Few Channels Draw The Whole Picture: Revealing Massive Activations in Diffusion Transformers

Evelyn Turri*[1] Davide Bucciarelli*[1,2]
Sara Sarto[1] Lorenzo Baraldi[1] Marcella Cornia[1]
[1]University of Modena and Reggio Emilia, Italy [2]University of Pisa, Italy
{name}.{surname}@unimore.it *Equal contribution

aimagelab.github.io/MAs-DiT

Abstract

Diffusion Transformers (DiTs) and related flow-based architectures are now among the strongest text-to-image generators, yet the internal mechanisms through which prompts shape image semantics remain poorly understood. In this work, we study *massive activations*: a small subset of hidden-state channels whose responses are consistently much larger than the rest. We show that, despite their sparsity, these few channels effectively *draw the whole picture*, in three complementary senses. First, they are functionally critical: a controlled disruption probe that zeroes the massive channels causes a sharp collapse in generation quality, while disrupting an equally-sized set of low-statistic channels has marginal effect. Second, they are spatially organized: restricting image-stream tokens to massive channels and clustering them yields coherent partitions that closely align with the main subject and salient regions, exposing a structured spatial code hidden inside an apparently outlier-like subspace. Third, they are transferable: transporting massive activations from one prompt-conditioned trajectory into another, shifts the final image toward the source prompt while preserving substantial content from the target, producing localized semantic interpolation rather than unstructured pixel blending. We exploit this property in two use cases: *text-conditioned* and *image-conditioned semantic transport*, where massive activations transport enables prompt interpolation and subject-driven generation without any additional training. Together, these results recast massive activations not as activation anomalies, but as a sparse prompt-conditioned carrier subspace that organizes and controls semantic information in modern DiT models.

1 Introduction

Text-to-image generation has shifted from convolutional denoisers to Transformer-based architectures [37]. Diffusion Transformers (DiTs) [2, 18, 25] replace the U-Net [6, 17, 29] with a Transformer backbone, enabling scalable, high-quality image synthesis. Recent models such as FLUX [18, 19], SANA [40], and Qwen-Image [39] highlight the flexibility of this paradigm. However, how these models internally represent and propagate semantic information from prompts is still poorly understood. A growing body of work shows that diffusion models contain rich internal semantic structure [13, 34, 43]. Prior approaches leverage attention maps for localization and editing [32, 36], while more recent studies suggest that DiTs exhibit emergent interpretability without supervision [7, 14]. In parallel, recent analyses have identified *massive activations* (MAs), defined as a small subset of hidden channels with exceptionally large magnitude [9, 10]. However, their functional role remains unclear. Some works treat them as outliers that degrade feature quality, particularly for dense visual correspondence tasks [9], while others argue that they primarily contribute to local detail synthesis and have limited influence on global semantics [33]. This raises a fundamental question: *do MAs merely reflect outliers, or do they control meaningful semantic information?*

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

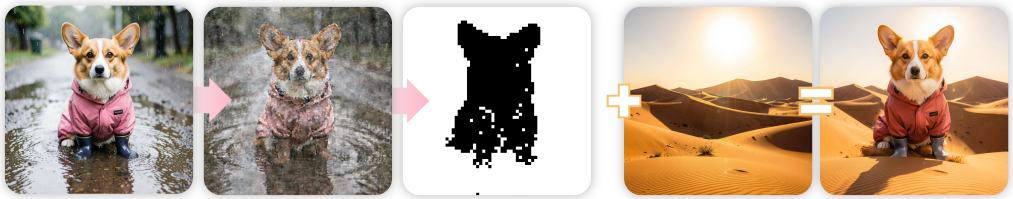

Figure 1: Overview of our findings. (Left) Disrupting the top-k massive activation channels severely degrades generation quality, revealing their functional importance. (Center) These channels exhibit structured spatial organization, which we capture as a mask. (Right) Transplanting the top-k activations within this mask from a source to a target generation enables localized semantic transfer, producing coherent compositions rather than pixel-level blending.

In this work, we examine the role of MAs in generation, identifying their concrete effects on semantic structure and controllability in the output. Rather than studying them only through representation quality or downstream tasks, we investigate whether they constitute a sparse subspace that mediates both semantic organization and controllability in diffusion models.

Our first finding is structural. While prior works show that class-conditioning vectors are sparse and exhibit properties of MAs [27], we demonstrate that this phenomenon is not limited to conditioning signals. Through a channel disruption analysis, we show the role of persistent MAs across layers, denoising steps, and both streams in MMDiT architectures.

Our second finding is spatial. Restricting representations to the top-k massive channels reveals structured activation patterns that consistently align with semantically meaningful regions, such as foreground objects and background. Despite their sparsity in channel space, these activations induce a compact representation in which semantic regions become more separable.

Our third finding is intervention-based. By transplanting top-k MAs from one prompt-conditioned trajectory into another, we achieve controlled and localized semantic shifts. The resulting images move toward the source prompt while preserving substantial content from the target, yielding coherent semantic compositions rather than pixel-level mixtures.

Building on this property, we demonstrate two practical use cases: ① *text-conditioned semantic transport*, where MAs transplanted across prompt-conditioned trajectories yield localized prompt combination, and ② *image-conditioned semantic transport*, where subject-specific activations extracted from a reference image are re-injected into new generations. Together, these results recast MAs as a *sparse semantic transport subspace*: privileged coordinates that stabilize prompt-conditioned information across denoising steps, localize it in space, and transmit it to the final image, reconciling prior views of these channels as mere outliers.

Our contributions are as follows: (i) We extend prior work with a channel-disruption probe tailored to MMDiT architectures, analyzing MAs across five state-of-the-art DiTs [18, 19, 39, 40]. (ii) We show that massive image-stream channels induce semantically meaningful spatial structure, validated segmenting GenAI-Bench [21] generated images. (iii) We introduce a masked activation-transport mechanism that, without training, enables controllable semantic composition on GenAI-Bench and matches dedicated editing backbones on DreamBench++ [26] personalization benchmark.

2 Related Work

Modern Flow-Matching DiTs for Text-to-Image. Recent advances in text-to-image generation have shifted from U-Net-based diffusion backbones [17, 29] to Transformer-based architectures, namely Diffusion Transformers (DiTs) [25]. In parallel, training objectives have evolved from denoising score matching to *flow matching* [1, 22, 24], which formulates generation as learning a velocity field that transports samples from a simple noise prior to the data distribution, enabling straighter and more efficient generation trajectories. Building on these, recent models explore different architectural instantiations of DiTs. Stable Diffusion 3 (SD3) [8] introduces the Multimodal Diffusion Transformer (MMDiT), where text and image tokens are processed with separate parameters and interact through joint attention, enabling bidirectional information flow between modalities. Subsequent models such as FLUX.1 and FLUX.2 [18, 19] scale this design and augment MMDiT blocks with single-stream

parallel-attention layers operating on concatenated multimodal sequences [5]. Alternative designs, such as Qwen-Image [39], incorporate a frozen multimodal LLM as the text encoder within an MMDiT-like framework, while SANA [40] departs from MMDiT altogether, adopting a single-stream linear-attention DiT paired with a deep-compression autoencoder for high-resolution efficiency. In parallel, recent work has focused on reducing sampling cost through distillation. Few-step variants such as FLUX.1-schnell [18], FLUX.2-klein [19], and SANA-Sprint [3] compress sampling to a handful of steps via trajectory or adversarial distillation. In this work, we study the phenomenon of *massive activations* across this spectrum of modern flow-matching DiTs, spanning MMDiT-based, parallel-attention, and linear-attention architectures, as well as their distilled counterparts.

Massive Activations in Transformers and DiTs. Massive activations (MAs) refer to a small subset of hidden-state entries whose values exceed the typical activation scale by several orders of magnitude, often concentrating in fixed channels and acting as implicit bias terms [33]. Originally identified in LLMs and Vision Transformers [4, 33], MAs have recently been observed in DiTs as well, where they concentrate in a handful of fixed channels across image tokens tied to AdaLN modulation [25]. Prior work suggests these channels encode limited spatial detail and can degrade dense correspondence unless explicitly controlled [9]. At the same time, MAs are not merely incidental: perturbations to these activations significantly impair *fine-grained detail synthesis* while leaving global semantics largely unchanged, motivating techniques such as Detail Guidance [10]. Complementary analyses of AdaLN conditioning embeddings further reveal a strong form of sparsity, where semantic information is concentrated in a small fraction of dimensions, effectively forming a low-dimensional bottleneck [27]. In contrast to prior work, which primarily studies MAs within individual forward passes or in terms of representation quality, we analyze their role in the generative process of modern MMDiTs. Specifically, we leverage the dual-stream structure of MMDiT architectures to examine image and text streams separately, and we focus on how massive activations behave *across multiple generation trajectories*, enabling a different perspective on their functional role.

3 Massive Activations as a Sparse Semantic Subspace

We study the role of massive activations (MAs) in DiTs and analyze their structure and effect on generation. Our analysis proceeds in three steps: (i) we identify their functional importance via channel disruption, (ii) we reveal their spatial organization through clustering, and (iii) we demonstrate their cross-generation role via activation transport. Together, these results provide a unified view of MAs as the main carriers of semantic information.

Preliminaries. MMDiTs process two parallel token streams at each layer: an *image stream* $\mathbf{X}_I^\ell \in \mathbb{R}^{N_I \times D}$, consisting of the N_I latent image tokens being denoised, and an *encoder stream* $\mathbf{X}_E^\ell \in \mathbb{R}^{N_E \times D}$, consisting of the N_E text tokens that carry the prompt conditioning. The two streams share the channel dimension D and interact through joint attention within each block. For the image stream, the tokens inherit the spatial layout of the latent image representation. Specifically, if the latent tensor has spatial dimensions $H_{\text{lat}} \times W_{\text{lat}}$, then, the number of image tokens is given by $N_I = H_{\text{lat}} W_{\text{lat}}$.

We denote by $\mathbf{X} \in \mathbb{R}^{N \times D}$ the hidden activations at a given layer and timestep, where N is the number of total tokens $N_I + N_E$ and D the number of channels. Let $\alpha \in \{I, E\}$ denote the stream, and let N_α be the corresponding number of tokens. Given a single sample, we write $\mathbf{X}_\alpha^\ell \in \mathbb{R}^{N_\alpha \times D}$ for the activations of stream α at layer $\ell \in \{1, \dots, L\}$, and $\mathbf{X}_\alpha^\ell[n, d]$ entry at token n and channel d.

