Registers Matter for Pixel-space Diffusion Transformers

Anonymous Authors[1]

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. Interestingly, register tokens significantly improve convergence and generation quality of pixel-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 insights, we propose a parameter-efficient, register-specialized dual-stream architecture that improves pixel-space generation quality with modest computational and memory overhead.

1. Introduction

Vision Transformers (ViTs) (Dosovitskiy et al., 2020; Liu et al., 2021; 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., 2025; Wang et al., 2024). As these artifacts lead to less interpretable attention maps and poorer 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 further moved beyond latent diffusion (Rombach et al., 2022; Podell et al., 2023) toward training directly in pixel space (Li & He, 2025; Yu et al., 2025; Lu et al., 2026), eliminating the need for pretrained autoencoders. This progress brings Diffusion Transformers (DiTs) closer to ViTs (Dosovitskiy et al., 2020; Liu et al., 2021; Touvron et al., 2021) and motivates two natural 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 related to a broader line of work on attention sink tokens in transformer models (Su et al., 2026), including several generative settings (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 research direction to App. A. In contrast to prior work, we focus on image diffusion transformers, where the presence and role of register-like tokens remain underexplored.

Contributions. We find that, unlike ViTs, diffusion transformers, both in latent and pixel spaces, do not exhibit 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 this absence of outliers, adding register tokens to DiTs leads to the emergence of high-norm tokens within the registers themselves.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

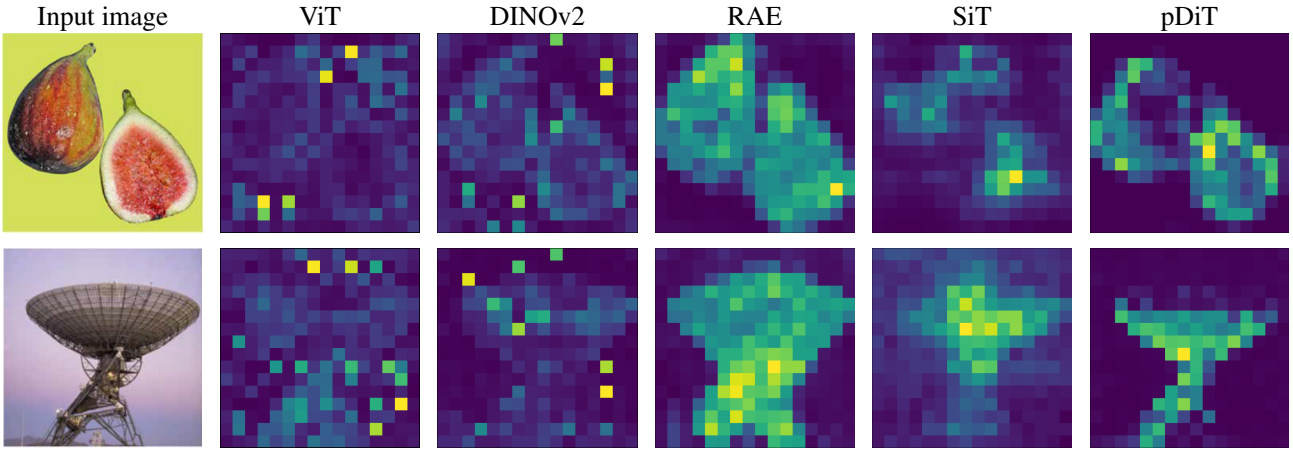

Figure 1. **Diffusion transformers do not exhibit attention-map outliers.** Unlike ViTs, where attention-map anomalies typically appear in low-information regions (e.g., background), DiT attention remains focused on the main objects.

Importantly, we find that the impact of register tokens differs across training spaces: pixel-space models benefit the most, whereas latent-space models exhibit only moderate improvements or even degraded performance.

Accordingly, we focus our study on pixel-space DiTs and find that registers benefit DiTs through a mechanism distinct from their role in ViTs. Our analysis shows that registers consistently reduce the norms of patch-token features and produce smoother intermediate feature maps, especially for 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 class tokens for in-context conditioning and achieve substantial performance gains. We find that these gains may arise from register-like behavior rather than from the additional class information. In particular, the in-context tokens exhibit behavior similar to register tokens: some encode global semantic information, while others become norm sinks.

From a practical standpoint, we introduce a parameter-efficient dual-stream architecture that treats register and patch tokens as distinct streams while sharing attention and most transformer parameters. This specialization improves generation quality at similar inference cost with only a $\sim$14% increase in model size.

2. Register Tokens for Image Diffusion Transformers

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

For diffusion models, 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; Albergo et al., 2025) on ImageNet (Deng et al., 2009) at resolution 256×256 with patch size 16. We additionally analyze latent-space architectures, SiT (Ma et al., 2024) and RAE (Zheng et al., 2026), using their original backbone designs and training pipelines. Generation 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 for different spaces by comparing their attention maps to those of ViTs. As shown in Figure 1, DiTs do not exhibit the artifacts observed in ViTs. In particular, their attention maps remain free of anomalies in low-information regions, which in ViTs are typically associated with unusually high-norm tokens.

This observation is further supported by Figure 2, which reports token-wise feature norms across layers for pDiT. In

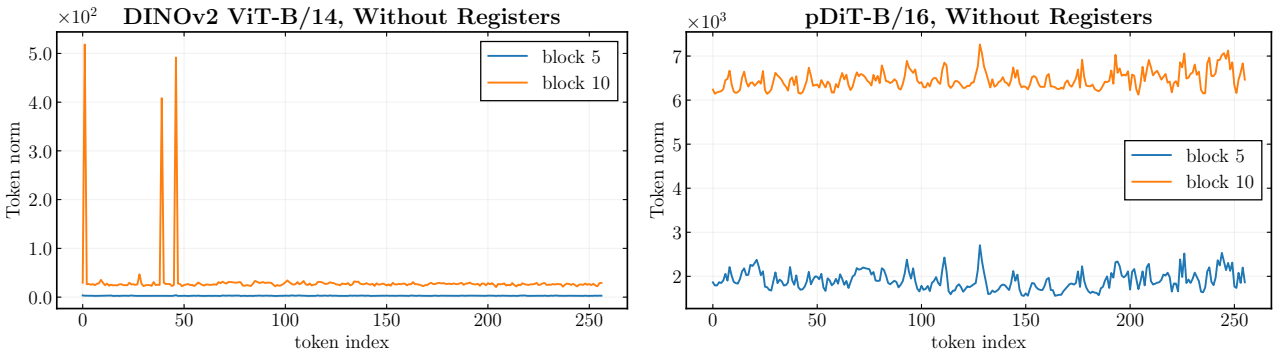

Figure 2. **Without Registers.** (a) In DINOv2, anomalies are localized to few image tokens, which exhibit significantly higher norms than others. (b) In contrast, no outliers are observed for pDiTs, suggesting that registers may be unnecessary in this case.

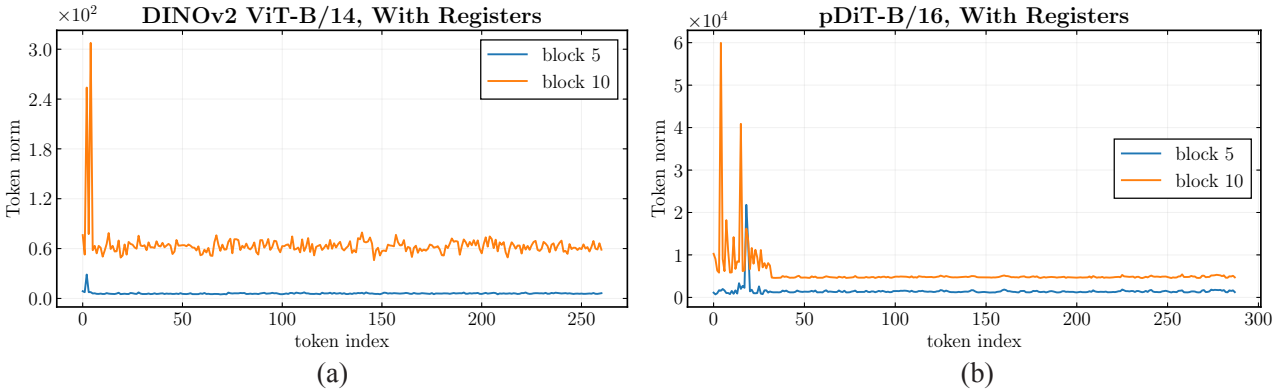

Figure 3. **With Registers.** (a) As expected, introducing register tokens in DINOv2 shifts high-norm outliers into these tokens. (b) Interestingly, pDiTs also exhibit high-norm tokens in the added registers, even though such outliers are absent without registers.

contrast to DINOv2, where a few tokens attain significantly larger norms, pDiTs exhibit a nearly uniform distribution of patch-token norms. As shown in Figure 10, 11 this behavior consistently holds across larger model variants and latent-space architectures.

