IT'S ALL ABOUT ONE DIRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion Transformers exhibit several sparse structures, including high-norm tokens, massive activation channels, and attention sinks, but their relationship remains unclear. Across FLUX.1-Dev, FLUX.1-Schnell, and PixArt-Σ, we find that high-norm register tokens collapse onto a direction $v^\star$ shared across prompts and seeds. This direction is dominated by a single massive activation channel that accounts for most of the register tokens' high norm. Controlled interventions show that direction, rather than magnitude, determines sink identity, while large norms stabilize the register direction across subsequent layers. Query–key geometry explains this directional preference. The $v^\star$-aligned state primarily affects image detail, and its effect depends on the Transformer layer at which it is present. Finally, we trace a coordinated lifecycle in which projection onto $v^\star$ grows as high-norm registers and attention sinks emerge, persists through intermediate layers, and is later attenuated as these sparse structures disappear. Together, these results link massive activation channels, high-norm tokens, attention sinks, and image refinement in Diffusion Transformers through a shared internal direction.

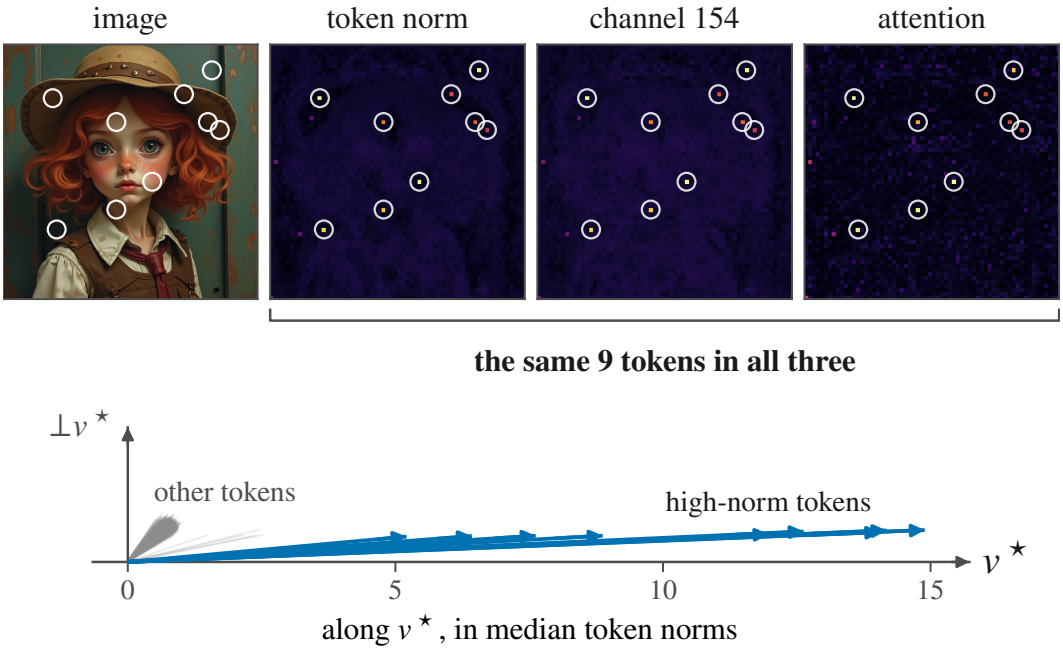


Figure 1: In FLUX.1-Dev at layer 21, High-norm register tokens coincide with massive activations and attention sinks. The same token positions dominate all three maps and their residual representations align with a shared direction $v^\star$. The same pattern holds for PixArt-Σ (Figure 11).

1 INTRODUCTION

Diffusion Transformers (DiTs) have become an increasingly important architecture for image generation, using transformer backbones in place of the convolutional U-Nets of earlier diffusion models

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.

(Peebles & Xie, 2023). At each denoising step, DiTs transform a sequence of image tokens through many attention and feed-forward layers before predicting an update to the image representation.

Recent work has uncovered several unusual structures within transformer representations. In vision transformers, a small number of patch tokens can develop exceptionally large norms, often in low-information image regions, and have been interpreted as implicit registers repurposed for internal computation (Darcet et al., 2024). Other work has identified *massive activations*, individual activation values whose magnitudes greatly exceed the model's typical scale and recur in a small number of feature dimensions (He et al., 2024; Sun et al., 2024). A third line of work studies *attention sinks*, tokens that receive a disproportionate share of attention despite not necessarily carrying corresponding semantic importance (Xiao et al., 2024). Related forms of all three structures have now also been reported in diffusion transformers (Wu et al., 2026; Gan et al., 2025; 2026; Wu & Summa, 2026).

Although these phenomena can appear similar, they describe different parts of the transformer computation. Token norm is a property of an entire representation vector, a massive activation is an extreme value at a particular token and feature dimension, and an attention sink is defined by how attention is routed. In language models, massive activations can concentrate attention on the tokens that contain them (Sun et al., 2024), yet the two phenomena can also be decoupled and serve distinct roles (Sun et al., 2026). In diffusion transformers, it remains unclear how these structures relate to one another: whether high-norm register states arise from the same feature-level structure as massive activations, what properties of these states make them preferred attention destinations, why they have such large norms, and whether their location in the transformer matters for generation.

In this work, we connect these observations in the image stream of diffusion transformers. We focus primarily on Flux.1-Dev, Flux.1-Schnell and PixArt-Σ. We study token representations, individual activation channels, and attention patterns across model depth, and use controlled interventions.

Our main contributions are:

- We show that high-norm register tokens, massive activation channel, and attention sinks are linked by a shared residual direction $v^\star$ across prompts, seeds.
- We disentangle the roles of direction and magnitude of tokens in attention. Direction is responsible for attention sink identity, whereas large norm helps that direction $v^\star$ to persist.
- We trace the register states across model depth to understand its lifecycle and contribution to image generation.

2 Related Work

High-norm tokens. High-norm tokens in Vision Transformers have been interpreted as implicit registers used for internal computation (Darcet et al., 2024). Jiang et al. (2025) further connect these outliers to a sparse set of activation dimensions. Related register-like states have since been observed in generative Transformers. We focus on naturally occurring image-stream registers and ask how their high norm relates to channel structure and attention routing.

Massive activations. Massive activations are rare hidden-state coordinates whose magnitudes greatly exceed typical features (Sun et al., 2024). In DiTs, Gan et al. (2026) connect such channels to timestep conditioning and local detail synthesis, while Turri et al. (2026) show that high-magnitude channels contain spatially structured and functionally important information. We study the token-wise structure within these channels and connect their sparse spikes to high-norm registers.

Attention sinks. Attention sinks are tokens that receive disproportionate attention from many queries (Xiao et al., 2024). Prior work has related sinks to massive activations and studied their functional importance in generative Transformers. Our question is complementary: *what property of an image-token representation makes it a sink?* We address this by separating residual direction from magnitude and linking both to the high-norm register state.

3 HIGH-NORM REGISTERS CO-LOCALIZE WITH ATTENTION SINKS

We first ask whether high-norm tokens and attention sinks, previously described as separate phenomena, occur at the same image-token positions.

3.1 MEASURING HIGH-NORM TOKENS AND ATTENTION SINKS

We measure the residual norm of each image token at the output of every Transformer block. A token is high-norm at a given layer when its residual norm exceeds three times the median image-token norm at that layer, with at most the eight largest retained as registers. Attention sinks are measured independently, with incoming attention renormalized over image keys only, then averaged over image queries, giving each image token a share of the image-to-image attention mass. We report this share relative to uniform, i.e. multiplied by the number of image tokens, so that a value of one is the uniform share. All measurements are taken at the final denoising step, intervention edits are applied at every step or early steps if stated and the resulting sink statistics are read at the final step.

3.2 HIGH-NORM REGISTERS MARK A DISTINCT SINK REGIME

In Flux, a sparse high-norm population appears abruptly near the transition from dual-stream to single-stream blocks (Figure 6). Token norms rise and remain elevated through the middle-layers which we refer to as the *register zone*. PixArt-Σ shows an analogous but shorter regime around layers 13-21.

Attention becomes strongly concentrated over the same interval. Throughout much of the Flux register zone, the strongest sink of approximately 90–100% of heads lies among the high-norm tokens. This overlap becomes stronger for increasingly extreme outliers (Figure 6).

The depth ordering is particularly informative. At the final Flux dual-stream block, sparse high-norm outliers are already present while incoming attention remains comparatively diffuse (Figure 2a). In the following single-stream blocks, attention concentrates onto these same sparse positions, often across multiple heads (Figure 2b). Thus, high-norm registers and attention sinks are strongly associated, with the high-norm state appearing before pronounced shared sink behaviour. This observation naturally motivates the next question: *what is special about the representation of these tokens that makes them preferred attention targets?*

4 REGISTERS COLLAPSE ONTO A SHARED DIRECTION $v^\star$

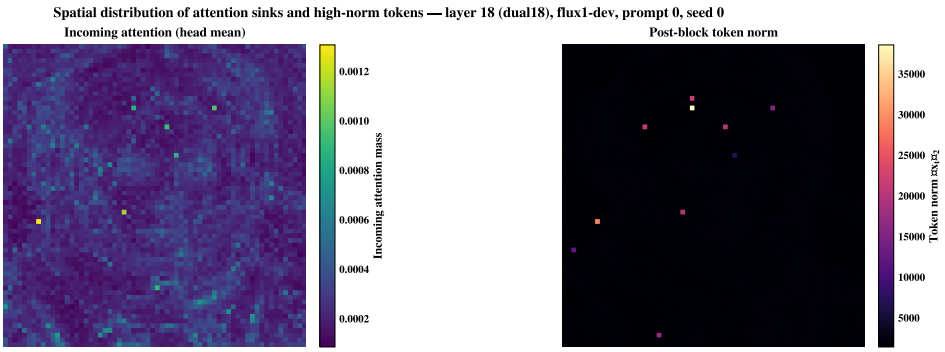
High residual norm alone does not specify whether different register tokens share a common direction. We therefore normalize their representations and test for directional alignment.

We define $v^\star$ separately for each model using only a fitting set of prompt–seed pairs, and evaluate the alignment results reported below on held-out prompt–seed pairs not used to define the direction. We unit-normalize every high-norm register representation in the fitting set, pool the resulting vectors across denoising steps and layers in the register zone, and stack them as the rows of a matrix $U \in \mathbb{R}^{M\times C}$. We define $v^\star$ as the top right-singular vector of U, without centering. Because U is not centered, this captures the common direction of the register representations rather than their variation around a mean. The sign of a singular vector is arbitrary, so we orient $v^\star$ using the fitting set such that the mean register projection onto it is positive. Comparisons between independently fitted directions use absolute cosine similarity.