Let us define $\{|\mu_d^\ell|\}_{d=1}^D$ the channel-wise mean distribution at layer ℓ. This distribution is highly concentrated, and there exists a small subset of channels that accounts for most of the total activation magnitude. To define this set, we compute a channel-wise importance score, instantiated as the absolute channel mean $|\mu_d^\ell|$ at layer ℓ. We then select the top-k and bottom-k according to this score for each layer independently as:

$$\mathcal{T}_k^\ell = \text{top-}k\left(\{|\mu_d^\ell|\}_{d=1}^D\right), \qquad \mathcal{B}_k^\ell = \text{bottom-}k\left(\{|\mu_d^\ell|\}_{d=1}^D\right). \tag{1}$$

We refer to the activations in $\mathcal{T}_k^\ell$ as *massive activations*, as they account for most of the total activation magnitude, while the complementary set $\mathcal{B}_k^\ell$ contributes negligibly. Further, we define $\mathcal{C}$ as a set of channels corresponding to either $\mathcal{T}_k^\ell$ or $\mathcal{B}_k^\ell$. All channel statistics are computed within a single sample, across the token dimension of the given stream layer ℓ in stream α. These are per-sample, per-layer, per-stream, per-channel quantities; no cross-sample averaging is performed.



Figure 2: Effect of channel disruption across models, metrics, and streams. Each plot reports performance as a function of the number of disrupted channels (k), expressed as a percentage relative to the undisrupted baseline (100). Solid lines correspond to disrupting the top-k channels ($\mathcal{T}_k$), while dashed lines correspond to disrupting the bottom-k channels ($\mathcal{B}_k$). Rows separate interventions on the encoder stream (top) and image stream (bottom), while columns report different evaluation metrics.

3.1 Channel Ablation as an Interventional Probe

Building on the concentration of channel statistics in Eq. 1, we investigate their role during generation. To this end, we extend prior work [27] with an interventional probe for modern MMDiT architectures that independently evaluates the role of their two information streams. We call this pipeline *channel disruption*: selectively zeroing a subset of channels to measure their impact on generation.

Formally, given activations $\mathbf{X} \in \mathbb{R}^{N \times D}$, we define the disrupted activations as $\widetilde{\mathbf{X}} = \mathbf{X} \odot \mathbf{m}$, where $\mathbf{m} \in \{0, 1\}^D$ is a binary mask that zeros the channels in $\mathcal{C}$. The intervention is applied independently at each layer, timestep, and stream.

Emergent Property. We evaluate channel disruption across five models, namely FLUX.1-schnell, FLUX.1-dev, FLUX.2-klein, Qwen-Image, and SANA1.5[1]. We apply disruption at all layers and timesteps, and report (Fig. 2) results on ImageNet-1k [30] using standard alignment (CLIP-T [15], ImageReward [41]), quality (Aesthetic Score [31]), and distributional (FID [16]) metrics. Each metric is reported as a percentage of its value on the corresponding un-disrupted generation. We refer the reader to the Appendix for further details and GenAI-Bench [21] results.

Two main findings emerge. First, we extend the observation of [27], originally reported on class-conditioned DiTs, to state-of-the-art text-conditioned MMDiT models: disrupting top-ranked channels causes a sharp degradation across models and metrics, whereas removing an equally-sized set of low-ranked channels has only a limited effect. This further supports the view that generative computation is concentrated in a small subset of high-statistic channels. More importantly, the effect of disruption differs across the two streams. In both streams, suppressing top-ranked channels leads to a comparable drop in CLIP-T and ImageReward, indicating that both streams carry information necessary to produce images aligned with the prompt. However, disrupting the image stream causes a larger deterioration in FID and Aesthetic score signaling heavier unnatural artifacts. By contrast, disrupting the encoder stream yields a milder degradation, suggesting that it primarily affects prompt alignment while leaving the ability of the models to generate plausible images comparatively intact.

Additional results with a magnitude-preserving token-shuffling control are provided in Appendix B.1.

3.2 Spatial Structure of Image-Stream Channels

While MAs are defined along the channel dimension, we now investigate how they induce structured patterns across the token axis. To expose this structure, we restrict the activations to the top-k selected

[1]Note that SANA1.5 is based on a linear DiT, and therefore uses a single generation stream.

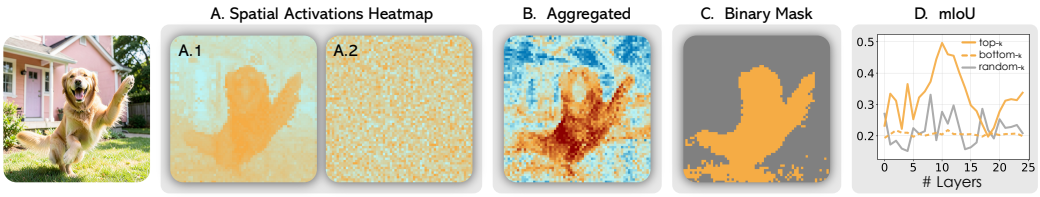

Figure 3: Spatial structure of MAs on FLUX.2-klein model. (A) Channel-wise activation maps for high- (A.1) and low-importance (A.2) channels: massive channels exhibit coherent, subject-aligned patterns, while low-importance channels are diffuse. (B) Aggregating activations over top-k channels yields a structured heatmap aligned with salient regions. (C) Clustering these activations produces a binary mask that closely captures the main subject. (D) Quantitative evaluation (mIoU) shows that masks derived from MAs (bold orange) are more aligned with reference segmentation across layers, differently from low-importance activations (dash orange) and random activations (grey).

channels and analyze their spatial organization, to see if salient concepts are highlighted by salient activations. Specifically, we focus on the image stream, whose tokens inherit the spatial layout of the latent image grid. Let $\mathbf{X}_{I,\mathcal{C}} \in \mathbb{R}^{N_I \times k}$ denote the activations restricted to a selected set of channels $\mathcal{C}$, where $\mathcal{C}$ corresponds to either $\mathcal{T}_k^\ell$ or $\mathcal{B}_k^\ell$ as in Eq. 1.

To extract spatial structure from these activations, we partition tokens based on their feature representations in the restricted subspace. Concretely, we apply K-means clustering with $K = 2$ to the channels of $\mathbf{X}_{I,\mathcal{C}}$, assigning each token to one of two clusters according to its feature vector in $\mathbb{R}^k$.

Each cluster is classified as salient or non-salient based on a score reflecting the strength of the selected activations across its tokens. Specifically, we apply min-max normalization on $\mathbf{X}_{I,\mathcal{C}}$ along the channel dimension so that the k spatial maps have comparable scale, denoting the result $\hat{\mathbf{X}}_{I,\mathcal{C}}$. Let $\{C_0, C_1\}$ denote the two clusters obtained by K-means, where each $C_j \subseteq \{1, \ldots, N_I\}$ is the set of tokens assigned to cluster j. For each token n we aggregate the normalized activations across the selected channels and compute the per-cluster average as:

$$\bar{s}_j = \frac{1}{|C_j|} \sum_{n \in C_j} s_n \quad \text{for} \quad j \in \{0, 1\}, \quad \text{where} \quad s_n = \sum_{c \in \mathcal{C}} \hat{X}_I[n, c]. \tag{2}$$

The cluster with the highest average score is assigned to the foreground class, the other to the background. This induces a binary mask $\mathbf{p} \in \{0, 1\}^{N_I}$ with $p_n = 1$ if token n belongs to the foreground cluster and $p_n = 0$ otherwise. Fig. 3 (B) shows a representative example of the normalized activation heatmap computed via Eq. 2, alongside the corresponding binary mask in Fig. 3 (C).

Emergent Property. To evaluate whether the clusters induced by massive activations correspond to semantically meaningful regions, we frame the task as *dichotomous segmentation*. We compute mIoU between the binary mask from the last denoising timestep and a pseudo-ground-truth from BiRefNet [44] on the final generated image, averaged over the 1,600 prompts of GenAI-Bench [21].

Fig. 3 (D) reports this metric on FLUX.2-klein for three channel-selection strategies with $k = 12$, kept fixed for all subsequent experiments. Top-k consistently dominates bottom-k and random-k at every layer, peaking at mIoU ≈ 0.5 at layer 10, while bottom-k remains flat around 0.2 and random-k falls in between, indicating that semantic spatial structure is concentrated in a small, identifiable subset of channels rather than being a generic property of the activations. Together with the qualitative visualization in Fig. 3 (A–C), these results reveal an emergent property: activations in $\mathcal{T}_k^\ell$ concentrate semantic information in space, producing structured maps that align with foreground objects and background regions, whereas $\mathcal{B}_k^\ell$ yields diffuse, noise-like patterns. This shows that spatial semantic structure is not uniformly distributed across channels, but is concentrated in a small, identifiable set of massive activations. We refer the reader to the Appendix for further quantitative evaluations.

3.3 Channel-Selective Activation Transport

Channel disruption shows that a small subset of high-absolute-mean channels has a disproportionate effect on the final generation. We now move from *removing* these channels to *transferring* them across prompts. The goal is to test whether MAs are not only indispensable to the generation process, but also carry prompt-specific semantic information that can be reused in another generation.

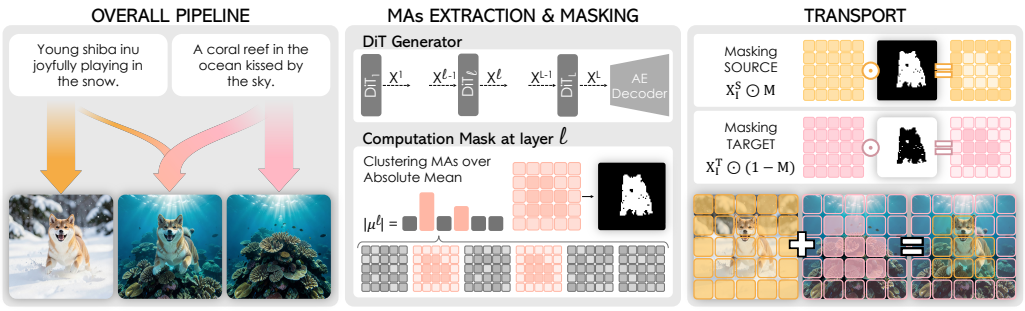

Figure 4: Activation transport via MAs. (Left) Overall pipeline: two prompts are used to generate a source and a target image from the same DiT. (Center) Top-k channels are identified in the source trajectory and used to derive a spatial mask. (Right) Source activations are injected into the target generation within the masked region, while the remaining channels are preserved.

