Diffusion Transformers Use Sink Registers

Amna Jamal[1,2], **Mika Tan**[1,3], **Clarissa Aurelia Nahid Saputra**[1,4], **Quân Huỳnh**[1,5], **Antonio Mari**[1,6,7]
Kevin Zhu[1],

[1]Algoverse AI Research
[2]COMSATS University Islamabad
[3]University of British Columbia
[4]University of California San Diego
[5]University of Economics and Law at Vietnam National University
[6]Swiss Federal Technology Institute of Lausanne
[7]Supervisor
antonio@algoverseairesearch.org

Abstract

Diffusion Transformers (DiTs) have recently replaced U-Net backbones as the dominant architecture in state-of-the-art text-to-image generative models, achieving remarkable visual fidelity. However, their internal mechanisms remain largely unexplored. In this work, we investigate the emergence of *high-norm activations* within DiTs—tokens with unusually large magnitudes that resemble the "outlier" tokens previously identified in Vision Transformers (ViTs). Through a systematic analysis of four DiT architectures, we find that only Flux-Schnell and PixArt-σ exhibit such activations in the image stream, primarily concentrated in the central transformer layers. Using linear probes and qualitative ablations, we show that these activations encode global or semantic image information, while their removal has negligible effect on the generation process. We refer to these as *sink registers*, reflecting their passive, semantic role. Our findings highlight an architectural divergence between ViTs and DiTs, and contribute to a deeper interpretability of diffusion-based generative models.

Code — https://github.com/AmnaJamalKhattak/Antonio-acwk

Datasets — https://huggingface.co/datasets/MikaTan2007/DiffusionTransformersUseSinkRegisters-LinearProbe

Introduction

Diffusion Transformers (DiTs) (Peebles and Xie 2023) have rapidly replaced U-Net–based diffusion models as the dominant paradigm in high-quality image generation. Despite their success, the internal mechanisms driving DiTs remain poorly understood. In particular, we observe the emergence of *tokens with unusually high activation norms* across layers. These artifacts are visually and statistically prominent in models such as Flux (Labs 2024), as shown in Figure 1.

Interestingly, these tokens resemble the "outliers" identified in ViTs, which (Darcet et al. 2024) showed they store and process global image information. Recent work (Darcet et al. 2024; Jiang et al. 2025) has demonstrated that managing such outliers through architectural interventions can substantially improve ViT stability and output consistency. However, the function and implications of high-norm activations in DiTs, models optimized for *generation* rather than *discrimination*, remain unclear.

While both DiTs and ViTs share the same underlying transformer architecture (Vaswani et al. 2017), they differ fundamentally in learning objectives and data flow. ViTs process a single image in one forward pass to produce a class prediction, whereas DiTs iteratively denoise a noisy input over multiple diffusion steps to generate an image. Consequently, we cannot directly infer the role of high-norm activations in DiTs from their behavior in ViTs. These activations might facilitate the generative process, interfere with it, or merely reflect architectural artifacts.

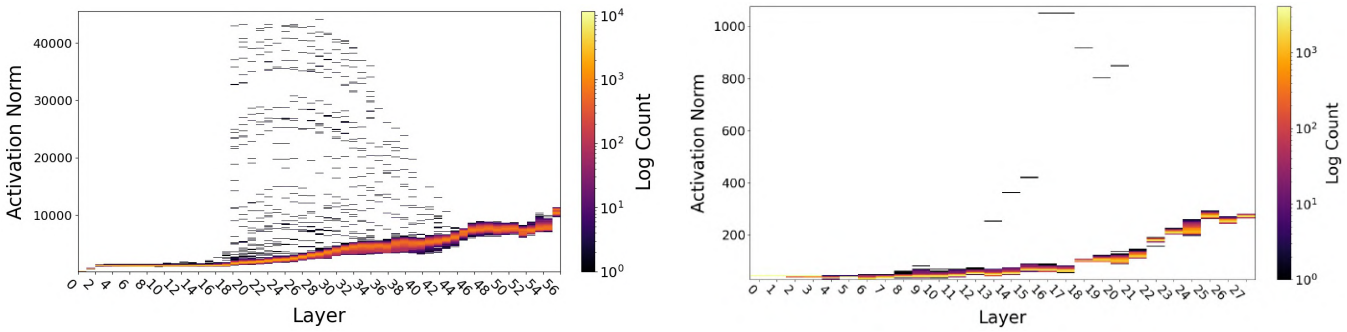
Understanding this phenomenon is essential for *interpreting how complex models represent and transform information during image synthesis*. In our work, we aim to investigate these activations as a window into the internal reasoning of diffusion-based generative models. By doing so, we hope to contribute to move beyond "black-box" understanding toward uncovering interpretable, possibly human-aligned mechanisms underlying image generation.

Specifically, we (1) systematically analyze the behavior, and consistency of high-norm activations in DiTs; (2) test their role with linear probe and ablations to reveal different behaviors than registers in DiTs, finding that they do encode semantic information, but can be can be removed without affecting the image generation process.

We believe the phenomenon we study opens pathways to understanding and re-engineering internal representations to enable scientific insight and control in generative AI.

Related Work

Registers (Darcet et al. 2024) first identified the emergence of "high-norm" outlier tokens in large Vision Transformers (ViTs), showing that a small fraction of patch embeddings develop norms an order of magnitude larger than the rest and disproportionately carry global image information. They propose adding learnable register tokens during training to absorb this global context, which both eliminates the outliers and significantly improves dense-



(a) Flux-Schnell (b) PixArt-σ

Figure 1: We plot the distribution of norm of latent activations across each layer in Flux-Schnell and PixArt-σ.

prediction tasks like segmentation and object discovery without harming classification accuracy. Building on this mechanistic insight, (Jiang et al. 2025) demonstrate a test-time intervention, without any retraining by detecting the few "register neurons" that drive the outliers and redirecting their activations into a single untrained token appended at inference. This "post-hoc register" approach matches or exceeds the benefits of trained registers, offering a lightweight, training-free path to cleaner attention maps and stronger interpretability across classification, zero-shot segmentation, and vision–language models.

Attention in Video Diffusion Transformers (Wen et al. 2025) extend the focus to Video Diffusion Transformers (VDiTs), performing the first in-depth analysis of self-attention in video diffusion backbones. They uncover three consistent patterns: Structure (stable spatial–temporal motifs that enable zero-shot video editing via attention-map transfer); Sparsity (most attention weights can be pruned if critical layers are preserved, yielding about 70% sparsity with no perceptual loss); and Sinks (later-layer heads that collapse queries into a single token and can be removed without harming video quality). These findings provide practical guidance for controlling and optimizing attention in generative video models.

Massive Activation The phenomena of high-norm activations has also been observed in Large Language Models (LLMs) as shown in (Sun et al. 2024). Therefore, it is worth noting that these are not unique only to computer vision transformers like DiTs and ViTs. Massive activations refer to activations with a magnitude around 10,000 times greater than the median. These activations in LLMs act as indispensable bias terms, as shown when setting just a few of these activations to zero caused catastrophic collapse in model performance.

The paper also investigates massive activations in ViTs, verifying their existence in many models, such as CLIP ViT-L and DINOv2 ViT-L, but not all, like MAE. The paper does not make a clear distinction between the definition of massive activations in LLMs versus ViTs. However, massive activations in ViTs are described as activations with significantly larger magnitudes than the median, and are very few in number (less than 4 observed per model). Just like in LLMs, setting massive activations to zero led to significant drop in accuracy, while setting them to the median had negligible effect, therefore indicating that massive activations in ViTs also act as fixed but crucial biases. Additionally, when the same analysis is applied to ViTs with register tokens as per the (Darcet et al. 2024) recommendation to improve interpretability and downstream performance, massive activations did not appear in patch tokens as was the case in regular ViTs, but rather, appeared almost exclusively within a fixed register token, specifically register 3, therefore suggesting that register tokens are effective at managing this outlier behavior.

Attention Sinks More recent work has explored complementary interventions on related architectures. (Wang, Zhang, and Salzmann 2024) diagnoses and corrects norm anomalies in self-supervised DINOv2 models by adaptively re-scaling defective patch embeddings, further improving object localization and interpretability. (Xiao et al. 2024) shows that certain Transformer heads in language models similarly act as "attention sinks," and that pruning or rerouting these heads at inference can dramatically reduce latency in streaming applications without sacrificing accuracy.

High-norms appear in central layers

We analyze four DiTs, namely Flux-Schnell (Labs 2024), PixArt-σ (Chen et al. 2024), Stable-Diffusion 3 (Peebles and Xie 2023), and NVIDIA-SANA (Xie et al. 2024), and find high-norm activations on the image latents only in the first two. Note also that both Flux-Schnell and Stable-Diffusion 3 exhibit high-norm activations in the text-sequence stream, but their analysis is left for future work.