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 3, 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 9, 10, 11 show that it consistently holds across different timesteps and extends to latent-space models as well.

Second, as shown in Table 1, introducing register tokens in pDiTs leads to consistent improvements in generation quality across model sizes. However, Table 2 shows that registers provide significantly smaller gains in latent-space models, a phenomenon we discuss in Section 2.4.

Model	Epoch	w/o reg.	w/ reg.	w/ in-context
pDiT-B/16, 131M	200	7.39	**5.30**	4.71
	600	4.80	**3.80**	3.71
pDiT-L/16, 459M	200	4.13	**3.17**	2.95
	600	2.80	**2.47**	2.43
pDiT-H/16, 953M	200	3.31	**2.61**	2.51
	600	2.62	**2.35**	2.23

Table 1. **Generation quality (FID) of pDiTs w/ and w/o register tokens.** Registers consistently improve FID across model sizes and training epochs on ImageNet 256×256. The shaded column reports in-context class conditioning.

Based on these observations, pDiTs benefit from the presence of outliers but lack a mechanism to accommodate them in the absence of register tokens. We attribute this to all patch tokens in DiTs contributing to the loss, leaving no capacity for specialized outlier tokens. In contrast, in ViTs, only a few tokens participate in the loss, allowing high-norm

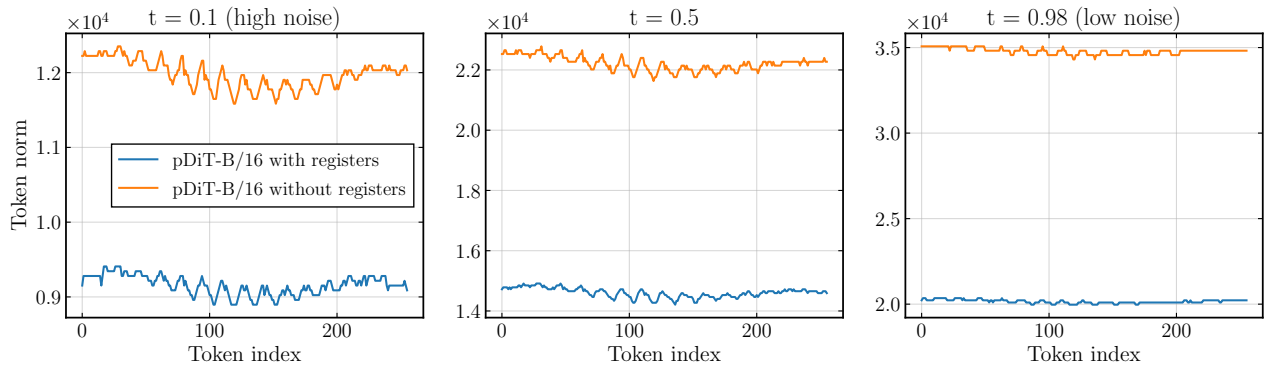

Figure 4. **Register tokens consistently reduce feature norms across patch tokens.** We measure feature norms for image tokens only (excluding registers) at three diffusion timesteps and observe a consistent reduction across all tokens when registers are used.

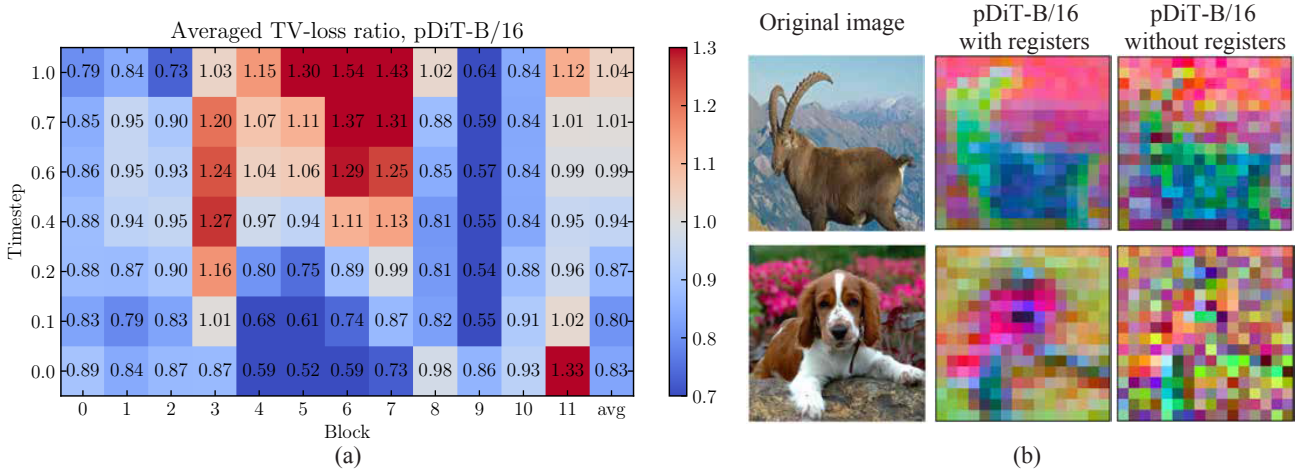

Figure 5. **Register tokens improve intermediate representations**. **(a)** We compute the Total Variation (TV) of intermediate features for models with and without register tokens. We report the ratio (with / without registers), where lower values indicate that the model with registers produces smoother features. We find that registers improve feature smoothness at high noise levels ($t \in [0, 0.2]$). **(b)** We visualize feature maps using PCA, which qualitatively confirms this effect.

outliers to emerge among image tokens. When register tokens are introduced, they similarly do not participate in the loss, thereby providing convenient slots for such high-norm outliers.

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 diffusion models, 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 4). Interestingly, DINOv2 register tokens do not exhibit this behavior for non-outlier patch tokens (see Figure 13).

A possible explanation for this behavior is that larger feature norms may reflect high local variability in hidden representations. In diffusion models, such variability may arise from the models' need 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), 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 ex-

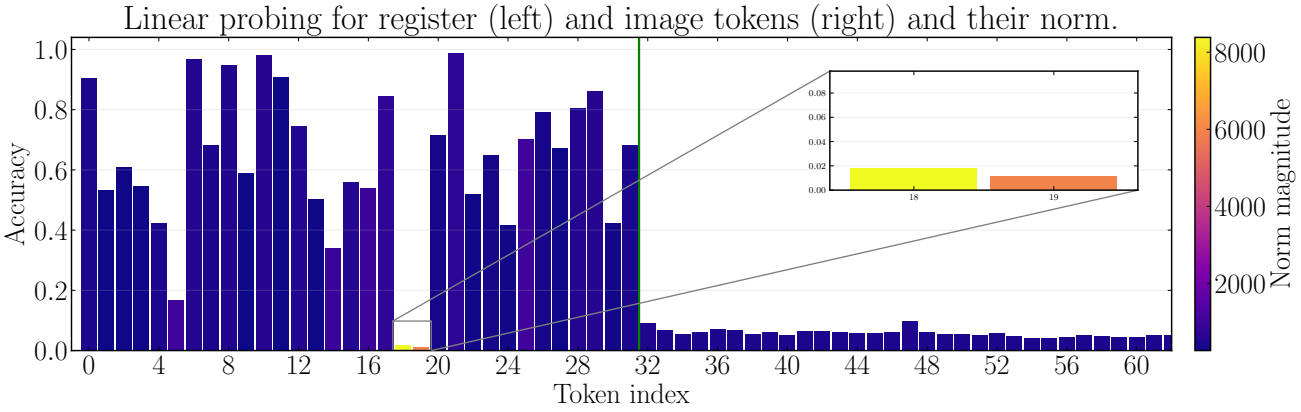

Figure 6. **Register tokens act as both global information carriers and norm sinks**. Linear probing reveals that low-norm register tokens encode meaningful global semantics and achieve strong classification accuracy, whereas low-accuracy registers exhibit extremely large norms, suggesting that they primarily function as norm sinks that absorb magnitude from patch tokens.