We consider two generations obtained from the same initial noise but different prompts, and denote their hidden activations at a given layer by $\mathbf{X}^S$ (source) and $\mathbf{X}^T$ (target), with $\mathbf{X}^S, \mathbf{X}^T \in \mathbb{R}^{N \times D}$.

Channel-Selective Replacement. Instead of zeroing out selected channels, we replace them in the target generation with those from the source. Specifically, using the channel mask $\mathbf{m} \in \{0,1\}^D$ defined in Sec. 3.1, the merged activations are defined as:

$$\widetilde{\mathbf{X}}^T = \mathbf{X}^T \odot (1 - \mathbf{m}) + \mathbf{X}^S \odot \mathbf{m}, \tag{3}$$

where $\mathbf{m}$ is broadcast across tokens. This channel-wise replacement applies to both encoder and image streams, injecting sparse source activations into the target while preserving other channels.

Spatially Selective Replacement in the Image Stream. The emergent spatial interpretation of activations (Sec. 3.2) enables localized transport using the structure induced by massive activations. We reuse the spatial mask $\mathbf{p} \in \{0,1\}^{N_I}$ from Sec. 3.2, where N_I is the number of image tokens.

We combine token mask $\mathbf{p}$ with the channel mask $\mathbf{m}$ to obtain a joint channel-spatial mask:

$$\mathbf{M} = \mathbf{p}\mathbf{m}^\top \in \{0,1\}^{N_I \times D}. \tag{4}$$

The image-stream activations are then updated as:

$$\widetilde{\mathbf{X}}_I^T = \mathbf{X}_I^T \odot (1 - \mathbf{M}) + \mathbf{X}_I^S \odot \mathbf{M}, \tag{5}$$

while the encoder stream is updated using channel-only replacement (Eq. 3). In the image stream, the joint mask $\mathbf{M}$ replaces the channel mask $\mathbf{m}$, ensuring that replacement occurs only on selected channels within selected spatial regions. We apply this at a selected set of layers at all denoising steps during generation. In practice, intervening at a subset of intermediate layers is already sufficient to induce consistent semantic effects. Fig. 4 depicts an overview of the transport mechanism.

Emergent Property. We analyze the effect of activation transport in embedding space. Fig. 5 reports the semantic shift induced by transporting top-k MAs across models, layers, and configurations. Each point corresponds to a transport setup. The horizontal axis, ΔCLIP-I = CLIP-I (S) $-$ CLIP-I (T), measures the relative similarity of the merged image to the source versus the target, with values near zero indicating merges that integrate both prompts rather than collapsing to one. The vertical axis captures the joint preservation of source and target semantics through CLIP-T. We conduct this analysis on 703 GenAI-Bench [21] prompts with a `scene` attribute to guarantee that a salient subject and a meaningful background are present, ensuring coherent merges. We randomly pair each prompt with five others to form 3,515 source-target pairs and report results across the generators used in the previous analyses. Across all models, three consistent patterns emerge. First, top-k controls the horizontal position: larger top-k transfers more activations from the source and shifts points rightward along ΔCLIP-I, while small top-k keeps generations closer to the target. Second, middle-layer configurations (orange) more often lie in the upper region than lower (green) or upper (purple) layers, suggesting that transport at mid-depths preserves better joint alignment; we therefore focus on mid-layer configurations. Third, the encoder stream contributes little; indeed, configurations with encoder top-$k = 0$ tend to match or outperform those that also transport encoder features.

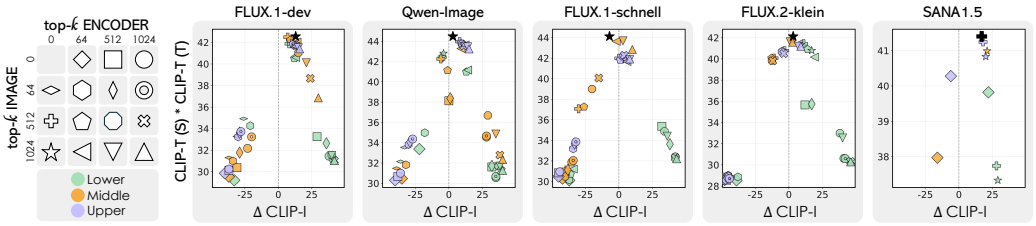

Figure 5: Semantic effect of activations transport across models and layer regimes. Each point represents a configuration, where ΔCLIP-I measures the shift toward the source prompt, and CLIP-T(S)*CLIP-T(T) captures joint semantic preservation. Colors indicate the layer regime where the intervention is applied (lower, middle, upper), while marker shapes encode the number of transported channels (k) and the stream (encoder or image). Interventions in the middle layers achieve better trade-offs, suggesting that semantic transport is most effective in middle stages of the network.

The configuration achieving ΔCLIP-I ≈ 0 and the highest CLIP-T(S)*CLIP-T(T) is applied at middle layers with encoder top-$k = 0$ and image top-$k = 1024$ on all models but SANA1.5, where the optimum shifts to image top-$k = 512$. We adopt these settings in subsequent experiments. These consistent trends suggest activation transport induces structured semantic directions, rather than random perturbations, with MAs operating in a controllable subspace.

4 Use Cases

The observations suggest that MAs are a lightweight, architecture-agnostic interface for semantic control in DiTs. We illustrate two practical use cases: *prompt-to-prompt semantic transport*, where activations are transplanted between different, text-conditioned generations and *image-conditioned semantic transport*, where salient subject semantics are extracted and re-injected into new generations.

4.1 Text-Conditioned Semantic Transport

We consider a setting where two images are generated from the same initial noise but different prompts, denoted x^S and x^T. Using the massive activations masks and channel selection described in the previous sections, we transplant a set of MAs from the trajectory of x^S into that of x^T, to transfer semantic attributes while preserving the structure of the target image.

Experimental Setup. Following the setup of Sec. 3.3, we evaluate on the 3,515 prompt pairs from GenAI-Bench across the five DiT-based models. Table 1 reports CLIP-T, CLIP-I and DINO-I, computed independently against x^S and x^T and combined as a product (S*T) to capture joint fidelity. We focus on the combined product, which rewards preserving both source and target.

Baselines. We compare our best configuration against four alternatives. First, we consider linear interpolation between the two prompts, applied either to DiT activations at every timestep – across all layers or restricted to those used in our identified setting – or to autoencoder latents at each timestep. Second, to assess the role of spatial selection, we include an ablation of our MAs substitution method without the spatial mask (Sec. 3.3). Moreover, as an upper bound that bypasses multi-prompt combination by operating in the natural regime of the generator, we report results obtained by merging the two prompts into a single description using Gemma3-4B [35] and generating one image from it. Finally, we compare with FreeFlux [38] on FLUX.1-dev in the background replacement setting.

Experimental Results. Across all backbones, MAs-based transport with spatial masking consistently outperforms interpolation baselines in terms of combined image fidelity. For instance, on FLUX.1-schnell MAs transport reaches CLIP-I 55.2 and DINO-I 20.1 (vs. 53.2 / 2.7 for vanilla interpolation). Similar margins hold on FLUX.2-klein (CLIP-I 55.2, DINO-I 19.1) and FLUX.1-dev (CLIP-I 58.1, DINO-I 19.2), where our method also surpasses FreeFlux across all joint metrics (Table 1). Notably, interpolation methods often collapse toward one of the two sources (despite an interpolation $\alpha = 0.5$), as reflected by near-zero DINO-I (S*T) scores, whereas our approach preserves meaningful information from both. The single-prompt upper bound achieves the best text alignment, as expected, since the LLM can fully orchestrate both prompts semantically, leaving the generator with the simpler task of single-prompt generation. However, it consistently underperforms on image-based metrics, as reducing the pair to one prompt sacrifices the identities of x^S and x^T.

Table 1: Prompt-to-prompt composition results. We compare strategies across models, reporting CLIP-T, CLIP-I, and DINO-I scores for source (S), target (T), and their joint preservation (S*T).

				CLIP-T			CLIP-I			DINO-I		
		Mask	Layers / Steps	S*T	S	T	S*T	S	T	S*T	S	T
FLUX.1-schnell	Single Prompt	-	-	46.5	70.3	66.3	52.7	74.5	70.8	15.5	45.4	34.2
	Interpolation	✓	all	31.7	83.2	38.1	53.2	99.5	53.5	2.7	99.0	2.7
	Latent Interpolation	-	all	42.2	63.7	66.2	52.2	71.3	73.2	18.5	39.8	46.5
	Interpolation	✓	mid	37.4	79.0	47.3	53.3	90.7	58.8	9.3	80.1	11.6
	MAs	-	mid	32.4	83.0	39.0	53.0	98.0	53.4	3.0	96.3	3.1
	MAs (Ours)	✓	mid	**44.1**	62.8	70.3	**55.2**	71.2	77.6	**20.1**	42.4	47.5
FLUX.1-dev	Single Prompt	-	-	44.2	68.7	64.4	54.9	76.0	72.2	14.4	43.9	32.9
	FreeFlux	-	-	39.5	54.3	72.8	56.9	68.2	83.5	17.6	30.2	58.3
	Interpolation	✓	all	31.2	74.0	42.2	43.4	79.7	54.5	2.4	62.4	3.8
	Latent Interpolation	-	all	30.6	80.2	38.1	57.1	100	57.1	4.4	100	4.4
	Interpolation	✓	mid	32.6	78.4	41.6	49.6	88.4	56.1	3.2	76.5	4.2
	MAs	-	mid	30.9	80.4	38.5	57.0	99.9	57.0	4.3	99.9	4.3
	MAs (Ours)	✓	mid	**42.5**	69.2	61.5	**58.1**	82.6	70.4	**19.2**	61.4	31.3
FLUX.2-klein	Single Prompt	-	-	43.7	67.3	65.0	53.3	74.0	72.1	16.9	46.1	36.7
	Interpolation	✓	all	30.2	81.0	37.3	52.7	99.1	53.2	3.2	98.7	3.2
	Latent Interpolation	-	all	31.1	81.0	38.4	52.4	98.0	53.5	3.6	95.4	3.7
	Interpolation	✓	mid	41.0	69.8	58.8	**55.2**	81.9	67.4	16.4	68.4	24.0
	MAs	-	mid	30.8	81.0	38.0	52.7	98.7	53.4	3.3	98.6	3.4
	MAs (Ours)	✓	mid	**41.7**	64.6	64.5	**55.2**	75.2	73.4	**19.1**	58.0	32.9
Qwen-Image	Single Prompt	-	-	44.0	68.2	64.5	51.8	73.7	70.3	12.5	40.5	30.9
	Interpolation	✓	all	23.2	49.5	46.9	27.9	53.7	52.1	0.0	2.7	1.3
	Latent Interpolation	-	all	31.2	82.8	37.7	**55.8**	100	55.8	3.7	99.9	3.7
	Interpolation	✓	mid	22.8	48.3	46.8	26.2	51.9	50.5	0.0	2.0	1.3
	MAs	-	mid	32.4	82.8	39.2	55.2	98.2	56.2	4.0	96.0	4.2
	MAs (Ours)	✓	mid	**44.5**	67.0	66.4	55.4	76.0	72.9	**18.0**	48.6	37.1
SANA1.5	Single Prompt	-	-	46.8	70.0	66.8	55.6	76.0	73.2	18.0	46.6	38.6
	Interpolation	✓	all	30.2	63.8	47.3	34.1	63.7	53.6	1.9	32.6	5.8
	Latent Interpolation	-	all	32.5	83.4	39.0	56.5	100.0	56.5	5.4	99.9	5.4
	Interpolation	✓	mid	32.7	59.2	55.3	35.5	61.4	57.9	4.00	23.6	17.0
	MAs	-	mid	33.0	83.0	39.8	56.1	98.3	57.1	5.30	96.9	5.50
	MAs (Ours)	✓	mid	**41.4**	72.8	56.8	**55.7**	83.7	66.6	**17.0**	65.0	26.1

