermediate representations in DiTs without registers using TV. As shown in Figure 22, pixel-space models produce the least smooth (i.e., noisiest) intermediate features compared to latent DMs. In addition, we find that pixel-space models exhibit the highest feature norms for all patch tokens (Figure 21), further supporting this observation.

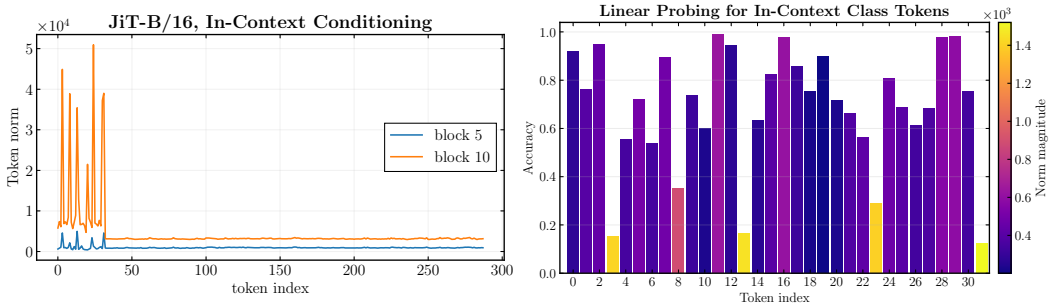
These findings suggest that training DMs in pixel space is inherently more challenging and requires stronger regularization. In contrast, latent-space encoders can compress imperceptible noise and fine-grained details. As a result, register tokens appear less critical for latent-space models, while serving as an effective mechanism for improving degraded representations in pixel-space DiTs.

	RAE-space DiT^{DH} backbone	VAE-space SiT backbone	Pixel-space pDiT backbone
	Base size		
w/ reg.	**6.34**	**9.40**	**5.30**
w/o reg.	6.53	10.40	7.39
	Large size		
w/ reg.	**1.76**	**2.38**	**2.69**
w/o reg.	1.78	2.53	3.52

Table 2: **Registers are more effective in pixel-space.** FID comparison with and without registers across training spaces and model sizes. Registers help pixel-space most, moderately help VAE-space, and provide no gains in RAE-space.

	Registers configuration			FID at Epoch		
	Size	Start	End	40	80	120
w/ reg.	32	4	11	37.7	**9.59**	**6.45**
	32	4	9	**36.9**	9.95	6.65
	32	0	11	59.7	19.3	11.9
	32	0	4	62.4	19.6	12.3
	16	4	11	40.4	10.2	6.80
	4	4	11	46.3	12.8	8.37
w/o reg.	–	–	–	60.6	18.4	11.1

Table 3: **Registers are effective in deeper layers.** Unlike DINOv2, pDiT-B/16 benefits from register tokens only when they are introduced after the first 4 layers. Increasing the number of registers further improves performance.



(a) (b)

Figure 4: **In-context class tokens act as registers.** (a) Certain tokens acquire disproportionately high feature norms, functioning as norm sinks. (b) Some tokens encode broad global information, rather than purely class-specific features as originally intended.

2.5 Registers Are Effective in Deeper Layers

Next, we analyze how many registers to use and where to insert them within the transformer backbone. We consider pDiT-B with 12 layers. Based on the results in Table 3, we observe two key differences compared to standard ViTs. We explore additional configurations in Table 18.

First, pDiTs benefit from enabling register tokens only in deeper blocks (4–11), whereas ViTs use them from the first layer (Darcet et al., 2023). For example, applying registers throughout all layers (0–11) provides performance comparable to the model without registers. A similar effect appears for early-layer registers (0–4). Moreover, the 4–9 setup suggests that the final layers contribute less.

Previously, we found that register tokens encode diverse semantic information about the input image and help form more structured representations. We hypothesize that their ineffectiveness in early layers stems from the lack of semantic structure at this stage. Specifically, early-layer registers primarily capture low-level or non-informative signals, providing weak conditioning to subsequent layers and ultimately degrading performance. To support this observation, Figure 20 compares linear probing results for models with register tokens introduced at layers 4 and 0. When registers are enabled from the beginning, many non-sink registers capture little semantic information after the 5th block, resulting in weak signals that are subsequently propagated to further layers.

Second, we observe that pDiTs require more registers than ViTs. While 4 registers are typically sufficient for ViTs (Darcet et al., 2023), pDiT-B achieves the best performance with 32 tokens.

2.6 Registers Are Implicitly Present in Existing Diffusion Transformers

Recent DiTs incorporate conditioning signals (e.g., text or class labels) by appending additional tokens to the sequence of image patches and processing them jointly through shared transformer layers. For example, JiT (Li & He, 2025), a pixel-space DiT, employs in-context conditioning by adding duplicated class embeddings to the input sequence, leading to notable improvements in generation quality. This raises the question of whether in-context tokens act as implicit registers.



Figure 5: **Registers improve image quality while preserving the overall image content.** We compare unconditional generations from the JiT-B/16 model without register tokens (top) and with register tokens (bottom), isolating the effect of registers from class conditioning. Using registers produces more coherent object structure and cleaner visual details across samples.

To address this, we train JiT-B/16 with in-context conditioning, using the same diffusion backbone and number of tokens as in the register setting. Thus, the only difference between JiT-B/16 and pDiT-B/16 lies in the additional token sequence (in-context vs register tokens).

Then, we measure the norms of in-context tokens (Figure 4a) and evaluate their representations using linear probing (Figure 4b). Interestingly, we observe similar behavior as for registers: (a) some tokens encode diverse global information rather than only class-specific features as originally intended, (b) while others act as norm sinks. This suggests that in-context tokens implicitly behave as registers.

Moreover, Table 1 compares models with registers, without registers, and with in-context conditioning. We find that most of the improvement over the baseline comes from the presence of pure register tokens rather than from the additional class information itself, helping explain the large quality gains from in-context conditioning. However, in-context conditioning further improves performance, suggesting that such tokens help the model form better initial representations.

We also analyze large-scale text-to-image models based on MM-DiT (Esser et al., 2024), where text tokens are appended to image tokens. Interestingly, we observe a similar phenomenon: some text tokens exhibit high-norm outlier behavior (Figure 24), suggesting they may act as implicit registers.

3 Register Guidance

In our method, register tokens are initialized from copied class embeddings, corresponding to in-context conditioning (Li & He, 2025; Lu et al., 2026), which gives slightly better performance. We nevertheless refer to them as registers because both variants play a similar role. Thus, our approach builds on JiT models (Li & He, 2025).

Intuition. Based on our analysis, registers improve generation quality, functioning as tokens that enhance internal representations and encode global information about the input image. So, we first investigate how generations specifically improve when registers are included.

To this end, we compare generations from a JiT-B/16 model with and without registers. Importantly, we perform this comparison in the unconditional setting to remove the influence of class information, which is encoded in the registers. This allows us to isolate the effect of registers themselves and examine how they improve generation quality beyond acting as an additional conditioning signal.

Figure 5 shows a clear effect of register tokens. With registers, generations have more coherent object structures and cleaner visual details, while the overall image content remains preserved.

These results motivate the following question: *can the influence of registers be enhanced to improve quality even further?* We draw inspiration from AutoGuidance (Karras et al., 2024), which improves sample quality by guiding a model with a weaker version of itself rather than with an unconditional model. Our setup follows a similar intuition: the same model without registers can be viewed as a weaker counterpart to the model with registers. This connection is possible because both variants

Snippet 1. RG, Inference

```
# z: current samples at t
# y: class label
# w_rg: register guidance scale

x_r  = net(z, t, y, regs=True)
x_nr = net(z, t, y, regs=False)

v_r  = (x_r  - z) / (1 - t)
v_nr = (x_nr - z) / (1 - t)

v_rg = v_nr + w_rg * (v_r - v_nr)

z_next = z + (t_next - t) * v_rg
```

Snippet 2. RG + CFG, Inference

```
# y0: null label
# w_cfg: CFG scale

x_r  = net(z, t, y,  regs=True)
x_nr = net(z, t, y,  regs=False)
x_u  = net(z, t, y0, regs=True)

v_r  = (x_r  - z) / (1 - t)
v_nr = (x_nr - z) / (1 - t)
v_u  = (x_u  - z) / (1 - t)

v = v_u + w_cfg * (v_r - v_u)
        + w_rg * (v_r - v_nr)

z_next = z + (t_next - t) * v
```

Snippet 3. RG, Training

```
# net(z, t, regs): network
# x: training batch
# p: prob. of dropping registers

t, e = sample_t(), randn_like(x)
z = t * x + (1 - t) * e
v = (x - z) / (1 - t)

# Drop registers during training
use_regs = rand() >= p

x_pred = net(z, t, regs=use_regs)
v_pred = (x_pred - z) / (1 - t)

loss = l2_loss(v - v_pred)
```

preserve similar image content, while the version without registers exhibits stronger artifacts and less coherent structure. Thus, their difference provides a meaningful direction for improving quality.

Our method. We call our approach *Register Guidance (RG)*. It is simple to implement and apply, as described in Snippets, 1, 2 and 3.

Inference. First, we describe the inference procedure in Snippet 1. RG follows a similar principle to CFG, however, instead of using an unconditional prediction, it uses the prediction of the model evaluated without registers. We use an additional hyperparameter `w_rg` to control the RG strength.

Moreover, in Snippet 2, we show how to combine RG with CFG. Since registers carry a source of information beyond the class label, RG can be combined with CFG in a complementary way. This strategy further improves generation quality.

Training. We apply RG (Snippet 1) in the unconditional setting using two separate models: JiT-B/16 models trained with and without registers. Interestingly, we observe quality improvements, but also the introduction of artifacts at larger guidance scales (Figure 25 top). This suggests that the model difference is meaningful, as it improves quality, but also imperfectly aligned, as reflected by the emerging artifacts. More examples are shown in Figure 26.

The artifacts arising from using two separate models suggest that better alignment between predictions is needed. This motivates us to train a single model that supports generation both with and without registers, improving alignment between the two predictions. We describe the training procedure in Snippet 3. We use the original training procedure (Li & He, 2025), but occasionally drop all register tokens from the sequence, enabling the model to operate both with and without registers. We introduce an additional hyperparameter `p`, which controls the probability of dropping registers.

With a single model for RG, the artifacts disappear, indicating better alignment between the with- and without-register predictions, while generation quality continues to improve (Figure 25 bottom).

Importantly, during training, we use small values of `p`, which preserve the performance of the original register-based model while enabling the model to operate without registers. Technically, we use separate batches for the two regimes, which does not degrade performance. Although the same behavior could be achieved by masking register tokens in self-attention.

Discussion. Our approach can be understood as a specific instance of AutoGuidance (Karras et al., 2024). However, our method gives a concrete way to obtain a weaker model, while in AG this choice can be challenging because it requires careful manual design (Zhou et al., 2026). Moreover, we find that AG does not improve results in the JiT setup, which is further confirmed by Baade et al. (2026).

Results. We test our approach on ImageNet 256 and 512 in the conditional setting. First, we ablate the influence of `w_rg` and `p`. We use JiT-B/16 and report results for different hyperparameter values in Figure 6 and Table 4. Under optimal parameters, RG improves performance across different model scales and image sizes (Table 5). More details on hyperparameter selection are in App. B.2.

Tables 6 and 7 show a system-level comparison, our method outperforms all JiT configurations in FID and FDr^6 (Yang et al., 2026). To achieve the best results, we use the original JiT checkpoints and fine-tune them with our training procedure. More experimental results are provided in App. D.1.

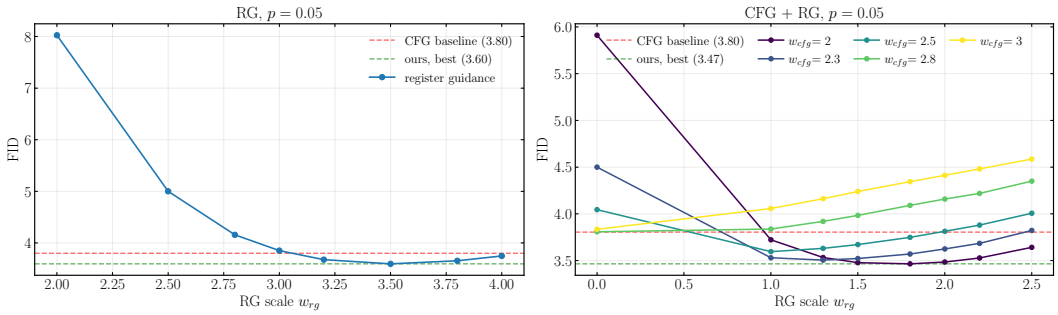


Figure 6: **Effect of `w_rg` under fixed `p=0.05`.** Left: RG improves FID as the scale increases, with the best result at `w_rg=3.5`. Right: RG remains complementary to CFG, further improving performance over the CFG baseline, with the best setting at `w_rg=1.8` and `w_cfg=2.0`.

p	0.0	0.1	0.05	0.03	0.01
CFG	3.71	3.96	3.80	**3.69**	3.86
RG	13.13	4.55	3.60	**3.46**	3.77
RG+CFG	8.02	3.85	3.47	**3.32**	3.64

Table 4: **Effect of `p` under fixed `w_rg`.** We vary `p` during training, where `p=0` corresponds to the original JiT setting. `p=0.03` gives the best performance without degrading the CFG baseline.