Figure 1 illustrates distributions of norms of latents for each layer activations of the DiTs. We observe how the two amount of high-norm activations is different between models but they consistently appear in central transformer layers, (18-39 for Flux-Schnell, 13-18 for PixArt-σ).

Flux-Schnell We work with 4-steps generations with Flux and consistently observe high-norm activations starting from layer 18 until layer 39 for all time steps in the image stream, as observed in Figure 1.

The positions of the outlier activations vary across prompts and seeds as observed in Appendix 5.

To further investigate the origin of high-norm tokens, we analyzed both the attention and MLP layers within the Single and Dual blocks. As shown in Figure 2a, the norm values immediately following the attention layers do not exhibit significant increases, indicating that attention does not produce high-norm activations. In contrast, Figure 2b demonstrates that the MLP layers in Dual Blocks consistently generate high-norm activations, suggesting that these layers are the primary contributors to extreme token norms in Dual Blocks. Further examples are given in appendix.

Interestingly, this pattern does not hold for Single Blocks. It does not produce high-norm activations after the MLP; instead, high-norms emerge only after the Attention+Residual operation as seen in 7. This discrepancy between Dual and Single blocks can be attributed entirely to structural differences in the blocks, which dictate how extreme activations propagate through the network. Furthermore, we observed the emergence of Attention sinks in the blocks where Attention layers were responsible for the high-norm activations i.e, Single blocks in Flux-Schnell as seen in 11 and not Dual Blocks in 10.

PixArt-σ We analyze 10-steps generations for this model, observing that the artifacts appear consistently over all time steps, prompts, and seeds. They first appear in layer 13 and persist in the image-stream until layer 20. Unlike Flux, PixArt-σ does not update a text-stream in its architecture, it only conditions the generation on fixed textual embeddings. Interestingly, according to Figure 1 there seem to be much fewer high-norm activations present in PixArt-σ compared to Flux-Schnell.

Notably, in PixArt-σ, we find a single high-norm activation which emerges consistently at the bottom right of the image regardless of prompt or seed used. Furthermore, we conducted attention and mlp analysis for all the blocks. As shown in Figure 9, the norm values immediately following the attention layers exhibit significant increases, indicating that attention is responsible for producing high-norm activations in PixArt-σ and not MLP.

Stable Diffusion 3 and Nvidia-SANA In SD3, high-norm activations appear only in the text stream. For the text stream of both the positive and negative prompts, high-norms are seen throughout the layers, starting from the first all the way to the last layer. One can observe that the high-norms get much more intense starting the 31st layer. Conversely, Nvidia-SANA works with fixed textual embeddings and does not show any high-norm activation at all.

Diffusion Transformers use sink registers

To investigate the role of high-norm activations in Flux-Schnell and PixArt-σ we perform both quantitative and qualitative experiments. Specifically, following (Darcet et al. 2024) we train linear probes on high-norm activations to predict global image information and compare results with the rest of activations, and following (Wen et al. 2025) we ablate high-norm activations during generation to check if the whole image gets corrupted or semantically change.

High-norm thresholds

We identify High-norm activations in both Flux and PixArt-σ by computing the L2 norm of each output activation vector. For a vector $\mathbf{h} \in \mathbb{R}^D$ ($D = 3072$ for Flux, 1152 for PixArt) at a given layer. Assuming approximate independence and Gaussianity of each vector component components, then the norm follows a chi distribution with D degrees of freedom. For Flux, high-norm activations were selected using layer-specific percentile thresholds. For middle layers 97.7th percentile was used, while for the first and last few layers 99.99th percentile was used. For PixArt-Σ, the separation between normal and high-norm activations was visually clear, so thresholds were manually set for each layer rather than computed via percentiles, shown in Appendix 6.

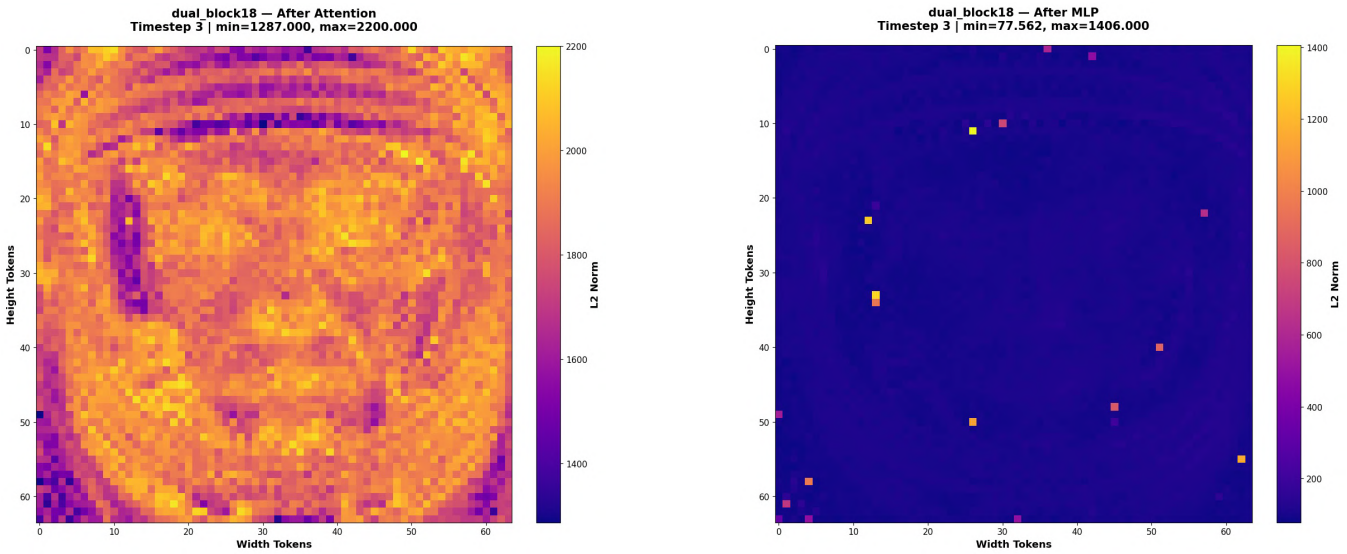
Linear Probe

We select high-norm activations as all activations whose norm exceeds the threshold computed for each specific time step and layer. We conduct a linear probe experiment on PixArt and Flux in order to determine whether high-norm activations encode semantically rich information compared to other activations. To do so, we generate 50 prompts using large language models for each CIFAR-10 (Krizhevsky 2009) class (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) and generate an image on each prompt, collecting model activations at all layers where high-norm activations arise. Activations are then split into two distinct datasets, one containing only high-norm activations and the other containing the rest. For each dataset, we train linear probes to predict the CIFAR-10 class of generated images. The probe's ability to accurately predict the class label serves as a proxy for how much semantic information is preserved in an activation.

We run all experiments using 10-fold cross validation and report average accuracy along with standard deviation. We fit all linear probes as logistic regressions with l2-regularization and perform subsample of "standard" norm activations to make sure all datasets have equal size. We report three sets of experiments:

1. Global probe: we compare accuracies between one linear probe trained on all high-norm activations, and one trained on the rest of activations.
2. Layer-Wise Probes: for each layer individually, we concatenate activations across all timesteps and probe the two sets of activations.
3. Temporal Probes: for each time step, we concatenate activations across all layers and probe the two sets of activations.

Results For PixArt-σ, activations are collected from layers 13-20 across for all timesteps, classes, and prompts, while For Flux-Schnell, we collect all activations from layers 18-39 in the same way. In total, we have $40,000$ latent activation maps for PixArt and $44,000$ for Flux, containing 4096 activations each (one for each latent image patch). For each generation, we split activations according to norm thresholds and compute average-pooling of all high-norm activations in



(a) After Attention

(b) After MLP

Figure 2: Comparison of norm values post Attention and post MLP for layer 18 in Flux.

a generation with all subsequent ones. This is repeated for other activations.

Table 1 contains results for the global probe, while appendix reports layer-wise and temporal probe results. Since the difference in accuracy is negligible (below 0.5% difference), we conclude that *high-norm activations contain an equal amount of semantic information about the image that is being generated compared to other activations, and do not play the "registers" role as reported in Vision Transformers (Darcet et al. 2024).* We baptize DiTs high-norm activations as **sink registers** for this reason.

	High-Norm		Others	
Model	**Mean**	**Std**	**Mean**	**Std**
Flux-Schnell	0.9993	0.0003	0.9999	0.0001
PixArt-σ	0.9665	0.0017	0.9711	0.0028

Table 1: Comparison of mean probe accuracy between high-norm and other activations across models.

Ablations

We visualize the effect of removing high-norm activations during image generation by reporting qualitative ablations. In practice, we select top-percentile activation values to mask, sweeping the percentile in 0.5, 1%, 5%, 10%, 50%, and 100%. As a reference intervention, we ablate randomly chosen activations instead of high-norm ones matching their numbers, as depicted in figures 3 and 4. We replace the selected activations, either with vectors sampled by a standard gaussian distribution, or with the average of activations for the same layer. The replacement happens at the first layer where the high-activation norms emerge (i.e. 18 for Flux-Schnell, 13 for PixArt-σ).