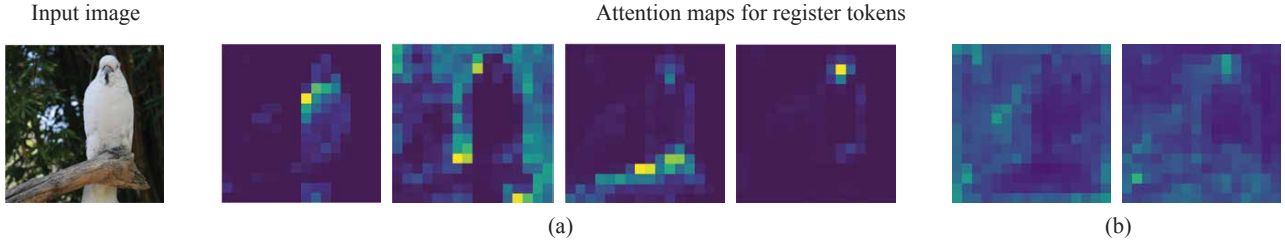

Figure 7. **(a) Registers with high probing accuracy encode diverse semantic information about an image, whereas (b) low-accuracy norm sinks do not.** We visualize attention maps for register tokens and observe that some attend to distinct semantic regions, such as foreground objects and background areas. In contrast, norm sinks with low probing accuracy do not exhibit meaningful semantic structure.

tract 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 5(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 5(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), as they play a central role in forming the main image content. Therefore, the improved representations induced by register tokens at these stages provide a plausible explanation for the observed quality gains.

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 6 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 further validate that non-sink tokens encode diverse information, we visualize attention maps of different register tokens (Figure 7a). 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 do not encode meaningful semantic information (Figure 7b).

	RAE-space DDT backbone	VAE-space SiT backbone	Pixel-space pDiT backbone
		Base size	
w/ reg.	7.48	**9.40**	**5.30**
w/o reg.	**6.58**	10.40	7.39
		Large size	
w/ reg.	4.44	**2.38**	**2.61**
w/o reg.	**3.91**	2.53	3.31

Table 2. **Registers are more effective in pixel-space.** Generation quality for models with and without registers across different training spaces and model sizes. Registers yield the largest improvements in pixel space, moderate gains in VAE space, and degraded performance 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 only 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.

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.

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, and, interestingly, degrade performance in RAE-based models.

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 6 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 17, 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, further supporting this observation.

These findings further suggest that training DMs in pixel space is inherently more challenging and requires stronger regularization. In contrast, latent spaces are more structured and lower-dimensional, where imperceptible noise and fine-grained details are compressed. 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.

2.5. Registers Are Effective in Deeper Layers

Next, we ablate both the number of register tokens and the transformer blocks in which they are introduced. 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 8.

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 configuration 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 15 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 register tokens than ViTs. While 4 registers are typically sufficient for ViTs (Darcet et al., 2023; Siméoni et al., 2025), pDiT-

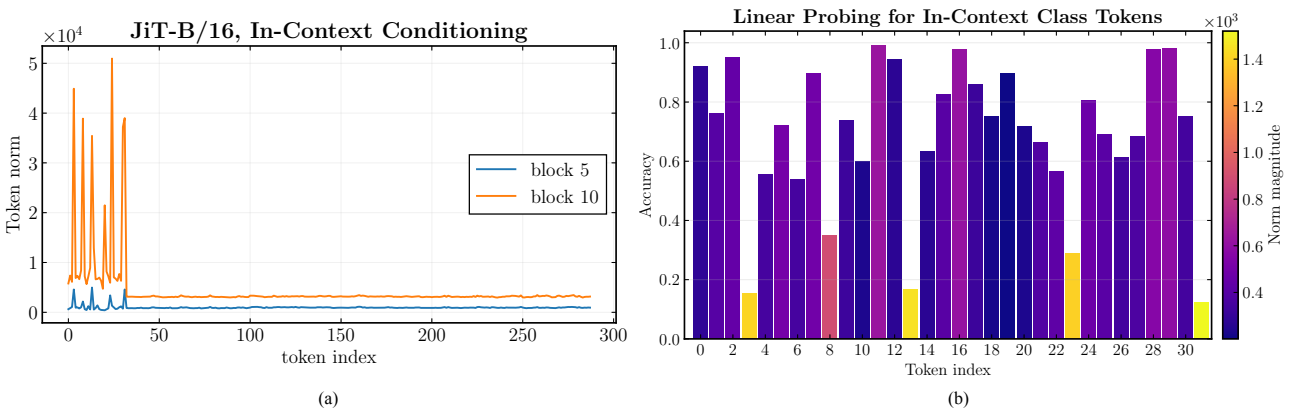

Figure 8. **In-context class tokens act as registers.** (a) Certain tokens acquire disproportionately high feature norms, functioning as norm sinks. (b) Some tokens encode broad global information, rather than purely class-specific features as originally intended.

B achieves the best performance with 32 registers. This suggests that registers in DiTs are required to solve a more challenging task, necessitating a larger number of 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 such in-context tokens implicitly function as register tokens.

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 8a) and evaluate their representations using linear probing (Figure 8b). 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 registers rather than from the additional class information, 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 note that JiT (Li & He, 2025) introduces in-context tokens only in deeper layers, which motivates our ablation study on the register starting layer in Section 2.5.

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 18), suggesting they may act as implicit registers.

3. Parameter-Efficient Dual-Stream Architecture

Method. Our analysis shows that register and patch tokens play distinct roles, yet pixel-space architectures (Li & He, 2025; Lu et al., 2026) share all parameters across them, which may be suboptimal. Large-scale models with long appended sequences, such as text, often use double-stream architectures (Esser et al., 2024) with modality-specific parameters and attention-based interaction, but naive duplication would double the parameter count. We therefore propose a parameter-efficient double-stream design that specializes register processing without full duplication.

Our approach builds on 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.

For `RMSNorm`, which has few parameters, we use separate parameters for register and patch tokens with negligible cost (Snippet 1).

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

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 4. **Ablation of compact dual-stream designs.** The best compact design (green), dualizes `adaLN`, `MLP`, and `RMSNorm`, but keeps `Attention` shared. (gray) denotes the single-stream JiT baseline.

adaLN	MLP	Attn	RMSNorm	Params (M)	FID
				149M	3.41
✓	✓	✓	✓	216M	**3.34**
✓				172M	3.49
	✓			174M	3.35

Table 5. **Compact vs fully dual-stream designs.** Starting from our best compact design in Table 4 (green), compact-dual modules are selectively replaced with fully duplicated ones. A fully duplicated `MLP` alone recovers nearly the same performance, suggesting that most benefits of full decoupling come from the `MLP` layers.

and patch tokens (Snippet 2).

For `adaLN` and `Attention`, we use lightweight LoRA adaptations (Hu et al., 2022; Marouani et al., 2026) applied only to register tokens, while shared parameters process all tokens. In `adaLN`, shared normalization parameters are refined for registers via a LoRA branch (Snippet 3). In `Attention`, QKV projections are shared, LoRA modifies only register-token representations, and 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 $\sim$14%, compared to $\sim$65% for naive duplication, while maintaining comparable performance with minimal runtime overhead.