	B/16	L/16	B/32	L/32
CFG	3.71	2.47	4.12	2.69
RG	3.46	2.18	4.01	2.47
RG+CFG	**3.32**	**2.16**	**3.69**	**2.41**

Table 5: **Register Guidance improves across model scales and image sizes.** With the best `p` and `w_rg`, RG improves over CFG, and RG+CFG gives the best FID in all settings.

ImageNet 256×256	**Params**	**FID↓**	**FDr^6↓**
Latent-space Diffusion			
SiT (Ma et al., 2024), XL/2	675M	2.06	8.44
REPA (Yu et al., 2024), SiT-XL/2	675M	1.42	3.04
RAE (Zheng et al., 2026), DiT^{DH}-XL/2	839M	1.13	3.26
Pixel-space Diffusion			
VDM++ (Kingma & Gao, 2023), UViT/2	2B	2.12	–
SiD2 (Hoogeboom et al., 2025), UViT/1	N/A	**1.38**	–
PixelFlow (Chen et al., 2025a), XL/4	677M	1.98	–
PixNerd (Wang et al., 2025), XL/16	700M	2.15	**5.01**
JiT (Li & He, 2025), B/16	131M	3.66	16.14
JiT (Li & He, 2025), L/16	459M	2.36	10.79
JiT (Li & He, 2025), H/16	953M	1.86	7.66
JiT-B/16, CFG + RG	131M	2.96	12.26
JiT-L/16, CFG + RG	459M	2.00	8.35
JiT-H/16, CFG + RG	953M	1.80	6.10

Table 6: **ImageNet** 256×256 **results.**

ImageNet 512×512	**Params**	**FID↓**	**FDr^6↓**
Latent-space Diffusion			
SiT (Ma et al., 2024), XL/2	675M	2.62	–
REPA (Yu et al., 2024), SiT-XL/2	675M	2.08	–
RAE (Zheng et al., 2026), DiT^{DH}-XL/2	839M	1.13	–
Pixel-space Diffusion			
VDM++ (Kingma & Gao, 2023), UViT/4	2B	2.65	–
SiD2 (Hoogeboom et al., 2025), UViT/2	N/A	**1.48**	–
PixNerd (Wang et al., 2025), XL/16	700M	2.84	–
JiT (Li & He, 2025), B/32	133M	4.02	15.78
JiT (Li & He, 2025), L/32	462M	2.53	9.73
JiT (Li & He, 2025), H/32	956M	1.94	6.42
JiT-B/32, CFG + RG	133M	3.34	12.62
JiT-L/32, CFG + RG	462M	2.23	6.90
JiT-H/32, CFG + RG	956M	1.87	**5.05**

Table 7: **ImageNet** 512×512 **results.**

4 Discussion

Registers in image DiTs represent a promising research direction, offering both analytical insights and practical implications. While this work has explored certain aspects of this topic, we believe considerable scope remains for further investigation, and we hope our findings encourage future work on methods that better exploit their potential.

One potential direction follows from the clearly distinct behaviors of register and patch tokens: developing dual-stream DiT architectures (Esser et al., 2024) that process them more effectively. We preliminarily explored several efficient dual-stream designs in App. D.2 but did not observe major improvements. Nevertheless, we believe this direction merits further exploration.

Moreover, our exploration of Register Guidance leaves room for further study. While we focus on pixel-space models in the class-conditional setting, extending Register Guidance to text-to-image models presents an interesting direction for future research (Gan et al., 2025).

AI USE STATEMENT

Generative AI assistants were used to support literature search, conducting the evaluation, result collection, language editing and formatting, and figure preparation. The authors take responsibility for the final content of the paper.

REFERENCES

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Hussein A Aly and Eric Dubois. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing*, 14(10):1647–1659, 2005.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. *arXiv preprint arXiv:2502.06415*, 2025.

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.2: Analyzing and enhancing the latent space of FLUX – representation comparison, 2025. URL `https://bfl.ai/research/representation-comparison`.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

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

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025a.

Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F Luo. Vision transformers with self-distilled registers. *arXiv preprint arXiv:2505.21501*, 2025b.

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

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.

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.

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.

Chaofan Gan, Zicheng Zhao, Yuanpeng Tu, Xi Chen, Ziran Qin, Tieyuan Chen, Mehrtash Harandi, and Weiyao Lin. Massive activations are the key to local detail synthesis in diffusion transformers. *arXiv preprint arXiv:2510.11538*, 2025.

Chaofan Gan, Yuanpeng Tu, Xi Chen, Tieyuan Chen, Yuxi Li, Mehrtash Harandi, and Weiyao Lin. Unleashing diffusion transformers for visual correspondence by modulating massive activations. *Advances in Neural Information Processing Systems*, 38:114432–114462, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026b.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

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.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion: 1.5 fid on imagenet512 with pixel-space diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18062–18071, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Amna Jamal, Mika Tan, Clarissa Aurelia Nahid Saputra, Quan Huynh, Kevin Zhu, and Antonio Mari. Diffusion transformers use sink registers. In *Second Workshop on XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*, 2026.

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

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

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

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models, 2024. URL https://arxiv.org/abs/2404.07724.

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

Alexander Lappe and Martin A Giese. Register and [cls] tokens induce a decoupling of local and global features in large vits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

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.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy V Vo. Revisiting [cls] and patch token interaction in vision transformers. *arXiv preprint arXiv:2602.08626*, 2026.

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.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

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

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

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

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

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

Zunhai Su, Hengyuan Zhang, Wei Wu, Yifan Zhang, Yaxiu Liu, He Xiao, Qingyao Yang, Yuxuan Sun, Rui Yang, Chao Zhang, et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

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

Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Belinkov. Padding tone: A mechanistic analysis of padding tokens in t2i models, 2025. URL `https://arxiv.org/abs/2501.06751`.

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.

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

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. Sinder: Repairing the singular defects of dinov2. In *European Conference on Computer Vision*, pp. 20–35. Springer, 2024.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. *arXiv preprint arXiv:2507.23268*, 2025.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *IEEE transactions on pattern analysis and machine intelligence*, 45(12):15790–15801, 2023.

Yuxin Wen, Jim Wu, Ajay Jain, Tom Goldstein, and Ashwinee Panda. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Jiawei Yang, Zhengyang Geng, Xuan Ju, Yonglong Tian, and Yue Wang. Representation fr\'echet 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. *Advances in Neural Information Processing Systems*, 37:55342–55369, 2024.

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.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025.

Junno Yun, Yaşar Utku Alçalar, and Mehmet Akçakaya. No alignment needed for generation: Learning linearly separable representations in diffusion models. *arXiv preprint arXiv:2509.21565*, 2025.

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11545, 2026.

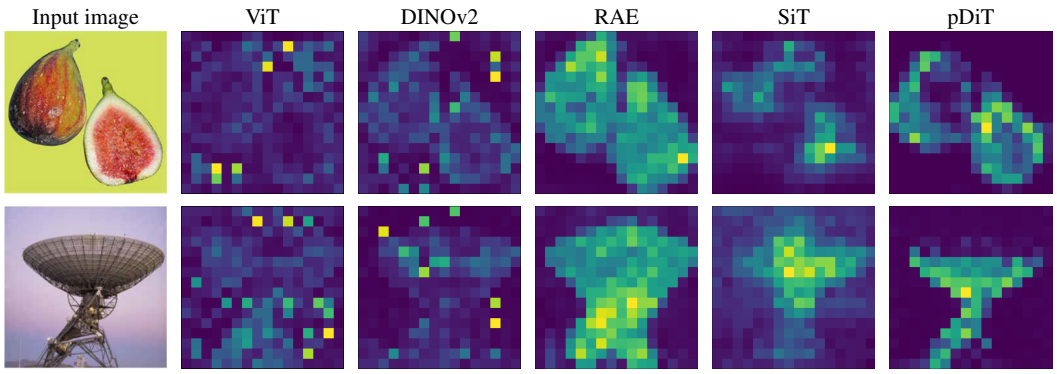


Figure 7: **Diffusion transformers do not exhibit attention-map outliers.** Unlike ViTs, where attention-map anomalies typically appear in low-information regions (e.g., background), DiT attention remains largely focused on the main objects.

A Related Work

Attention Sinks in Large Language Models. In autoregressive LLMs, attention sinks are a well-explored area (Xiao et al., 2023; Sun et al., 2024; Gu et al., 2024; Yona et al., 2025; Qiu et al., 2025). Xiao et al. (2023) first analyzes anomalies in attention and finds that a large portion of the attention score is allocated to the initial tokens, making them attention sinks. The authors propose preserving these initial register tokens during inference, which leads to a significant quality boost. Sun et al. (2024) continues the sink analysis and finds that outliers appear in a few fixed feature dimensions, regardless of the input, as well as in two types of special tokens. Gu et al. (2024) analyzes why and how attention sinks emerge during training. Yona et al. (2025) explores the functional role of attention sinks, showing their connection to the repeated token divergence phenomenon. Qiu et al. (2025) demonstrates that sparse gating can eliminate attention sinks.

Attention Sinks in Diffusion Language Models. Rulli et al. (2025) extends the study of attention sinks to DLMs, showing that attention sinks persist, but with different behavior: in DLMs, the positions of attention sinks tend to shift during the generation process as tokens are progressively unmasked. Moreover, these sinks can be masked without significant degradation. Zhang et al. (2026) addresses this moving-sink behavior by adding an additional sink token that is globally visible to all other tokens while attending only to itself. Such a token is shown to stabilize DLM inference.

Attention Sinks in Video Diffusion Transformers. Wen et al. (2025) presents the first analysis of attention sinks in video DiTs, highlighting similarities and differences with LLMs. Notably, they find that these sinks are concentrated in the first frame, analogous to the <BOS> token in LLMs. Liu et al. (2025); Shin et al. (2025) consider an autoregressive approach to video generation, predicting multiple frames simultaneously. Similar to LLMs (Xiao et al., 2023), they find that it is important to retain the first frame, which acts as an attention sink. Yi et al. (2025) extends the idea of keeping the initial frame by proposing Deep Sink, which aims to stabilize global context during long rollouts. Note that video DiTs use attention sinks mainly to enable long-horizon AR sampling, rather than to study their effect on denoising performance.

Outliers in Feature Dimensions. Another line of work studies outliers that emerge not across tokens, but within feature (channel) dimensions of learned representations (An et al., 2025; Siméoni et al., 2025; Gan et al., 2026; 2025). Initially studied in LLMs (An et al., 2025), this phenomenon was later analyzed in DiTs (Gan et al., 2026; 2025). In particular, (Gan et al., 2026) links massive activations to layer normalization, while Gan et al. (2025) shows that manipulating them can improve fine-grained image details.

Importantly, concurrent work by Gan et al. (2025) is closely related to our Register Guidance, but uses a model with channel outliers removed as the weak model. They show that this mainly improves fine-grained details. In contrast, our approach uses a model without register tokens as the weak model, leading to substantially larger changes that improve not only fine details but also global structure and visual coherence. We hypothesize that removing register tokens produces a stronger weak counterpart than removing only a subset of channels, resulting in a more informative guidance signal.

Register Tokens in Vision Transformers. Several works study the role and behavior of register tokens in ViTs (Darcet et al., 2023; Jiang et al., 2025; Lappe & Giese, 2025; Shi et al., 2026; Chen et al., 2025b; Wang et al., 2024; Marouani et al., 2026). Darcet et al. (2023) introduces register tokens to avoid sink artifacts in attention maps, improving ViT internal representations. Subsequent studies further analyze their functional role, showing that they can influence feature aggregation (Lappe & Giese, 2025). Jiang et al. (2025); Chen et al. (2025b) study post-hoc or self-distilled ways to add registers to pretrained ViTs without full retraining. Shi et al. (2026) argue that register tokens alone do not fully explain or resolve all ViT artifacts. Recent large-scale ViTs also retain explicit outlier-handling mechanisms: DINOv3 adopts register tokens after comparing them with attention-bias and value-gating alternatives inspired by LLM outlier analyses (Siméoni et al., 2025).