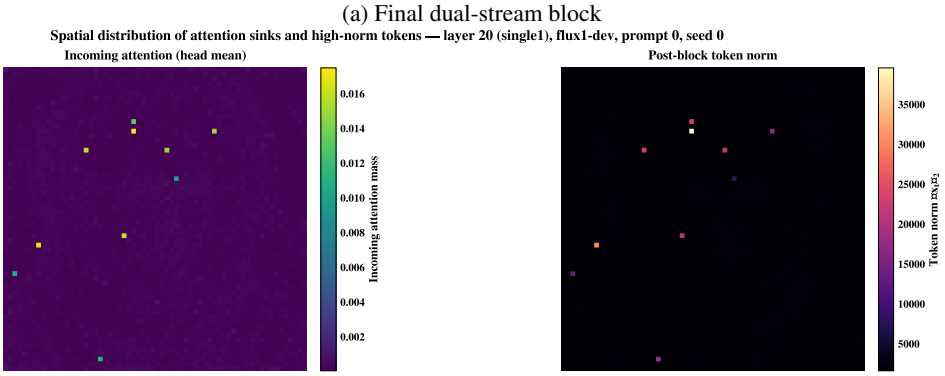
On held-out prompt–seed pairs, high-norm register representations remain nearly collinear with $v^\star$, while ordinary image tokens form a substantially broader distribution. Independently fitting $v^\star$ on non-overlapping subsets recovers nearly the same axis in both FLUX and PixArt-Σ (Figure 1). We can write each register representation exactly as

$$x_{\mathrm{reg}} = \alpha v^\star + r, \qquad \alpha = x_{\mathrm{reg}}^\top v^\star, \qquad r \perp v^\star,$$

where the near-unit cosine similarity shows that the component along $v^\star$ dominates the direction of the register representation. Across prompts and seeds, carrier position and magnitude vary while this direction remains stable.



(a) Final dual-stream block



(b) Early single-stream block

Figure 2: **High-norm outliers precede pronounced sink behaviour.** At the final dual-stream block, sparse high-norm tokens are already present while incoming attention remains relatively diffuse. Two layers later, attention concentrates strongly onto the same sparse positions.

Ordinary tokens already contain a weaker component along $v^\star$ before the register zone. Register formation therefore appears as selective amplification of an existing direction rather than creation of an unrelated one. We next ask what this direction consists of.

5 One Massive Channel Accounts for the High-Norm Signature

The previous section showed that high-norm register tokens align with a shared residual-stream direction $v^\star$. We now examine how this direction is distributed across individual residual-stream channels and how those channels contribute to the register-token norms. In Flux.1-Dev, $v^\star$ is nearly axis-aligned with channel 154, whose sparse token-wise peaks occur at the high-norm register positions and account for most of the squared-norm difference between these tokens and the remaining image tokens.

5.1 One Channel Dominates the Shared Direction

We first measure the fraction of $\|v^\star\|_2^2$ contributed by each residual-stream channel c, given by $\frac{(v_c^\star)^2}{\|v^\star\|_2^2}$.

In Flux.1-Dev, channel 154 contributes approximately 99%. Channel 1446 has the second-largest absolute coefficient, with sign opposite to that of channel 154; however, channel 1446 and all remaining channels together contribute only approximately 1%. Thus, $v^\star$ is nearly axis-aligned with channel 154.

We independently rank channels by their mean absolute activation across image tokens. For a generation with N image tokens, the statistic for channel c is

$$\frac{1}{N}\sum_{i=1}^{N}|H_{i,c}|.$$

We compute this statistic at layer 18 and the analyzed denoising step for 100 Flux.1-Dev generations spanning 20 prompts and five random seeds per prompt. Channel 154 ranks first in all 100 generations, while channel 1446 ranks second. Channel 154 therefore ranks first under both the direction-based and activation-based measurements.

5.2 Sparse Spikes in the Dominant Channel Mark Register Tokens

The high mean absolute activation of channel 154 does not reflect uniformly large values across image tokens. Instead, the token-wise distribution of $|H_{i,154}|$ contains a lower-magnitude bulk and a sparse upper tail, with the largest values reaching tens of thousands. The indices of these peaks coincide with the tokens having the largest residual norms $\|H_i\|_2$ (Figure 3).

This overlap alone does not establish that channel 154 accounts for most of the register-token norm difference, because $H_{i,154}^2$ contributes directly to

$$\|H_i\|_2^2 = H_{i,154}^2 + \sum_{c\neq 154} H_{i,c}^2.$$

Section 5.3 therefore recomputes the token norms after omitting this coordinate.

Channel 1446 has a different token-wise activation distribution. Although it ranks second by mean absolute activation, its absolute activations are distributed across many image tokens and do not contain isolated peaks comparable in magnitude to those of channel 154. A high mean absolute activation therefore does not by itself imply sparse, register-aligned token structure.

The indices of the smaller channel-154 peaks vary across random seeds. Large channel-154 activations therefore coincide with the register-token indices in each generation rather than with a fixed set of spatial locations.

5.3 Excluding the Dominant Channel Removes the High-Norm Signature

Let $c^\star$ denote the channel with the highest mean absolute activation. For each recorded token representation H_i, we zero this coordinate by subtracting its contribution,

$$H_i^{(-c^\star)} = H_i - H_{i,c^\star} e_{c^\star},$$

where $e_{c^\star}$ is the basis vector corresponding to channel $c^\star$. We then compute the norm $\|H_i^{(-c^\star)}\|_2$. Equivalently,

$$\left\|H_i^{(-c^\star)}\right\|_2 = \left(\sum_{c\neq c^\star} H_{i,c}^2\right)^{1/2}.$$

The zeroed representations are used only for this analysis and are not propagated through subsequent model blocks.

Figure 3 shows this analysis at layer 18 for all three models. Taking Flux.1-Schnell as a worked example ($c^\star = 154$): before coordinate exclusion, the tokens with the largest values of $\|H_i\|_2$ coincide with those having the largest values of $|H_{i,154}|$; after channel 154 is omitted, the recomputed norms at these indices no longer separate from the remaining image-token norms on the shared absolute color scale. The same pattern holds for Flux.1-Dev ($c^\star = 154$) and PixArt-Σ ($c^\star = 293$), whose register zone also contains layer 18. The dominant channel therefore contributes most of the squared-norm difference between the register tokens and the remaining image tokens at this layer.

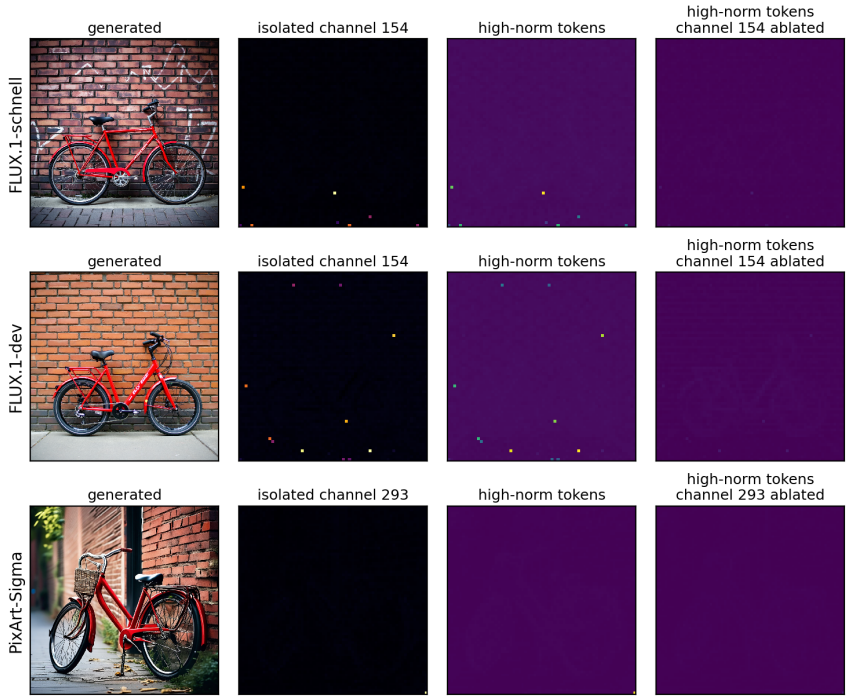


Figure 3: The top-ranked massive channel accounts for the high-norm token pattern. Columns show the generated image, the activation map of the top-ranked channel, the full token ℓ_2 norm, and the norm recomputed after omitting that channel. Activations at Layer 18 are shown. Within each row, the full and coordinate-excluded norm maps use the same absolute color scale.

6 Direction, Not Norm, Determines Sink Identity

The register's large magnitude and its direction $v^\star$ arise together, so observational data cannot tell us which property attention uses. We separate them with complementary interventions.

Sink identity survives norm reduction. We compare two ablations to FLUX registers. In one, we reduce a register to an ordinary-token norm while preserving its direction. In the other, we retain the original large norm but replace its direction. Removing the extreme magnitude has little effect: approximately 89% of affected heads continue to select the same register after its norm is reduced to the ordinary-token scale, and nearby downstream blocks continue to recognize the ablated token as the sink (Appendix C.2). The same holds far beyond the natural register. When we hold a FLUX.1-dev register on $v^\star$ at every layer and let its norm decay as it naturally does, it remains the top sink of at least half of the heads, with key rank 1 and 39–82 times the uniform share, at layers 39, 45, and 50 in all 100 generations, extending sink persistence 11 layers beyond the unmodified sink endpoint, although by layer 46 its norm has fallen to half that of a typical token (Appendix C.4). Thus, an extreme residual norm is not necessary for sink selection.

Sink identity depends on the register direction. Changing direction while preserving the original large norm produces the opposite effect: approximately 96% of affected heads abandon the original sink. Removing only the $v^\star$ component of a FLUX.1-dev register at unchanged norm likewise reduces its attention share at the next layer by 84%, from 62 to 10 times uniform. A large residual norm therefore cannot maintain sink identity once the register direction is removed. We observe the same pattern at fixed norm in FLUX.1-schnell (Figure 4). Removing or reversing the register tokens' $v^\star$ component takes them below the sink threshold at layers 20 and 24, and rotating the state 15° off $v^\star$, leaving its cosine above 0.95, removes it as the sink of 82–100% of the heads that selected it.

Together, the interventions resolve the ambiguity from Section 3:

The register direction determines sink identity.