In contrast, our identified setting remains within $\sim$2 CLIP-T points of this bound, while substantially improving image fidelity (*e.g.*, $+2.5$ CLIP-I and $+4.6$ DINO-I on FLUX.1-schnell), with similar gains across other models. Consistent with our earlier observations, spatial masking is crucial for properly localizing the intervention. Removing it causes MA transplantation to propagate beyond the target region, resulting in diminished alignment and image fidelity. On FLUX.1-schnell, adding the mask increases DINO-I from 3.0 to 20.1 and CLIP-T from 32.4 to 44.1. This highlights the importance of spatially restricting the intervention to preserve coherent structure. Overall, these results are consistent with the view that MAs define a prompt-conditioned semantic subspace. Transplanting them steers the generative trajectory along meaningful semantic directions, enabling targeted attribute transfer rather than uncontrolled mixing. Fig. 6 (left) shows a qualitative example of prompt-to-prompt semantic transport on FLUX.1-dev. More baselines and ablations are shown in the Appendix.

4.2 Image-Conditioned Semantic Transport

We extend our analysis to image-conditioned generation, examining whether MAs capture subject-specific semantics not only in text-driven trajectories, but also when conditioning on real images. Here the source trajectory x^S is derived from a reference subject image rather than a text prompt, while the transport of Sec. 3.3 is unchanged. To obtain x^S, we rely on the reconstruction technique proposed by TokenVerse [11], which learns per-example MLPs to predict latent prompt adjustments that steer the generation toward the reference image. MAs are extracted along this reconstruction trajectory and transported into the target-prompt generation. For FLUX.2-klein, which natively supports image conditioning, we additionally analyze MAs obtained by pairing the subject image with an empty prompt. This allows us to compare representations derived from reconstruction-based inversion with those obtained through direct image conditioning.

Table 2: Image-conditioned transport results. We compare strategies across models, reporting CLIP-T, CLIP-I, and CLIP-I-personalized scores and their joint composition

Generator	Features	CLIP-T*CLIP-I	CLIP-T*CLIP-I_{pers}	CLIP-I_{pers}	CLIP-I	CLIP-T
Qwen-Image-Edit	-	65.5	-	-	74.8	87.5
FLUX-Kontext	-	66.0	-	-	79.3	83.1
FLUX.2-klein	-	66.8	-	-	75.8	88.2
Qwen-Image	TokenVerse	58.2	58.2	68.1	68.0	85.5
	MAs (Ours)	**62.6**	**65.9**	80.5	76.5	81.9
FLUX.1-dev	TokenVerse	55.3	55.3	67.0	67.0	82.5
	MAs (Ours)	**58.1**	**64.7**	85.9	77.2	75.3
FLUX.2-klein	TokenVerse	63.4	62.2	71.6	73.0	86.9
	MAs (TV)	65.7	65.3	77.1	77.5	84.7
	MAs (EDIT)	**66.8**	**66.8**	79.6	79.6	83.9
FLUX.1-schnell	TokenVerse	57.5	58.1	67.6	67.0	85.9
	MAs (Ours)	**60.8**	**62.2**	75.5	73.8	82.4
SANA1.5	TokenVerse	56.7	55.6	64.3	65.4	86.6
	MAs (Ours)	**56.8**	**63.3**	80.0	71.9	79.1
FLUX-Kontext	**MAs (EDIT)**	64.3	64.3	81.1	81.1	79.2



Figure 6: Qualitative examples prompt-to-prompt (left) and image-conditioned (right) semantic transport. We refer the user to the Appendix for more qualitative results on all models for both tasks.

Experimental Setup. We evaluate MA-based transport in the image-conditioned setting on DreamBench++ [26], a standard personalization benchmark consisting of 150 real subject images paired with prompts in diverse contexts. To enable reconstruction, each subject image is first captioned using Qwen-2.5-VL-3B [42], providing the initial prompt representation on which TokenVerse learns its latent adjustments. Personalization methods typically involve a trade-off between subject fidelity and prompt alignment. To capture this, we report CLIP-I and CLIP-T as measures of image fidelity and text alignment, respectively. Since both TokenVerse and our MA-based approach rely on a reconstructed trajectory, we also report CLIP-I_{pers} against the reconstructed subject, separating injection from reconstruction quality, and the joint metrics CLIP-T*CLIP-I and CLIP-T*CLIP-I_{pers}.

Baselines. We compare our approach against the native transfer mechanism of TokenVerse, which propagates subject-token activations through the encoder stream across layers to guide the target generation. Moreover, we include dedicated image-editing models designed for personalization as reference upper bounds, namely FLUX.1-kontext [20], Qwen-Image-Edit [39] and Flux.2-klein [19].

Experimental Results. As shown in Table 2, MAs-based feature injection consistently outperforms TokenVerse on the joint personalization metric CLIP-T*CLIP-I_{pers}. This improvement is primarily driven by significantly stronger subject preservation: CLIP-I_{pers} improves by $+12.4$ on Qwen-Image, $+18.9$ on FLUX.1-dev, $+7.9$ on FLUX.1-schnell, and $+15.7$ on SANA1.5. Importantly, CLIP-T remains in a comparable high range (75-82), indicating that improved subject fidelity does not

Table 3: Entity-selective transport with external masks on two-object scenes. Each metric reports Chosen and Rejected source object. CLIP-I uses masked source regions; presence rates (%) are assessed by Qwen3.6-27B. Across both settings, the chosen entity is favored on all backbones.

	CLIP-T		**Masked CLIP-I**		**Presence**		**CLIP-T**		**Masked CLIP-I**		**Presence**	
Generator	Chosen	Rejected	Chosen	Rejected	Chosen	Rejected	Chosen	Rejected	Chosen	Rejected	Chosen	Rejected
	Text-Conditioned						*Image-Conditioned*					
FLUX.1-schnell	**60.0**	47.3	**86.9**	62.1	**76.0**	8.0	**52.3**	50.0	**78.3**	57.8	**69.4**	15.7
FLUX.1-dev	**62.5**	47.5	**92.1**	64.5	**88.0**	6.0	**57.3**	49.5	**93.4**	65.0	**89.8**	16.3
FLUX.2-klein	**60.8**	45.0	**89.9**	56.7	**91.0**	1.0	**60.5**	46.5	**85.8**	56.9	**84.7**	3.1
Qwen-Image	**66.5**	48.3	**92.8**	64.6	**96.0**	2.0	**59.5**	50.0	**89.1**	62.5	**85.7**	8.8
SANA1.5	**64.3**	48.3	**90.5**	64.8	**89.0**	5.0	**58.0**	49.0	**89.7**	64.6	**89.6**	7.3

come at the expense of prompt alignment. On FLUX.2-klein, the two extraction modes (*edit* and *TV*) yield comparable scores, and applying MAs transport to native image-conditioned trajectories on FLUX-Kontext similarly reaches 64.3 CLIP-T*CLIP-I (vs. 66.0 for native editing), confirming compatibility with both reconstruction-based inversion and direct image conditioning. Overall, this use case reaches scores comparable to dedicated editing backbones, matching FLUX.2-klein on CLIP-T*CLIP-I (66.8) and approaching Qwen-Image-Edit (62.6 vs. 65.5), while operating on top of unmodified text-to-image generators and confirming that the properties of MAs established in earlier sections transfer to a downstream personalization setting. Fig. 6 (right) shows a qualitative example of image-conditioned semantic transport on FLUX.2-klein.

4.3 Entity-Selective Semantic Transport with External Masks

We evaluate whether the transport mechanism can select one entity from a multi-object source in both the text-conditioned and image-conditioned settings introduced above. The operator is not restricted to the foreground/background masks obtained by K-means: an external entity mask can be used as $\mathbf{p}$ without changing the transported channels or replacement operator.

Experimental Setup. We use prompts from the two-object category of GenEval [12] and pair them with GenAI-Bench background prompts. GroundingDINO [23] localizes each prompt-specified object, and SAM2 [28] segments the corresponding source region. We randomly select one object to transport and discard the other; the selected object's mask is downsampled to the DiT token grid and used to restrict activation replacement. From these prompts, each evaluated backbone generates its own source image in the text-conditioned setting (Sec. 4.1), whereas in the image-conditioned setting (Sec. 4.2), reference images are synthesized using an independent model FLUX.2-dev.

Evaluation Metrics. For each output, we evaluate the chosen and rejected source objects separately. CLIP-T measures alignment between the generated image and the description of each object. Masked CLIP-I measures similarity to the corresponding source-object region: the chosen-object score assesses preservation, while the rejected-object score measures leakage of the discarded entity. Qwen3.6-27B additionally acts as a binary visual judge, assessing whether each object is present in the generated image. The chosen and rejected presence rates are the percentages of outputs in which the corresponding object is judged present.

Experimental Results. Table 3 reports higher CLIP-T, masked CLIP-I, and presence rates for the chosen object than for the rejected one on every backbone and setting. Chosen-object presence ranges from 69.4–89.8% with image conditioning and 76.0–96.0% with text conditioning. The rejected object remains present in at most 16.3% and 8.0% of outputs, respectively. Overall, external masks enable precise entity-selective transport, ensuring high target exclusivity and low non-target leakage.