Attention Sinks in Text-to-Image DiTs. For text-to-image diffusion, (Yi et al., 2024) highlights the special role of text tokens across different denoising steps, while Jamal et al. (2026) studies high-norm activations in pretrained text-to-image DiTs. Padding Tone (Toker et al., 2025) further shows that auxiliary text tokens, such as padding tokens, can influence text-to-image generation. However, these works do not fully explain why such outliers can improve generation quality or what role they play within the model. Our work addresses this gap by systematically studying high-norm tokens and connecting them to register tokens, showing that registers play an important role in generation quality.

To summarize, tokens with unusually large norms have been studied across different domains (Su et al., 2026). In this work, we extend this research direction to image diffusion transformers and show that they are particularly important for pixel-space models.

B Implementation Details

B.1 Analysis Implementation Details

In our implementation of pDiTs, we largely follow the JiT setup (Li & He, 2025). The model uses flow matching with x-prediction and the forward process $x_t = tx + (1-t)\epsilon$, where $t = 1$ corresponds to clean data. We adopt the same diffusion backbone, but remove in-context conditioning and instead introduce register tokens implemented as trainable parameters without additional layers (Darcet et al., 2023).

We train models of three sizes $-$ B (131M), L (459M), and H (953M) $-$ on ImageNet 256×256 and ImageNet 512×512. For the B and L models, we use a batch size of 1024, following JiT (Li & He, 2025). Due to limited computational resources, the H model on ImageNet 512×512 is trained with a smaller batch size of 512. All other training and inference settings follow JiT (Li & He, 2025).

For PixelDiT (Yu et al., 2025), we consider the XL model (800M) and train three models for 120 epochs: with registers, without registers, and with in-context conditioning. We use the same configuration except for gradient clipping, where we use a smaller value of 0.5 instead of 1.0 from the original implementation for stability.

For the latent-space models presented in Table 2, we follow their original setups (Zheng et al., 2026; Ma et al., 2024). Importantly, for RAE, we use the DINOv2 encoder with registers. For both models, we use the same batch size as in the original implementations and train SiTs and RAE for 300 and 600 epochs, respectively. Register tokens are implemented in the same manner as in the pixel-space models.

B.2 Register Guidance Implementation Details

As discussed in the main text, we train a JiT model from scratch while periodically dropping registers with probability p. We compare three settings: CFG (baseline), RG, and RG+CFG. We train all models for 600 epochs. For all settings, we use the interval guidance strategy (Kynkäänniemi et al., 2024). We present the optimal hyperparameters in Table 8.

Besides training from scratch, we also consider a fine-tuning setup (Tables 6 and 7). In this case, we take pretrained JiT models (Li & He, 2025), which cannot generate without registers, and fine-tune them for 50 epochs using a lower learning rate of $2\text{e}{-5}$. Moreover, we use a larger register-drop probability, $p = 0.3$, while keeping the other parameters the same as in Table 8. This setup gives the

		RG		RG+CFG		
Model	CFG	RG	p	RG	CFG	p
JiT-B/16	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 3.0$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 3.5$	0.03	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.8$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.0$	0.03
JiT-L/16	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.4$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 2.8$	0.03	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.4$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 1.6$	0.03
JiT-H/16	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.2$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 2.6$	–	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.1$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 1.4$	–
JiT-B/32	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 3.0$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 3.2$	0.03	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.8$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.0$	0.03
JiT-L/32	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.5$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 2.8$	0.03	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.4$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 1.6$	0.03
JiT-H/32	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 2.3$	$t^{\text{rg}}_{\text{min}} = 0.03$ $t^{\text{rg}}_{\text{max}} = 0.9$ $w_{\text{rg}} = 2.6$	–	$t^{\text{rg}}_{\text{min}} = 0.1$ $t^{\text{rg}}_{\text{max}} = 1.0$ $w_{\text{rg}} = 1.1$	$t^{\text{cfg}}_{\text{min}} = 0.1$ $t^{\text{cfg}}_{\text{max}} = 1.0$ $w_{\text{cfg}} = 1.4$	–

Table 8: **Training and inference configurations for CFG, RG, and RG+CFG.** For RG and RG+CFG, we separately report the guidance parameters and register-drop probability p.

best performance and shows that Register Guidance can be applied to already trained models with little additional effort.

C Additional Analysis Results

C.1 Analysis on ImageNet

Outliers and registers in DiTs. In the main text, we show that DiTs are free from the artifacts observed in ViTs. However, introducing register tokens leads to the emergence of high-norm tokens within the registers themselves. While the main analysis focuses on the small-scale pDiT-B/16 model at a single timestep ($t = 0.5$), here we present corresponding results for additional timesteps, larger model variants, higher resolution, latent-space counterparts, and few-step models (Lu et al., 2026; Geng et al., 2026a).

First, Figure 11 shows that this effect consistently holds across different timesteps.

Second, Figure 12 shows that the observations made for pDiT-B/16 also hold for larger pixel-space variants, namely pDiT-L/16 and pDiT-H/16. Models without registers maintain relatively uniform patch-token norms across layers, without pronounced outliers. Once registers are introduced, however, large-norm tokens systematically emerge within the register tokens.

In Table 9, we show that registers also improve performance on ImageNet 512×512.

Additionally, we analyze where outliers emerge within each block. In Figure 10, we show the register-token norms after different sub-layers of several pDiT-B/16 blocks. We find that the outliers are mainly formed after the MLP layer rather than after the attention layer.

	pDiT-B/32			pDiT-L/32			pDiT-H/32		
Epoch	w/o regs.	w/ regs.	IC	w/o regs.	w/ regs.	IC	w/o regs.	w/ regs.	IC
200	8.67	**6.29**	5.84	4.41	**3.33**	3.28	3.47	**2.70**	2.77
600	5.14	**4.20**	4.12	3.12	**2.72**	2.69	2.68	**2.08**	2.08

Table 9: **Generation quality (FID ↓) of pDiTs with and without register tokens on ImageNet 512.** We compare generation quality across model sizes and training epochs. Here, "w/o regs." and "w/ regs." denote training without and with register tokens, respectively. The shaded IC column reports performance with in-context class conditioning (Section 2.6).

	RAE-space pDiT backbone	**VAE-space pDiT backbone**	**Pixel-space pDiT backbone**
		Base size	
With registers	10.12	**7.17**	**5.30**
Without registers	**9.40**	7.20	7.39

Table 10: **Registers are more effective in pixel-space.** We compare the generation quality (FID) of models with and without register tokens across different training spaces (DINOv2, VAE, and pixel space) using the same pDiT backbone. Register tokens provide the largest improvements in pixel space, moderate gains in VAE space, and no performance improvement in DINOv2 space (RAE).

We analyze the latent-space models SiT (Ma et al., 2024) and RAE (Zheng et al., 2026) in terms of outlier behavior. Figures 13 and 14 show that the same phenomenon also holds for latent-space diffusion models.

Finally, we consider MeanFlows (Geng et al., 2026b) and pMF (Lu et al., 2026), since they also use in-context conditioning in their implementations, and find that they also exhibit outliers within the in-context tokens. Figures 15 and 16 show that the same phenomenon holds for few-step models in both latent and pixel spaces, across model sizes and image resolutions.

Improving Feature Map Quality. Next, we find that register tokens consistently reduce the feature norms of patch tokens. In Figure 17, we show this effect for larger pixel-space variants of pDiT. Interestingly, we do not observe the same behavior in SSL ViTs such as DINOv2, as shown in Figure 23.

We show that register tokens improve feature quality across larger pDiT variants. Figure 18 reports the TV ratio for pDiT-L/16 and pDiT-H/16. In addition, Figure 19 analyzes the correlation decay slope (Singh et al., 2025), where lower values indicate stronger spatial organization. Both metrics show the same trend: register tokens improve internal representations at high noise levels, beginning at block 4, where the registers are introduced.

Effect of Register Injection Layer and Number of Register Tokens. Our analysis shows that, unlike ViTs where register tokens are effective from the first layer (Darcet et al., 2023), pDiT-B/16 benefits primarily from their delayed introduction. For instance, introducing registers across all layers (0–11) results in performance similar to a model without registers.

Here, we provide a hypothesis for why this effect occurs. In Figure 20, we present linear probing results for register tokens introduced from layers 0 and 4. Importantly, we observe a notable difference between these two configurations. Specifically, models with registers introduced from the first layer produce substantially less informative register tokens. That is, we observe many tokens with moderate feature norms but very low probing accuracy. This suggests that these tokens serve neither as norm sinks nor as carriers of semantic information. Since we measure these results after the 5th block, we hypothesize that this poor signal from register tokens can negatively affect later layers and consequently degrade model performance.

We hypothesize that this poor register signal arises because, in the early layers of pDiTs, the model has not yet formed meaningful semantic structure. As a result, register tokens cannot capture diverse semantic information and instead propagate uninformative features.

Pixel-space versus latent-space. In the main text, we show that register tokens provide substantially larger gains for pixel-space models. In Table 10, we further report results for additional backbones in RAE and VAE spaces and observe the same trend: models in both spaces perform similarly with and without registers. This suggests that the effect is not backbone-specific.

Next, we provide an explanation for this behavior through an analysis of token feature norms and intermediate representations across different model types. In Figures 21 and 22, we observe that pDiTs consistently exhibit the largest feature norms and the noisiest intermediate representations compared to latent-space counterparts. Specifically, Figure 21 shows that patch-token norms in pDiTs are substantially higher across all timesteps, while Figure 22 demonstrates that their intermediate features have significantly larger TV values.

These observations suggest that pixel-space diffusion produces substantially noisier intermediate features, which may explain why register tokens are especially beneficial in this setting.

C.2 Analysis on Text-to-Image Models

In addition to ImageNet-based DiTs, we consider large-scale text-to-image approaches. We do not train these models with registers, but instead analyze pretrained open-source versions. Specifically, we analyze FLUX (Labs, 2024), SD3.5 Large (Esser et al., 2024), Qwen-Image (Wu et al., 2025), and Z-Image (Cai et al., 2025), which propagate textual information through an auxiliary token sequence appended to the image tokens. Importantly, this sequence does not directly participate in the diffusion loss, raising the possibility that it may partially serve a register-like role. Moreover, we consider PixArt-α (Chen et al., 2024), which does not introduce an additional sequence into the image-token sequence and instead follows the classical DiT (Albergo et al., 2025) regime with cross-attention.

We present the results in Figure 24. For SD3.5, Qwen-Image, and Z-Image, we observe that high-norm outliers emerge within the text-token sequence, while image-token norms remain comparatively uniform. This behavior closely resembles the role of register tokens in ImageNet-based DiTs, suggesting that auxiliary text tokens may implicitly act as repositories for high-norm representations.

Interestingly, for FLUX and PixArt-α, we observe outliers not only in text tokens but also in several image tokens. However, these image-token outliers disappear in later layers, suggesting that patch tokens cannot sustain high norms due to the loss constraint. Moreover, in FLUX, their norms are smaller than those of the text-token outliers. In PixArt-α, we find a single image-token outlier in an intermediate layer (block 16), which also disappears later.

These results suggest that the model needs space for outliers, but patch tokens are not convenient slots for this role. This further supports the importance of registers as dedicated outlier tokens.

D Additional Experiments

D.1 Register Guidance

In the main text, we find that Register Guidance produces artifacts when using two separate models, despite the fact that they generate similar images. However, when using a single model that can generate both with and without registers, this mismatch disappears and no artifacts are introduced. Here, we provide additional visual results for the two approaches in Figures 26 and 27.

In our experiments, we find that the best configuration combines CFG and RG. Here, we provide visual results illustrating this behavior in Figure 28.

To achieve the best results, we run our approach in the fine-tuning regime. Specifically, we take pretrained JiT checkpoints and fine-tune them for several epochs. We use a larger register-drop probability p than in the training-from-scratch setup. In Table 11, we ablate different values of p for JiT-B/16 and report FID at different fine-tuning epochs. As expected $p = 0$ does not enable effective RG, while $p = 0.1$ requires longer fine-tuning to converge. A large value, $p = 0.7$, improves RG but degrades the CFG baseline. The intermediate value $p = 0.3$ provides the best trade-off, yielding

$p = 0.0$	epoch $= 0$	10	20	30	50
CFG	3.60	3.46	3.41	3.44	3.38
RG	13.13	13.36	13.88	14.03	14.19
RG+CFG	8.02	8.17	8.20	8.43	8.41

$p = 0.1$	epoch $= 0$	25	30	50	85
CFG	3.60	3.55	3.51	3.44	3.41
RG	13.13	5.33	4.89	3.94	3.48
RG+CFG	8.02	3.51	3.29	3.10	2.90

$p = 0.3$	epoch $= 0$	10	20	30	50
CFG	3.60	3.59	3.61	3.59	3.52
RG	13.13	6.96	4.74	3.65	3.25
RG+CFG	8.02	4.41	3.38	3.15	2.96

$p = 0.7$	epoch $= 0$	10	20	30	50
CFG	3.60	3.79	3.81	3.79	3.68
RG	13.13	6.72	4.24	3.50	3.95
RG+CFG	8.02	4.45	3.51	3.29	3.21

Table 11: **Fine-tuning results for JiT-B/16 with different register-drop probabilities p.** We fine-tune a pretrained JiT-B/16 checkpoint and report FID across epochs for CFG, RG, and RG+CFG. $p = 0$ corresponds to fine-tuning the original model with the basic setup, which does not improve RG. A low value, $p = 0.1$, requires longer convergence, while a high value, $p = 0.7$, degrades the original performance. Therefore, we use the intermediate value $p = 0.3$.