Results. We first study which components should be decoupled in the 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 4 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 performance from 3.71 to 3.48. However, keeping `Attention` shared slightly improves performance while also reducing the parameter count, suggesting that `Attention` decoupling is not necessary in this setting. In contrast, `MLP` appears essential: sharing it degrades FID to 3.81. Finally, dual `adaLN` and `RMSNorm` are also important, as removing them hurts performance, increasing FID to 3.56 and 3.53, respectively. Based on these results, our final compact architecture dualizes `MLP`, `adaLN`, and `RMSNorm` modules, while sharing `Attention` (green-shaded row).

Next, we compare our compact design with a naive fully dual-stream architecture, where all layers are duplicated. Table 5 shows that the fully dual design provides only a modest improvement (3.41 $\rightarrow$ 3.34) while requiring much more parameters (149M $\rightarrow$ 216M).

To identify which components benefit most from full duplication, we selectively replace compact-dual modules with fully duplicated ones. For example, the 174M-parameter model uses a fully dual `MLP` while keeping other layers in their compact-dual form. Interestingly, this setup achieves nearly the same performance as the fully dual model (3.35 vs 3.34) while using significantly fewer parameters. This further highlights that most of the gains from full decoupling come from the `MLP` layers.

Finally, 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 9 and 10. Our approach consistently improves generation quality across model scales and image resolutions while introducing only a small parameter increase and negligible runtime overhead.

We believe our compact strategy can provide greater benefits at larger scales, where dual-stream architectures achieve significant quality improvements. In these settings, the appended sequence, e.g., text, becomes substantially larger than the small set of register tokens. However, existing dual-stream designs still rely on largely naive duplication, leading to significant parameter overhead. We provide more discussion in App. F.

References

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

Amir, S., Gandelsman, Y., Bagon, S., and Dekel, T. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

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

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, Y., Yan, Z., Zhou, C., Dai, B., and Luo, A. F. Vision transformers with self-distilled registers. *arXiv preprint arXiv:2505.21501*, 2025.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Gu, X., Pang, T., Du, C., Liu, Q., Zhang, F., Du, C., Wang, Y., and Lin, M. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., and Freeman, W. T. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., and Salimans, T. 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.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jamal, A., Tan, M., Saputra, C. A. N., Huynh, Q., Zhu, K., and Mari, A. Diffusion transformers use sink registers. In *Second Workshop on XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*, 2026.

Jiang, N., Dravid, A., Efros, A., and Gandelsman, Y. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025.

Labs, B. F. Flux. https://github.com/black-forest-labs/flux, 2024.

Lappe, A. and Giese, M. A. 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.

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

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

Liu, K., Hu, W., Xu, J., Shan, Y., and Lu, S. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lu, Y., Lu, S., Sun, Q., Zhao, H., Jiang, Z., Wang, X., Li, T., Geng, Z., and He, K. One-step latent-free image generation with pixel mean flows. *arXiv preprint arXiv:2601.22158*, 2026.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Marouani, A., Siméoni, O., Jégou, H., Bojanowski, P., and Vo, H. V. Revisiting [cls] and patch token interaction in vision transformers. *arXiv preprint arXiv:2602.08626*, 2026.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*, 2025.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

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

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

Rulli, M. E., Petruzzi, S., Michielon, E., Silvestri, F., Scardapane, S., and Devoto, A. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

Shi, C., Yu, Y., and Yang, S. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

Shin, J., Li, Z., Zhang, R., Zhu, J.-Y., Park, J., Shechtman, E., and Huang, X. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

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

Siméoni, O., Puy, G., Vo, H. V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., and Ponce, J. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., and Xie, S. What matters for representation alignment: Global information or spatial structure? *arXiv preprint arXiv:2512.10794*, 2025.

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

Su, Z., Zhang, H., Wu, W., Zhang, Y., Liu, Y., Xiao, H., Yang, Q., Sun, Y., Yang, R., Zhang, C., et al. Attention sink in transformers: A survey on utilization, interpretation, and mitigation. *arXiv preprint arXiv:2604.10098*, 2026.

Sun, M., Chen, X., Kolter, J. Z., and Liu, Z. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Wang, Y., Shen, X., Yuan, Y., Du, Y., Li, M., Hu, S. X., Crowley, J. L., and Vaufreydaz, D. 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.

Wen, Y., Wu, J., Jain, A., Goldstein, T., and Panda, A. Analysis of attention in video diffusion transformers. *arXiv preprint arXiv:2504.10317*, 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

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

Yi, J., Jang, W., Cho, P. H., Nam, J., Yoon, H., and Kim, S. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

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

Yona, I., Shumailov, I., Hayes, J., Barbero, F., and Gandelsman, Y. Interpreting the repeated token phenomenon in large language models. *arXiv preprint arXiv:2503.08908*, 2025.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Yu, Y., Xiong, W., Nie, W., Sheng, Y., Liu, S., and Luo, J. Pixeldit: Pixel diffusion transformers for image generation. *arXiv preprint arXiv:2511.20645*, 2025.

Zhang, Z., Xie, Z., Zhong, L., Liu, H., Hu, Y., and Cao, S. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026.

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

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 their counterparts in 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 (Xiao et al., 2023) in LLMs, 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 explore attention sinks to enable long-horizon AR sampling rather than investigating their effect on denoising performance.

Attention Sinks in Text-to-Image DiTs. For text-to-image diffusion, (Yi et al., 2024) highlight the special role of text tokens across different denoising stages, while (Jamal et al., 2026) study high-norm activations in pretrained text-to-image DiTs. However, these works do not fully characterize the nature and functional role of such tokens. Our work addresses this gap by providing a systematic investigation of high-norm tokens and establishing their connection to register tokens.

Register Tokens in Vision Transformers. In ViTs, multiple works have investigated the role and behavior of register tokens (Darcet et al., 2023; Jiang et al., 2025; Lappe & Giese, 2025; Shi et al., 2026; Chen et al., 2025; Wang et al., 2024; Marouani et al., 2026). (Darcet et al., 2023) introduces register tokens as a mechanism to avoid sink artifacts in attention maps, improving the internal representations of ViTs. Subsequent studies further analyze their functional role, showing that these tokens can influence feature aggregation (Lappe & Giese, 2025). (Jiang et al., 2025) and (Chen et al., 2025) study post-hoc or self-distilled ways to add registers to pretrained ViTs without full retraining. (Shi et al., 2026) further 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).

To summarize, tokens with unusually large norms have been studied across different domains (Su et al., 2026). In this work, we extend this line of research 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.

We train models of three sizes − B (131M), L (459M), and H (953M) − on ImageNet 256×256. 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 is trained with a smaller batch size of 512. All other training and inference settings follow JiT (Li & He, 2025).