5 Conclusion

We have shown that a small subset of channels, called *massive activations*, plays a central role in diffusion transformer generation, exhibiting structured spatial organization and supporting controllable semantic transfer across prompts. These results suggest that massive activations define a sparse subspace through which semantic information is represented and manipulated during generation, reconciling prior views of them as outliers or detail drivers. Broadly, these findings open questions about the emergence and exploitation of such low-dimensional control structures in generative models.

Acknowledgments

We acknowledge CINECA for the availability of high-performance computing resources under the ISCRA initiative, and for funding Evelyn Turri's PhD. This work has been supported by the EU Horizon projects "ELIAS" (GA No. 101120237) and "ELLIOT" (GA No. 101214398), and by the EuroHPC JU project "MINERVA" (GA No. 101182737). The authors declare no competing interests.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

[2] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are Worth Words: A ViT Backbone for Diffusion Models. In *CVPR*, 2023.

[3] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. SANA-Sprint: One-Step Diffusion with Continuous-Time Consistency Distillation. In *ICCV*, 2025.

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. In *ICLR*, 2024.

[5] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling Vision Transformers to 22 Billion Parameters. In *ICML*, 2023.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021.

[7] Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit Haim Bermano. Attention (as Discrete-Time Markov) Chains. In *NeurIPS*, 2025.

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[9] 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 *NeurIPS*, 2025.

[10] 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 *ICLR*, 2026.

[11] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM TOG*, 2025.

[12] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems*, 2023.

[13] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised Semantic Correspondence Using Stable Diffusion. In *NeurIPS*, 2023.

[14] Alec Helbling, Tuna Han Salih Meral, Benjamin Hoover, Pinar Yanardag, and Duen Horng Chau. ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features. In *ICML*, 2025.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 2017.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020.

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

[19] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

[20] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*, 2025.

[21] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. GenAI-Bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[24] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*, 2023.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[26] Yuang Peng, Yuxin Cui, Haomiao Tang, Zekun Qi, Runpei Dong, Jing Bai, Chunrui Han, Zheng Ge, Xiangyu Zhang, and Shu-Tao Xia. DreamBench++: A Human-Aligned Benchmark for Personalized Image Generation. In *ICLR*, 2025.

[27] Trung X Pham, Kang Zhang, Ji Woo Hong, and Chang D Yoo. A Hidden Semantic Bottleneck in Conditional Embeddings of Diffusion Transformers. In *ICLR*, 2026.

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[32] Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring Multi-modal Diffusion Transformers for Enhanced Prompt-based Image Editing. In *ICCV*, 2025.

[33] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive Activations in Large Language Models. In *COLM*, 2024.

[34] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *ACL*, 2023.

[35] Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

[36] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation using Stable Diffusion. In *CVPR*, 2024.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[38] Tianyi Wei, Yifan Zhou, Dongdong Chen, and Xingang Pan. FreeFlux: Understanding and Exploiting Layer-Specific Roles in RoPE-Based MMDiT for Versatile Image Editing. In *ICCV*, 2025.

[39] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025.

[40] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, et al. SANA 1.5: Efficient Scaling of Training-Time and Inference-Time Compute in Linear Diffusion Transformer. In *ICML*, 2025.

[41] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In *NeurIPS*, 2023.

[42] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[43] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A Tale of Two Features: Stable Diffusion Complements DINO for Zero-Shot Semantic Correspondence. In *NeurIPS*, 2023.

[44] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral Reference for High-Resolution Dichotomous Image Segmentation. *CAAI Artificial Intelligence Research*, 2024.

A Additional Implementation Details

Model Setup. For the disruption experiment (Sec. 3.1), we generate 5 images per ImageNet class, following [27]. This results in a total of 5000 images for each model. In text-based experiments, resolution is set to 1024x1024 for all models except Qwen-Image, where it is set to 512x512 for computational constraints, while in image-based ones (Sec. 4.2) also FLUX.1-dev and FLUX.1-dchnell are setup at 512x512 resolution. As for the experiments of spatial structures (Sec. 3.2), metrics are obtained extracting activations at the last step of generation.

Depending on the depth of each model, we define the Lower, Middle, and Upper layer groups as blocks 0–18, 19–37, 38–56 for FLUX.1-schnell and FLUX.1-dev; 0–7, 8–15, 16–24 for FLUX.2-klein; and 0–19, 20–39, 40–59 for Qwen-Image and SANA1.5, partitioning layers in equal parts.

To provide a more in-depth analysis, we portray both few step models, and classical multi-step ones, setting FLUX.1-schnell to 1 step, FLUX.2-klein to 4 steps, SANA1.5 to 20 steps, FLUX.1-dev to 28 steps and Qwen-Image to 28 steps.

Embedding-based Metrics. For CLIP-based similarity metrics, we employ CLIP ViT-B/32, and CLIP-T metrics are scaled with a $w = 2.5$, as in [15]. For DINO-based similarity metrics, the model used is DINOv2-Base.

TokenVerse Reconstruction Setup. We optimize each concept for 1400 steps with learning rate 10^{-4}, a cosine schedule preceded by 50 linear warmup steps, zero weight decay, and gradient clipping at norm 1.0. The training objective is the diffusion loss, with timesteps drawn from a logit-normal distribution. For 50% of the steps we additionally apply the concept isolation loss of [11].

B Additional Analysis

B.1 Additional Disruption Analysis

Magnitude-Preserving Token Shuffling. Zeroing top-k channels removes much more activation magnitude than zeroing bottom-k channels, leaving open whether the stronger degradation reflects channel identity or simply the amount of magnitude removed. To disentangle these factors, we randomly shuffle token values within each selected channel, preserving every activation value, channel-wise means, variances, and total L_2-energy. Under the same evaluation setting, Fig. 7 reproduces the top-/bottom-ranked channel and encoder/image-stream asymmetries despite removing no magnitude at all. The degradation therefore depends on which channels are perturbed and on the token-wise organization of their values, rather than simple activation magnitude.

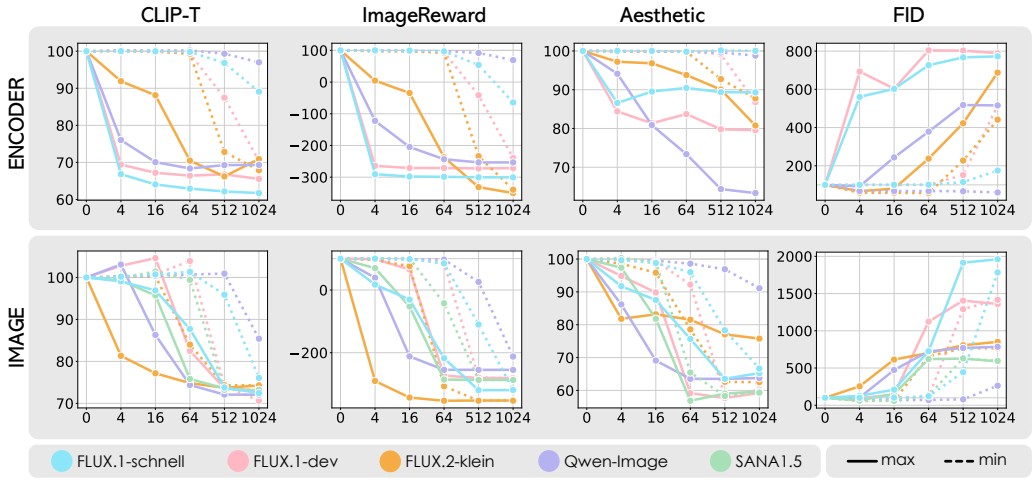

Figure 7: Magnitude-preserving token shuffling across models and streams. Metrics are percentages of the unperturbed baseline (100). Solid and dotted lines shuffle top-k and bottom-k channels, respectively; rows show encoder and image streams.

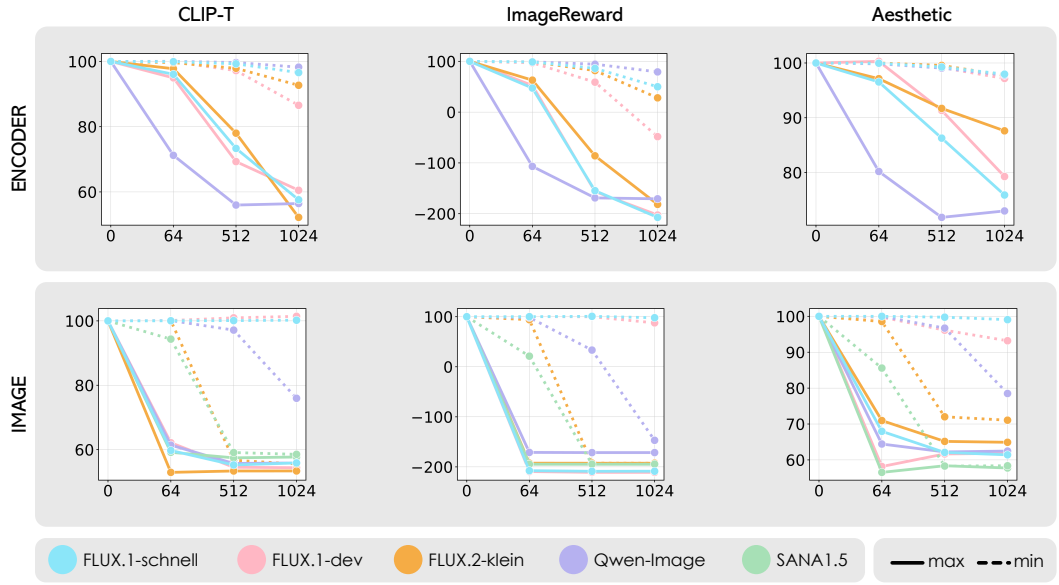

Figure 8: Effects of channel disruption on GenAI Bench, with top-k (solid) and bottom-k (dashed) channel selection, with k on the horizontal axis Results are reported relative the un-disrupted baseline.

Disruption on GenAI-Bench. We provide an additional insight into channel disruption by performing it on images generated on prompts from GenAI-Bench. Fig. 8 shows the obtained results, confirming the observations from Sec 3.1. The same considerations about different behaviours in the image stream and encoder stream hold also when considernig complex prompts, that go beyond simple class-like conditioning of ImageNet.

B.2 Additional Spatial Analysis

Fig. 9 shows an extended analysis of MAs-induced segmentation (as in Sec. 3.2) for the five selected models. In particular, we show metrics for dichotomous image segmentation, with pseudo ground truth clusters obtained through [44]. We report standard segmentation metrics: IoU, the intersection over union on the foreground class; mIoU, the mean IoU averaged over foreground and background; MAE, the pixel-level Mean Absolute Error; and bIoU, the IoU restricted to a narrow band around class boundaries.