	$p = 0.0$			$p = 0.3$		
Model	CFG	RG	RG+CFG	CFG	RG	RG+CFG
JiT-B/32	3.77	12.61	7.75	3.90	3.62	**3.34**
JiT-L/16	2.30	7.82	4.50	2.39	**2.00**	2.13
JiT-L/32	2.55	12.06	8.05	2.57	2.24	**2.23**
JiT-H/16	1.90	7.67	3.68	1.92	1.85	**1.80**
JiT-H/32	1.95	13.37	6.76	1.94	2.01	**1.87**

Table 12: **Best fine-tuning results for different JiT models.** We report the best FID across fine-tuning epochs for each setting. Register-drop fine-tuning with $p = 0.3$ enables effective RG and RG+CFG, outperforming the original fine-tuning setup with $p = 0.0$ using CFG. As expected, $p = 0.0$ does not enable effective RG.

strong RG performance while preserving the original model quality. In Table 12, we report results for other models using the selected value $p = 0.3$ and their optimal RG scales.

We report this tuned results in Table 6, 7, where we calculate FID and FDr^6 (Yang et al., 2026) over 50K samples generated with an equal number of images per class. FDr^6 averages Fréchet distances measured in six representation spaces instead of the Inception space alone, where each distance is divided by the distance that the same space assigns between real ImageNet data and the reference statistics, so that spaces with different feature scales contribute comparably. The six spaces are InceptionV3 (299px, 2048-d pooled features), ConvNeXt-V2-B (224px, 1024-d spatially averaged features), DINOv2 ViT-L/14 (256px, 1024-d class token), CLIP ViT-L/14 (256px, 1024-d class token), MAE ViT-L/16 (224px, 1024-d class token), and SigLIP2 SO400M/16 (224px, 1152-d attention-pooled features), with the reference statistics and normalisation constants released by Yang et al. (2026). Generated images are resized to each encoder's input resolution with bicubic interpolation and antialiasing, and normalised with the mean and standard deviation of the corresponding pretrained model.

D.2 Decoupled Processing of Register and Patch Tokens

Dual-stream architecture. Our analysis reveals that register and patch tokens play distinct roles, yet existing pixel-space architectures (Li & He, 2025; Lu et al., 2026) process them identically using fully shared parameters. Given their differing behaviors, such parameter sharing may be suboptimal.

In contrast, large-scale models with appended auxiliary sequences (e.g., text tokens) often employ dual-stream architectures (Esser et al., 2024; Labs, 2024), where separate parameters are used for different token types while interactions are maintained through attention. Motivated by these

Snippet 4. RMSNorm Dual

```
# registers + patches
# x: [B, n_reg + n_patch, h]

# dual RMSNorm
(x_reg, x_patch) = split(x)

# separate normalization
x_patch = RMSNorm(x_patch, w1,
      eps1)
x_reg = RMSNorm(x_reg, w2, eps2)

# merge streams
x = concat(x_reg, x_patch)
```

Snippet 5. SwiGLU MLP Dual

```
# shared hidden latent
x1, x2 = chunk(Linear(x))
hidden = silu(x1) * x2

# split output projection
(h_reg, h_patch) = split(hidden)
y_reg   = Linear(h_reg)
y_patch = Linear(h_patch)

# merge streams
y = concat(y_reg, y_patch)
```

Snippet 6. adaLN Dual

```
# condition
# c: [B, h]

# shared modulation
m = Linear(silu(c))

# dual branch params
m_r = m + LoRA(c)
(shift, scale, gate) = split(m)
(shift_reg, scale_reg, gate_reg)
      = split(m_r)
```

adaLN	MLP	Attn	RMSNorm	Params (M)	FID
				131M	3.71
✓	✓	✓	✓	230M	**3.14**
✓				173M	3.44
	✓			168M	3.44
		✓		150M	3.61
			✓	131M	3.70

Table 13: **Ablation of dual-stream designs.** We separately duplicate transformer components for register and patch tokens. Duplicating all components gives the best performance but significantly increases parameters. Among individual components, `adaLN` and `MLP` contribute the most, while `Attention` and `RMSNorm` provide limited gains. gray denotes the single-stream JiT-B/16 baseline.

adaLN	MLP	Attn	RMSNorm	Params (M)	FID
				131M	3.71
✓	✓	✓	✓	161M	3.48
✓	✓		✓	149M	**3.41**
✓			✓	136M	3.81
	✓		✓	143M	3.56
✓	✓			149M	3.53

Table 14: **Ablation of compact dual-stream designs.** We evaluate parameter-efficient dual-stream configurations by selectively dualizing transformer components for register and patch tokens. The best compact design (green) dualizes `adaLN`, `MLP`, and `RMSNorm`, while keeping `Attention` shared, achieving the best trade-off between parameter count and generation quality.

Model	Params (M)	GFLOPs	Epoch	FID
JiT-B/16	131	23.8	200	4.71
			600	3.71
JiT-L/16	459	84.2	200	2.95
			600	2.47
Ours-B/16	149	23.8	200	**4.25**
			600	**3.41**
Ours-L/16	518	84.2	200	**2.76**
			600	**2.32**

Table 15: Comparison of our compact dual-stream architecture with the JiT baseline on ImageNet 256×256.

Model	Params (M)	GFLOPs	Epoch	FID
JiT-B/32	133	24.3	200	5.84
			600	4.12
JiT-L/32	461	84.8	200	3.28
			600	2.69
Ours-B/32	151	24.3	200	**4.98**
			600	**3.92**
Ours-L/32	520	84.8	200	**3.08**
			600	**2.57**

Table 16: Comparison of our compact dual-stream architecture with the JiT baseline on ImageNet 512×512.

differences, we investigate dual-stream designs that enable specialized processing of register tokens in pixel-space DiTs.

We consider the JiT architecture (Li & He, 2025), which uses in-context conditioning. JiT blocks consist of `RMSNorm`, `adaLN`, `Attention`, and `MLP` layers. Thus, we selectively introduce token-specific specialization in these components.

In Table 13, we evaluate naive dual-stream designs that separately duplicate transformer components for register and patch tokens. Duplicating all components gives the best performance but significantly increases parameters. Among individual components, `adaLN` and `MLP` contribute the most, while `Attention` and `RMSNorm` provide limited gains.

Compact dual-stream design. As we find, naive duplication leads to a significant increase in parameters. We therefore explore parameter-efficient strategies for selective duplication.

For `RMSNorm`, which has few parameters, we use separate parameters for register and patch tokens with negligible cost (Snippet 4).

For parameter-intensive modules (`adaLN`, `Attention`, and `MLP`), we avoid full duplication. In `MLP`, we compute a shared `SwiGLU` (`Linear` projection followed by `SiLU` gating) and apply separate output projections for register and patch tokens (Snippet 5).

For `adaLN` and `Attention`, we use parameter-efficient LoRA adaptations (Hu et al., 2022), following Marouani et al. (2026). Shared parameters are computed for all tokens, while lightweight LoRA branches are applied only to register tokens. In `adaLN`, normalization parameters are first shared across all tokens, then refined for register tokens via a LoRA branch (Snippet 6). In `Attention`, QKV projections are shared, with LoRA applied only to register-token representations, while the output projection remains shared.

Our final architecture combines single-stream and dual-stream transformer layers. Motivated by the ablation results in Table 3, which show that register tokens are ineffective in early layers, we first process the image sequence using standard single-stream layers. Register tokens are then introduced only in later stages, where the model transitions to dual-stream processing. The overall design increases the parameter count by only $\sim 14\%$, compared to $\sim 77\%$ for naive duplication.

Results. We consider models of two sizes, B and L, and train them on ImageNet at resolutions of 256 and 512. The proposed architecture introduces an additional parameter overhead of approximately 14%. We train the models using the same training and inference configuration as in JiT (Li & He, 2025). For both configurations, we use LoRA with rank 128 in AdaLN.

We first study which components should be decoupled in the compact dual-stream design. We use JiT-B/16 adapted to a dual-stream architecture and train it on ImageNet 256×256 for 600 epochs, following the training and sampling setups of Li & He (2025). We enable in-context tokens at layer 4, resulting in single-stream layers 0–3 and dual-stream layers 4–11.

Table 14 reports the results for several parameter-efficient dual-stream designs, compared against the single-stream JiT-B/16 baseline (gray-shaded row).

Applying the dual architecture to all layers improves FID from 3.71 to 3.48. Sharing `Attention` further improves performance while reducing parameters, suggesting that attention decoupling is unnecessary. In contrast, sharing `MLP` significantly degrades performance, while removing dual `adaLN` or `RMSNorm` also hurts results. Based on these findings, our final compact design dualizes `MLP`, `adaLN`, and `RMSNorm`, while keeping `Attention` shared (green-shaded row).

Next, we conduct a system-level comparison between our compact dual architecture and the JiT baseline on ImageNet 256×256 and 512×512 across different model sizes. For all variants, we follow the training setup and baseline configurations of Li & He (2025).

We report the results in Tables 15 and 16. The parameter-efficient dual-stream architecture consistently improves generation quality across model scales and image resolutions, while introducing only a small increase in parameters and no additional runtime overhead in terms of GFLOPs.

We believe our compact strategy can provide greater benefits at larger scales, where dual-stream architectures (Esser et al., 2024) achieve significant quality improvements. In these settings, the appended sequence, e.g., text, becomes substantially larger than the small set of register tokens.

D.3 Other results

In Table 17, we evaluate whether register-like tokens remain beneficial when combined with representation alignment (Yu et al., 2024). We adopt the recent REPA adaptation (Shin et al., 2026), which was specifically proposed for JiT-based architectures. We observe that register tokens consistently improve the performance of REPA-enhanced models. Moreover, our compact dual-stream design remains effective with REPA.

In addition, we ablate register-token configurations in Table 18, varying both the number of registers and the blocks where they are enabled. The results show that delayed introduction is crucial: enabling registers only in deeper layers (4–11) consistently outperforms both the no-register baseline and using registers from the first layer (0–11). Increasing the number of registers also improves performance,

Method	200 epochs	600 epochs
pDiT-B/16 + PixelREPA	**4.15**	**3.38**
pDiT-B/16 (w/o registers) + PixelREPA	4.78	3.56
JiT-B/16 + PixelREPA + compact dual	**3.68**	**3.23**
JiT-B/16 + PixelREPA	4.00	3.40

Table 17: **REPA remains complementary to register-like tokens and further benefits from the proposed compact dual-stream design.** We evaluate pDiT models with and without register tokens using PixelREPA (Shin et al., 2026) on ImageNet 256×256. Register-like tokens consistently improve performance, and combining them with the proposed compact dual-stream architecture further improves the results.

	Registers configuration			FID at Epoch		
	Size	Start	End	40	80	120
w/ reg.	32	4	11	**37.7**	**9.59**	**6.45**
	32	4	9	36.9	9.95	6.65
	32	0	11	59.7	19.3	11.9
	32	0	4	62.4	19.6	12.3
	16	4	11	40.4	10.2	6.80
	16	0	11	62.4	18.6	11.3
	4	4	11	46.3	12.8	8.37
	4	0	11	54.8	16.2	10.4
w/o reg.	–	–	–	60.6	18.4	11.1

Table 18: **Registers are effective in deeper layers.** pDiT-B/16 benefits from registers only after the first 4 layers; early-layer registers match the no-register baseline. More registers improve quality.

with 32 registers performing best. Comparing the 4–11 and 4–9 settings further suggests that the latest layers contribute relatively little to register effectiveness.

In Figures 29 and 30, we present qualitative generation results of our approach, where CFG and RG are used together for the JiT-H/16 model, achieving an FID of 1.80.