We conducted this experiment across seeds and prompts as can be referenced from Table 6, and from Figure 3 we also qualitatively conclude that ablating Flux's high-norm activations does not cause significant changes than ablating random activations, thus confirming that they are not vital for generation and representing high-level image semantics. Figure 4, shows a different pattern for PixArt. Masking top 0.02% percentile of high-norm activations barely changes the image while replacing random activations, which are unlikely to be high-norms, completely degrades the generated image. This suggests that high-norm latents' role has negligible impact in the generation, while replacing ones with lower norm (randomly) causes extremely sensitive degradation.

Conclusions and Limitations

In this work, we presented the first systematic investigation of *high-norm activations* in Diffusion Transformers, comparing their behavior to the outlier tokens previously identified in Vision Transformers. Across four representative architectures, we found that these activations appear consistently in the middle layers of certain models (Flux-Schnell and PixArt-σ) but are absent or restricted to the text stream in others (Stable Diffusion 3 and NVIDIA-SANA).

Through linear probes, we demonstrated that these high-norm activations encode an equal amount of global or semantic information about the generated image as compared to other activations. Furthermore, qualitative ablations confirmed that masking or replacing these activations has negligible or no impact on the resulting images. We therefore term these activations *sink registers*, reflecting their semantic contribution and statistical prominence.

Our findings shed light on a subtle but important difference between discriminative and generative transformer-

Figure 3: Ablation performed on high-norm activations at layer 18 for Flux-Schnell generation.

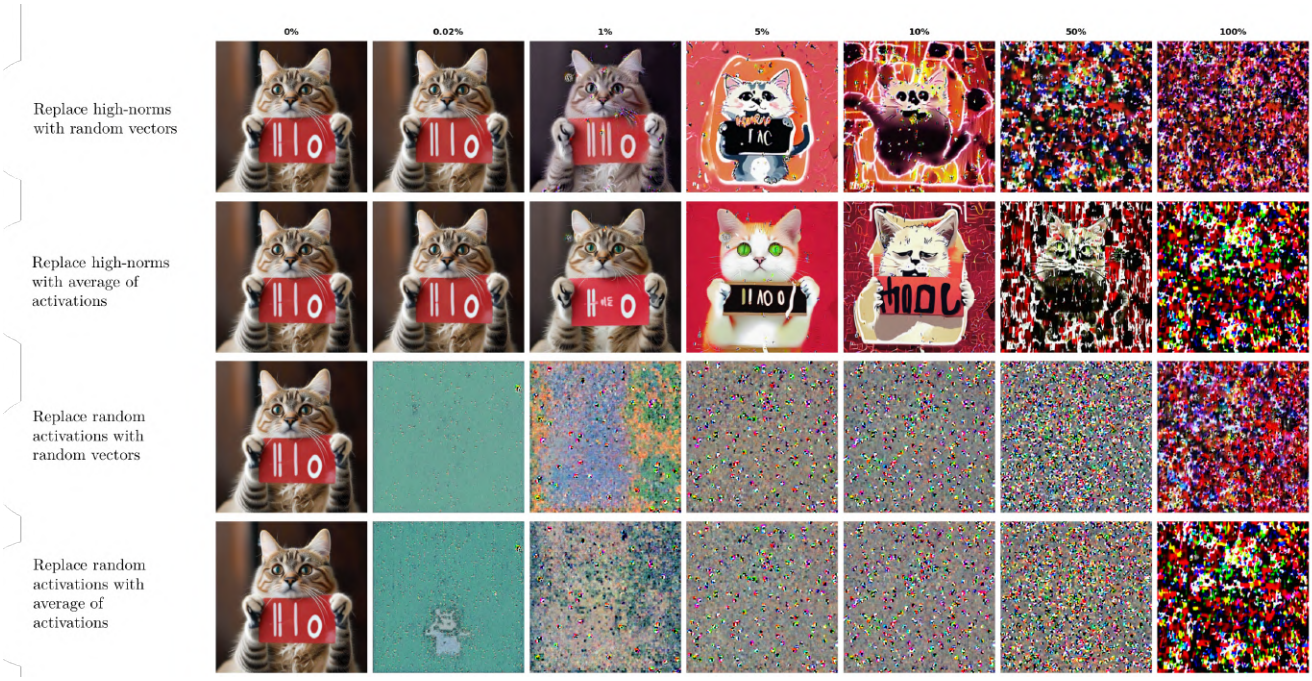


Figure 4: Ablation performed on high-norm activations at layer 13 for PixArt-σ generation.

based models: while ViTs leverage high-norm activations as structural components for global information flow, DiTs appear to develop them as architectural by-products without explicit functional purpose. Understanding such emergent behaviors is crucial for building more interpretable, controllable, and reliable generative systems—particularly as diffusion-based architectures continue to scale.

Limitations and Future Work. Our analysis focuses primarily on static activation statistics and perturbation-based evaluations. We do not yet visualize how these high-norm activations influence attention maps or specific neurons within DiT layers. Future work should examine their causal role within the attention mechanism, explore whether similar patterns emerge in text-conditioning streams, and assess

whether such artifacts impact generation stability or fairness across prompts. More broadly, identifying and characterizing these emergent structures contributes to the long-term goal of *explainable and scientifically grounded AI*, helping bridge the gap between performance-driven generative modeling and mechanistic understanding of large-scale neural systems.

Author's Contribution

Amna Jamal was responsible for running the high norm activation analysis on Flux, and prepared code for plots and visualizations. She also conducted Attention and MLP analysis to see the emergence of High norm activations in both Flux and Pixart-σ. Furthermore, Amna carried out attention sink analysis to look for any attention sink behavior in the attention heads. Clarissa Aurelia Nahid Saputra was responsible for running the high-norm activation analysis on SD3. She implemented and performed the high-norm activations ablations on both Flux and PixArt. Mika Tan was responsible for running high-norm activation analysis on PixArt-σ. He implemented and performed the linear-probe experiments on CIFAR-10 for Flux and PixArt, and is also responsible for editing the final version of the paper. Quân Huỳnh ran the high-norm activation analysis on NVIDIA-SANA. Antonio Mari, working in his capacity as the research supervisor, closely supervised and actively advised the research team, proposing the research idea and finding preliminary evidence of high-norm activations in Flux. He is also responsible for writing and polishing most sections of the paper.

References

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2024. PixArt-Σ : $Weak-to-StrongTrainingofDiffusionTransformerfor4KText-to-ImageGeneration.arXiv$: 2403.04692.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. arXiv:2309.16588.

Jiang, N.; Dravid, A.; Efros, A.; and Gandelsman, Y. 2025. Vision Transformers Don't Need Trained Registers. arXiv:2506.08010.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

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

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. arXiv:2212.09748.

Sun, M.; Chen, X.; Kolter, J. Z.; and Liu, Z. 2024. Massive Activations in Large Language Models. arXiv:2402.17762.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.

Wang, H.; Zhang, T.; and Salzmann, M. 2024. SINDER: Repairing the Singular Defects of DINOv2. arXiv:2407.16826.

Wen, Y.; Wu, J.; Jain, A.; Goldstein, T.; and Panda, A. 2025. Analysis of Attention in Video Diffusion Transformers. arXiv:2504.10317.

Xiao, G.; Tian, Y.; Chen, B.; Han, S.; and Lewis, M. 2024. Efficient Streaming Language Models with Attention Sinks. arXiv:2309.17453.

Xie, E.; Chen, J.; Chen, J.; Cai, H.; Tang, H.; Lin, Y.; Zhang, Z.; Li, M.; Zhu, L.; Lu, Y.; and Han, S. 2024. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers. arXiv:2410.10629.

Appendix

Layer-Wise Localization of High-Norm Activations

As seen in Figure 5, the position of high-norm activations change with prompts and seeds. We tried with 3 different prompts, across 3 different seeds: 0, 42, and 1234.

Attention and MLP Analysis

We further carried out Attention and MLP analysis for single blocks in Flux-Schnell and PixArt-σ. It can clearly be seen in 7 and 9, the high-norm activations emerge after Attention and before MLP for blocks in PixArt-σ and Single blocks in Flux-Schnell. Whereas for dual blocks in Flux-Schnell, they do not emerge after Attention as seen in 8.

Attention Sink Analysis

We conducted an Attention Sink Analysis for both dual and single blocks in Flux-Schnell. As illustrated in Figure 11, the attention sink phenomenon emerges exclusively in the single blocks, which exhibit high-norm activations in the attention layers, in contrast to the dual blocks in Figure 10.