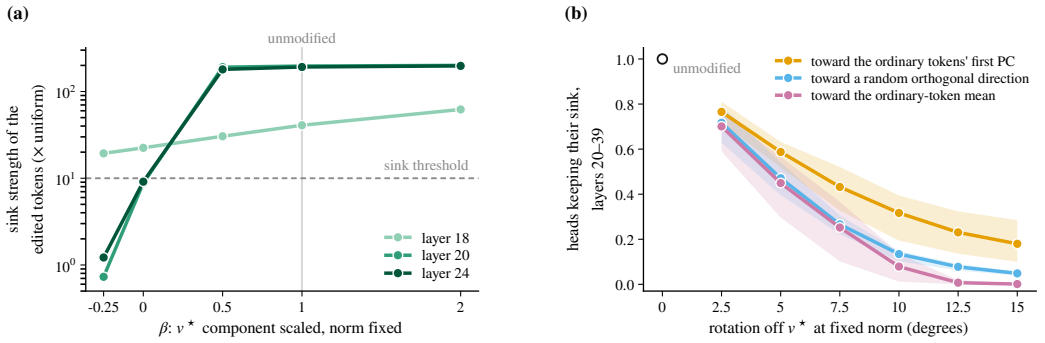


Figure 4: At fixed norm, alignment with $v^\star$ determines sink identity. (a) Scaling the $v^\star$ component changes sink strength; (b) rotating away from $v^\star$ reduces downstream sink persistence.

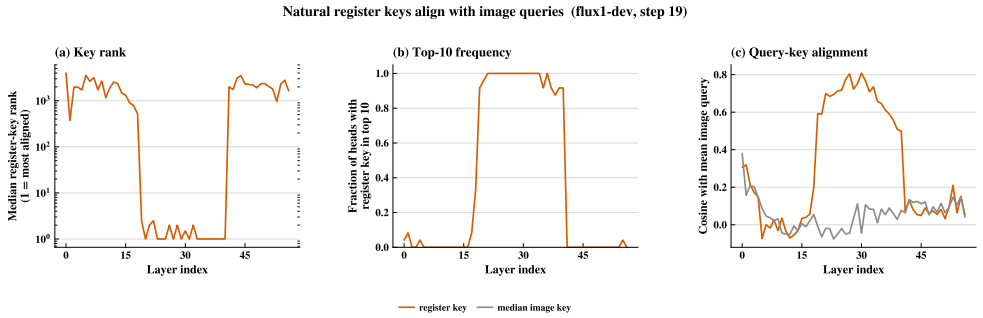


Figure 5: Register keys become highly ranked and strongly aligned with image queries during the FLUX.1-Dev register regime. PixArt-Σ results are shown in Appendix B.4.

7 Why Attention Prefers the Shared Direction

In the FLUX register zone, the natural register key changes from poorly ranked to one of the image keys most aligned with the mean image query (Figure 5). It lies in the top ten for roughly 90–100% of heads and reaches a query–key cosine of approximately 0.6–0.8, substantially above that of typical image keys. PixArt-Σ exhibits the same pattern over its shorter register interval.

This geometry explains the intervention result. Attention normalization largely removes raw residual scale, while representation direction determines the key presented to the head. Preserving $v^\star$ therefore preserves the favourable key geometry even after norm reduction; changing the direction removes that advantage despite retaining the original magnitude. When the register later dissolves, its query–key advantage disappears as well.

8 High Norm Stabilizes the Register State

If extreme norm does not determine sink identity, why does the register become so large? Holding the direction fixed while varying its magnitude shows that high norms do not make an arbitrary recipient a universal sink; similarly, after register damage, restoring the norm alone does not recover the sink, even though restoring the full clean register does (Figure 17). Magnitude is therefore not the primary determinant of immediate selection.

Its effect instead appears across depth. Larger realizations of the same direction are more likely to retain register-like behaviour after subsequent blocks, while reducing the norm of a natural register increasingly weakens its sink identity farther downstream. Geometrically, this is expected: the same additive residual update rotates a large vector less than a small vector pointing in the same direction.

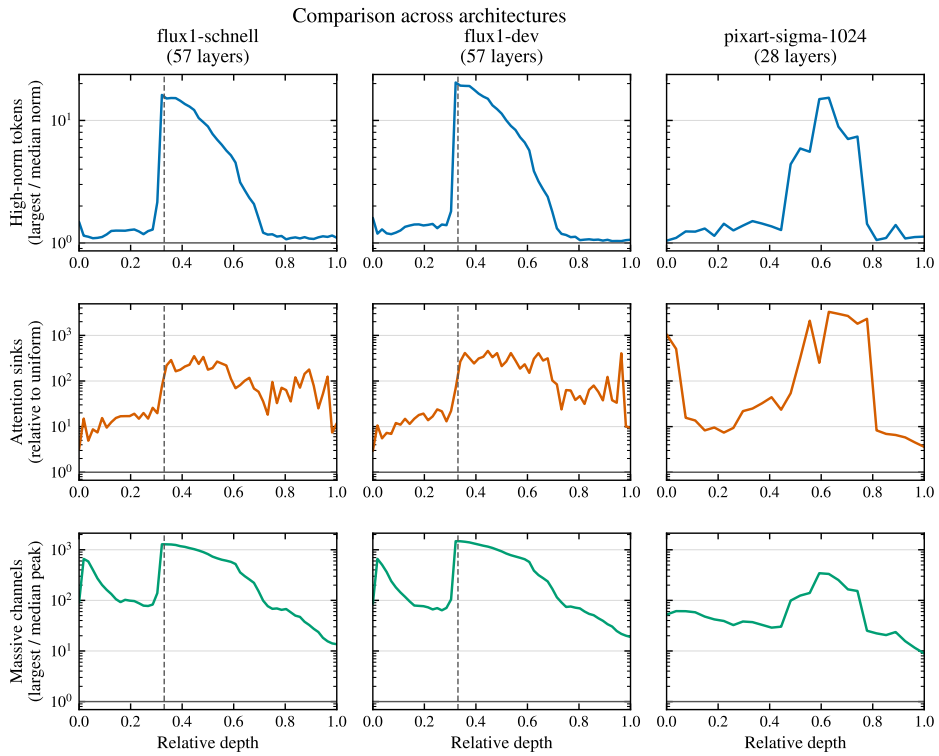


Figure 6: **The one-direction state has a lifecycle.** High-norm, attention-sink, and massive-channel structures emerge and later dissolve over an interval in Flux.1-Schnell, Flux.1-Dev, and PixArt-Σ.

We therefore assign the two properties complementary roles: *direction determines sink identity, whereas magnitude supports persistence*. This does not imply that the network creates high norm specifically for this purpose, only that large magnitude stabilizes an already-established register direction.

9 The Sparse State Has a Lifecycle

The register state follows a consistent lifecycle. First, the eventual register direction $v^\star$ is already weakly present in ordinary image tokens before register formation. Near the register boundary, a small subset of tokens is selectively amplified along this axis, producing both a large projection onto $v^\star$ and a high residual norm.

In Flux, this transition is localized to the final dual-stream block, where the MLP produces a strong update in channel 154 at a sparse set of tokens. The high-norm register state therefore appears before the pronounced shared attention-sink behaviour that develops in the following single-stream blocks. We restrict this sublayer-level observation to Flux because PixArt-Σ has a different block structure.

The state then gradually dissolves. In Flux, it is strongest around layers 19–35 and largely disappears by layer 40; PixArt-Σ shows the same qualitative pattern over a shorter interval around layers 13-22. As the state fades, projection onto $v^\star$, dominance of the register-carrying channel, query-key advantage, high norm, and sink behaviour all decline together. Across models, these quantities therefore trace a coordinated rise-and-fall trajectory (Figure 6). This dissolution follows a depth-specific schedule, with layer 21 in PixArt-Σ removing approximately 94% of the register's $v^\star$ projection, while in FLUX.1-Dev removal is distributed across layers 32–46, with layer 35 alone removing about one quarter. Manipulating the competing massive channel leaves the high-norm register endpoint unchanged within the pre-declared ± 1-layer margin in both models.



Figure 7: Representative FLUX.1-Dev and PixArt-Σ generations when the state is introduced early or late, or the natural register is removed. Numbers report LPIPS to the corresponding original image; aggregate results are given in Table 5.

10 The Shared Direction Has Layer-Specific Effects on Image Detail

We test whether the same register state has the same effect at different Transformer Layers. We write a state matched to the natural register's $v^\star$ projection and norm before, within, or after its usual layer range. We also remove the natural register or maintain it beyond its normal endpoint. All conditions use matched prompts, seeds, and denoising steps, more details are given in Appendix C.5.

In FLUX.1-Dev, introducing the state before or after its natural depth produces comparatively small changes, largely confined to image detail. PixArt-Σ is substantially more sensitive: late introduction, extending the state beyond its natural endpoint, and removing the natural register produce much larger changes and broke the image (Figure 7).

Introducing the same state earlier or later does not recover the original generation. These results show that the effect of the register state depends on where in the Transformer it is present, and contributes to local details in the image mainly, not global composition.

11 Future Work and Limitations

Our empirical analysis is limited to the image stream of the DiTs considered in this work. Accordingly, the text-stream counterparts of the relationships among high-norm tokens, massive activations, and attention sinks are not systematically characterized in our experiments. Extending the same analysis to text-stream representations would provide a more complete characterization of these sparse structures across modalities. Furthermore, our analysis is focused on Flux.1-Dev, Flux.1-Schnell, and PixArt-Σ, so the extent to which the mechanism generalizes across other DiT architectures remains open.

Although we identify how the register state is represented, read, and maintained, we do not yet explain why training produces this mechanism. Future work should also investigate whether this low-dimensional register structure can be exploited for activation-aware quantization or compression.

12 Conclusion

High-norm tokens, massive activation channels, and attention sinks have so far been studied as separate sparse structures in Transformers. We show that, in Diffusion Transformers, they are organized around a shared representation direction $v^\star$: this direction becomes selectively amplified into a high-norm register state, is itself dominated by a single massive activation channel, and produces keys that are unusually well aligned with image queries, causing the register tokens to become attention sinks. Controlled interventions separate the roles of direction and magnitude, showing that direction determines sink identity while large magnitude helps the state persist across later layers. As the dominant register channel loses its relative prominence, the high-norm and register-aligned sink signatures also decline. Thus, these sparse phenomena are linked through a model-specific register direction and share an overlapping transient regime across model depth.

AI USE STATEMENT

In this work, we used generative AI tools (Chatgpt and Claude) to assist with literature review and improving writing. We reviewed and approved all AI-assisted edits before including them in this manuscript. All intellectual contributions, experimental design, data analysis, and validation are our own work. We take responsibility of our work.

ETHICS STATEMENT

This work does not raise specific ethical concerns. It studies internal representations in publicly available pretrained text-to-image generative models and does not involve human subjects, private data, or new data collection.

REPRODUCIBILITY STATEMENT

The experimental setup, model checkpoints, generation settings, definitions, intervention procedures, and statistical analyses are described in the main text and Appendices A–C. The supplementary material provides the exact prompts and seeds, implementation details, and anonymized code and configuration files needed to reproduce the main experiments and figures.