E LIMITATIONS

Our study focuses mainly on ImageNet DiTs and a set of pretrained text-to-image models. While the observed behavior is consistent across architectures, model sizes, resolutions, and few-step models, broader evaluation across more training recipes and datasets would further strengthen the conclusions. Although we show that registers improve intermediate representations and are closely connected to generation quality, we do not provide a comprehensive explanation of why they help.

Register Guidance also introduces additional hyperparameters, such as the register-drop probability p and guidance scale w_{rg}, which may need to be tuned for new tasks, datasets, or architectures. Moreover, the best-performing RG+CFG setup requires additional model evaluations, increasing the number of function evaluations and limiting sampling speed compared to the original model.

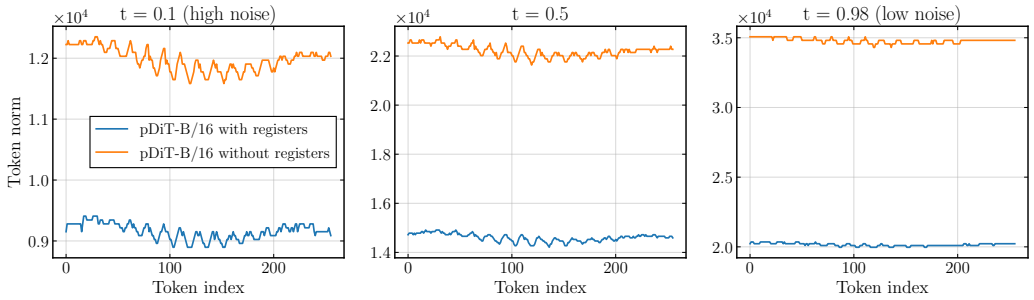


Figure 8: **Register tokens consistently reduce feature norms across patch tokens.** We measure feature norms for patch tokens only, excluding registers, at different diffusion timesteps.

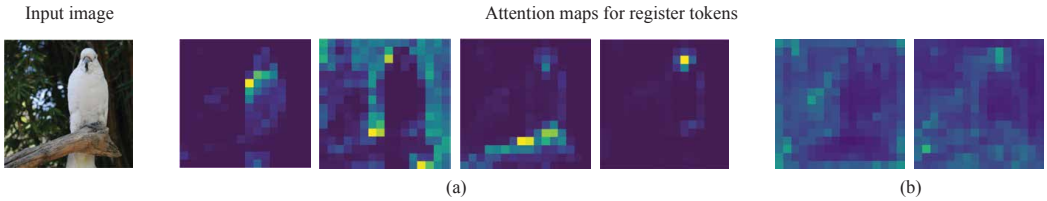


Figure 9: **(a) Registers with high probing accuracy encode diverse semantic information, whereas (b) low-accuracy norm sinks do not.** Visualizing register attention maps shows that some registers attend to distinct semantic regions, such as foreground objects and background areas. In contrast, low-accuracy norm sinks do not show meaningful semantic structure.

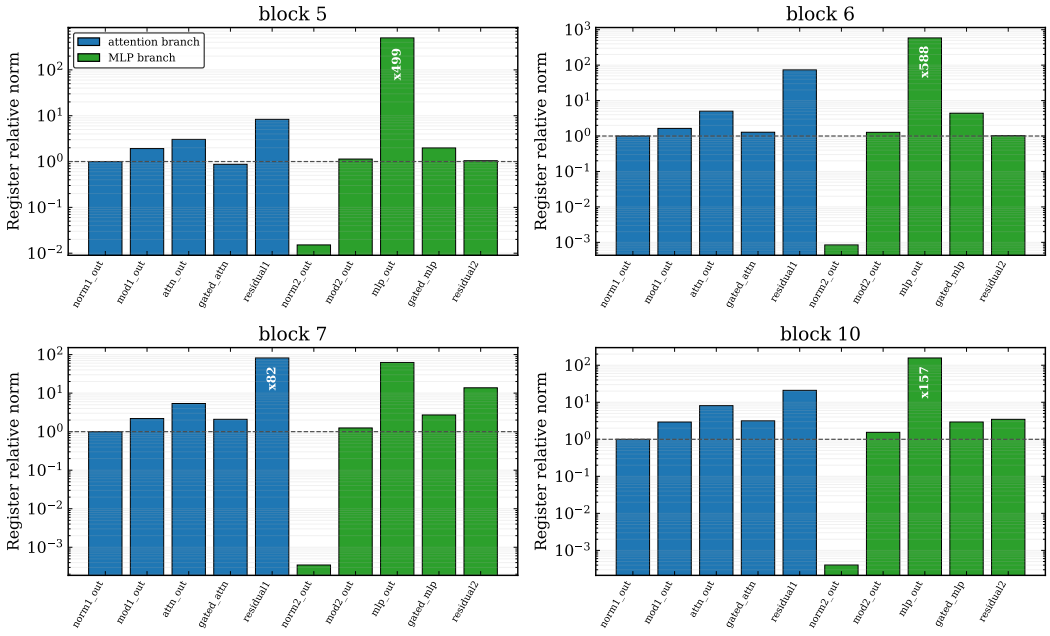


Figure 10: **High-norm register tokens are amplified mainly inside the MLP branch.** For blocks 5, 6, 7, 10 of pDiT-B/16 on ImageNet 256×256, we show the relative change in register-token feature norms across intra-block sub-layers. Attention and MLP sub-layers are shown separately. The largest norm amplification occurs at the MLP output, reaching up to ×499 in block 5, whereas the attention branch produces much smaller changes. This indicates that register-token outliers are primarily formed by the MLP rather than by attention.

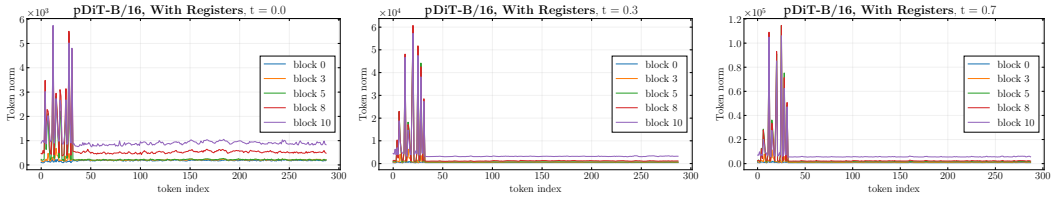


Figure 11: **High-norm outliers consistently emerge within register tokens across timesteps.** We visualize token-wise feature norms of pDiT-B/16 with registers for $t = 0.0$, 0.3, and 0.7, and observe the same behavior in all cases.

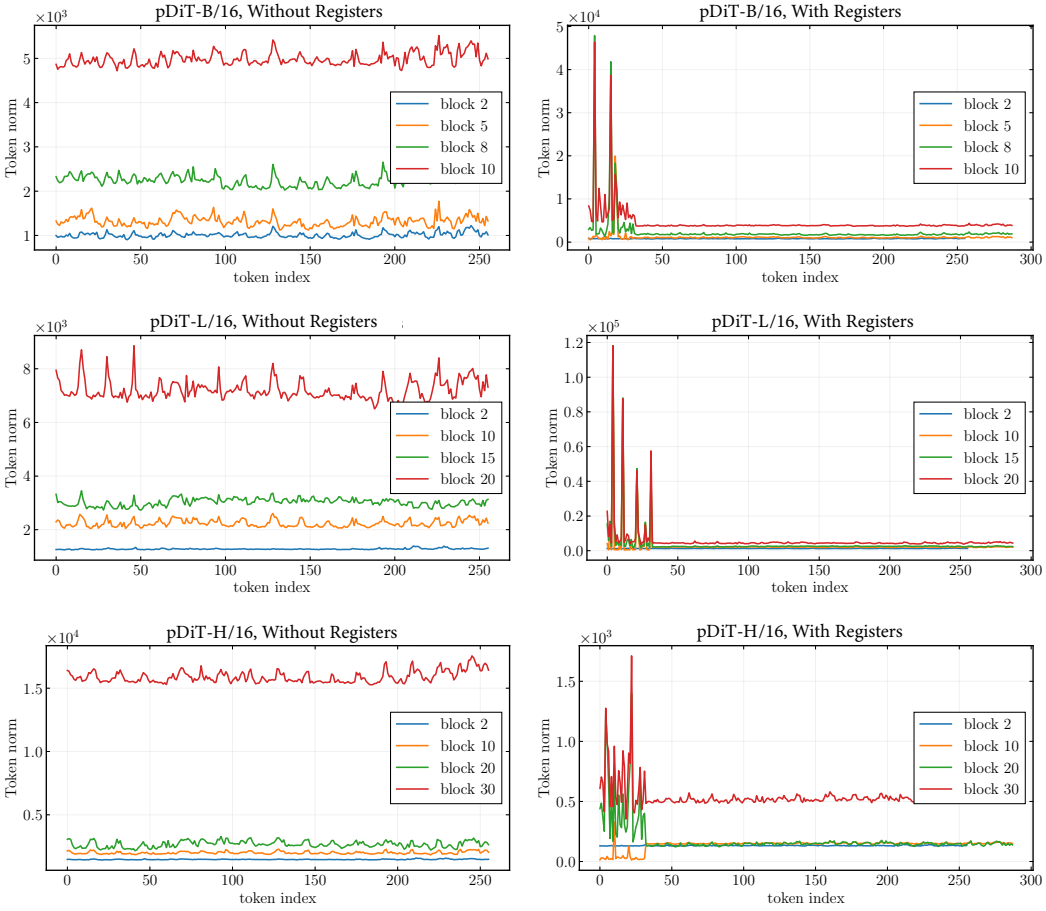


Figure 12: **Token-wise feature norms for pDiTs of varying scales on ImageNet** 256×256**, with and without registers.** Without registers, patch-token norms remain uniform across scales. Introducing registers leads to the emergence of high-norm outliers within the register tokens.

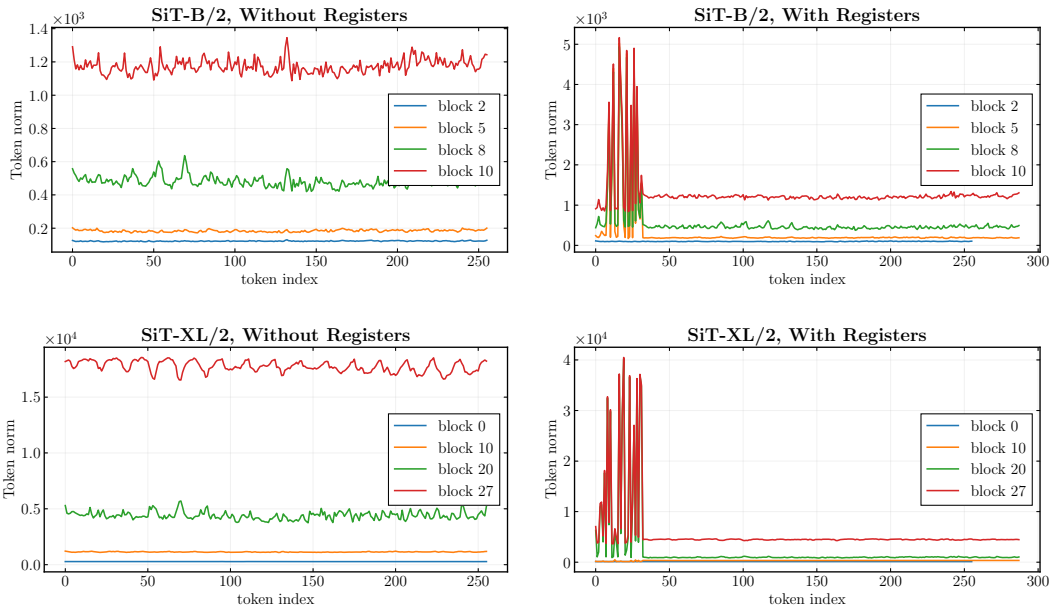


Figure 13: **Token-wise feature norms for VAE-space SiTs of varying scales on ImageNet** 256×256**, with and without registers.** SiTs without registers exhibit uniform patch-token norms across scales, while adding registers produces high-norm register tokens.