Linear Probe Results

Layer	high-norm Mean	high-norm Std	other Mean	other Std
13	0.6620	0.0878	0.7100	0.0531
14	0.6720	0.0840	0.7160	0.0747
15	0.6560	0.0880	0.6820	0.0642
16	0.6760	0.0618	0.6780	0.0887
17	0.6780	0.0610	0.6500	0.0835
18	0.6600	0.0780	0.6660	0.0633
19	0.6480	0.0676	0.6640	0.0496
20	0.6540	0.0732	0.6420	0.0555

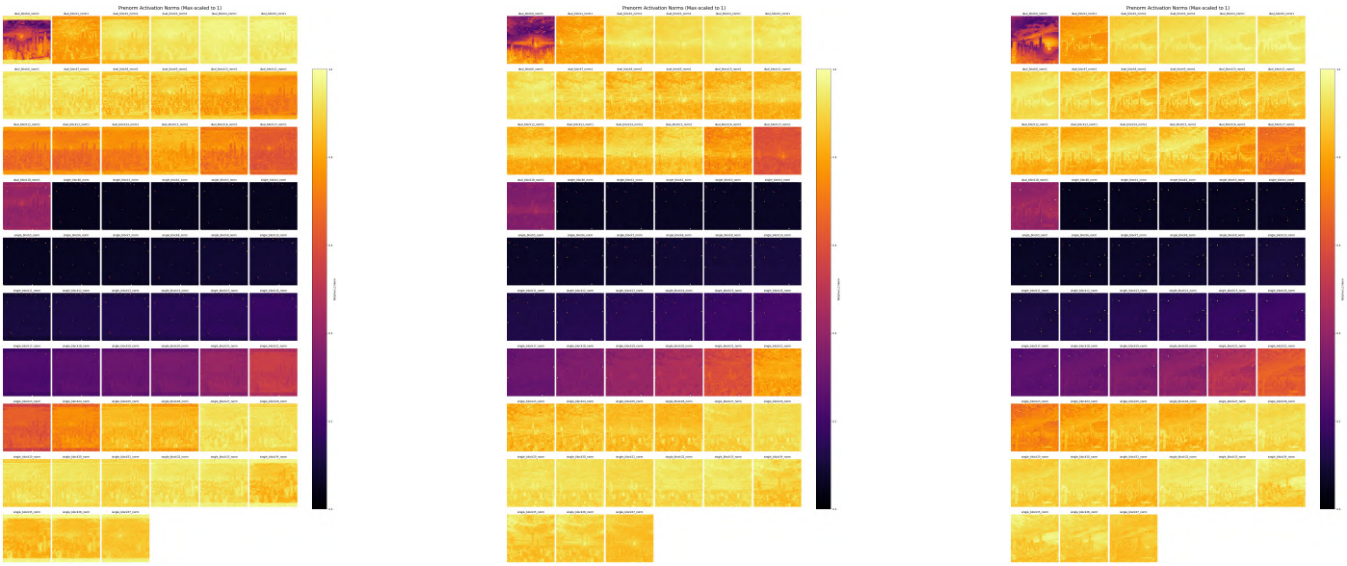
Table 2: Probe performance across layers in PixArt-σ

Ablation Results and Post-ablation Norms

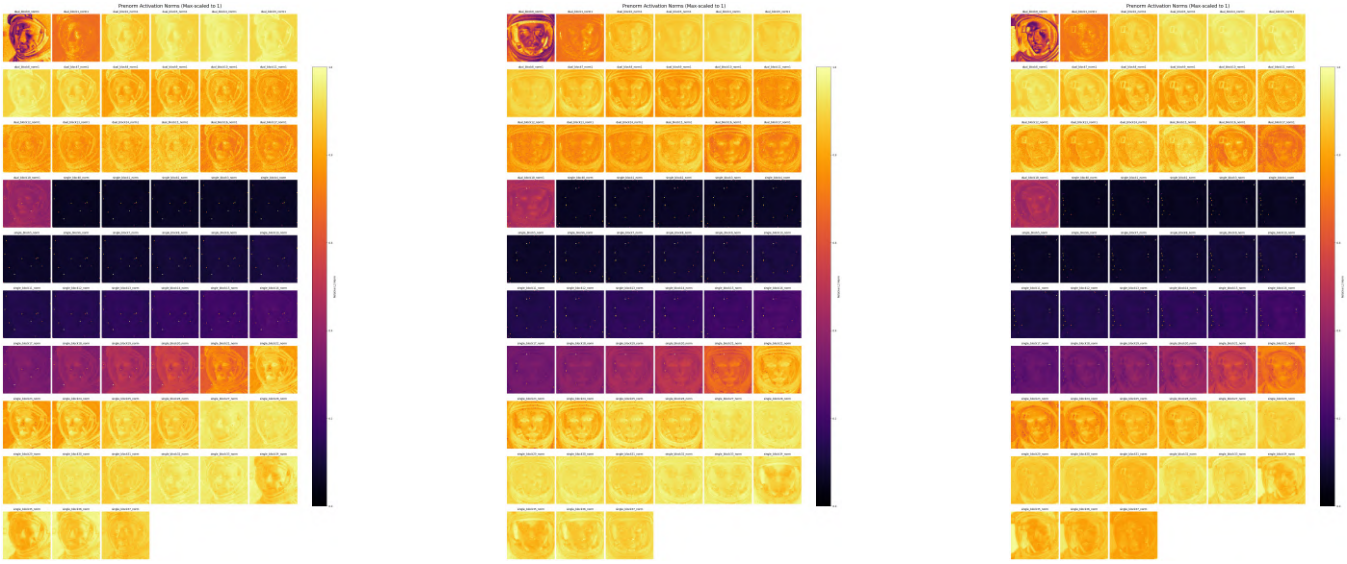
Some example of experiment configurations with prompts and seeds

As seen in Table 6, we have five configurations for both Flux-Schnell and PixArt-σ. The following figures show generations of each configuration.

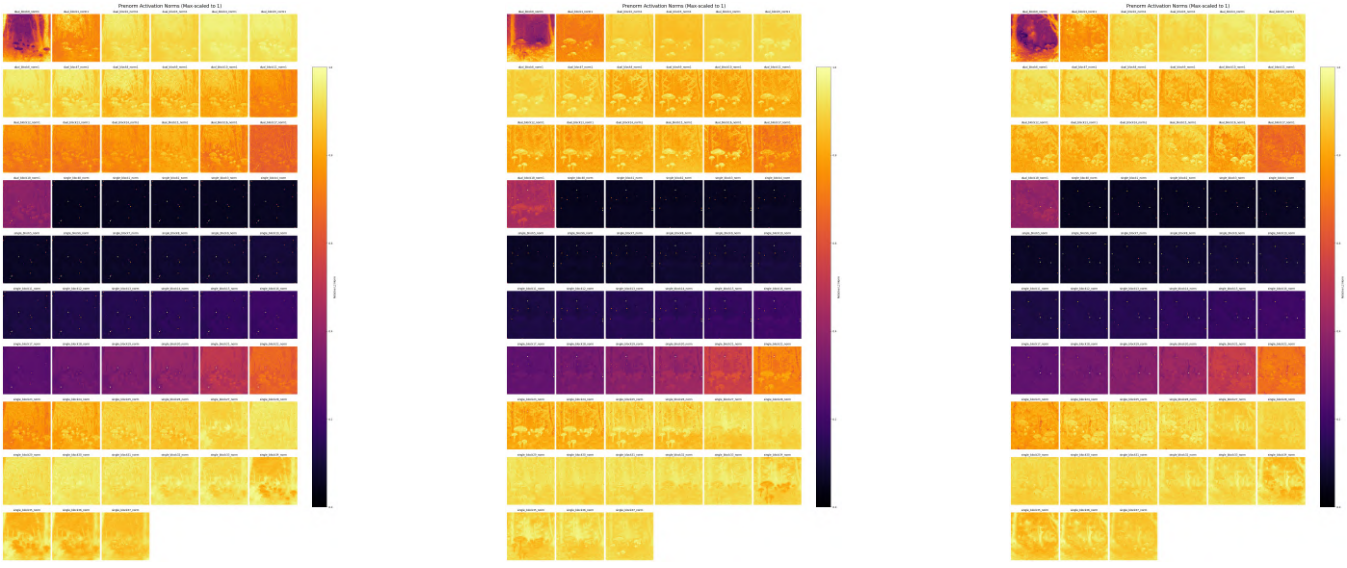
We can also plot the layer-wise distributions after the replacements to verify whether replacements are actually conducted and see the effect of changing the activations in the layers after. We have plotted some samples of 0-50-100 progressions for certain configurations and modes.



(a) Prompt 1 Seed 0 (b) Prompt 1 Seed 42 (c) Prompt 1 Seed 1234



(d) Prompt 2 Seed 0 (e) Prompt 2 Seed 42 (f) Prompt 2 Seed 1234



(g) Prompt 3 Seed 0 (h) Prompt 3 Seed 42 (i) Prompt 3 Seed 1234

Figure 5: **Flux-Schnell:** Activations across Multiple Prompts and Multiple Seeds.