The pattern is regular: best scores are consistently achieved by top-k activations, while we observe high variability with respect to layer selection. In particular, the best results are achieved at layer 11 for FLUX.1-dev, layer 44 for Qwen-Image, layer 10 for FLUX.2-klein and layer 56 for FLUX.1-schnell. This can be attributed to the architectural and training differences among the considered backbones, which differ in depth and in the way semantic information is consolidated along the network, so that the layer at which subject-level attributes emerge most prominently is model-specific. In contrast, the top-k activation selection consistently identifies the relevant units regardless of where this consolidation occurs. Finally, Fig. 19 shows qualitative examples of MAs-extracted masks, as described in Sec. 3.2.

B.3 Additional Transport Analysis

Baselines. We compare MAs against additional interpolation baselines that operate at different stages of the generation pipeline. Interpolation (α=0.5) is applied to DiT activations either across all transformer layers (*all*) or restricted to the middle layers, as identified as best in Sec. 3.3 as (*mid*).

Tab. 4 presents results over interpolation baselines. Latent Interpolation performs the analogous (α=0.5) mixing on the autoencoder latents, applied at either the first (*1st*) or last (*last*) denoising step. MAs (Ours) follows the masked, mid layer-step setup described in Sec. 3.3.

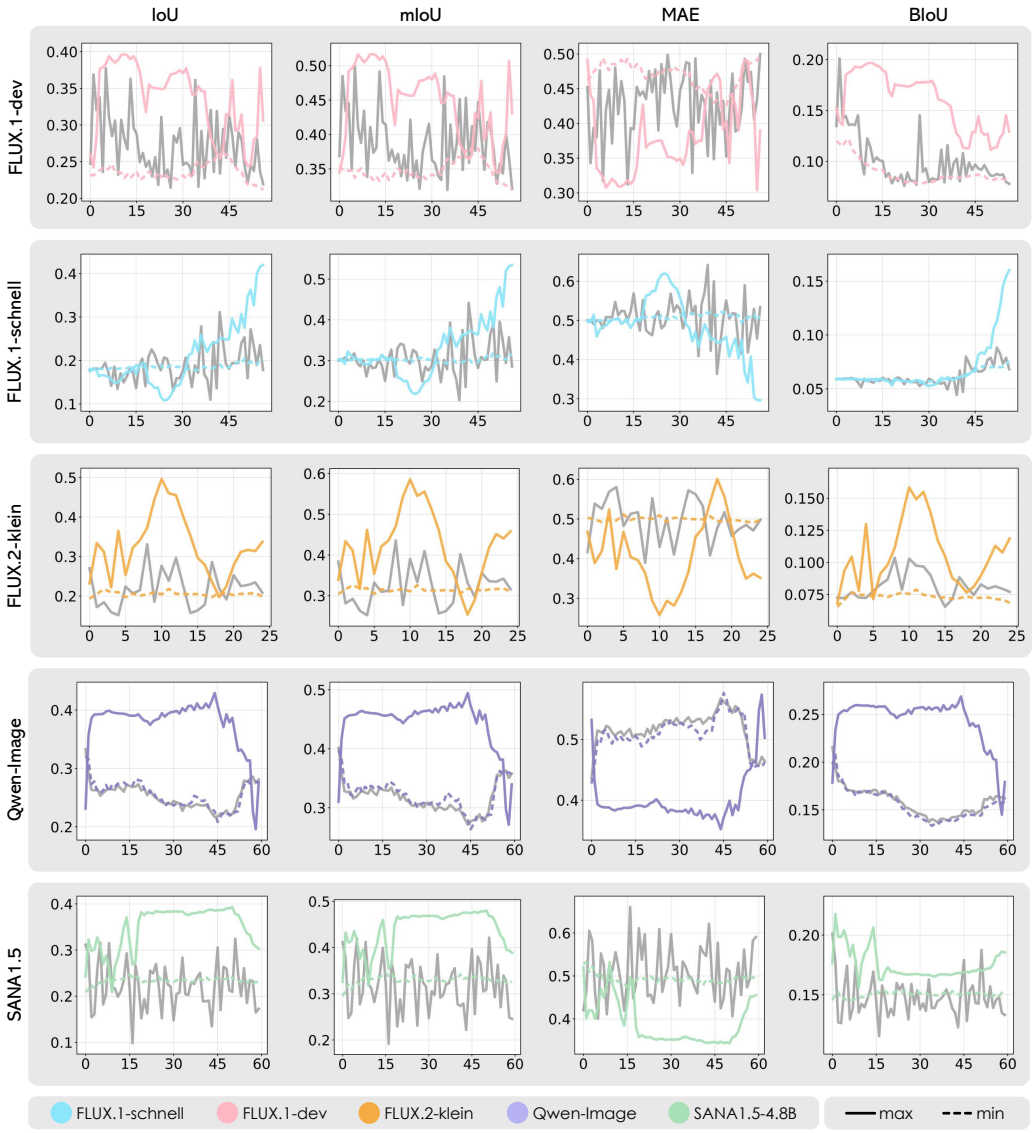

Figure 9: Metrics for dichotomous image segmentations across layers for various generative models. Solid lines refer to the use of top-k activations, grey lines to random-k activations, and dotted lines refer to bottom-k activations. Layer indexes are reported in $[0, L-1]$ range.

Interpolation collapses toward the source on all five backbones, with CLIP-T (S$*$T) between 22.1 and 33.9 and DINO-I (S$*$T) never above 3.3. Latent interpolation behaves differently: one denoising window (*1st*) yields a balanced regime competitive with MAs, while the other (*last*) collapses to a near-copy of source or target, evidenced by S or T above 80% paired with the complementary score under 40%. MAs is the only method that consistently balances source and target alignment.

Different Feature Selection Criteria. Following the same experimental protocol of Sec. 3.3, and starting from our best configuration (mid layers, 1024 shifted activations), we report in Tab. 5 an extended ablation that jointly varies the activation selection criterion (random, bottom, top) and the mask criterion (random, min, max) across five backbones (FLUX.1-schnell, FLUX.1-dev, FLUX.2-klein, Qwen-Image, SANA1.5) and three similarity metrics (CLIP-T, CLIP-I, DINO). For each metric we report the joint score S*T alongside the source-only (S) and target-only (T) components, the latter shown in gray as they are reported for completeness as their are not relevant on their own. The pattern is consistent across all backbones and metrics: Top activation selection dominates over bottom and random, and combining Top activations with the Max mask criterion (highlighted) yields the best joint score in nearly all cases, confirming that high-importance activations carry the signal most relevant to subject disentanglement, and that a precise mask is essential for a balanced transport.

Table 4: Additional comparison in prompt-to-prompt semantic transport with additional baselines.

		Mask	Layers / Steps	CLIP-T			CLIP-I			DINO-I		
				S*T	S	T	S*T	S	T	S*T	S	T
FLUX.1-schnell	Interpolation	-	all	31.6	83.2	38.0	53.4	100.0	53.4	2.7	99.9	2.7
	Interpolation	-	mid	31.6	83.2	38.0	53.3	99.8	53.4	2.7	99.7	2.7
	MAs (Ours)	✓	mid	**44.1**	62.8	70.3	**55.2**	71.2	77.6	**20.1**	42.4	47.5
FLUX.1-dev	Interpolation	-	all	32.4	82.0	39.5	47.8	91.0	52.5	2.5	86.2	2.9
	Interpolation	-	mid	33.9	82.9	40.8	45.4	88.3	51.5	2.6	84.7	3.0
	Latent Interpolation	-	1st	41.6	63.1	65.9	55.2	73.3	75.3	**20.4**	42.3	48.3
	Latent Interpolation	-	last	30.4	37.7	80.8	53.9	57.4	93.9	4.0	4.8	83.1
	MAs (Ours)	✓	mid	**42.5**	69.2	61.5	**58.1**	82.6	70.4	19.2	61.4	31.3
FLUX.2-klein	Interpolation	-	all	30.4	81.2	37.4	52.8	99.6	53.0	3.2	99.5	3.2
	Interpolation	-	mid	30.2	81.1	37.3	52.9	99.7	53.1	3.2	99.5	3.2
	Latent Interpolation	-	1st	29.8	37.2	80.0	50.4	54.0	93.4	3.6	4.2	86.1
	Latent Interpolation	-	last	39.1	61.2	63.9	47.7	68.5	69.6	13.3	33.4	39.7
	MAs (Ours)	✓	mid	**41.7**	64.6	64.5	**55.2**	75.2	73.4	**19.1**	58.0	32.9
Qwen-Image	Interpolation	-	all	22.1	47.1	47.0	25.1	50.6	49.6	0.0	0.8	0.7
	Interpolation	-	mid	22.8	47.8	47.7	25.6	51.1	50.1	0.0	0.7	0.6
	Latent Interpolation	-	1st	42.0	63.8	65.8	52.0	71.7	72.6	15.9	37.9	41.9
	Latent Interpolation	-	last	30.2	36.5	82.8	54.2	55.8	97.3	3.5	3.7	93.1
	MAs (Ours)	✓	mid	**44.5**	67.0	66.4	55.4	76.0	72.9	**18.0**	48.6	37.1
SANA1.5	Interpolation	-	all	31.4	66.3	47.3	34.8	64.9	53.7	1.4	34.1	4.2
	Interpolation	-	mid	32.6	69.3	47.0	36.8	67.9	54.3	2.0	43.4	4.6
	Latent Interpolation	-	1st	31.4	37.9	82.9	55.0	56.7	96.9	5.1	5.4	93.1
	Latent Interpolation	-	last	40.2	62.2	64.6	53.2	72.2	73.7	14.9	35.7	41.8
	MAs (Ours)	✓	mid	**41.4**	72.8	56.8	**55.7**	83.7	66.6	**17.0**	65.0	26.1

B.4 User Studies on Semantic Transport

We evaluate both transport settings with an image rating study. Annotators rate each generated image independently on a scale from 1 to 5 along three axes: concept preservation, prompt alignment, and visual quality. The user study presents participants with a generated image alongside two descriptions: Caption A (target concept) and Caption B (background). Here, concept preservation measures the retention of the subject defined by Caption A, while prompt following evaluates adherence to the background described in Caption B. Tables 6 and 7 report the ratings for the two settings.