REFERENCES

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

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 126899–126929. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-3816.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Bobby He, Lorenzo Noci, Daniele Paliotta, Imanol Schlag, and Thomas Hofmann. Understanding and minimising outlier features in transformer training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 83786–83846, 2024.

Nicholas Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*, 2024.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Evelyn Turri, Davide Bucciarelli, Sara Sarto, Lorenzo Baraldi, and Marcella Cornia. Few channels draw the whole picture: Revealing massive activations in diffusion transformers. *arXiv preprint arXiv:2605.13974*, 2026.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 893–911, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.51. URL https://aclanthology.org/2023.acl-long.51/.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Xiaoyu Wu, Yifei Wang, Tsu-Jui Fu, Liang-Chieh Chen, Zhe Gan, and Chen Wei. Taming outlier tokens in diffusion transformers. *arXiv preprint arXiv:2605.05206*, 2026.

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

APPENDIX

Appendix A gives the models, prompts, seeds, and measurements that all experiments share. Appendix B extends the observational results of Sections 3 to 5 and 7. Appendix C gives the protocol and the full results of every intervention, in the order of the main text.

A EXPERIMENTAL SETUP

A.1 MODELS AND GENERATION

We study the image stream of FLUX.1-Dev and FLUX.1-Schnell (Black Forest Labs, 2024) and of PixArt-Σ (Chen et al., 2024), using their public checkpoints.[1] Table 1 lists the settings that differ between them. Each model uses the default sampler of its public pipeline.

Table 1: **Models and generation settings.** Every image is 1024×1024, which gives 4,096 image tokens (a 64×64 grid) in every model. Layers and denoising steps are numbered from 0; in FLUX, layer 18 is the last dual-stream block and layer 19 the first single-stream block.

	FLUX.1-Dev	**FLUX.1-Schnell**	**PixArt-Σ**
Transformer blocks	19 dual-stream, then 38 single-stream		28
Text conditioning	joint attention with 512 text tokens		cross-attention
Denoising steps	28	4	20
Guidance	distilled, scale 3.5	none	classifier-free, scale 4.5

A.2 PROMPTS AND SEEDS

All prompts come from DiffusionDB (Wang et al., 2023). We draw them from 20,000 prompts sampled across the dataset's metadata table, keeping at most one prompt per user and only prompts with at least five words and DiffusionDB safety scores below 0.2, and rejecting any prompt that shares half or more of its content words with an earlier pick. The draw is deterministic given a selection seed. Two draws are used (Table 2): set F, on which $v^\star$ is fitted, and set H, which shares no prompt with set F. Generations use the seeds 0, 42, 777, 1234, and 3407. The channel analysis of Appendix B.3 (20 prompts crossed with five seeds) and the fixed-norm experiment of Figure 4 (five prompts at seed 0) use DiffusionDB prompts that are not listed here.

A.3 MEASUREMENTS

High-norm and register tokens. Token norms are ℓ_2 norms of the image-token residual stream at the output of a block. A token is high-norm when its norm exceeds three times the median image-token norm of its layer; Sections 3 to 5 keep at most the eight largest as registers. The interventions add an alignment condition, which Appendix C states for each experiment.

Attention sinks. For every head, incoming attention is renormalized over image keys and averaged over image queries, which gives each image token a share of the image-to-image attention. Shares are reported relative to the uniform share $1/4096$. A head's sink is the image key with the largest share, and a token counts as a sink when a head gives it more than 10 times the uniform share.

Shared direction and channels. $v^\star$ is fitted once per model and then kept fixed. A channel's magnitude is its mean absolute activation over image tokens (Section 5), and a channel is massive when this exceeds 100 times the median over channels.

[1] `https://huggingface.co/black-forest-labs/FLUX.1-dev` (FLUX.1 [dev] Non-Commercial License), `https://huggingface.co/black-forest-labs/FLUX.1-schnell` (Apache 2.0), and `https://huggingface.co/PixArt-alpha/PixArt-Sigma-XL-2-1024-MS` (CreativeML Open RAIL++-M), as listed on their model cards.

Table 2: **Prompts of sets F and H**, with the generations that use each set. Seeds are 0, 42, 777, 1234, and 3407.

Set F (selection seed 0). Fitting $v^\star$: F0–F23 at seed 0 (FLUX.1-Dev, PixArt-Σ) and F0–F9 at all five seeds (FLUX.1-Schnell). Dissolution and restoration: F0–F19 at all five seeds. Removing one structure: F0 at seed 0.	
F0	photo the last us part 2 kitchen : 5 4 k rtx hdr volume light concept studio matte painting environement digital illustration fanart artstation by eytan zana and balazs agoston, jad saber, aaron limonick, pablo dominguez and florent lebrun and robby johnson
F1	in game footage of the hindenberg crashing into the pitch during a game of football manager 2 2
F2	textless logo. a giant human fist crushes the planet earth. logo design. graphic design. clean lines, solid colours. textless
F3	edwardian photograph of a mix of billie eilish and henry cavill, male, 1 9 0 0 s, 1 9 1 0 s, grainy, slightly blurry, faded, realistic face
F4	i think this just might be my favourite door ever. st saviour's, dartmouth
F5	minimalist boho style art of colorful halloween, illustration, vector art
F6	futuristic solarpunk city with happy people and animals, happy scenes, late afternoon light, wispy clouds in a blue sky, unreal engine, intricate, highly detailed, elegant, trending on artstation, sharp focus, cinematic lighting, art by greg rutkowski and alphonse mucha
F7	a glass bubble space utopia where the path waves are filled with fauna, dynamic lighting, photorealistic fantasy concept art, trending on art station, stunning visuals, creative, cinematic, ultra detailed, ray tracing, sun rays
F8	centered oil painting portrait of demon king, realistic character concept, gazing eyes, artstation, cgsociety, 8 k, high resolution, well rendered, golden ratio, art by john howe and keith parkinson and larry elmore, trending on artstation, featured on cgsociety
F9	a portrait painting of rosemarie konkel
F10	elephant picks up a painting canvas, and places it on his backpack with his trunk, fantasy, cinematic, octane render, 8 k
F11	a digital painting of underground car racing, jdm, initial d, rainy night, a giant storm about to come, 8 k, unreal engine, highly detailed, 3 d art, perfect lighting, digital illustration
F12	a portrait of emad on ama an ultrafine detailed painting, detailed painting, detailed eyes!!, final fantasy octopath traveler lovecraft ghibly
F13	ultra realistic illustration, handsome kira. intricate, elegant, highly detailed, digital painting, artstation, concept art, smooth, sharp focus, illustration, art by artgerm and greg rutkowski and alphonse mucha and wlop
F14	a young girl holds hands with a demon, atlantis, webtoon art by michelle czajkowski fus! dream
F15	the pilot at the helm of the plane, view inside the cockpit, in the window the planet saturn in space, sketch a pen
F16	whimsical tadpole creatures shaped like yinyangs from the depths of the collective unconscious, dramatic lighting, surreal darkly painting by ronny khalil
F17	hyperbolic spacetime vortex powers the void. by mc escher, hyperrealistic photorealism acrylic on canvas, resembling a high resolution photograph
F18	piano inside a steampunk glass jar, hourglass, lightning, intricate detail, volumetric lighting, epic composition, hyper detailed, ultra realistic, sharp focus, octane render, candle, volumetric, ray tracing, artstation trending, cgsociety, sense of awe, swirling mist, 4 k
F19	the mythical scottish creature known as the haggis. cute. photo realistic. octane render
F20	detailed portrait of a teenage boy, around 1 9 yo. looking sad. shirtless, tattoo. ominous and eerie looking forest i'm background.
F21	a beautiful painting of old golden samovars of russian empire, by beeple and roger dean, character concept art, matte painting, heavy metal magazine, trending on artstation
F22	product photo of 2 0 2 3 maserati concept pickup truck
F23	oil painting by georgia o'keefe of a llama.
Set H (selection seed 1, disjoint from set F). Calibrating the retiming windows: C0–C3 at seed 0. Retiming experiment and the register state: H0–H9 at all five seeds. Figures 1 and 11 show H3 at seed 3407.	
C0	closeup portrait of a handsome man, bouguereau : 1 0, leyendecker : 2 5, alma tadema : 1 0, john singer sargent : 1 5
C1	epic face portrait polaroid of john dwyer live on stage, 1 9 8 0 s
C2	mass effect forged armor in unknow world wallpaper 3 d pixar disney digital cgi rtx hdr painting bioluminance alena aenami artworks in 4 kby jesper ejsing, feng zhu and loish beeple, by thomas kinkade 2 2 2 2
C3	a knight in armor riding a bicycle horse with training wheels on the back legs, high quality digital art trending on artstation
H0	the painting depicts multicolored images of large, elaborate flowers, a cubist painting by james rosenquist, polycount, orphism, soviet propaganda, fauvism, cubism
H1	vfx film, destiny 2 irl, low - key lighting award winning photography arri alexa cinematography, hyper real photorealistic cinematic beautiful, atmospheric
H2	ice meets fire, painting by mc escher, very detailed!!, illusion, surreal!!!, art by greg rutkowski, trending on artstation, beautiful color art!
H3	photograph of young steampunk orphan annie, city character portrait, in the style of moebius, tom bagshaw, and waterhouse, funny, freckles, cinematic lighting artstation, oil painting,
H4	mythical war between inner and outer forces in the visceral anatomical human heart imaginal realm of the collective unconscious, in a dark surreal mixed media oil painting by ronny khalil, dramatic lighting from inner fire, 3 0 s cartoons by haeckel
H5	mechanical butterfly inside a glass jar buried in sand, intricate detail, volumetric lighting, epic composition, hyper detailed, ultra realistic, sharp focus, octane render, ray tracing, sense of awe, swirling mist, 4 k, moon
H6	studio ghibli anime, adorable woman walking in the forest, romantic magical, fairytale, fantasy, pastel colors
H7	paperback book cover by leonardo da vinci. pure colors, melting clouds, accurately drawn details, a sunburst above a receding road with the light reflected in furrows and ruts, after rain. photorealistic. octane render. cinematic. trending on artstation. textless.
H8	a snowing field of black flowers and white flowers in front, white mountains in background, wide angle, sad feeling, in style of john harris s, watercolor, artstation
H9	a silver constellation necklace pendant, 3 d rendering, in the style of pandora, tiffany, swarovski, van cleef & arpels, cartier, boucheron, bulgari, chaumet, elegant, noble, stylish

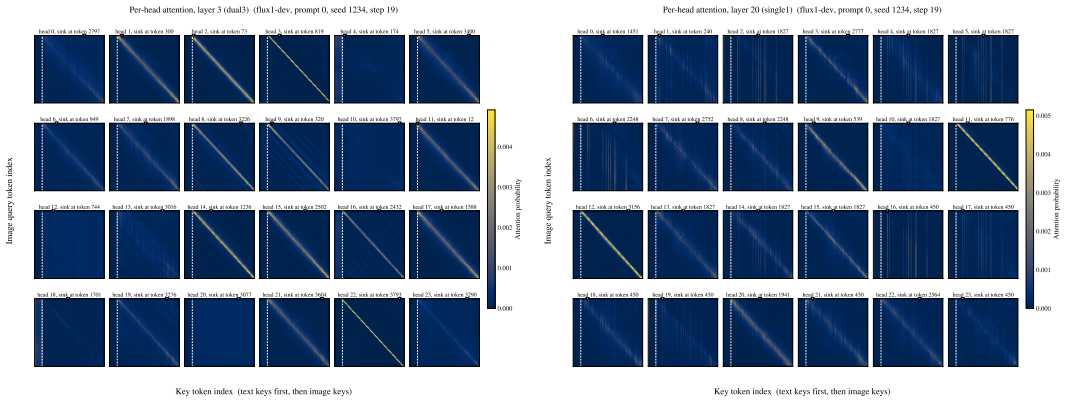
Image metrics. We compare an edited image with the unmodified image of the same prompt and seed by LPIPS with AlexNet features, by the cosine similarity of their CLIP ViT-L/14 image embeddings (CLIP-I), and by PSNR.