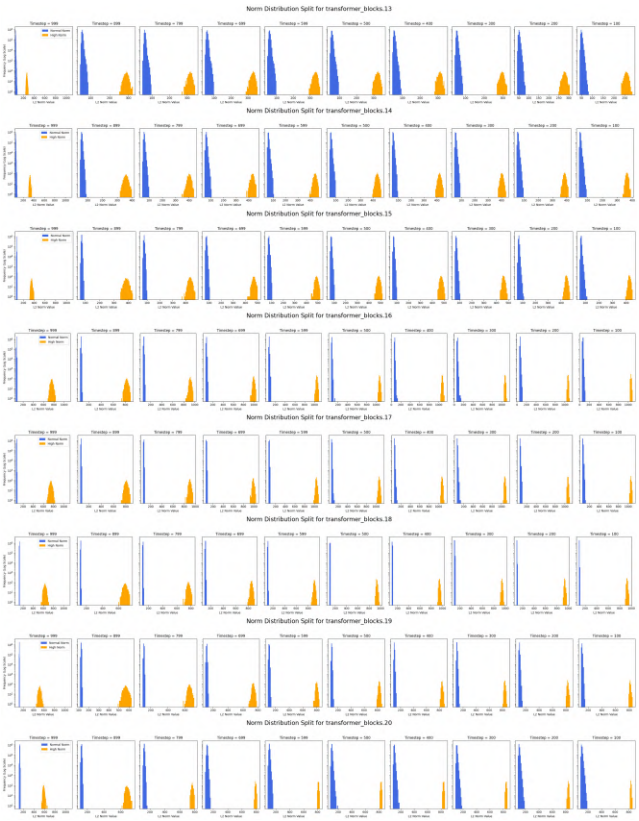


Figure 6: Histogram of activations per condition aggregated across all 500 prompts in PixArt-σ. Orange is high-norm activations, blue is other activations

Diffusion Timestep	High-Norm Mean	High-Norm Std	Other Mean	Other Std
999	0.6720	0.0688	0.6700	0.0581
899	0.6800	0.0632	0.7060	0.0770
799	0.6800	0.0710	0.7120	0.0786
699	0.6560	0.0662	0.7040	0.0852
599	0.6420	0.0740	0.6940	0.0853
500	0.6180	0.0860	0.6800	0.0876
400	0.5960	0.0656	0.6680	0.0652
300	0.5820	0.0836	0.6480	0.0553
200	0.6020	0.0583	0.6220	0.0494
100	0.5720	0.0671	0.5980	0.0404

Table 3: Probe performance across diffusion timesteps in PixArt-σ

Layer	High-Norm Mean	High-Norm Std	Other Mean	Other Std
18	0.9990	0.0016	0.9994	0.0009
19	0.9996	0.0008	0.9994	0.0009
20	0.9996	0.0008	0.9994	0.0009
21	0.9996	0.0008	0.9994	0.0009
22	0.9998	0.0006	0.9996	0.0008
23	0.9996	0.0008	0.9996	0.0008
24	0.9992	0.0013	0.9996	0.0008
25	0.9994	0.0009	0.9996	0.0008
26	0.9994	0.0009	0.9998	0.0006
27	0.9996	0.0008	0.9998	0.0006
28	0.9994	0.0009	0.9998	0.0006
29	0.9996	0.0008	0.9998	0.0006
30	0.9994	0.0009	0.9996	0.0008
31	0.9994	0.0013	0.9998	0.0006
32	0.9996	0.0008	0.9992	0.0010
33	0.9994	0.0009	0.9998	0.0006
34	0.9996	0.0008	0.9996	0.0008
35	0.9994	0.0009	0.9998	0.0006
36	0.9986	0.0016	0.9998	0.0006
37	0.9986	0.0013	0.9996	0.0012
38	0.9978	0.0028	0.9998	0.0006
39	0.9980	0.0024	0.9998	0.0006

Table 4: Probe performance across layers in Flux-Schnell

Diffusion Timestep	High-Norm Mean	High-Norm Std	Other Mean	Other Std
-1	0.9940	0.0092	0.9980	0.0060
200	0.9680	0.0349	0.9720	0.0256
300	0.9820	0.0209	0.9680	0.0256
400	0.9840	0.0196	0.9720	0.0256
500	0.9820	0.0166	0.9780	0.0227
600	0.9900	0.0100	0.9780	0.0244
700	0.9860	0.0156	0.9800	0.0237
800	0.9840	0.0174	0.9820	0.0209
900	0.9960	0.0080	0.9880	0.0204
1000	0.9920	0.0098	0.9900	0.0134

Table 5: Probe performance across diffusion timesteps in Flux-Schnell

Table 6: Experiment Configuration

#	Prompt	Seed
1	a cat holding hello world sign	0
2	a cat holding hello world sign	42
3	a cat holding hello world sign	100
4	an astronaut in the moon holding the USA flag	42
5	a protestor in a city square holding a 'change' sign	42

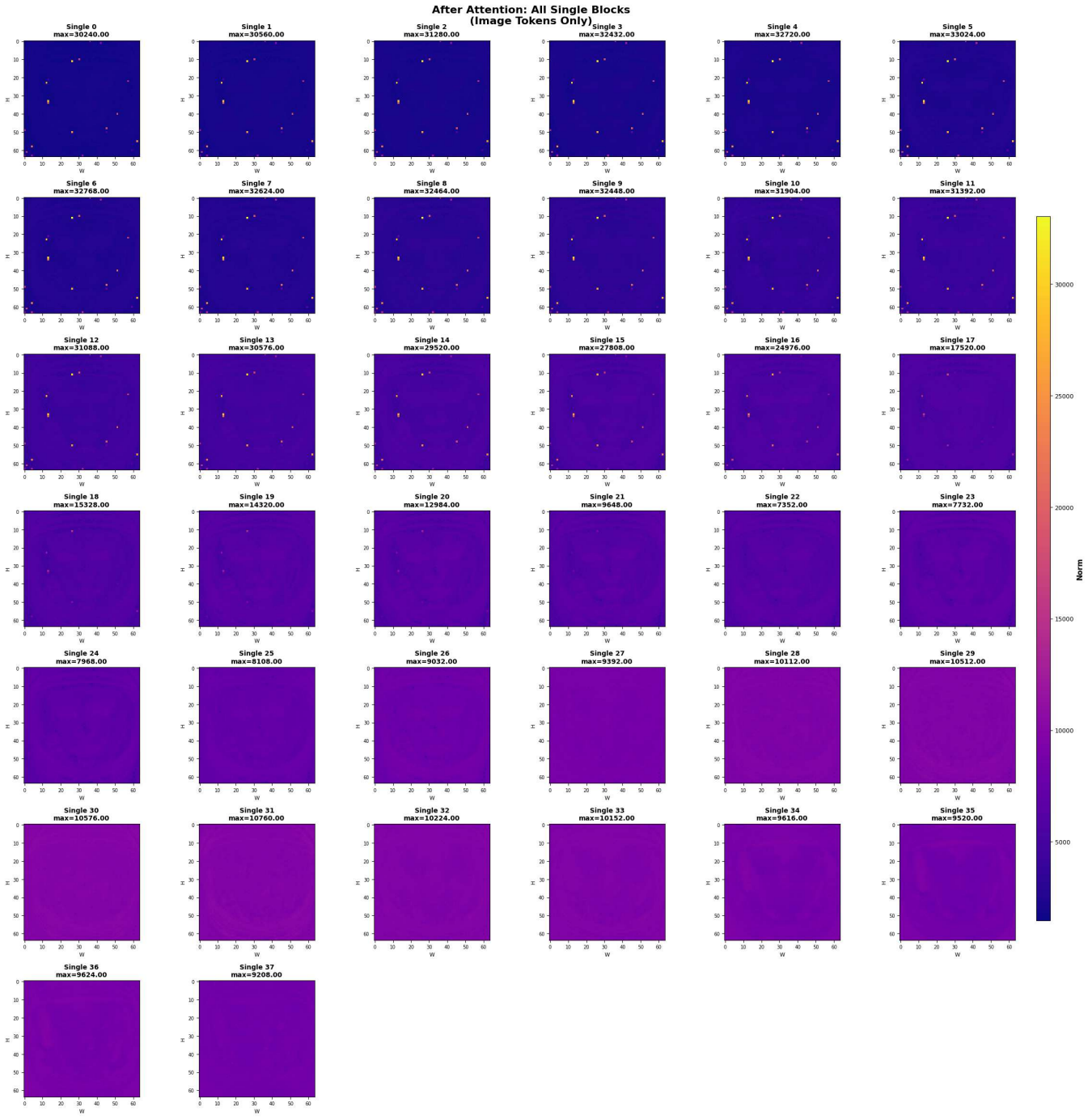


Figure 7: Heatmap of activations of all Single blocks in Flux-Schnell. high-norm activations are observed right after Attention layers.