Text-Conditioned Transport. We compare MAs with latent interpolation, the strongest interpolation baseline in the main paper. We sample 50 prompt pairs from the GenAI-Bench pair set for each of the five backbones, yielding 250 method comparisons and 500 generated images to rate. The human ratings corroborate the automated evaluation: across all five backbones, MAs achieve higher prompt alignment and overall scores than latent interpolation, while maintaining high subject faithfulness. This supports the improved balance between source-subject preservation and target-prompt alignment observed in the main paper.

Image-Conditioned Transport. On DreamBench++, we compare MAs with TokenVerse across all five backbones and include FLUX-Kontext, FLUX.2-klein, and Qwen-Image-Edit as dedicated editing references. The human ratings confirm the improvements over TokenVerse, with higher subject faithfulness and overall scores on every backbone, and show performance comparable to dedicated editing models. FLUX.1-dev with MAs matches the overall rating of FLUX-Kontext (4.3), while FLUX.2-klein with native MAs transport matches its editing baseline (4.7); Qwen-Image with MAs scores 4.3, compared with 4.4 for Qwen-Image-Edit. These findings support the conclusions of the automated evaluation with direct human assessments of the generated images.

C Limitations and Societal Impacts

Limitations. Our study is empirical in nature: we characterize the role of massive activations through interventional probes and demonstrate their use for semantic transport across generations, but we do not investigate the underlying reasons behind the emergence of this sparse subspace, nor what individual channels encode. Furthermore, the spatial masks induced by MAs reliably separate the salient subject from the background, but do not characterize what lies inside the salient region: they are agnostic to object identity and part-level structure. The transport mechanism inherits this granularity, transferring information within the salient region without distinguishing the entities it contains; finer, concept-aware masks would be required to disentangle and selectively manipulate

Table 5: Prompt-to-prompt composition under different activation (random, bottom-k, top-k) and masking strategies. Scores are reported for source (S), target (B), and joint preservation (S*T) using CLIP-T, CLIP-I, and DINO similarity.

	Criterion		CLIP-T			CLIP-I			DINO		
	Activations	**Mask**	**S*T**	**S**	**T**	**S*T**	**S**	**T**	**S*T**	**S**	**T**
FLUX.1-schnell	Random	Random	34.6	42.3	81.8	52.2	56.5	92.4	6.9	8.4	82.3
		Min	34.9	42.9	81.5	52.4	57.0	91.9	7.1	8.7	81.5
		Max	39.0	50.1	77.9	53.3	61.3	87.0	14.8	22.6	65.5
	Bottom	Random	38.1	48.1	79.2	53.0	60.0	88.4	11.9	16.4	72.6
		Min	38.9	49.5	78.6	53.3	61.0	87.4	12.8	18.1	70.4
		Max	42.3	57.4	73.8	54.5	66.6	81.7	18.6	34.0	54.8
	Top	Random	42.1	55.7	75.6	53.9	65.0	82.8	15.6	25.1	62.2
		Min	42.4	56.8	74.8	54.0	65.9	81.9	16.1	26.5	60.7
		Max	**44.1**	62.8	70.2	**55.2**	71.2	77.6	**20.1**	42.4	47.5
FLUX.1-dev	Random	Random	40.2	55.1	73.1	55.4	68.8	80.5	14.8	27.4	54.0
		Min	37.4	48.9	76.5	54.5	64.5	84.6	10.0	15.5	64.7
		Max	40.9	57.0	71.8	56.6	70.2	80.6	18.5	36.7	50.5
	Bottom	Random	37.7	50.0	75.3	55.9	66.3	84.4	14.0	23.1	60.4
		Min	34.5	44.1	78.1	55.7	62.6	89.0	9.3	13.0	71.7
		Max	38.8	52.4	74.1	56.9	67.5	84.2	17.4	30.3	57.6
	Top	Random	40.7	71.4	57.0	56.8	85.4	66.5	16.0	62.1	25.8
		Min	39.0	74.0	52.7	56.3	88.6	63.5	14.6	68.0	21.5
		Max	**42.5**	69.2	61.5	**58.1**	82.6	70.4	**19.2**	61.4	31.3
FLUX.2-klein	Random	Random	40.4	59.8	67.6	53.7	70.5	76.1	17.7	39.5	44.8
		Min	40.6	60.7	66.9	53.5	71.5	74.9	17.6	41.7	42.2
		Max	40.9	61.1	66.9	54.5	71.6	76.1	19.1	50.7	37.7
	Bottom	Random	40.3	61.2	65.9	53.3	71.7	74.3	16.6	37.5	44.4
		Min	40.4	61.5	65.6	52.8	71.9	73.5	16.2	37.1	43.8
		Max	41.2	63.2	65.3	54.9	73.5	74.7	**19.2**	51.2	37.5
	Top	Random	41.2	63.7	64.7	54.4	74.5	73.0	18.7	48.1	38.9
		Min	41.5	65.6	63.3	54.2	76.3	71.0	18.2	52.0	34.9
		Max	**41.7**	64.6	64.5	**55.2**	75.2	73.4	19.1	58.0	32.9
Qwen-Image	Random	Random	32.4	39.0	82.9	52.3	56.2	93.0	3.5	4.1	84.5
		Min	32.3	38.9	82.8	52.2	56.1	93.2	3.5	4.1	84.2
		Max	32.8	39.7	82.6	52.1	56.5	92.2	4.0	4.9	81.1
	Bottom	Random	30.9	37.3	82.8	53.9	56.3	95.7	3.8	4.3	89.2
		Min	30.9	37.4	82.8	53.9	56.3	95.7	3.9	4.4	88.9
		Max	31.5	38.2	82.5	53.5	56.6	94.4	4.5	5.2	85.3
	Top	Random	43.3	66.0	65.7	52.1	73.7	70.7	14.9	40.1	37.1
		Min	44.1	65.4	67.5	53.8	73.6	73.1	16.7	42.1	39.7
		Max	**44.5**	67.0	66.4	**55.4**	76.0	72.9	**18.0**	48.6	37.1
SANA1.5	Random	Random	39.8	73.1	54.4	53.8	83.4	64.5	14.9	56.6	26.4
		Min	40.2	71.3	56.4	53.6	81.2	66.0	15.3	50.5	30.4
		Max	40.7	73.1	55.6	55.4	83.8	66.1	**16.0**	65.1	24.6
	Bottom	Random	38.9	71.3	54.5	52.3	80.7	64.8	13.6	48.2	28.1
		Min	39.6	70.6	56.2	52.7	80.0	65.9	14.6	47.6	30.6
		Max	40.2	72.5	55.5	54.6	82.8	66.0	15.7	61.8	25.4
	Top	Random	39.8	74.6	53.4	54.4	85.2	63.8	14.7	61.2	24.1
		Min	40.7	73.7	55.1	54.7	84.3	65.0	15.9	59.4	26.8
		Max	**41.0**	74.9	54.7	**56.2**	86.4	65.1	15.7	71.3	22.1

individual semantic components. We leave both directions to future work. Finally, the transport mechanism operates on activations extracted from a specific generation instance, so artifacts and errors produced by the underlying generator on the source is inherited by the transported output.

Societal Impacts. Our analysis operates on pretrained DiTs without additional training, improving interpretability and controllability of text-to-image generators at low compute cost. The image-conditioned transport mechanism shares the risks common to personalization methods (non-consensual imagery, copyright infringement, misleading content); since we build on pretrained backbones, we inherit their safety mechanisms without weakening them, and recommend pairing downstream deployments with standard mitigations such as prompt filtering and watermarking.

Table 6: Human ratings for text-conditioned semantic transport. Annotators score each image from 1 to 5; higher is better. MAs are compared with latent interpolation on all five backbones.

Generator	Method	Subject Faithfulness	Prompt Alignment	Visual Quality	Overall
FLUX.1-dev	Latent Interpolation	4.6	1.3	**4.7**	3.5
	MAs	**4.7**	**4.6**	4.2	**4.5**
FLUX.1-schnell	Latent Interpolation	3.8	2.8	2.2	3.0
	MAs	**4.7**	**4.2**	**3.4**	**4.1**
FLUX.2-klein	Latent Interpolation	3.9	1.1	**4.3**	3.1
	MAs	**4.8**	**4.4**	3.6	**4.3**
Qwen-Image	Latent Interpolation	**4.6**	1.4	**4.7**	3.55
	MAs	4.2	**4.0**	4.4	**4.2**
SANA1.5	Latent Interpolation	**4.3**	1.1	**4.4**	3.2
	MAs	**4.3**	**3.2**	3.8	**3.8**

Table 7: Human ratings for image-conditioned semantic transport on DreamBench++. Scores range from 1 to 5; higher is better. Gray rows are dedicated editing baselines. MAs (TV) and MAs (EDIT) use reconstructed and native image-conditioned source trajectories, respectively.

Generator	Method	Subject Faithfulness	Prompt Alignment	Visual Quality	Overall
Qwen-Image-Edit	Edit	4.5	4.5	4.1	4.4
FLUX-Kontext	Edit	4.4	4.1	4.5	4.3
FLUX.2-klein	Edit	4.8	4.6	4.8	4.7
FLUX.1-dev	TokenVerse	2.1	3.7	**4.3**	3.4
	MAs	**4.3**	**4.4**	4.0	**4.3**
FLUX.1-schnell	TokenVerse	1.9	3.7	3.2	2.9
	MAs	**4.2**	**4.4**	**4.0**	**4.2**
FLUX.2-klein	TokenVerse	2.3	3.3	4.2	3.3
	MAs (TV)	4.4	**4.7**	4.7	4.6
	MAs (EDIT)	**4.5**	**4.7**	**4.8**	**4.7**
Qwen-Image	TokenVerse	2.2	3.3	**4.4**	3.3
	MAs	**4.8**	**4.2**	3.8	**4.3**
SANA1.5	TokenVerse	2.2	**4.2**	**4.0**	3.5
	MAs	**3.9**	3.9	3.4	**3.7**

D Additional Qualitative Results

We provide additional qualitative results for prompt-to-prompt semantic transport across models in Figs. 10-13. Each example shows a source, a target, and the merged output. Figs. 14-17 present image-conditioned semantic transport. Across both settings, the results exhibit consistent compositional behavior: salient elements from the source are integrated into the target scene while preserving its structure, lighting, and perspective. This is not a pixel-level mixture, the outputs remain visually coherent and semantically consistent, indicating that the transported information corresponds to high-level features. The effect is robust across multiple composition types, including object transfer, attribute transfer, and scene recomposition. In many cases, higher-level attributes such as artistic style, lighting, and semantic roles are preserved or reinterpreted. Notably, transferred subjects are not simply pasted into the scene, but are adapted to match the geometry, scale, and visual context of the target generation.