Paired design and statistics. Every edited generation is paired with an unmodified generation of the same prompt, seed, and initial noise, so every effect is a within-pair difference. In every intervention experiment, a hooks-only condition that reads the edited states and writes them back unchanged reproduces the unmodified image exactly (99 dB PSNR, LPIPS 0). Intervals are 95% percentile bootstrap intervals that resample prompts together with all of their seeds.

B Additional Observational Results

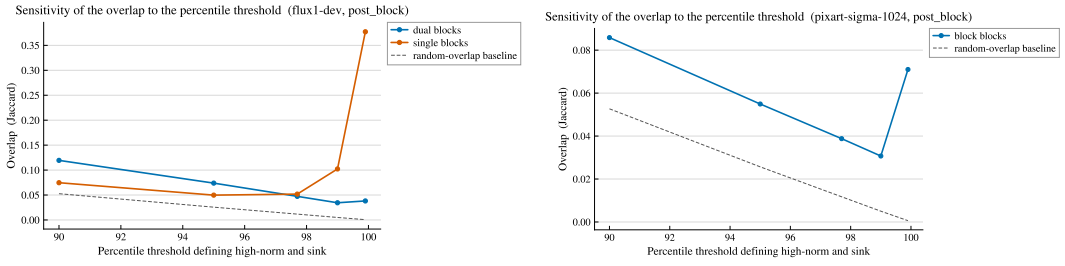
B.1 High-Norm Tokens and Attention Sinks

Figure 8 shows the attention matrices behind the sink statistics of Section 3. Before the register zone, each head spreads its attention over many keys; in the register zone, many heads concentrate on the same few image keys, which appear as vertical columns. Figure 9 varies the percentile that defines both the high-norm set and the sink set. In both models, the two sets overlap more than independent sets of the same size, and most for the most extreme tokens, which is why the main analyses use a strict threshold.



(a) Layer 3, before the register zone (b) Layer 20, in the register zone

Figure 8: **Attention sinks appear as columns.** Attention of every image query (rows) to every key (columns; text keys first) for individual heads of FLUX.1-Dev, in one generation at denoising step 19.



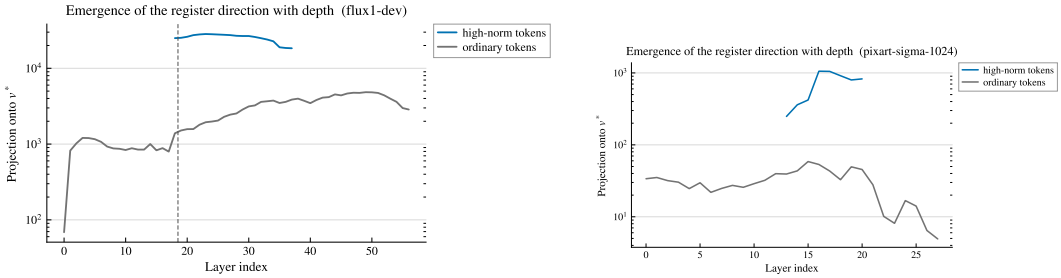
(a) FLUX.1-Dev (b) PixArt-Σ

Figure 9: **High-norm tokens and sinks overlap most in the extreme tail.** Overlap (Jaccard index) between the image tokens above a norm percentile and those above the same percentile of received attention, against the percentile. The dashed line is the overlap expected for independent sets of the same size.

B.2 The Shared Direction and the Register State

In every model, $v^\star$ captures 98.7 to 99.7% of the squared length of the unit-normalized register representations, and one channel carries 97 to 99% of $\|v^\star\|_2^2$. The FLUX.1-Dev and FLUX.1-Schnell fits use different checkpoints, prompts, seeds, and denoising steps, yet recover the same axis ($|\cos| = 0.9999$, and 0.988 after channel 154 is removed from both). On held-out prompts, FLUX.1-Dev carries about 37 register tokens per image and PixArt-Σ a single one, at the bottom-right token, which by itself receives 8.2% of all attention.

Figure 10 follows the projection onto $v^\star$ through depth. Ordinary tokens carry a weak component along $v^\star$ before the register zone, the high-norm tokens separate sharply along it inside the zone, and the separation closes as the state dissolves. Figure 11 repeats Figure 1 for PixArt-Σ.



(a) FLUX.1-Dev (b) PixArt-Σ

Figure 10: **The register direction is amplified, not created.** Projection onto $v^\star$ of the high-norm tokens and of ordinary image tokens at every layer.

B.3 Channel Structure

This analysis scores every channel by its mean absolute activation over image tokens, $\mathrm{mean}\,|x|$, in 100 generations per model (20 DiffusionDB prompts crossed with five seeds). Figure 12 shows the score of every channel at every layer: a few channels, including channel 154 of FLUX.1-Dev and channel 293 of PixArt-Σ, stay elevated over a contiguous range of layers. Table 3 ranks the channels at one layer and step inside the register zone, chosen by a pilot rule (layer 22 of FLUX.1-Dev; Section 5.1 ranks the same channels at layer 18). Only two channels of FLUX.1-Dev and one of PixArt-Σ exceed 100 times the median channel, and the dominant channel is ranked first in at least 90% of generations. Figure 13 shows that the same few channels recur across generations and keep their magnitude across denoising.

Table 3: **Channel statistics in the register zone** (100 generations per model). The first block is at the analysis layer and step; the per-layer ranking uses the five pilot generations at every layer of the register zone (18–39 and 13–20).

	FLUX.1-Dev	**PixArt-Σ**
Analysis layer and denoising step	22, step 21 of 28	13, step 15 of 20
Median $\mathrm{mean}\,\lvert x\rvert$ over channels	6.3	0.4
Channels above 100 times the median	154 (1,945), 1446 (880)	293 (47)
Channel ranked first, by layer	154, except 1446 at layer 35	293 (13–17), 295 (18–20)
Channels in the top 10 of any generation	41 of 3,072	23 of 1,152
In the top 10 of $\geq 90\%$ of generations	154, 1446, 1831, 2566	293, 295, 856, 884, 900
Ranked first in $\geq 90\%$ of generations	154	293

B.4 Query–Key Geometry in PixArt-Σ

Figure 14 repeats Figure 5 for PixArt-Σ. During its register interval, the register key ranks among the image keys best aligned with the mean image query and is more aligned with it than the median image key, as in FLUX.1-Dev.

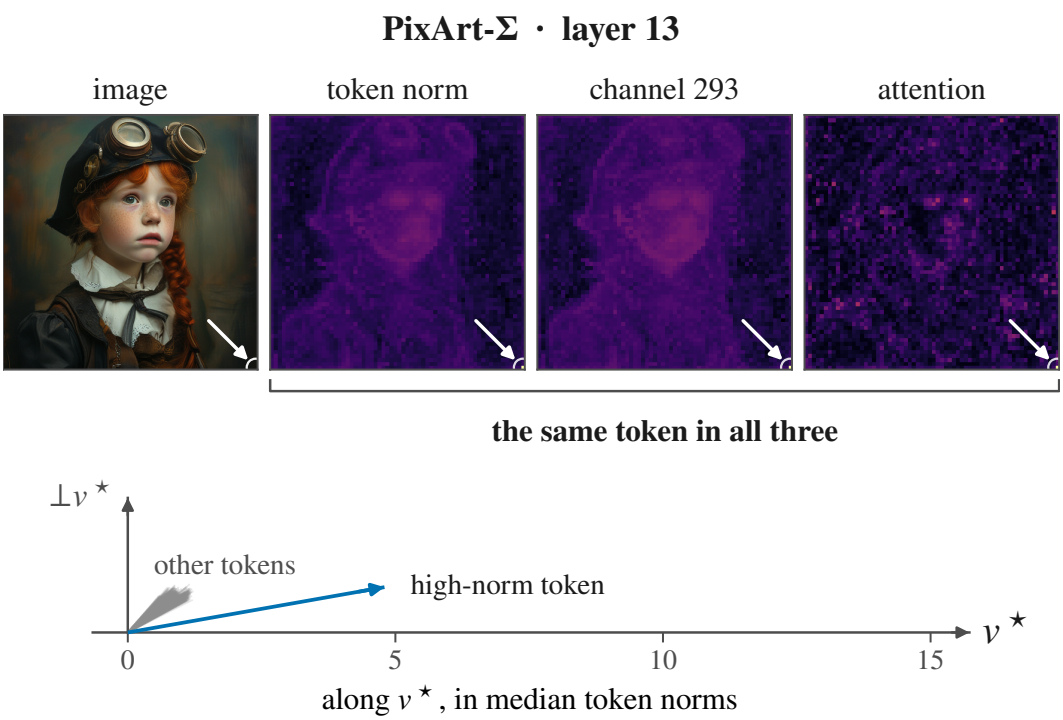
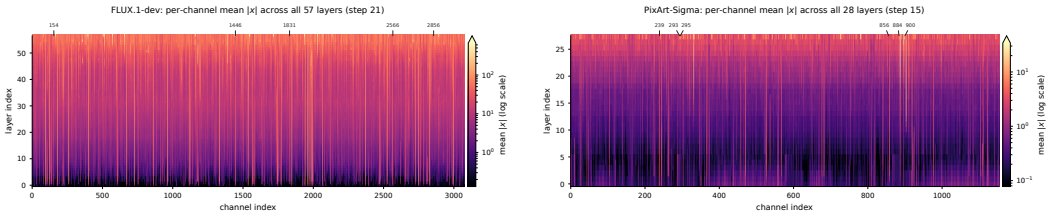


Figure 11: **PixArt-Σ: the same state in a single register token.** As Figure 1, for PixArt-Σ at layer 13 and the final denoising step (prompt H3, seed 3407). One token, at the bottom-right corner (white arrows), leads the token norm, channel 293, and attention. Its state is 4.9 times the median token norm and lies 10° from $v^\star$ ($|\cos| = 0.985$, against a median of 0.785 for the other tokens). The arrows share the scale of Figure 1.