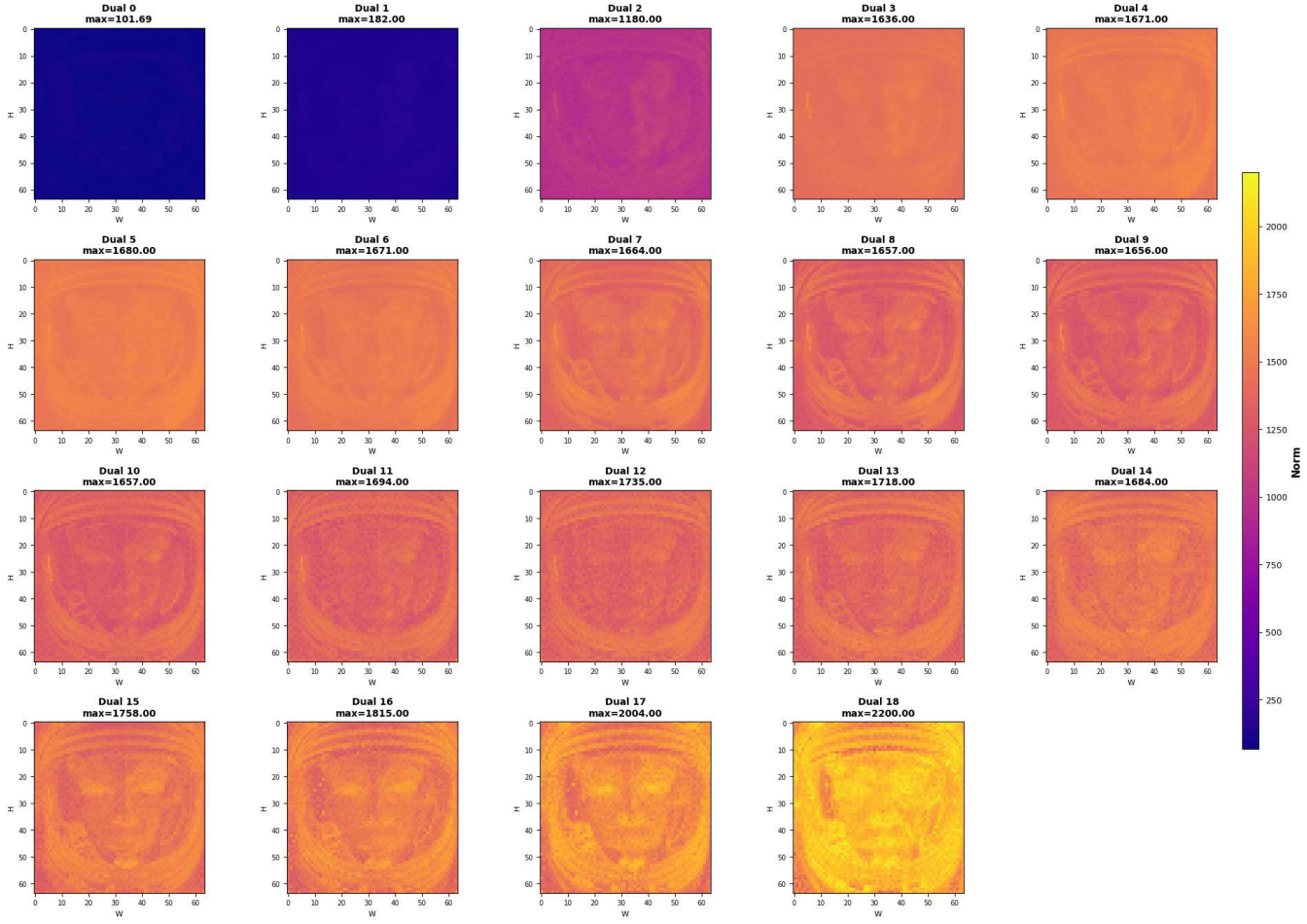
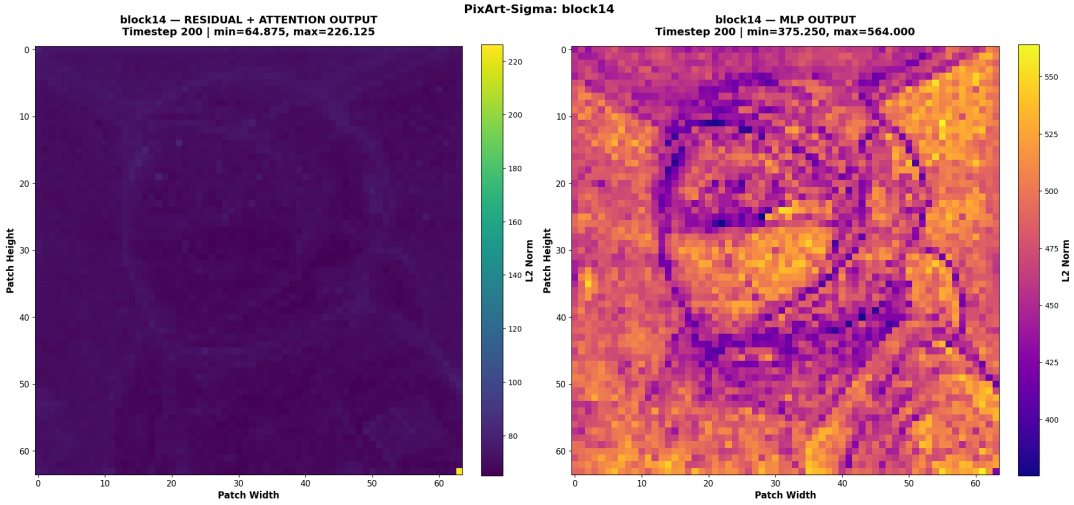
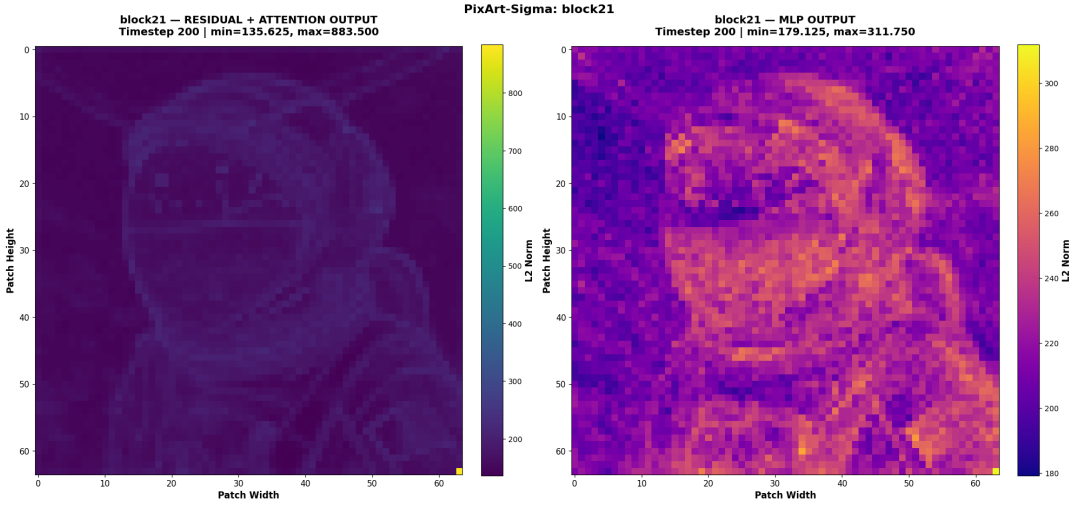


Figure 8: Heatmap of activations of all Dual blocks in Flux-Schnell. high-norm activations are NOT observed right after Attention layers.



(a) Block 14 in PixArt-σ



(b) Block 21 in PixArt-σ

Figure 9: **Attention v MLP in PixArt-σ:** high-norm activations are observed to emerge in the Attention layers, and not MLP.