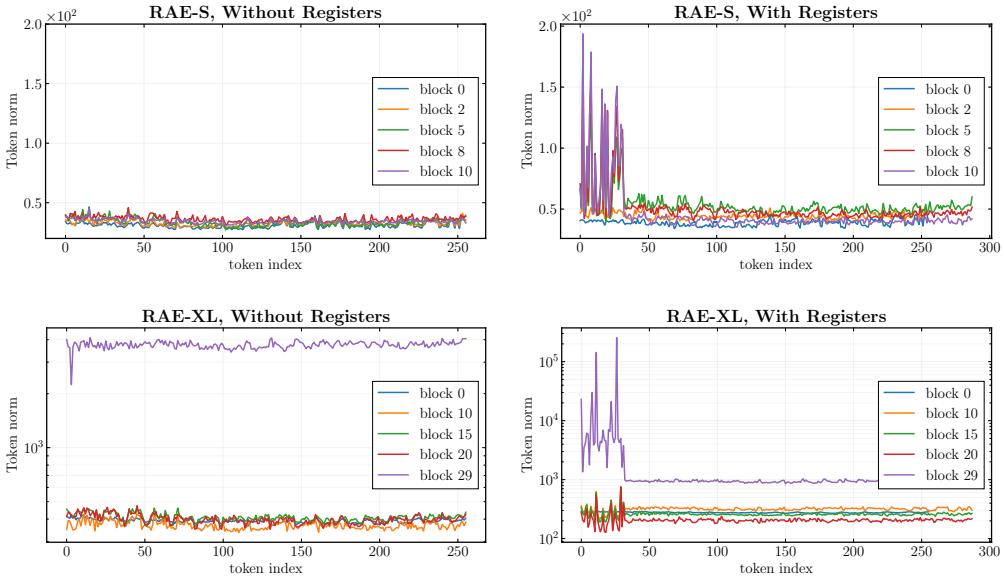


Figure 14: **Token-wise feature norms for DINOv2-space RAEs of varying scales on ImageNet** 256×256**, with and without registers.** RAEs without registers exhibit uniform patch-token norms across scales, while adding registers produces high-norm register tokens.

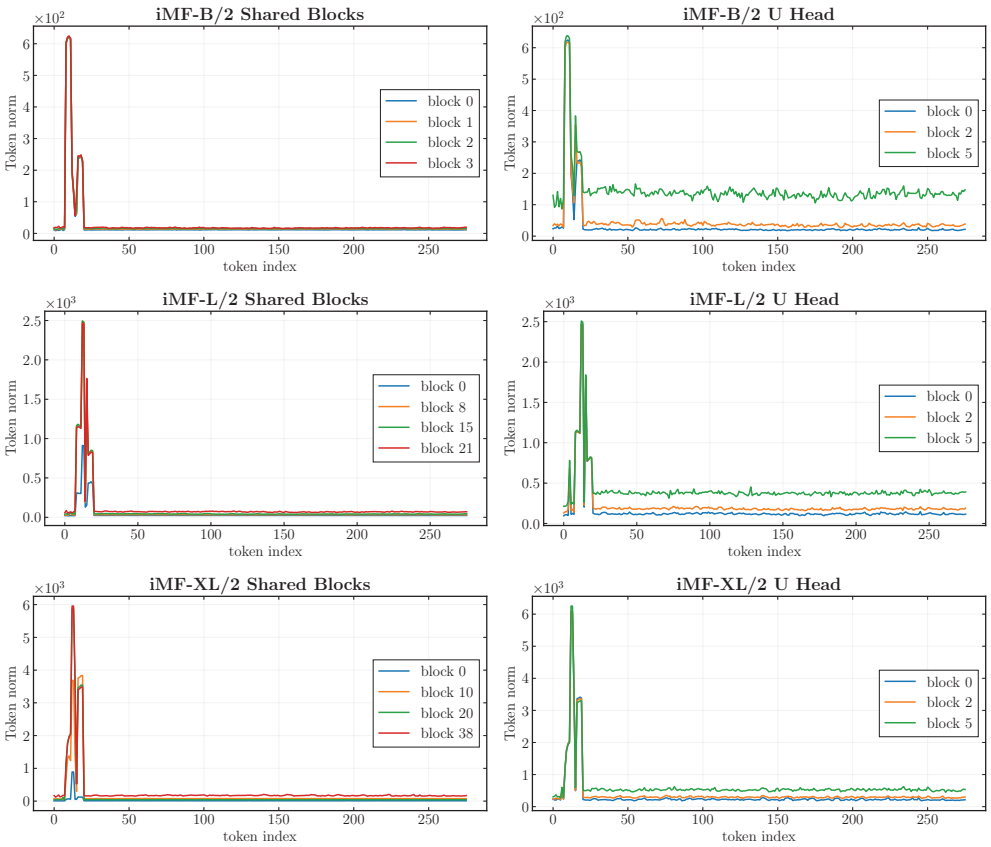


Figure 15: **Token-wise feature norms for improved MeanFlows (Geng et al., 2026b) of varying scales on ImageNet** 256×256 **with in-context conditioning.** We consider both shared blocks and blocks that predict the average velocity, and find high-norm tokens within the in-context tokens.

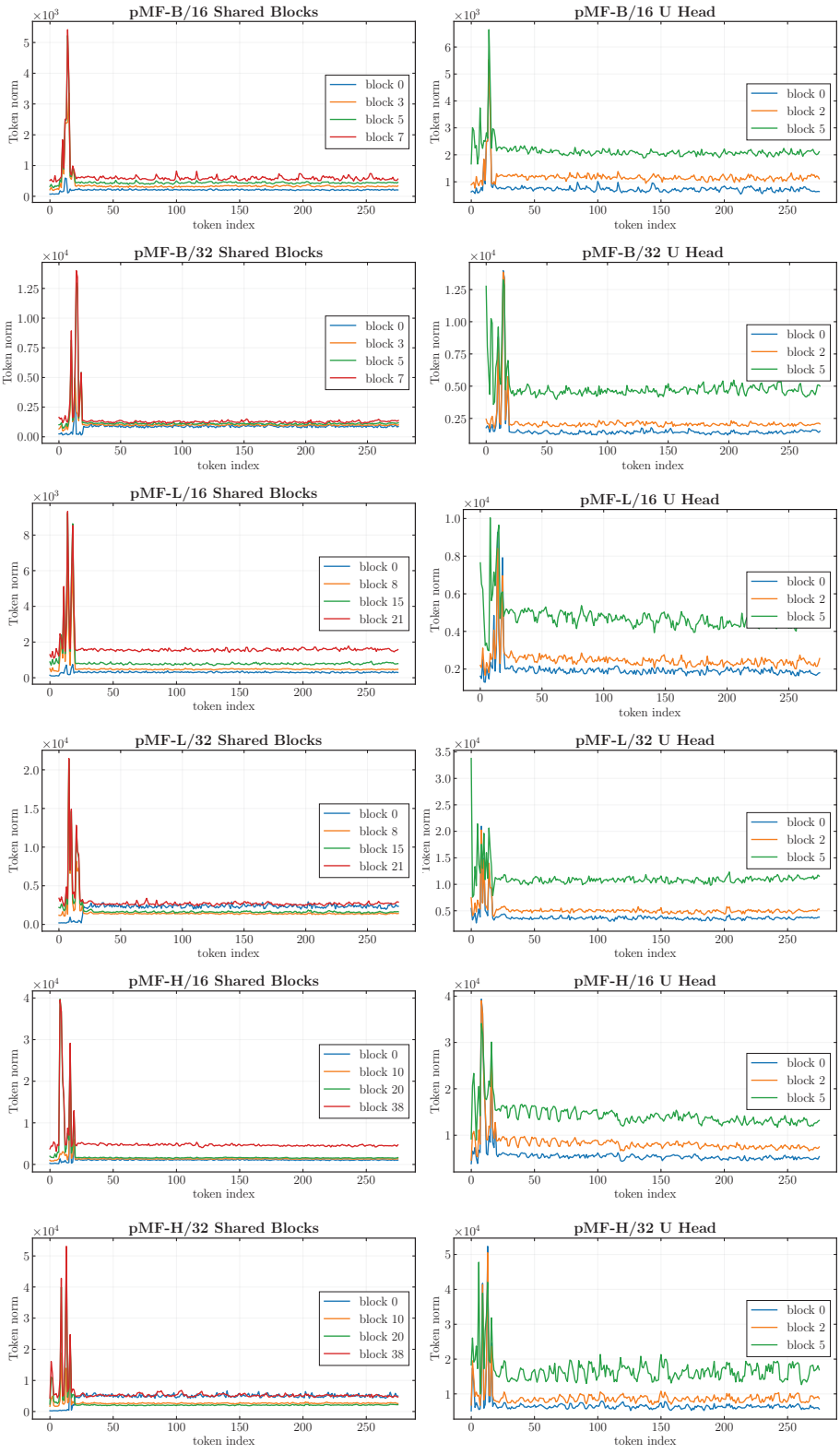


Figure 16: **Token-wise feature norms for pixel MeanFlows (Lu et al., 2026) of varying scales on ImageNet** 256×256 **and** 512×512 **with in-context conditioning.** We consider both shared blocks and blocks that predict the average velocity, and find high-norm tokens within the in-context tokens.

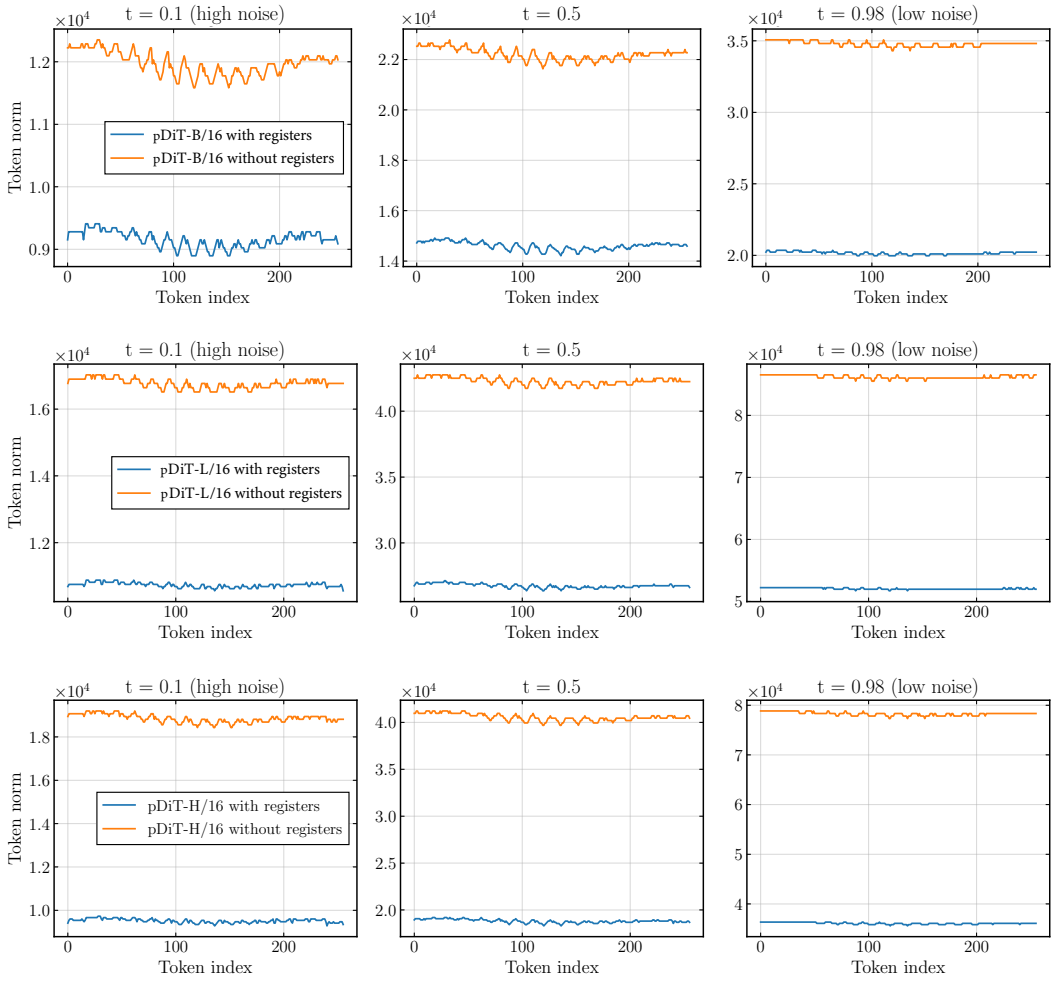


Figure 17: **Register tokens consistently reduce feature norms across patch tokens.** We measure feature norms for image tokens only (excluding register tokens) at three diffusion timesteps for pDiT models of different scales, and observe a consistent reduction in feature norms across nearly all tokens when register tokens are used.