(a) FLUX.1-Dev (57 layers)

(b) PixArt-Σ (28 layers)

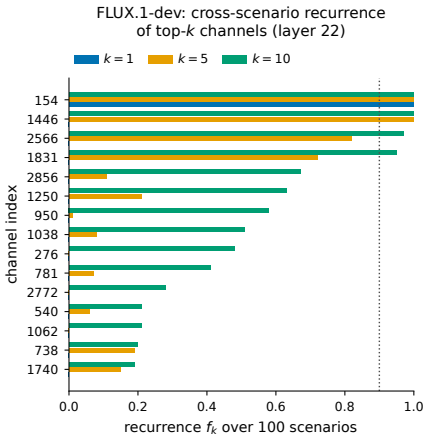
Figure 12: **A few channels stay elevated over a range of depth.** $\mathrm{mean}\,|x|$ of every channel (columns) at every layer (rows), log color scale, median over five pilot generations at the analysis step. Labels mark the largest channels and the channels that recur across generations.

C Intervention Experiments

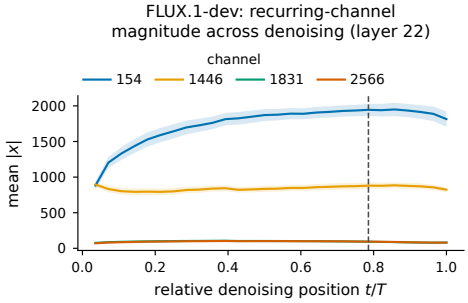
Each subsection gives the protocol of one intervention and the results that the main text does not show. Every intervention follows the paired design of Appendix A.3, with a hooks-only control.

C.1 Removing One Structure at a Time

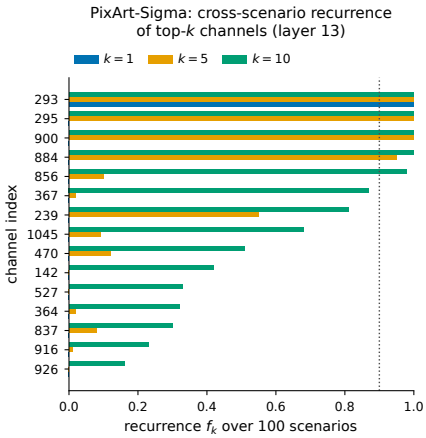
High norm, channel 154, and sink behaviour co-occur on the same register tokens (Sections 3–5). To test how they depend on one another, we remove each structure separately in one FLUX.1-Schnell generation and then measure all three quantities on the same register-token set (selected at layer 20 and denoising step 2 of 4; at most eight tokens).



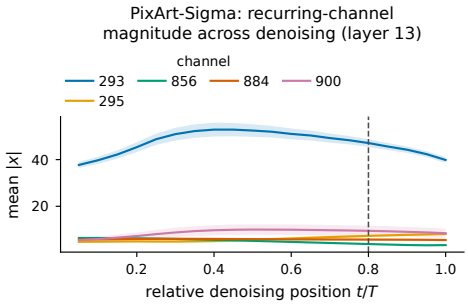
(a) FLUX.1-Dev: recurrence



(b) FLUX.1-Dev: across denoising



(c) PixArt-Σ: recurrence



(d) PixArt-Σ: across denoising

Figure 13: **The top channels are the same across generations and steps.** (a, c) Fraction of the 100 generations in which a channel ranks in the top k by $\mathrm{mean}\,|x|$; the dotted line marks 0.9. (b, d) $\mathrm{mean}\,|x|$ of the recurring channels across denoising at the analysis layer (mean and one standard deviation over the 100 generations); the dashed line marks the analysis step.

Replacing the register tokens with matched ordinary tokens removes all three signatures: only 17% of the original high norm remains, channel 154 falls to 13%, and sink strength disappears. Suppressing channel 154 removes the channel itself and reduces the token norm to 16% of its original value, but leaves most sink strength intact (92% remaining). By contrast, suppressing sinks by replacing the register keys removes the sink almost completely (1% remaining) while leaving both high norm and channel activation essentially unchanged (101% of each).

Thus, the dependence is one-way: the sink does not create the register state, but removing the register state removes the sink. Channel 154 contributes to the high-norm state, but suppressing it at these layers does not by itself eliminate sink behaviour.

C.2 Norm Reduction and Direction Replacement

The two ablations of Section 6 edit the high-norm register tokens of FLUX.1-Dev at the input of one layer, at every denoising step, and let the edited state propagate through all later layers. *Norm reduction* rescales each register token to the ordinary-token scale and keeps its direction; *direction replacement* gives it a random direction and keeps its norm. A head is *affected* if its top-1 image key in the hooks-only run is an edited token, and it *keeps* its sink if its top-1 image key after the edit is the

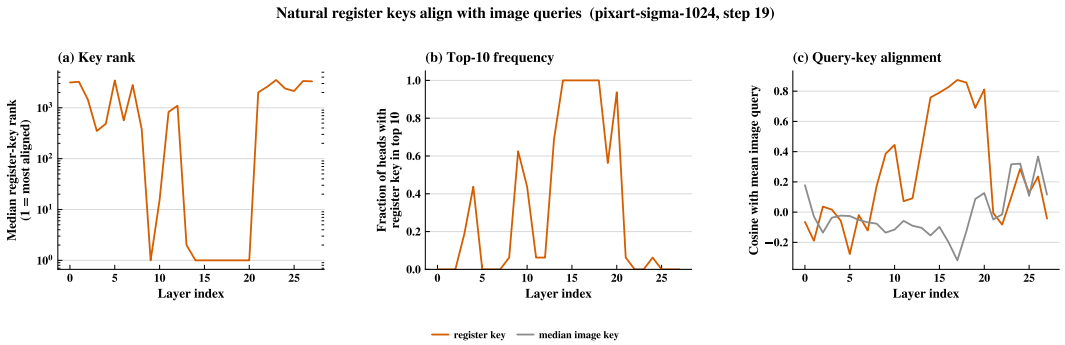


Figure 14: **Register keys are aligned with image queries in PixArt-Σ.** As Figure 5: (a) median rank of the register key among image keys by alignment with the mean image query (1 is most aligned), (b) fraction of heads with the register key in their top 10, and (c) cosine with the mean image query, for the register key and the median image key.

same token. Heads whose sink is not an edited token cannot be changed by the edit and are excluded. Figure 16 shows both ablations.

C.3 Alignment with $v^\star$ at Fixed Norm

Figure 4 changes only the alignment of the register tokens with $v^\star$ and keeps the norm of every token exactly. It uses FLUX.1-Schnell with five DiffusionDB prompts at seed 0, at denoising step 2 of 4, and edits the tokens that are both high-norm and aligned with $v^\star$.

(a) Scaling the $v^\star$ component. Each register token is written as $x = \alpha v^\star + r$ with $r \perp v^\star$. Its $v^\star$ component is scaled by β, and the result is rescaled to $\|x\|_2$: $\beta = 1$ leaves the token unchanged, $\beta = 0$ removes the component, and $\beta = -0.25$ reverses it. The edit is applied at layer 18, the last dual-stream block, where the register is written; at layer 20, after the transition to single-stream blocks; or at layer 24, inside the register zone. Sink strength is the largest share of attention that any head gives the edited tokens at the edited layer, relative to the uniform share. At layer 18 the block itself restores most of the removed component (cosine 0.98 with $v^\star$ at $\beta = 0$), so sink strength only halves there.

(b) Rotating the state off $v^\star$. At layer 18, each register state is rotated by θ in the plane spanned by $v^\star$ and a second unit vector: the mean of the ordinary tokens, their first principal component, or a random direction orthogonal to $v^\star$. The panel reports the share of affected heads that keep their sink (Appendix C.2), averaged over layers 20 to 39. No edit in this experiment changes the image by more than 0.05 LPIPS.

C.4 Holding the Direction or the Norm through Depth

This experiment asks how the register state ends. It uses the same generations as previous and follows the register, the largest-norm image token at the first register layer, at denoising step 21 of 28 (FLUX.1-Dev) and 15 of 20 (PixArt-Σ). Each condition acts at every denoising step and at every layer from 18 (FLUX.1-Dev) or 14 (PixArt-Σ) to the last. *Direction held*: the register is reset onto $v^\star$ at every layer, with its norm unchanged. *Norm held*: its norm is kept at or above its unmodified peak, with its direction unchanged. The second massive channel, which grows as the register fades (1446 in FLUX.1-Dev, 295 in PixArt-Σ), is set to zero in every token, kept from growing, or has its growth doubled; an unrelated channel (774 and 239) is set to zero as a control. Each endpoint is the deepest layer at which a property holds: the largest token's norm is at least four times the median token norm (*end*; the largest token rather than the tracked one, so that a register that moves is still followed), the register's cosine with $v^\star$ is at least 0.9 (*alignment*), or the register is the top sink of at least half of the heads at a probe layer (*sink*). A pre-declared margin of ± 1 layer defines no change. Table 4 reports every condition.