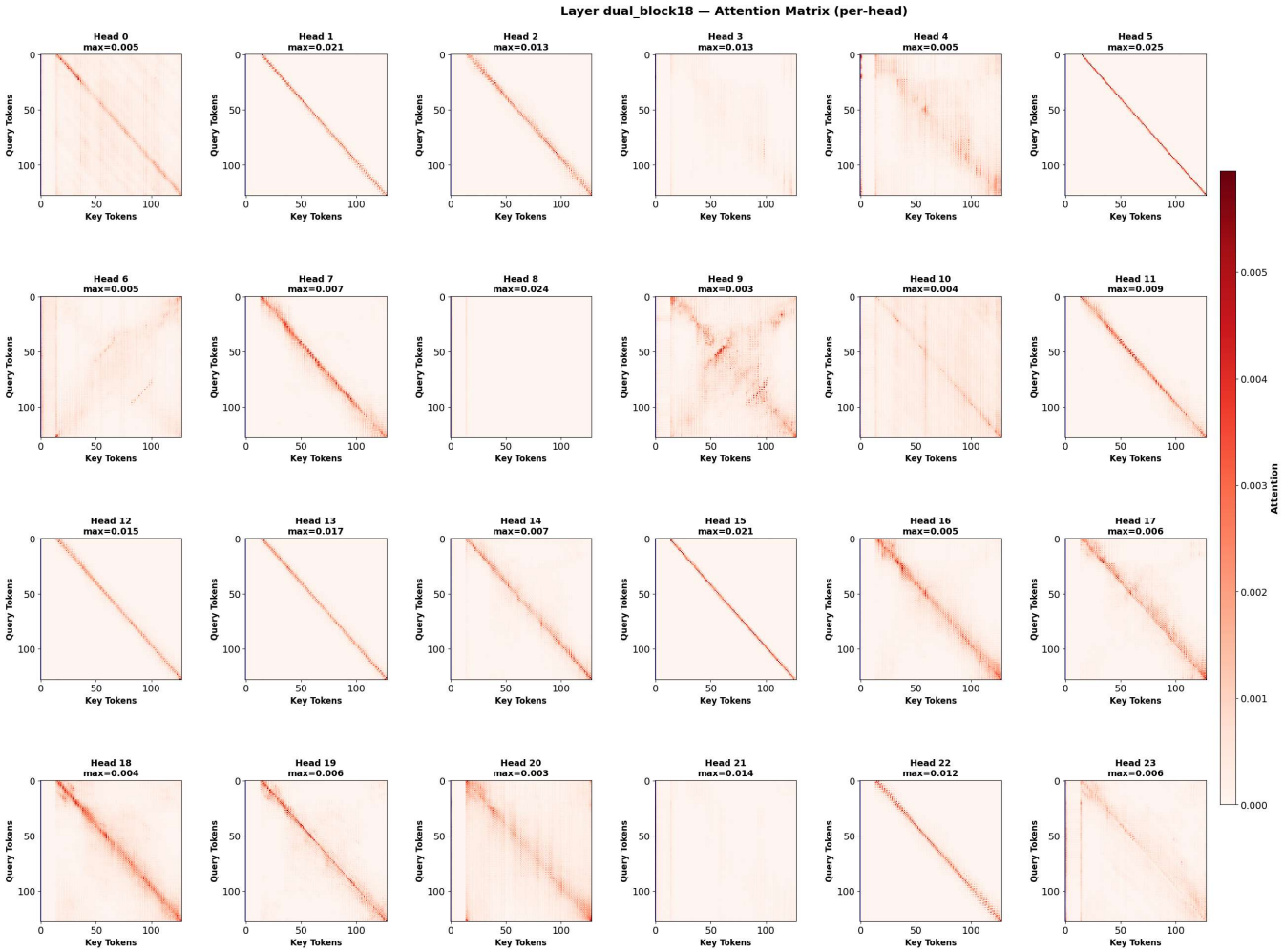


Figure 10: In Flux-Schnell Layer 18 (Dual Block 18), no Sink behaviour observed.

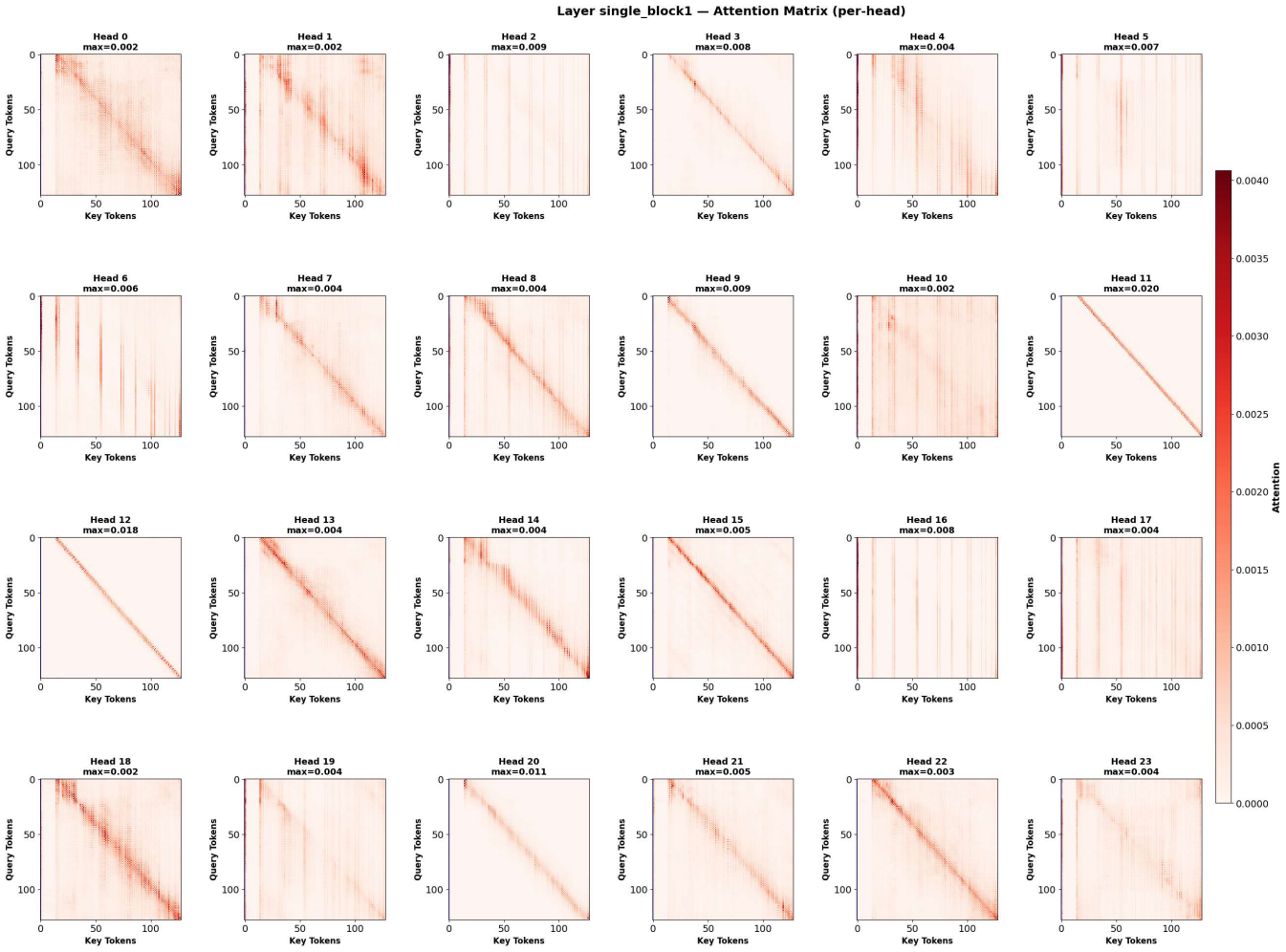


Figure 11: In Flux-Schnell Layer 20 (Single Block 1), Attention sink behavior is clearly visible in these vertical lines that are emerging in various Attention heads.