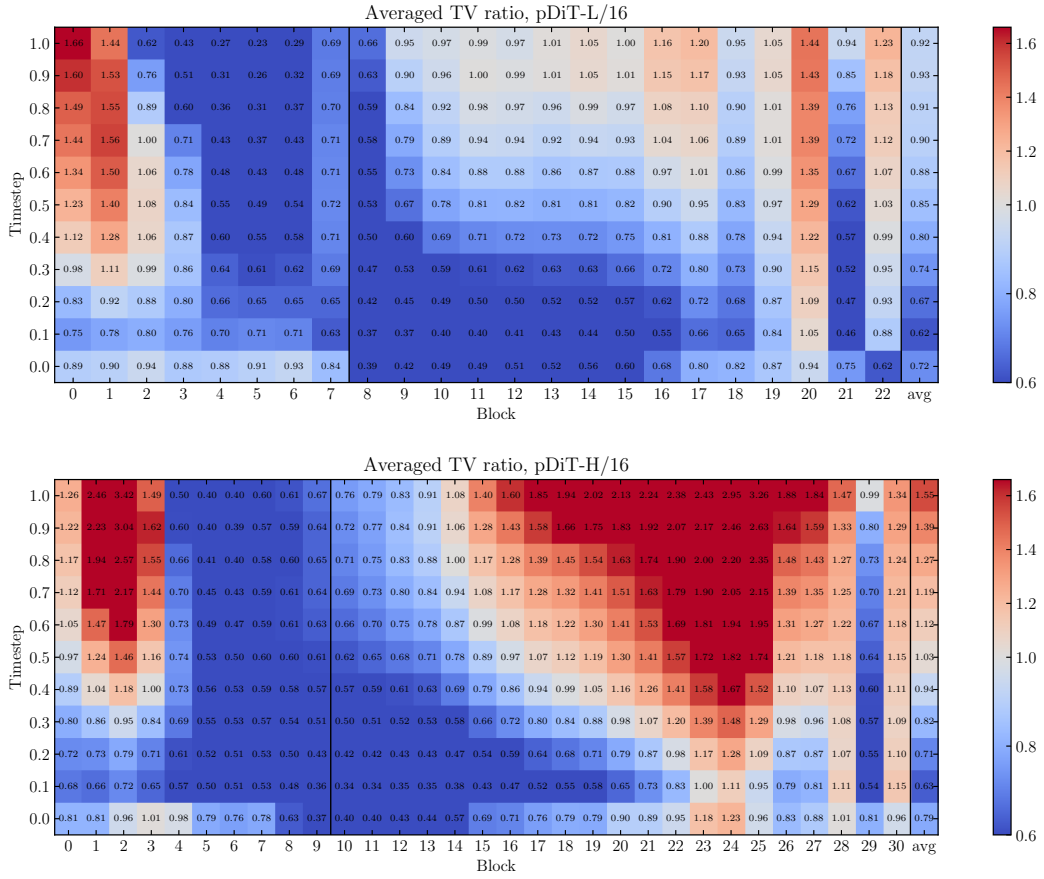


Figure 18: **Register tokens make intermediate representations cleaner by reducing noise.** We compute the Total Variation of intermediate features for models with and without register tokens. We report the ratio (with registers / without registers), where lower values indicate that models with registers produce smoother feature representations. We find that register tokens improve feature smoothness at high noise levels ($t \in [0, 0.2]$) for both pDiT-L/16 (top) and pDiT-H/16 (bottom) models.

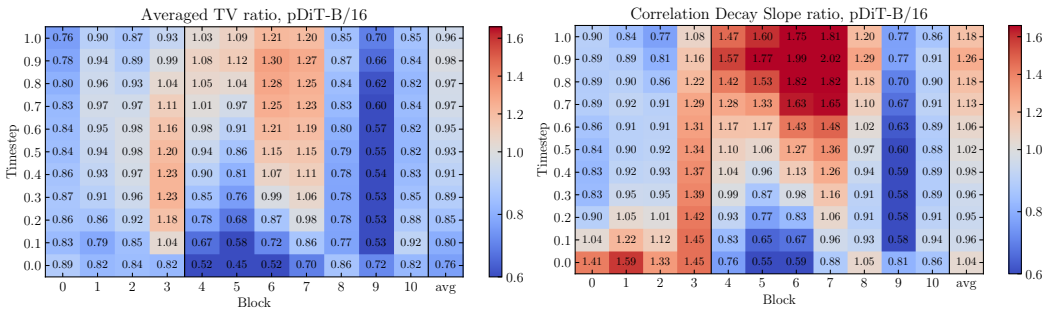


Figure 19: **Registers improve spatial organization at high noise levels.** In addition to the TV ratio (left), we also analyze the correlation decay slope (Singh et al., 2025) (right), where lower values indicate stronger spatial organization. Both metrics show the same trend: register tokens improve internal representations at high noise levels, starting from block 4, where the registers are introduced.

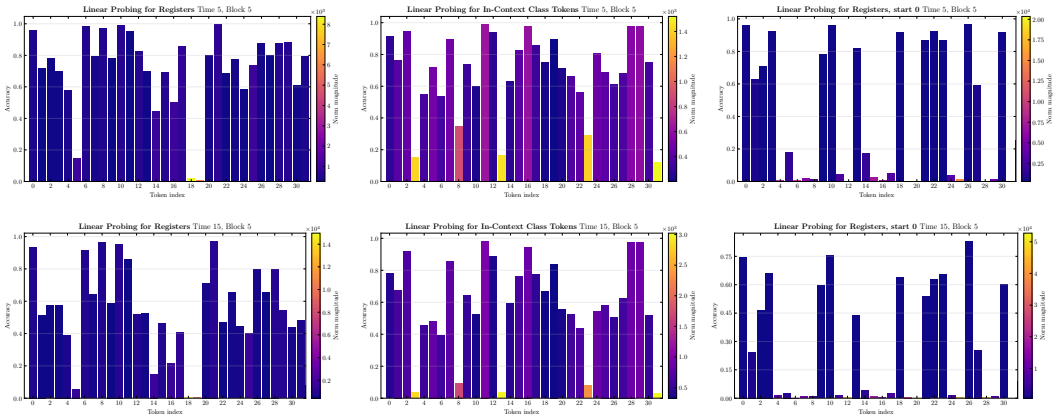


Figure 20: **Linear probing of register tokens under different configurations.** (Left) Standard register tokens introduced from the 4th layer; (Middle) Register tokens used as in-context class embeddings introduced from the 4th layer; (Right) Standard register tokens introduced from the 0th layer. Across different timesteps, we find that introducing registers from the earliest layers produces substantially less informative register tokens. In particular, we observe more low-norm (non-sink) tokens with poor linear probing accuracy.

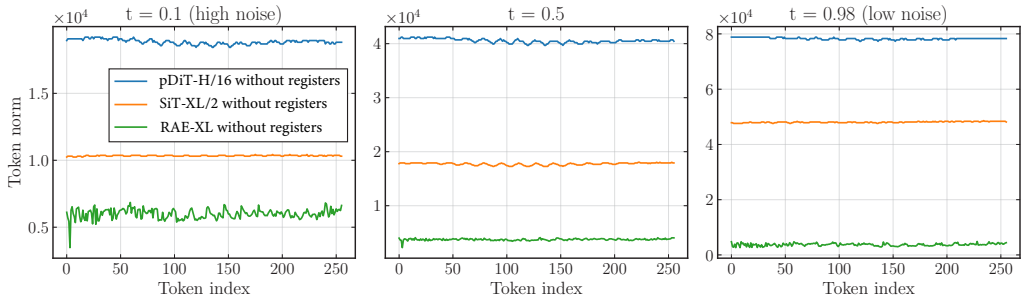


Figure 21: **Pixel-space pDiTs have the highest feature norms across all tokens for different timesteps compared to latent-space counterparts.** We compare token-wise feature-map norms for pDiT, SiT, and RAE models, all without register tokens.

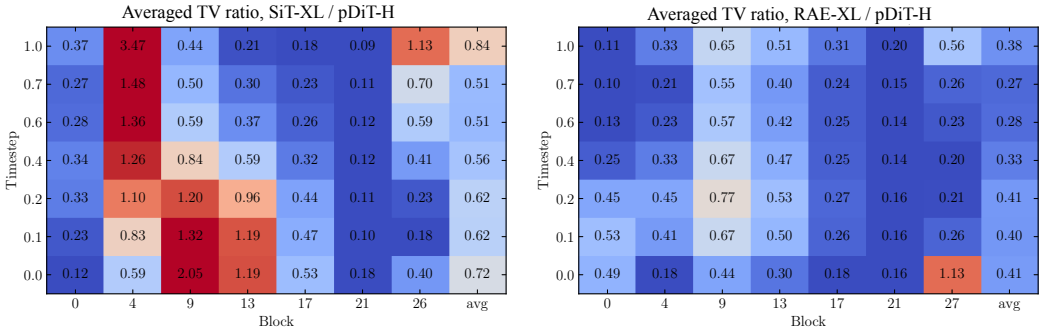


Figure 22: **Pixel-space pDiTs exhibit substantially higher Total Variation (TV) values than latent-space counterparts.** We compare the TV ratio of intermediate feature maps across timesteps and transformer blocks for pixel-space pDiTs (pDiT-H) and latent-space models (SiT-XL and RAE-XL) without registers. Pixel-space pDiTs consistently produce noisier intermediate representations.

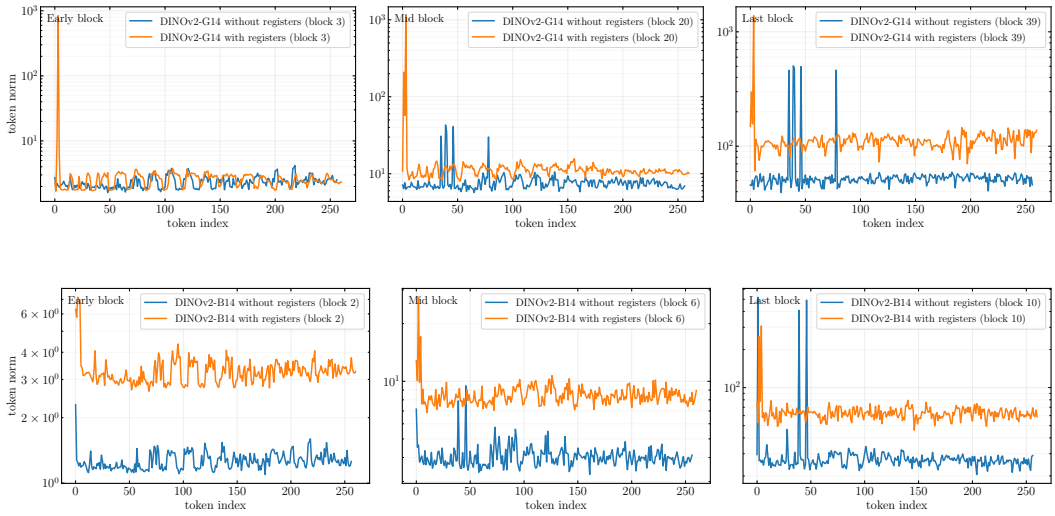


Figure 23: **For SSL ViTs such as DINOv2, register tokens do not reduce patch-token feature norms, unlike in DiTs.** We measure feature norms across all tokens for different blocks and model sizes of DINOv2, and observe that register tokens do not consistently reduce feature norms of patch tokens.

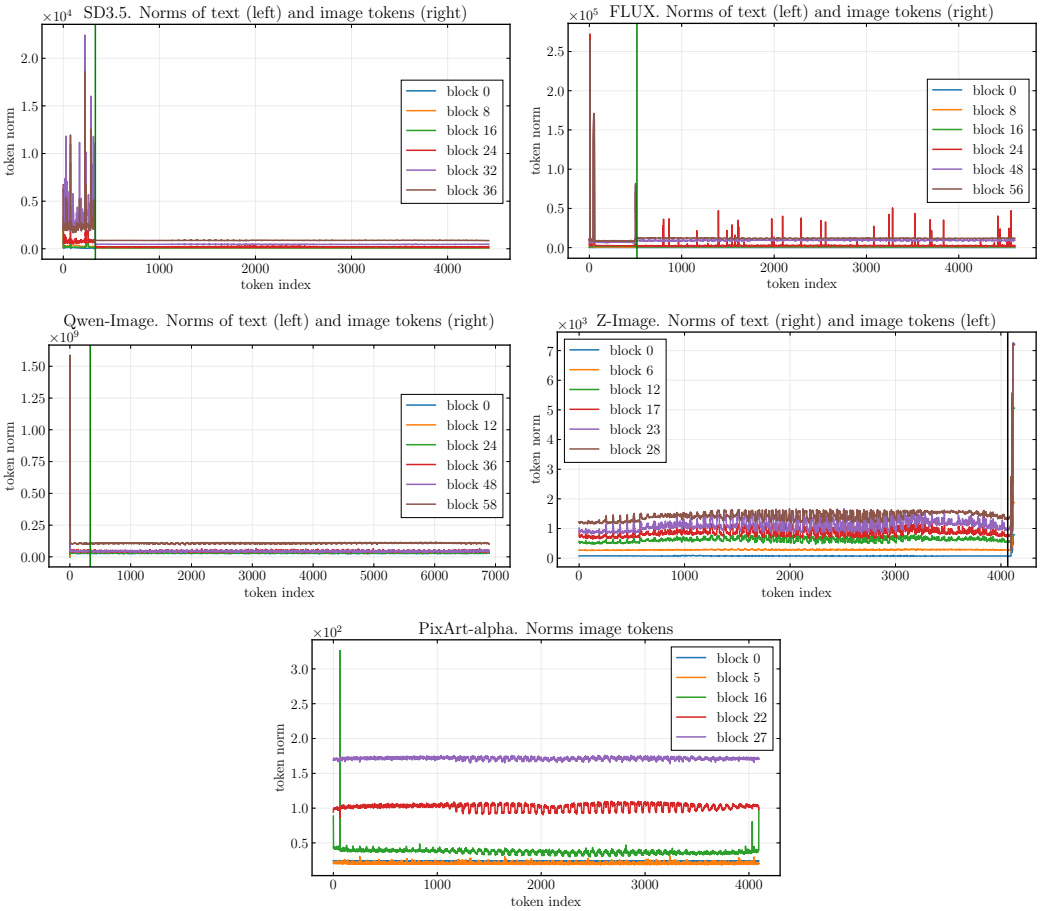


Figure 24: **Text sequences in text-to-image diffusion models exhibit behavior similar to register tokens in ImageNet-based DiTs: some tokens become high-norm outliers and potentially act as registers.** We measure token-wise feature norms in SD3.5 (left) and FLUX (right) for both text and image tokens. We observe that the outliers primarily emerge within the text sequence.