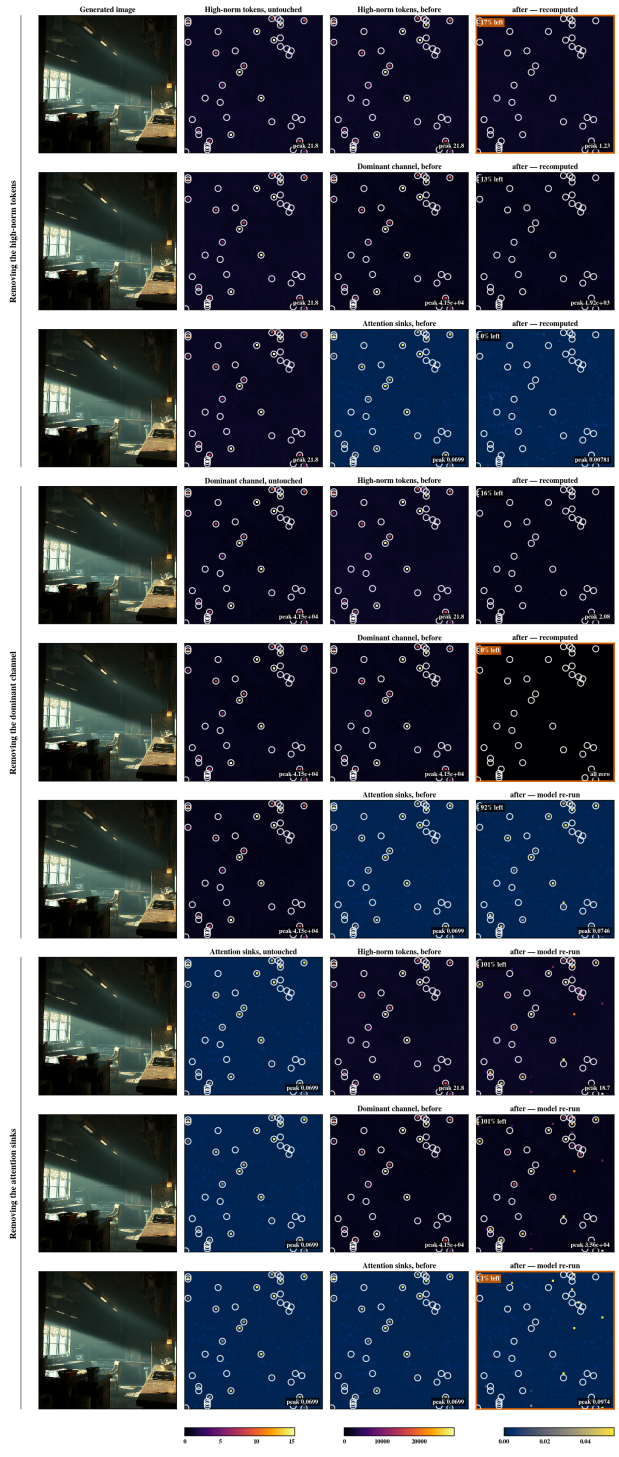


Figure 15: Removing each structure separately reveals a one-way dependence. Replacing the register tokens removes high norm, channel 154, and sink behaviour together; suppressing channel 154 reduces high norm but leaves most sink strength as attention is redirected to other tokens; suppressing sink behaviour leaves high norm and channel activation essentially unchanged.

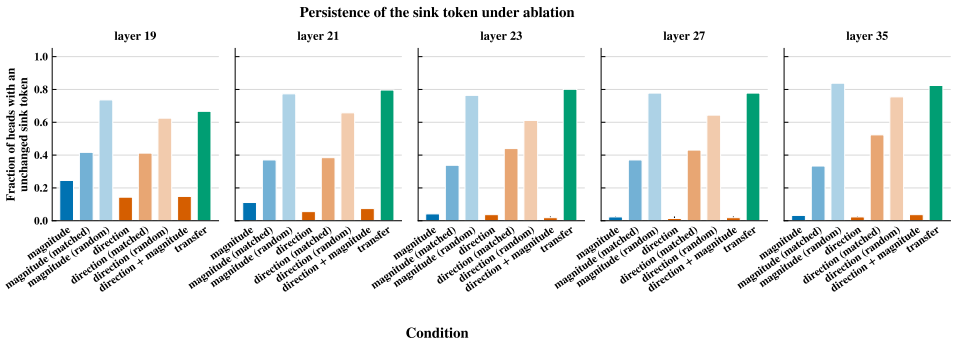


Figure 16: **Direction, rather than extreme norm, determines sink identity.** Reducing the norm of a register while preserving its direction leaves much of its sink identity intact, whereas changing its direction while preserving its large norm causes the sink to move.

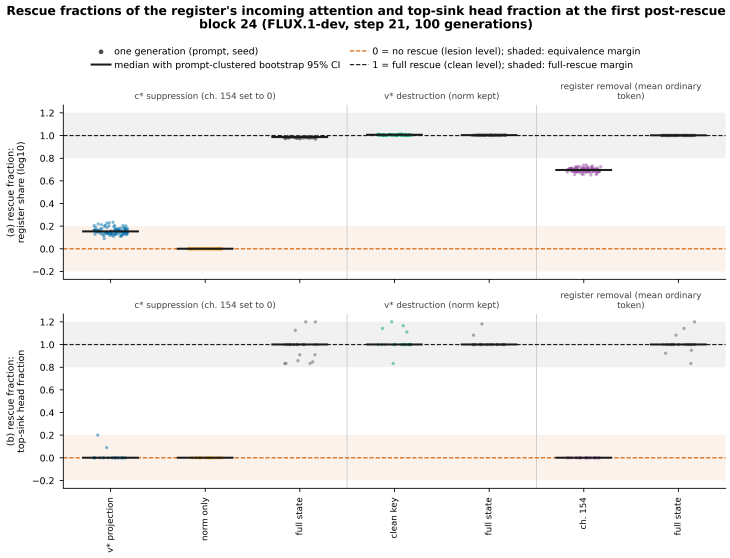


Figure 17: **Restoration after one-layer register damage in FLUX.1-Dev.** Rescue fractions of the register's incoming attention share and top-sink head fraction at the first post-rescue layer. Restoring the norm alone does not recover sink behaviour, whereas restoring the full clean register produces near-complete rescue.

Results. Holding the FLUX.1-Dev register on $v^\star$ keeps its sink (Section 6) but does not slow the loss of its norm. Holding its norm keeps its direction ($\cos = 0.99$ through layer 50) and its sink (85 of 100 generations at layer 50), the division of labor of Section 8. The competing channel does not end the register: setting it to zero, keeping it from growing, or doubling its growth leaves the end within the margin in both models. These conditions change the typical token, which also carries the competing channel (without it, the median token norm is 7 to 15% lower in FLUX.1-Dev and up to 40% lower in PixArt-Σ), while the register follows its unmodified schedule. The register ends because specific layers remove its $v^\star$ component: layer 21 in PixArt-Σ (between the outputs of layers 20 and 21, its norm falls from 850 to 107 and its cosine with $v^\star$ from 0.95 to 0.45), and layers 32 to 46 in FLUX.1-Dev, where layer 35 alone removes about a quarter of the projection.

Caveats. The conditions act at every denoising step, so later steps see edited latents. In PixArt-Σ, every condition but the hooks-only control changes the image drastically (6.6 to 12.4 dB), so these conditions probe the lifecycle rather than normal generation. The sink that the held conditions keep

Table 4: **Holding the direction or the norm through depth** (100 generations per model and condition). First row of each model: the endpoints of the unmodified run. Other rows: median paired shift of each endpoint in layers, with the number of generations in which it moved up / did not move / moved down, and the PSNR of the final image to the unmodified image. *By construction of the condition. †Dominated by the register moving to another token.

Condition	End	Alignment	Sink	PSNR (dB)
FLUX.1-Dev, step 21 of 28				
Unmodified: endpoint layer	36	42	39	–
Direction held	0 (20/54/26)	$+14^{*}$	$+11$ (100/0/0)	35.4
Norm held	$+19^{*}$	$+14$ (94/0/6)	$+11$ (87/3/10)	36.7
Channel 1446 set to zero	0 (38/48/14)	-3	$-11.5^{\dagger}$	18.5
Channel 1446 kept from growing	0 (38/50/12)	-3	$0^{\dagger}$	19.1
Channel 1446 growth doubled	-1 (8/34/58)	-2.5	$-18^{\dagger}$	18.5
Control channel 774 set to zero	0 (6/86/8)	0	0	30.6
Hooks only	0	0	0	99
PixArt-Σ, step 15 of 20				
Unmodified: endpoint layer	20	20	20	–
Direction held	$+1$ (62/38/0)	$+7^{*}$	$+6$ (89/11/0)	10.8
Norm held	$+4^{*}$	$+3$ (96/0/4)	$+6$ (86/13/1)	7.0
Channel 295 set to zero	0 (0/100/0)	0	0	6.8
Channel 295 kept from growing	0 (0/100/0)	0	0	6.9
Channel 295 growth doubled	0 (0/100/0)	-1	0	12.4
Control channel 239 set to zero	0 (0/100/0)	0	0	6.6
Hooks only	0	0	0	99

in PixArt-Σ is carried by a subset of queries: at layer 23, 19% of the heads select it, and its key ranks last against the mean image query.

C.5 Retiming the Register State

Section 10 writes the register state at other depths than its own, removes it, or keeps it past its end, and measures the effect on the final image.

Generations. Each model generates prompts H0–H9 with all five seeds, 50 generations per condition. The layer windows, conditions, and contrasts were fixed beforehand on the calibration prompts C0–C3. Every edit acts on the first third of the denoising steps (0–8 of 28 and 0–6 of 20), where composition is set. In PixArt-Σ, each edit is applied to both the conditional and the unconditional evaluation, at the register positions of the conditional one, which the unconditional evaluation shares in all 50 generations.

Conditions. The register interval runs from formation, the first layer at which the register state reaches half of its peak, to the natural end, the last layer (18–41 in FLUX.1-Dev and 13–20 in PixArt-Σ). To *induce* the state, we write the natural register's own $v^{\star}$ projection and norm, as multiples of each layer's median token norm, at the natural register positions over one window of layers: three windows before formation (E1–E3), one inside the interval (N, a rewrite in place that measures the write itself), and three after the natural end (L1–L3), all of equal length. To *remove* the state, we subtract the $v^{\star}$ component of every token that passes the register test, from one layer before formation to the natural end. To *maintain* it, we keep writing the natural state from two layers before it begins to decline through the last layer of L1. Three controls use the layers of E3: a random direction orthogonal to $v^{\star}$ with the same projection and norm, the same state at positions that never carry the register, and $v^{\star}$ at the largest projection that tokens of the recipient layer already hold.

Results. Table 5 reports every condition, and Table 6 the pre-declared paired contrasts. In FLUX.1-Dev, inducing the state before formation or after the natural end changes the image by the same small amount (LPIPS 0.07 to 0.09), about twice an in-place rewrite (0.04), and CLIP-I stays above 0.98:

these edits change detail rather than content. The same projection and norm along a random direction change the image about as much as removing the register (0.17 and 0.18). In PixArt-Σ, an early write is tolerated as well (0.09), but inducing the state after the natural end, maintaining it, or removing it changes the image substantially (LPIPS 0.70 to 0.94). The single PixArt-Σ register receives 8.2% of all attention, against 0.57% for one of the 37 FLUX.1-Dev register tokens, and no second register can take over its role. In neither model does inducing the state after removing it bring the image back toward the unmodified one (last two rows of Table 6). Figures 18 and 19 show three prompt–seed pairs per model.

Table 5: **Retiming the register state: effect on the image.** Layers: the edited layers of each condition; after a removal, the induction window that follows it. LPIPS (AlexNet) between each edited image and the unmodified image of the same prompt and seed (mean over 50 prompt–seed pairs, with 95% intervals), and the cosine similarity of their CLIP ViT-L/14 image embeddings (CLIP-I, mean).