Text Stream Flux-Schnell

Image Stream Flux-Schnell

Both Streams Flux-Schnell

Figure 12: Flux-Schnell Configuration 1

Text Stream Flux-Schnell

Image Stream Flux-Schnell

Both Streams Flux-Schnell

Figure 13: Flux-Schnell Configuration 3

Text Stream Flux-Schnell

Image Stream Flux-Schnell

Both Streams Flux-Schnell

Figure 14: Flux-Schnell Configuration 4

Text Stream Flux-Schnell

Image Stream Flux-Schnell

Both Streams Flux-Schnell

Figure 15: Flux-Schnell Configuration 5

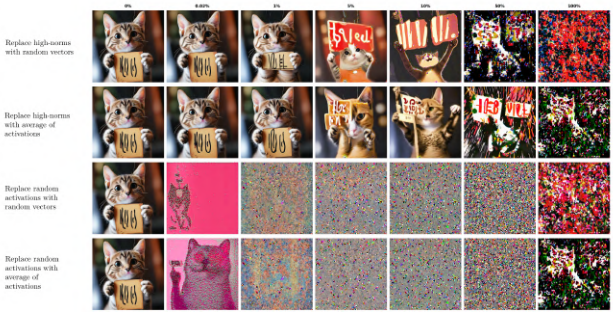


Figure 16: PixArt-σ Configuration 1

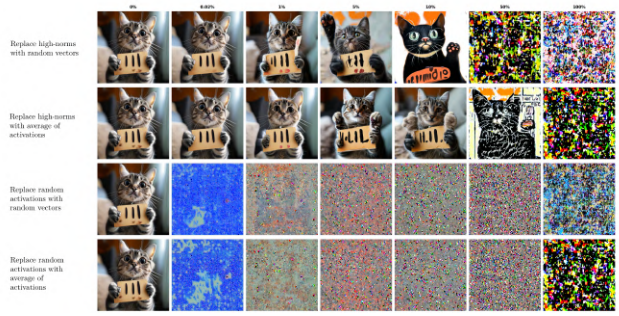


Figure 17: PixArt-σ Configuration 3

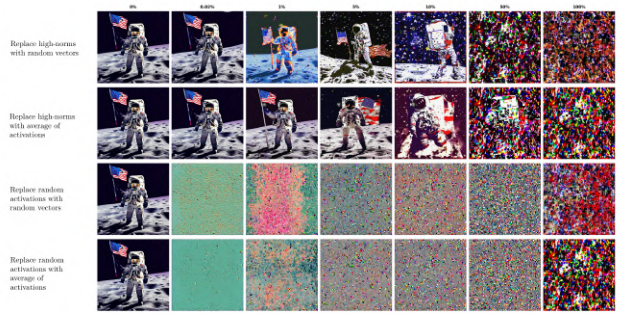


Figure 18: PixArt-σ Configuration 4

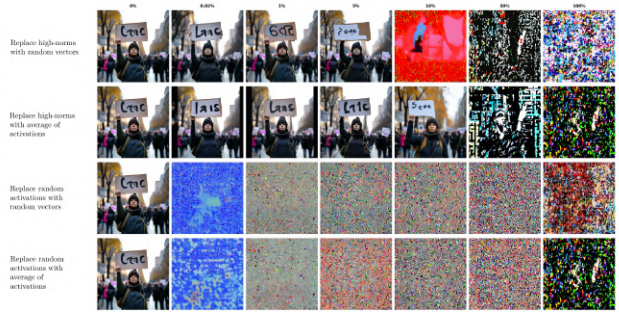


Figure 19: PixArt-σ Configuration 5

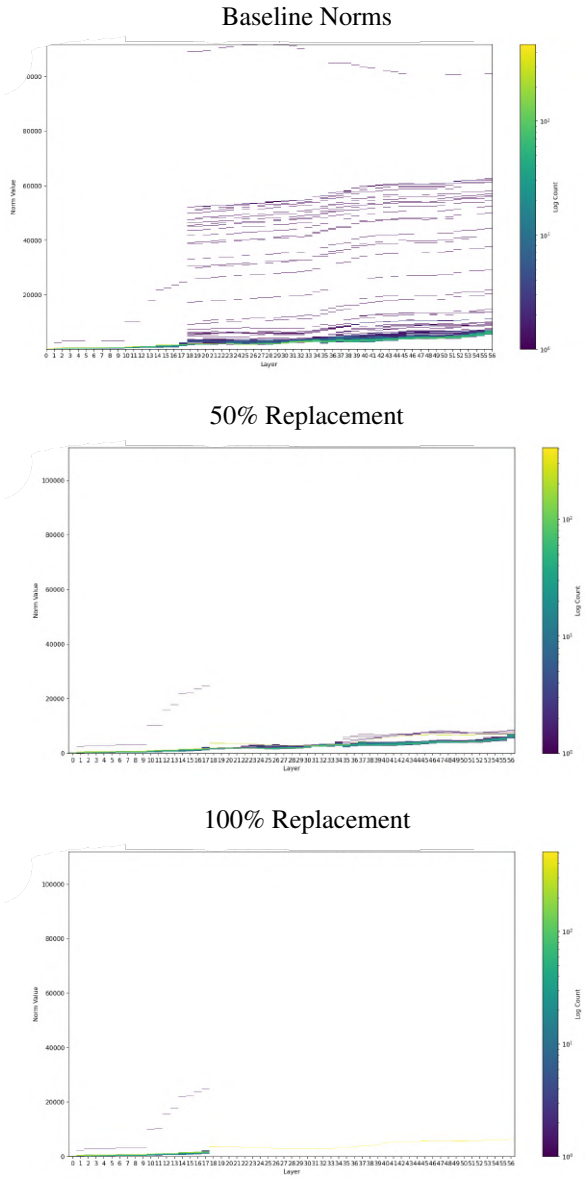


Figure 20: Flux-Schnell Configuration 2 Text Stream, in which high-norm activations are replaced with the mean activation values.

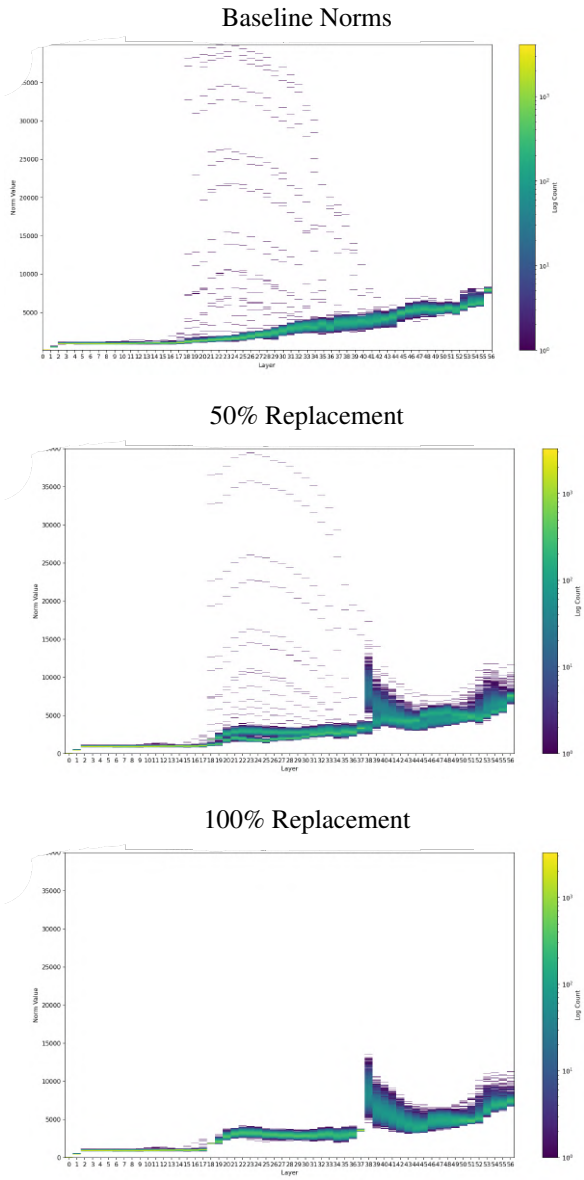


Figure 21: Flux-Schnell Configuration 1 Image Stream, in which random activations get replaced with random vectors.

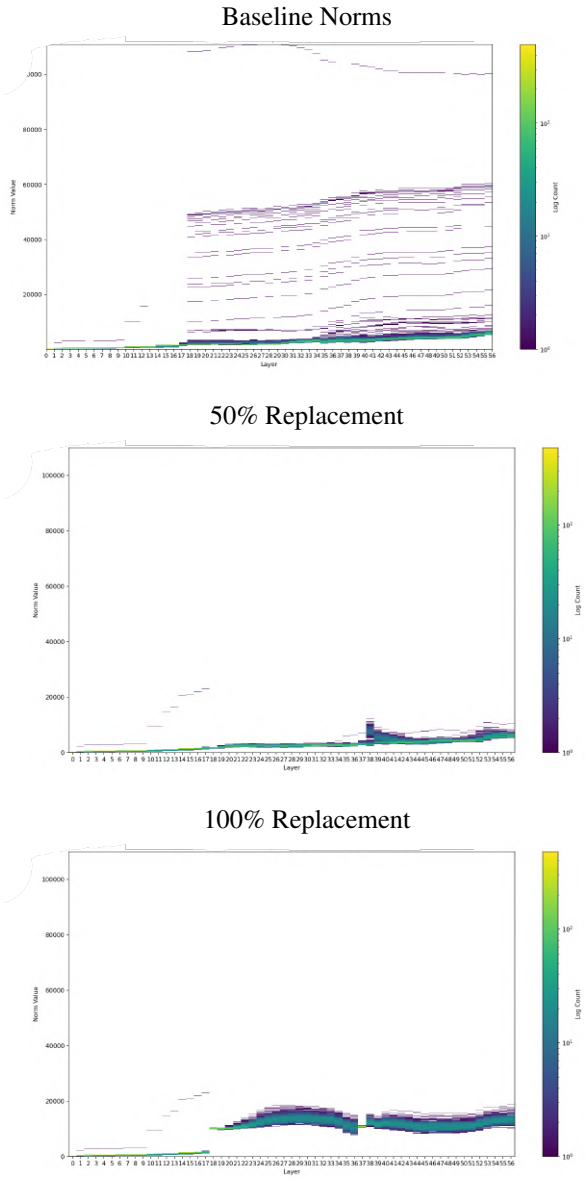


Figure 22: Flux-Schnell Configuration 5 Text Stream, in which high-norm activations get replaced with random vectors.