Snippet 1. 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 2. 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 3. 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)
```

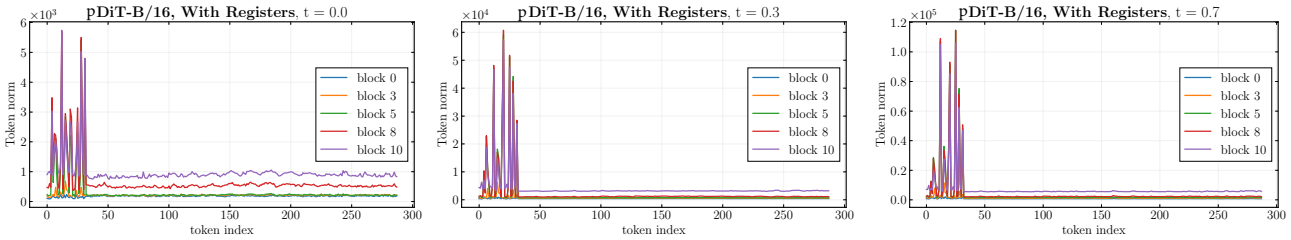

Figure 9. **High-norm outliers consistently emerge within register tokens across timesteps.** We visualize token-wise feature norms of pDiT-B/16 with registers for $t = 0.0$, 0.3, and 0.7, and observe the same behavior in all cases.

For the latent-space models presented in Table 2, we follow their original setups (Zheng et al., 2026; Ma et al., 2024), with the only modification being the addition of register tokens implemented in the same manner as in the pixel-space models.

B.2. Method Implementation Details

In our method, we propose a parameter-efficient dual-branch architecture built on top of the JiT model with in-context conditioning, as it demonstrates better performance than pure register tokens. Through ablation studies, we find that the most critical components for effective separation are the RMSNorm, SwiGLU MLP, and adaLN layers. We present their implementations in Snippets 1, 2, and 3.

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.

C. Additional Analysis Results

C.1. Analysis on ImageNet

Outliers 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 the corresponding results for additional timesteps, larger model variants, and latent-space counterparts.

First, Figure 9 shows that this effect consistently holds across different timesteps.

Second, Figure 10 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.

Second, we analyze the latent-space models SiT (Ma et al., 2024) in terms of outlier behavior. Figure 11 shows that the same phenomenon also holds for latent-space diffusion models.

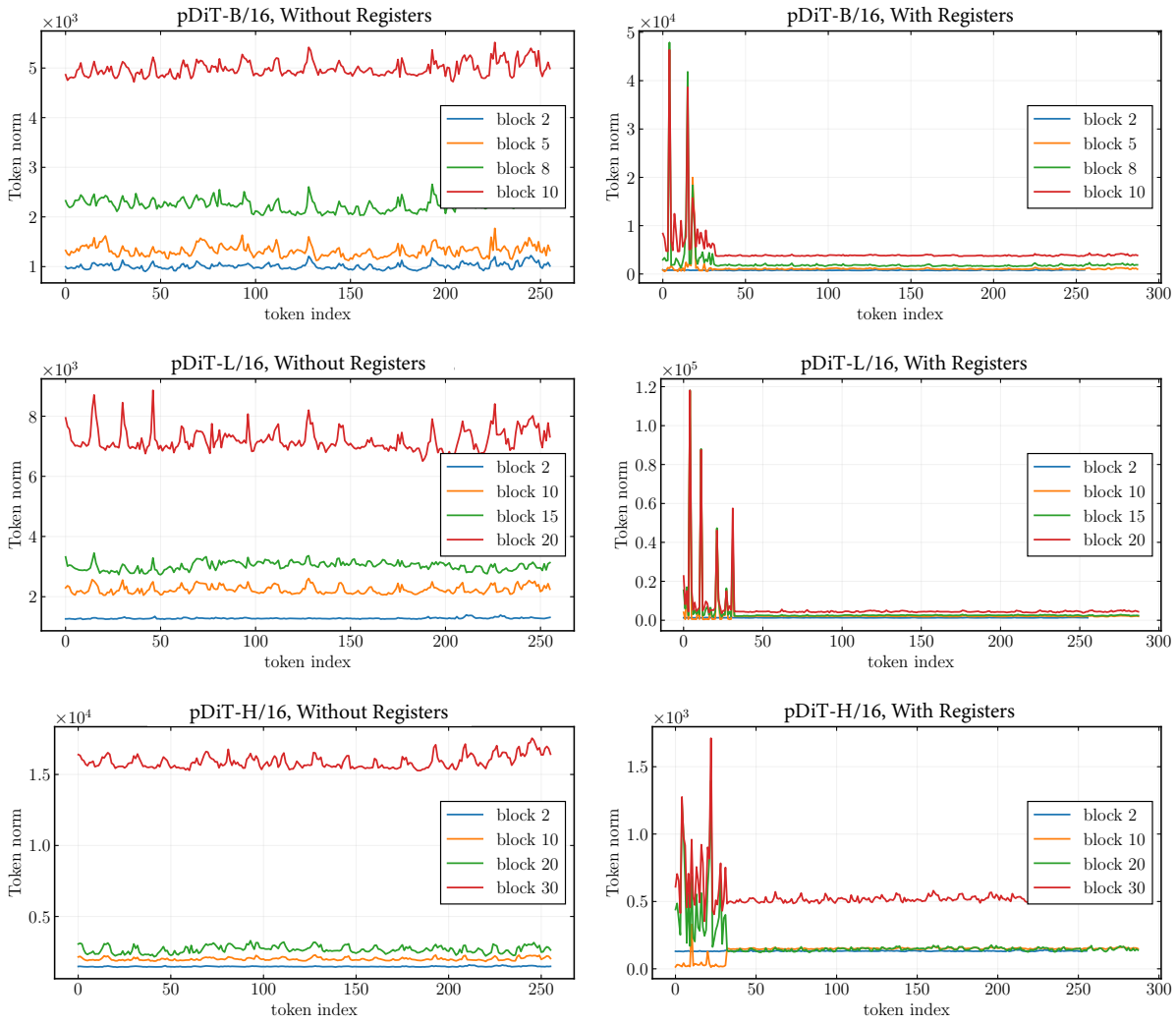

Figure 10. **Token-wise feature norms for pDiTs of varying scales on ImageNet 256×256, with and without registers.** Without registers, patch-token norms remain uniform across scales. Introducing registers leads to the emergence of high-norm outliers within the register tokens.

Improving Feature Map Quality. Next, we find that register tokens consistently reduce the feature norms of patch tokens. In Figure 12, 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 13.

Second, we show that register tokens improve feature quality across other variants of the pDiT model. Specifically, Figure 14 presents results for pDiT-L/16 and pDiT-H/16.

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), pDiTs benefit 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 15, 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.

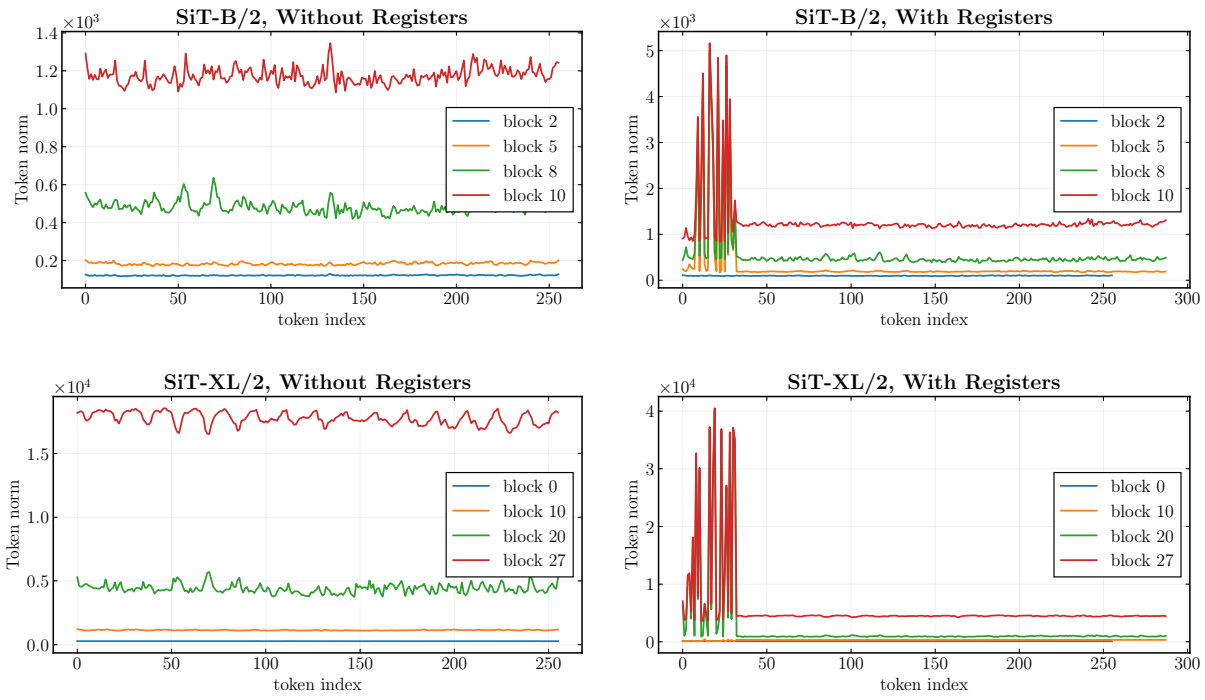