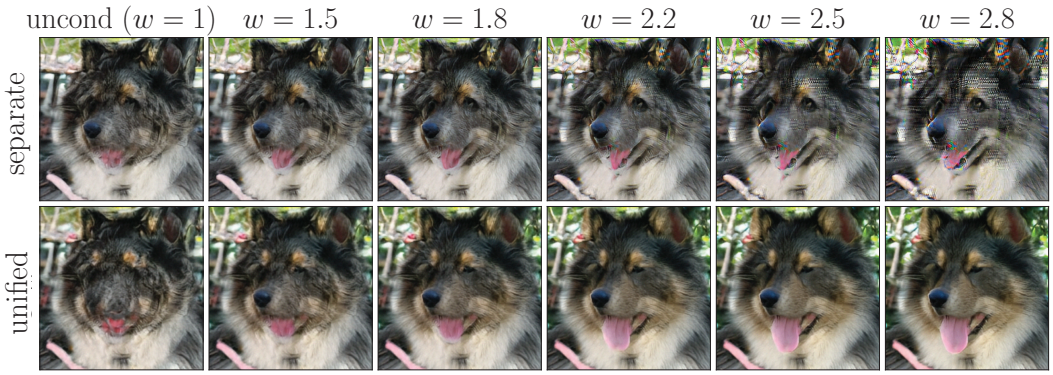


Figure 25: **The influence of registers on improving image quality can be enhanced.** We apply Register Guidance using predictions with and without registers for different guidance scales w. Using two separate models introduces artifacts at large w (top), whereas the unified model preserves the overall content and improves structure and details without introducing artifacts (bottom).

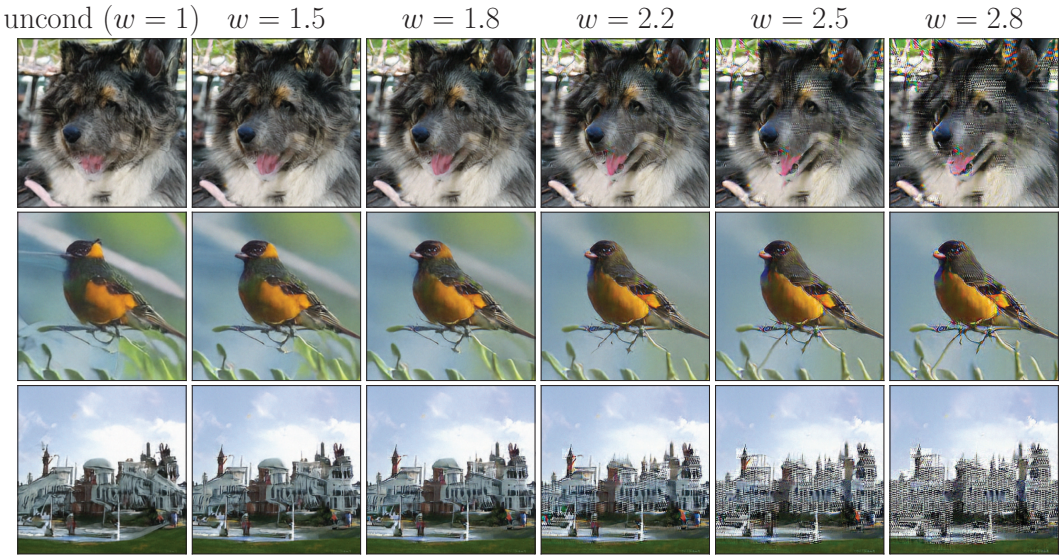


Figure 26: **Artifacts from Register Guidance with separate models.** We apply Register Guidance using predictions from separately trained models with and without registers. As the guidance scale increases, visible artifacts appear, suggesting imperfect alignment between the two models.

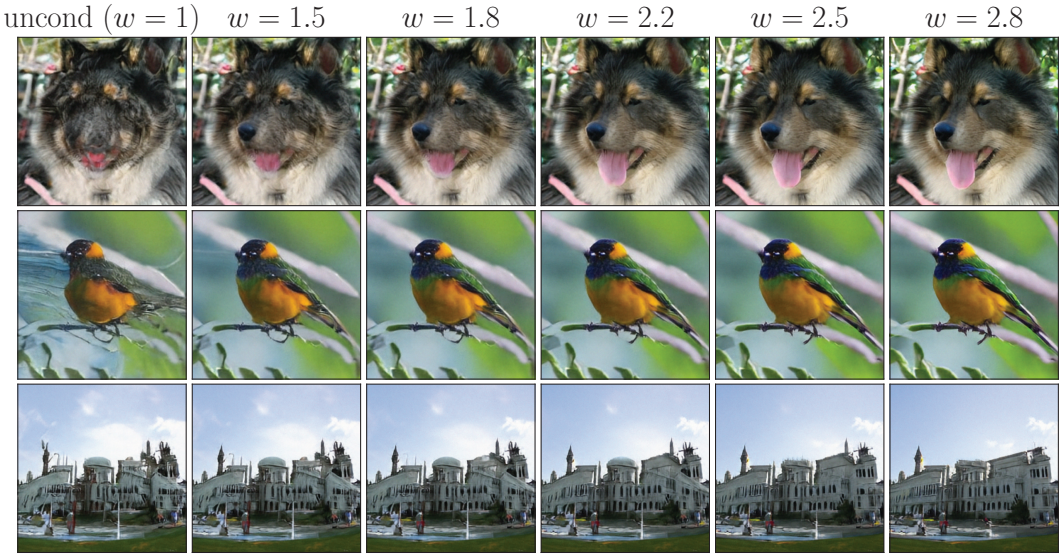


Figure 27: **Register Guidance with a single model removes artifacts.** We apply Register Guidance using a single model trained to operate both with and without registers. Unlike guidance with separate models, increasing the guidance scale improves structure and details without introducing visible artifacts.

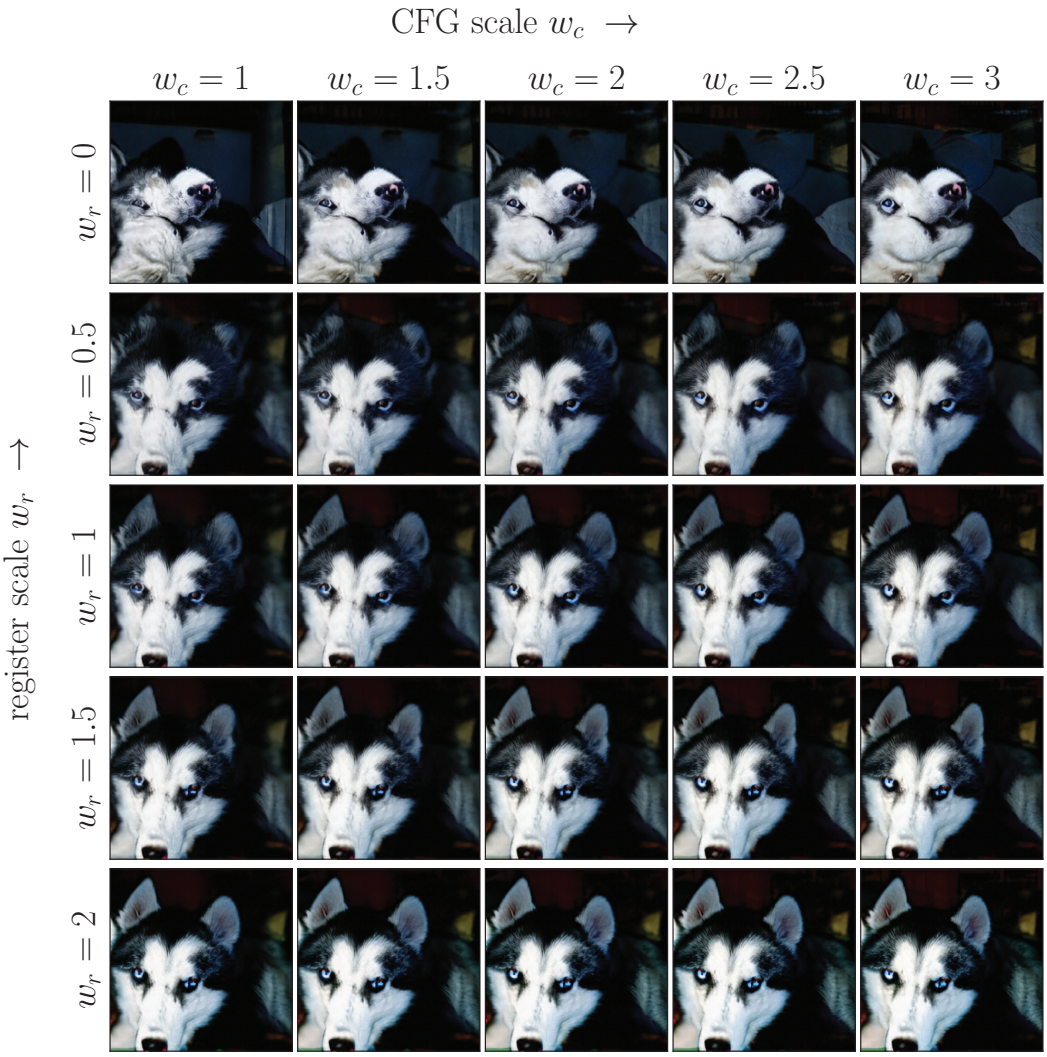


Figure 28: **Visual effect of combining CFG and RG.** We vary the CFG scale w_c horizontally and the RG scale w_r vertically. CFG strengthens class conditioning, while RG improves structure and visual details, showing that the two guidance signals are complementary.


class 012: house finch, linnet, Carpodacus mexicanus


class 014: indigo bunting, indigo finch, indigo bird, Passerina cyanea


class 042: agama


class 081: ptarmigan


class 107: jellyfish


class 108: sea anemone, anemone


class 110: flatworm, platyhelminth


class 117: chambered nautilus, pearly nautilus, nautilus


class 130: flamingo

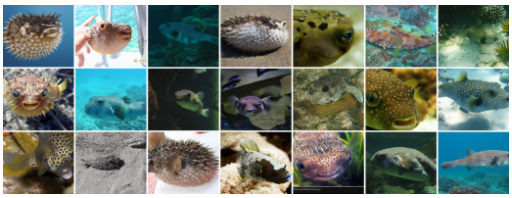

class 279: Arctic fox, white fox, Alopex lagopus

Figure 29: **Uncurated samples on ImageNet** 256×256 **using JiT-H/16.** We show images generated with RG+CFG, which achieves the reported FID of 1.80.


class 288: leopard, Panthera pardus


class 309: bee


class 349: bighorn, bighorn sheep, cimarron, Rocky Mountain bighorn


class 397: puffer, pufferfish, blowfish, globefish


class 425: barn


class 448: birdhouse


class 453: bookcase


class 458: brass, memorial tablet, plaque


class 495: china cabinet, china closet


class 500: cliff dwelling

Figure 30: **Uncurated samples on ImageNet** 256×256 **using JiT-H/16.** We show images generated with RG+CFG, which achieves the reported FID of 1.80.