	FLUX.1-Dev			PixArt-Σ		
Condition	**Layers**	**LPIPS ↓**	**CLIP-I ↑**	**Layers**	**LPIPS ↓**	**CLIP-I ↑**
Induced before the register forms						
E1	5–8	0.089 [0.065, 0.113]	0.987	6–7	0.095 [0.074, 0.119]	0.981
E2	9–12	0.086 [0.063, 0.110]	0.988	8–9	0.088 [0.071, 0.108]	0.986
E3	13–16	0.087 [0.062, 0.112]	0.988	10–11	0.091 [0.071, 0.114]	0.986
Rewritten at the natural depth						
N	19–22	0.042 [0.025, 0.061]	0.995	14–15	0.082 [0.067, 0.101]	0.988
Present after the natural end						
L1	43–46	0.089 [0.072, 0.106]	0.986	22–23	0.700 [0.603, 0.779]	0.753
L2	47–50	0.084 [0.066, 0.102]	0.989	24–25	0.769 [0.699, 0.823]	0.704
L3	51–54	0.071 [0.058, 0.083]	0.989	26–27	0.801 [0.739, 0.848]	0.682
Maintained	29–46	0.096 [0.078, 0.114]	0.985	16–23	0.755 [0.681, 0.814]	0.724
Natural register removed						
Removed	17–41	0.181 [0.135, 0.227]	0.968	12–20	0.913 [0.867, 0.958]	0.542
Removed, then E3	13–16	0.173 [0.131, 0.217]	0.969	10–11	0.915 [0.888, 0.939]	0.525
Removed, then L1	43–46	0.183 [0.143, 0.223]	0.966	22–23	0.943 [0.910, 0.977]	0.526
Controls at the layers of E3						
Random direction	13–16	0.168 [0.129, 0.211]	0.973	10–11	0.869 [0.835, 0.899]	0.676
Non-register positions	13–16	0.098 [0.072, 0.128]	0.986	10–11	0.082 [0.065, 0.101]	0.989
In-distribution $v^\star$	13–16	0.050 [0.034, 0.069]	0.993	10–11	0.078 [0.061, 0.098]	0.989

Table 6: **Paired contrasts of the retiming experiment.** Difference in LPIPS to the unmodified image between two conditions on the same prompt–seed pairs (mean over 50 pairs, with 95% intervals). °The interval includes zero.

Contrast	**FLUX.1-Dev**	**PixArt-Σ**
E3 − N	$+0.045$ [$+0.029$, $+0.064$]	$+0.009$ [$+0.002$, $+0.017$]
L1 − N	$+0.048$ [$+0.033$, $+0.062$]	$+0.618$ [$+0.528$, $+0.694$]
E3 − random direction	-0.081 [-0.105, -0.062]	-0.777 [-0.815, -0.740]
E3 − non-register positions	-0.012 [-0.021, -0.003]	$+0.010$ [$+0.002$, $+0.018$]
E3 − in-distribution $v^\star$	$+0.036$ [$+0.020$, $+0.056$]	$+0.013$ [$+0.005$, $+0.021$]
Removed, then E3 − removed	-0.008 [-0.022, $+0.004$]°	$+0.002$ [-0.032, $+0.039$]°
Removed, then L1 − removed	$+0.002$ [-0.011, $+0.014$]°	$+0.030$ [-0.001, $+0.057$]°

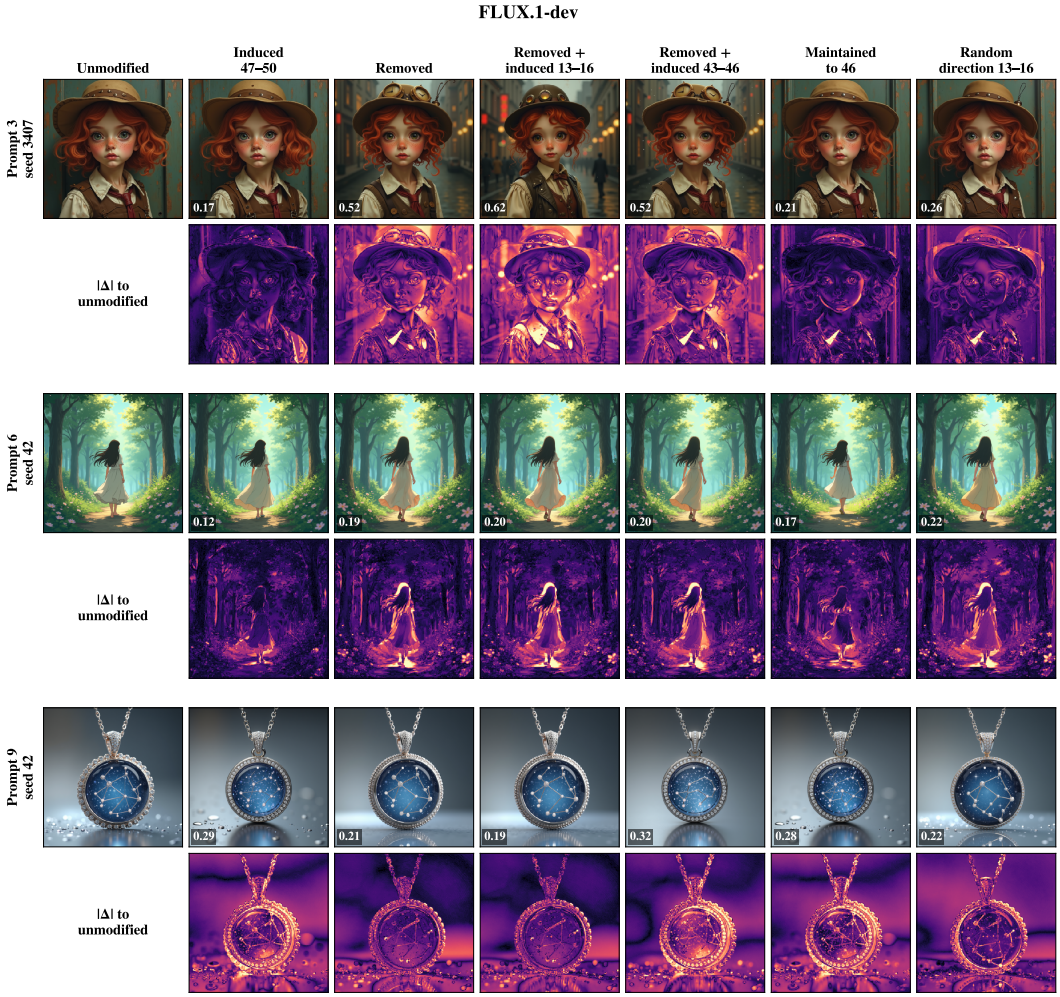


Figure 18: **Retiming examples in FLUX.1-Dev.** Rows: prompts H3, H6, and H9 of Table 2, with the seeds shown, generated unmodified and under the six conditions that change these three generations most (columns). Below each generation: its absolute difference to the unmodified image (mean over color channels, pixel values in $[0, 1]$, one square-root color scale from 0 to 0.42 for the whole figure). Corner numbers: LPIPS to the unmodified image. We chose the rows; each changes more than the average pair under every condition shown (Table 5). The first row is the generation of Figure 1 and Figure 7.

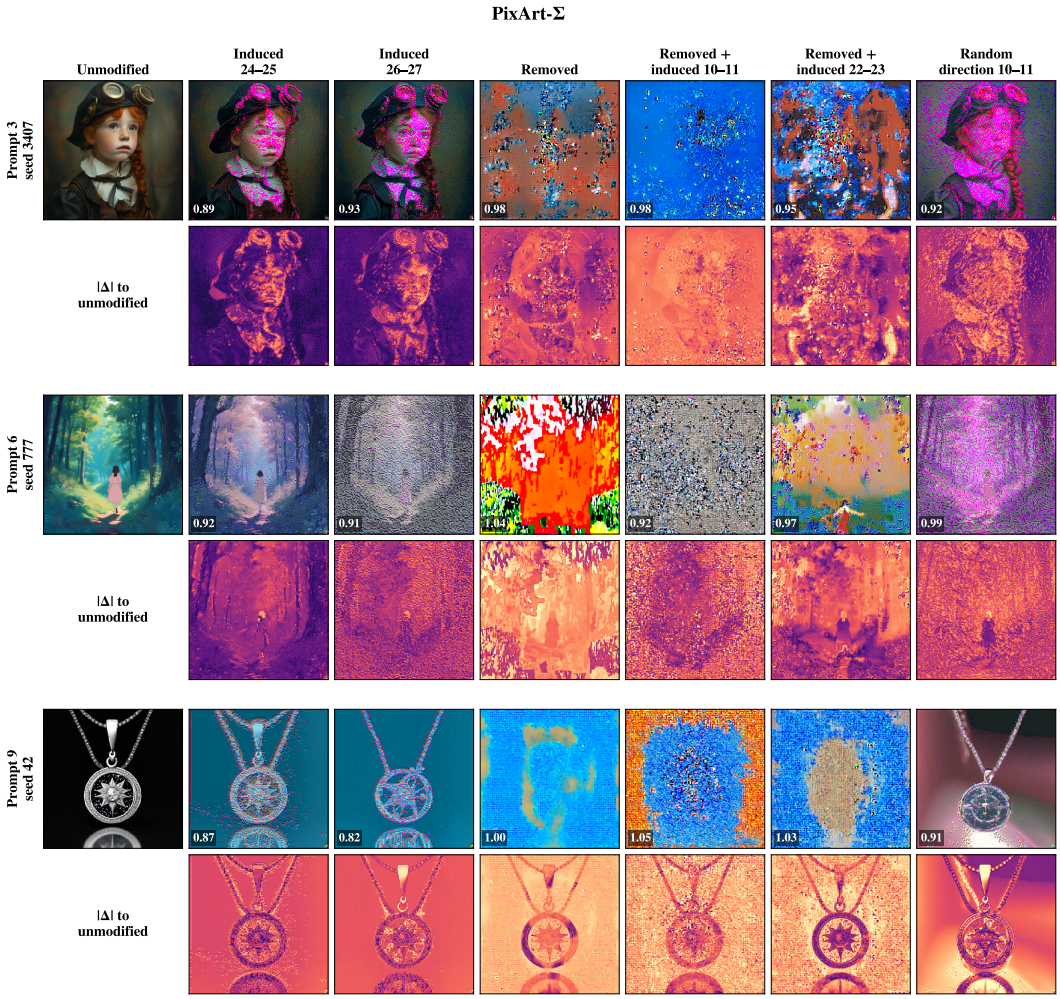


Figure 19: **Retiming examples in PixArt-Σ.** As Figure 18, with a color scale from 0 to 0.75. The first row is the generation of Figure 11 and Figure 7. Inducing the state before formation (LPIPS 0.09 on average; Table 5) is not among the columns shown.