Figure 11. **Token-wise feature norms for VAE-space SiTs of varying scales on ImageNet 256×256, with and without registers.** SiTs without registers exhibit uniform patch-token norms across scales, while adding registers produces high-norm register tokens.

	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 6. **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 degrade performance in DINOv2 space (RAE).

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 6, we further present results for additional backbones in RAE and VAE spaces and observe the same trend: performance degrades for RAE-based models, while VAE-space models show similar performance 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 16 and 17, we observe that pDiTs consistently exhibit the largest feature norms and the noisiest intermediate representations compared to latent-space counterparts. Specifically, Figure 16 shows that patch-token norms in pDiTs are substantially higher across all timesteps, while Figure 17 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.

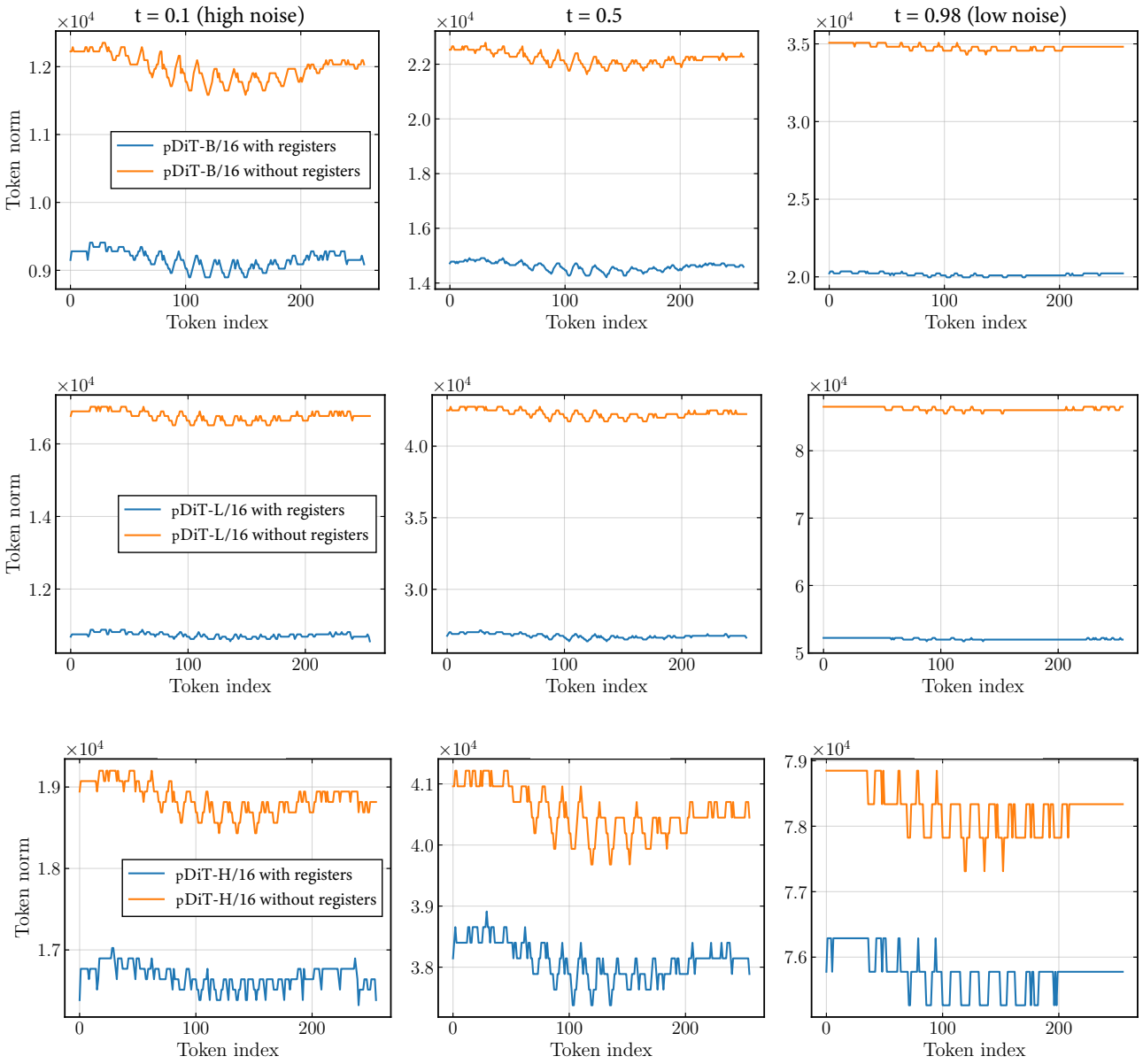

Figure 12. **Register tokens consistently reduce feature norms across patch tokens.** We measure feature norms for image tokens only (excluding register tokens) at three diffusion timesteps for pDiT models of different scales, and observe a consistent reduction in feature norms across nearly all tokens when register tokens are used.

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 consider pretrained open-sourced versions. Specifically, we analyze FLUX (Labs, 2024) and SD3.5 Large (Esser et al., 2024), 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.

For SD3.5 (Figure 18, left), we observe that high-norm outliers predominantly 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.

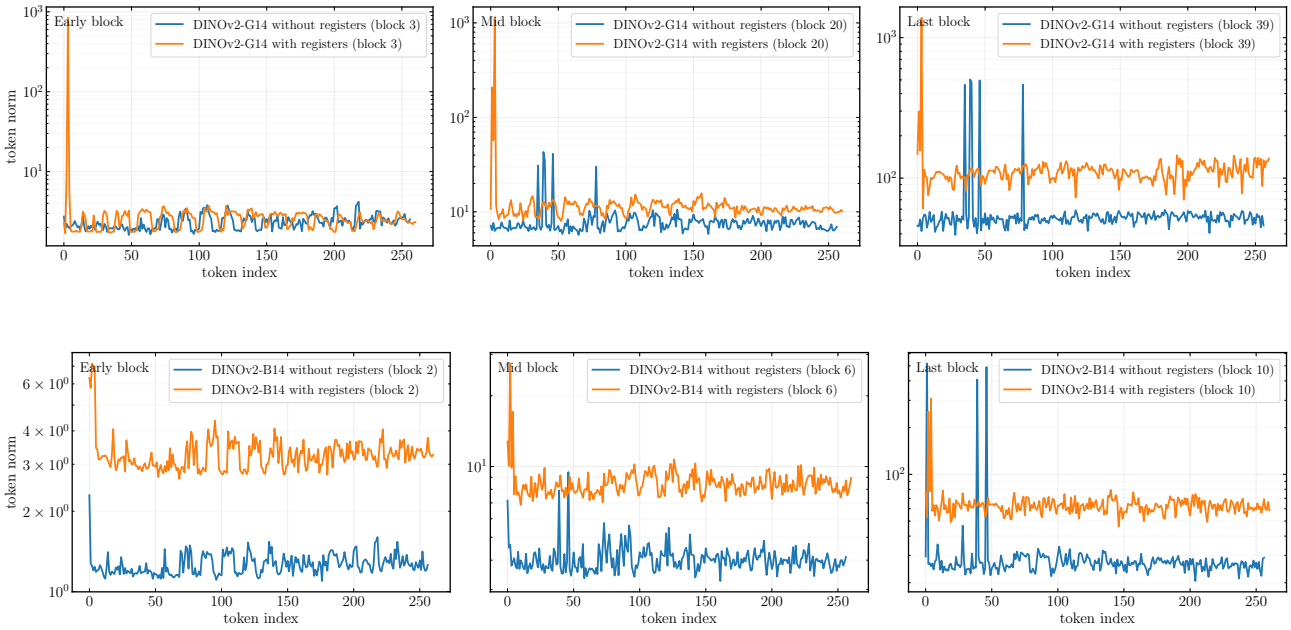

Figure 13. **For SSL ViTs such as DINOv2, register tokens do not reduce patch-token feature norms, unlike in DiTs.** We measure feature norms across all tokens for different blocks and model sizes of DINOv2, and observe that register tokens do not consistently reduce feature norms of patch tokens.

Method	200 epochs	600 epochs
JiT-B/16 + PixelREPA + compact dual	**3.7**	**3.2**
JiT-B/16 + PixelREPA	4.0	3.4
JiT-B/16 + PixelREPA (w/o in-context)	4.8	4.1

Table 7. JiT evaluation with and without in-context tokens using PixelREPA (Shin et al., 2026) on ImageNet 256×256. REPA is complementary to register-like tokens and further improves under the proposed compact dual-stream design.