Figure 10: Qualitative examples of MAs-based text-conditioned transport on FLUX.1-dev.



Figure 11: Qualitative examples of MAs-based text-conditioned transport on Qwen-Image.

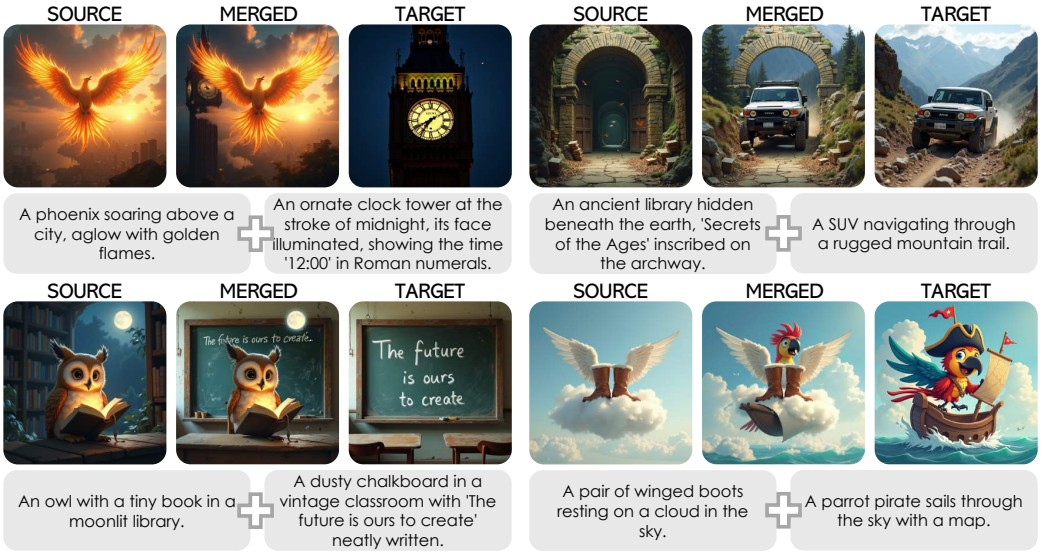

Figure 12: Qualitative examples of MAs-based text-conditioned transport on FLUX.2-klein.

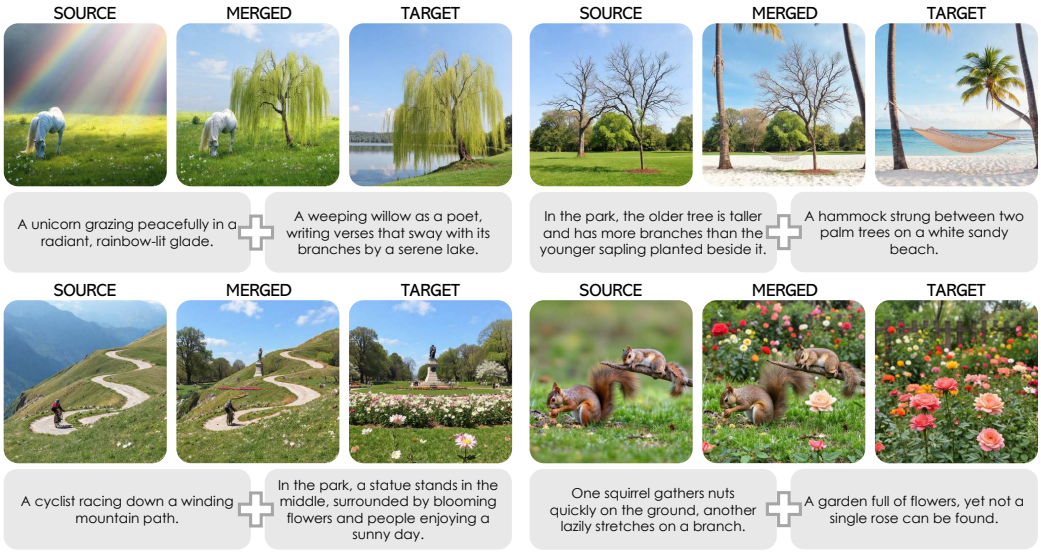

Figure 13: Qualitative examples of MAs-based text-conditioned transport on FLUX.1-schnell.

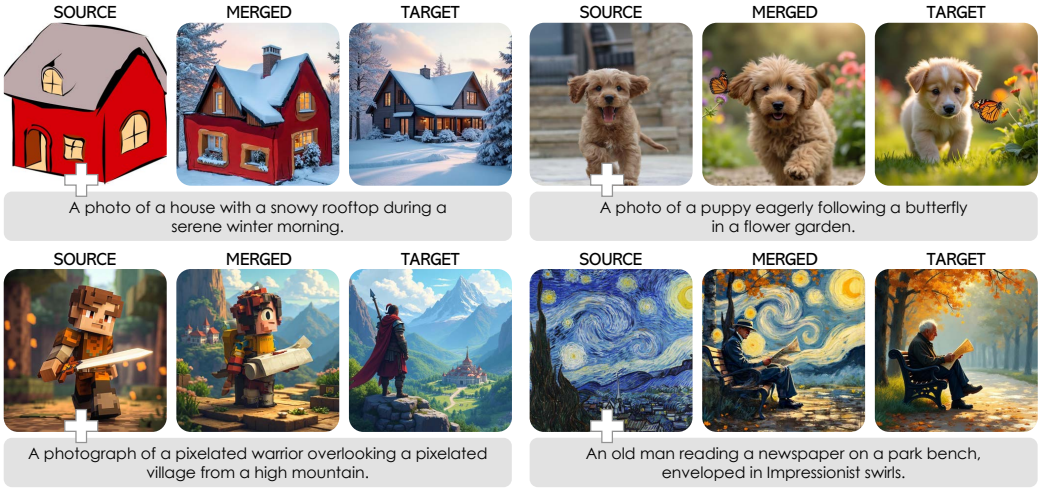

Figure 14: Qualitative examples of MAs-based image-conditioned transport on FLUX.1-dev.



Figure 15: Qualitative examples of MAs-based image-conditioned transport on Qwen-Image.



Figure 16: Qualitative examples of MAs-based image-conditioned transport on FLUX.2-klein.

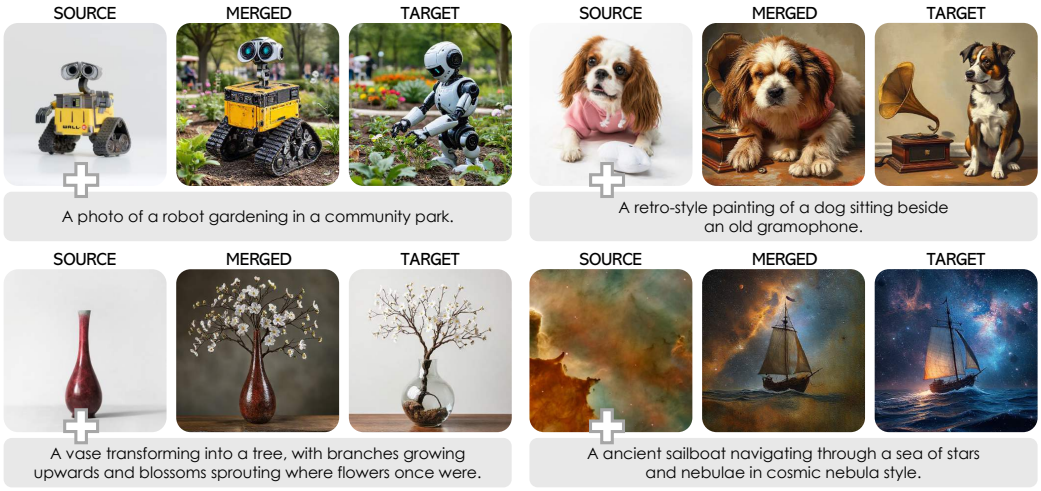

Figure 17: Qualitative examples of MAs-based image-conditioned transport on FLUX.1-schnell.

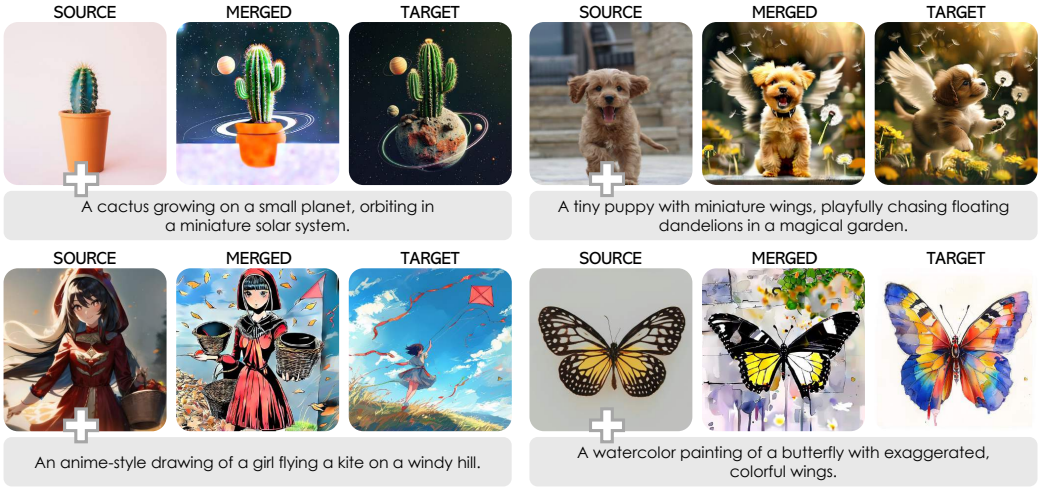

Figure 18: Qualitative examples of MAs-based image-conditioned transport on SANA1.5.



Figure 19: Qualitative examples for MAs-based masking extraction, on FLUX.1-dev (top row) and FLUX.2-klein (bottom row)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction accurately previews the contributions and define the scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a dedicated section in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: We do not provide any theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental settings are disclosed in the main paper and detailed in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Full code will be available at a later time.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All implementation details are provided in the main paper and in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report experiments conducted on statistically relevant sets, and do not require error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: the paper does not introduce any large-scale trainings, thus does not introduce non-trivial computational overhead with respect to prior cited literature.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We provide a discussion of societal impacts in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release any new model or dataset that pose additional risks with respect to prior cited literature.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All original owners are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release any new asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: A user study was conducted and a proper description of the human evaluation protocol and the corresponding results are provided in Appendix B.4.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: No IRB was required for this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not involved in the core methodology employed by this work, but have been employed for grammar checking and rephrasing during the writing of the manuscript.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.