Interestingly, for FLUX (Figure 18, right), we observe not only outliers in text tokens but also several high-norm image tokens. We hypothesize that this difference stems from architectural design choices: FLUX employs dual-stream layers only in the early stages of the network, whereas SD3.5 maintains dual-stream processing throughout all layers.

D. Additional Experimental Results

In Table 7, 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. In addition, our compact dual-stream architecture remains effective with REPA.

In addition, we ablate different register-token configurations in Table 8, varying both the number of registers and the transformer blocks in which they are enabled. The results show that delayed introduction is crucial: configurations that activate registers only in deeper layers (4–11) consistently outperform both the no-register baseline and configurations where registers are introduced from the first layer (0–11). We also observe that increasing the number of registers improves performance, with 32 registers yielding the best results. Finally, comparing the 4–11 and 4–9 settings suggests that the latest layers contribute relatively little to the effectiveness of registers.

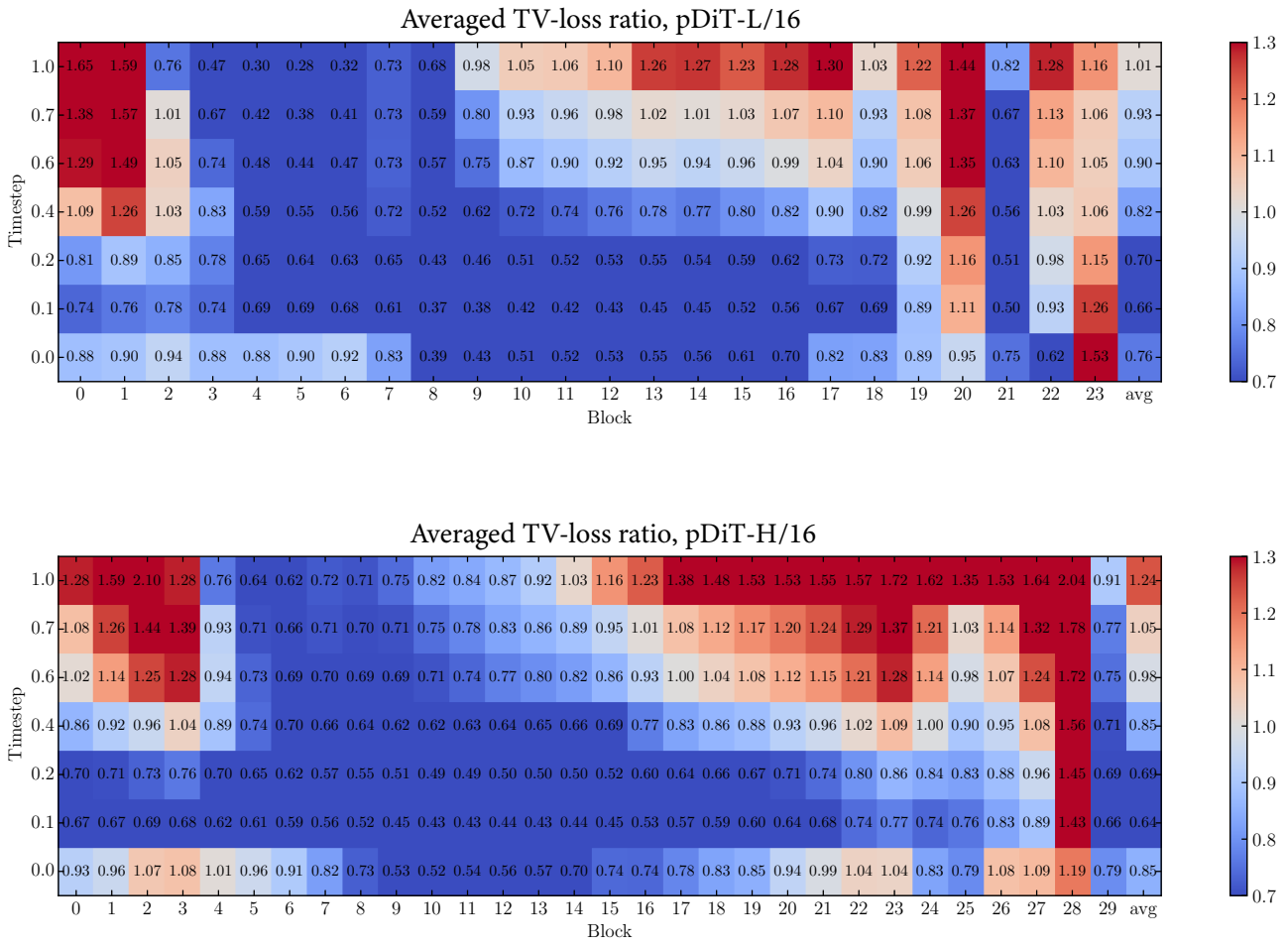

Figure 14. **Register tokens make intermediate representations cleaner by reducing noise.** We compute the Total Variation of intermediate features for models with and without register tokens. We report the ratio (with registers / without registers), where lower values indicate that models with registers 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.

E. Limitations

In this work, we study register tokens in a standard class of image diffusion transformers, with a primary focus on pixel-space models. A natural direction for future work is to extend this analysis to other pixel-space architectures (Yu et al., 2025; Hoogeboom et al., 2025) and to provide a more comprehensive study of latent-space models. This would help clarify which properties of the diffusion space determine the effectiveness of register tokens.

Also, our experiments are conducted only in the standard ImageNet setting. It remains important to study whether the observed role of register tokens transfers to other image distributions, especially datasets with different levels of diversity, structure, and fine-grained detail.

F. Discussion

This work provides a systematic study of the role of register tokens in image diffusion transformers. We show that registers are particularly important for pixel-space models, where they facilitate cleaner and more structured intermediate representations. These findings suggest that register-aware architectural design can be an effective route toward higher-quality pixel-space diffusion transformers. We hope our findings encourage future work on pixel-space diffusion architectures that reduce reliance on pretrained autoencoders, mitigating VAE-induced artifacts, reconstruction bottlenecks, and training–inference mismatches while preserving the simplicity of end-to-end generation.

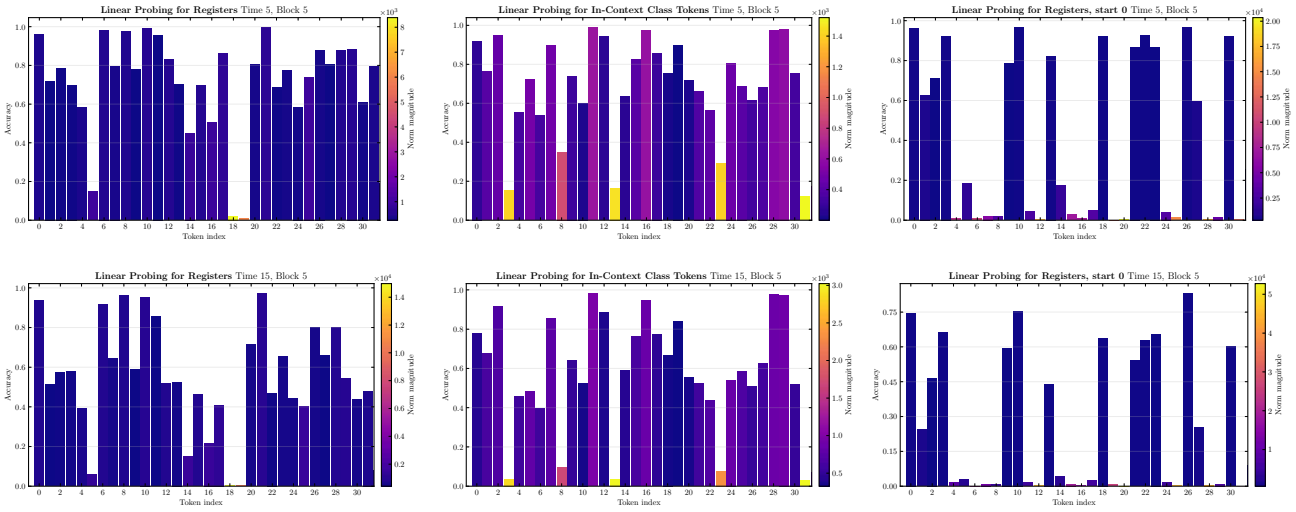

Figure 15. **Linear probing of register tokens under different configurations.** (Left) Standard register tokens introduced from the 4th layer; (Middle) Register tokens used as in-context class embeddings introduced from the 4th layer; (Right) Standard register tokens introduced from the 0th layer. Across 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.

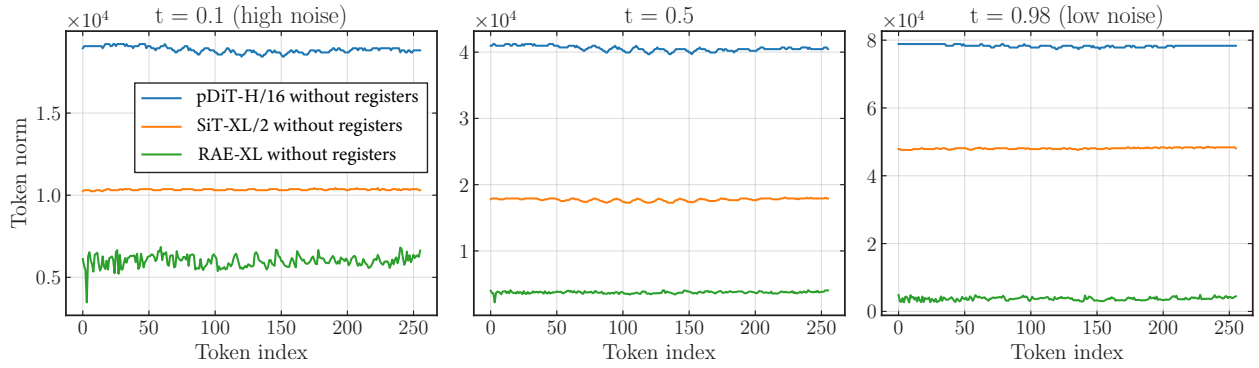

Figure 16. **Pixel-space pDiTs have the highest feature norms across all tokens for different timesteps compared to latent-space counterparts.** We compare token-wise feature-map norms for pDiT, SiT, and RAE models, all without register tokens.

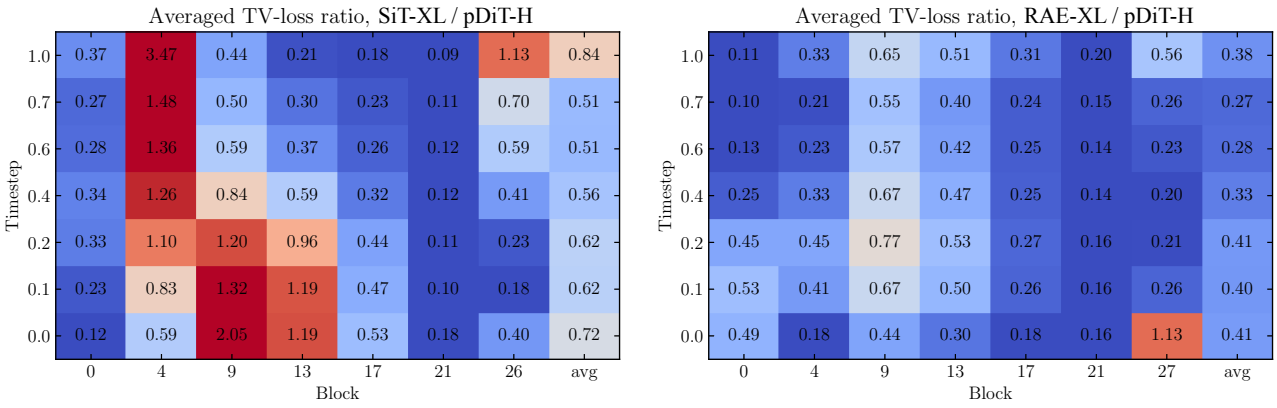

Figure 17. **Pixel-space pDiTs exhibit substantially higher Total Variation (TV) values than latent-space counterparts.** We compare the TV-loss ratio of intermediate feature maps across timesteps and transformer blocks for pixel-space pDiTs (pDiT-H) and latent-space models (SiT-XL and RAE-XL) without registers. Pixel-space pDiTs consistently produce noisier intermediate representations.

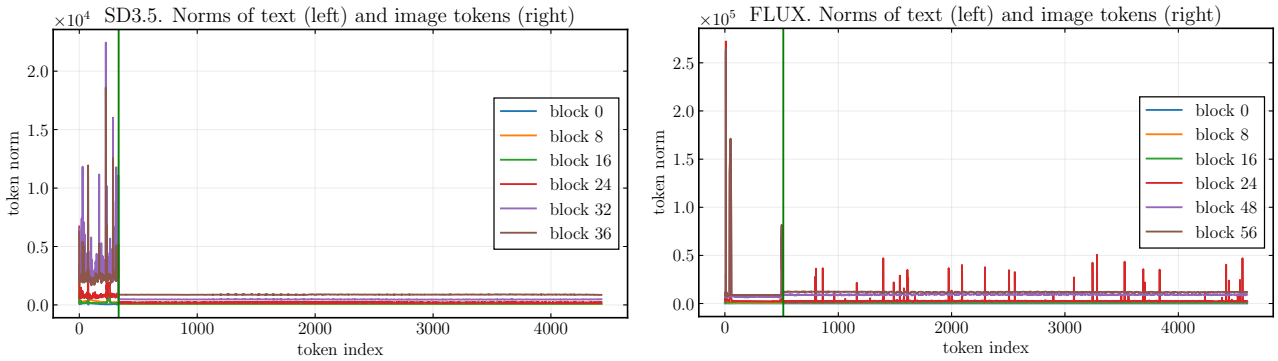

Figure 18. **Text sequences in text-to-image diffusion models exhibit behavior similar to register tokens in ImageNet-based DiTs: some tokens become high-norm outliers and potentially act as registers.** We measure token-wise feature norms in SD3.5 (left) and FLUX (right) for both text and image tokens. We observe that the outliers primarily emerge within the text sequence.

	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 8. **Registers are effective only in deeper layers.** Unlike DINOv2, pDiT-B/16 benefits from register tokens when they are introduced after the first 4 layers. Configurations with early-layer registers perform similarly to the no-register baseline, while increasing the number of registers consistently improves performance.

Model	Epoch	Params (M)	FID
JiT-B/16	200	131	4.71
	600	131	3.71
JiT-L/16	200	459	2.95
	600	459	2.43
Ours-B/16	200	149	**4.25**
	600	149	**3.41**
Ours-L/16	200	518	**2.76**
	600	518	**2.32**

Table 9. Comparison of our compact dual-stream architecture with the JiT baseline on ImageNet 256×256.

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

Table 10. Comparison of our compact dual-stream architecture with the JiT baseline on ImageNet 512×512.