SHIFT: Steering Hidden Intermediates in Flow Transformers

Nina Konovalova [1 2] **Ibragim Idrisov** [3] **Aibek Alanov** [1 2]

Abstract

Diffusion models have become leading approaches for high-fidelity image generation. Recent DiT-based diffusion models, in particular, achieve strong prompt adherence while producing high-quality samples. We propose **SHIFT**, a simple but effective and lightweight framework for *concept removal* in DiT diffusion models via targeted manipulation of intermediate activations at inference time, inspired by activation steering in large language models. **SHIFT** learns steering vectors that are dynamically applied to selected layers and timesteps to suppress unwanted visual concepts while preserving the prompt's remaining content and overall image quality. Beyond suppression, the same mechanism can shift generations by adding or changing target objects. We demonstrate that **SHIFT** provides effective and flexible control over DiT generation across diverse prompts and targets without time-consuming retraining.

1. Introduction

Diffusion models (Ho et al., 2020; Rombach et al., 2022) have established a new state-of-the-art in high-fidelity text-to-image synthesis (Podell et al., 2023). However, as model capabilities scale, so do the risks associated with the generation of harmful, copyrighted, or prohibited content. This has motivated the development of Concept Erasure (CE): the task of reliably suppressing specific semantic concepts within a model's generative manifold.

Existing erasure strategies for text-to-image diffusion models mostly rely on model optimization. Approaches such as ESD (Gandikota et al., 2023), Concept Ablation (CA) (Kumari et al., 2023), and Erase-and-Preserve (EAP) (Bui et al., 2024) require gradient-based tuning to achieve robust concept removal. While these methods were computationally viable for earlier UNet architectures (typically < 1B parameters), they become prohibitively expensive when applied to modern Diffusion Transformers (DiTs) like Flux (Labs et al., 2025), which contains 12B parameters. Furthermore, methods like ESD or EAP frequently utilize negative guidance during the optimization phase to steer the model distribution away from the target concept. This dependency presents a significant challenge for distilled models such as Flux.1[schnell], because these models are guidance-distilled to operate at a fixed guidance scale. Non-optimization alternatives, such as Unified Concept Editing (UCE) (Gandikota et al., 2024), propose closed-form weight edits that avoid retraining. However, these edits are often insufficient for concrete object erasure in unified architectures.

In parallel, activation steering has emerged as a lightweight yet potent control mechanism for Large Language Models (LLMs) (Liu et al., 2025; Turner et al., 2024; Beaglehole et al., 2026). Frameworks such as Representation Engineering (Zou et al., 2023) demonstrate that high-level semantic behavior can be modulated by injecting linear directions into the latent representation space at inference time, without any weight modification. While recent work like CASteer (Gaintseva et al., 2025) has sought to adapt these principles to diffusion models, such implementations remain fundamentally dependent on the cross-attention mechanisms typically present in UNet-based architectures.

However, the field is currently shifting toward Diffusion Transformers (DiTs) (Peebles & Xie, 2023). Architectures like Flux (Labs et al., 2025) or SD3.5 (Esser et al., 2024) utilize Multimodal Diffusion Transformer (MM-DiT) blocks, where the traditional decoupling of cross-attention and self-attention is replaced by a unified attention mechanism that processes text and image tokens in a shared latent space. Consequently, erasing concepts in DiTs requires a more precise intervention capable of navigating this integrated representation manifold without compromising the global structure of the generated image. This raises a natural question for DiT: can we achieve robust, inference-time concept removal in DiTs by directly steering these unified internal activations?

We introduce **SHIFT**, a steering framework specifically engineered for the Multimodal Diffusion Transformer (MM-

[1]FusionBrain Lab [2]HSE University [3]Lomonosov Moscow State University. Correspondence to: Nina Konovalova <nina.konovalova.k@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.

DiT) architecture. In contrast to existing methodologies that necessitate complex, timestep-specific interventions, we demonstrate that SHIFT identifies temporally invariant steering vectors that remain semantically stable, providing a concept control mechanism that does not require retraining the base model.

While our primary evaluation focuses on concept erasure, we establish that this activation-level modulation extends to object-specific biasing. Importantly, SHIFT is positioned as a generation control framework rather than a traditional image-editing tool; it focuses on the global redirection of the model's generative manifold rather than the preservation of spatial layouts or background consistency from a specific input image. This distinction allows SHIFT to achieve robust semantic shifts while maintaining the high-fidelity synthesis inherent to the Flux backbone.

Our contributions are following:

- **Unified steering framework:** we introduce SHIFT, the first comprehensive framework for steering both original and distilled DiT-based models (e.g., Flux.1[dev] and Flux.1[schnell]) through latent activation shifts, enabling efficient and scalable concept manipulation without retraining.
- **Spatial and temporal dynamics analysis:** we conduct an extensive ablation study on the spatial and temporal aspects of steering, uncovering a remarkable temporal consistency in DiT activation spaces. Specifically, we show that a single, time-independent steering vector can be applied effectively across all diffusion timesteps, streamlining the steering process and indicating that semantic concepts are encoded as stable, global directions in the latent manifolds of DiTs.
- **Cross-distillation vector transfer:** we investigate the transferability of steering vectors across distillation boundaries (e.g., from Flux.1[schnell] to Flux.1[dev]) and validate the efficacy of a unified steering vector shared across multiple timesteps, demonstrating robust performance in resource-constrained settings.

2. Related works

2.1. Concept erasure and safety

Recent state-of-the-art text-to-image diffusion models trained on large, imperfectly filtered datasets such as LAION (Schuhmann et al., 2022) can generate inappropriate or copyrighted content. One mitigation is dataset filtering (Rombach et al., 2022) or post-generation filtering (Rando et al., 2022). However, these approaches do not fully prevent harmful content and require additional processing. Many works therefore focus on concept erasure (CE) while preserving overall generation quality. Methods such as ESD (Gandikota et al., 2023) employ fine-tuning to unlearn concepts using negative guidance from a frozen teacher model, while Concept Ablation (CA) (Kumari et al., 2023) optimizes cross-attention layers to redirect target concepts toward neutral anchors. Subsequent approaches like Unified Concept Editing (UCE) (Gandikota et al., 2024) introduce closed-form edits to attention projections for efficient, training-free erasure, though they may struggle to fully remove concrete objects. For modern architectures, EraseAnything (Gao et al., 2025) adapts erasure to rectified flow models via LoRA tuning and attention regularization, and Erasing with Adversarial Preservation (EAP) (Bui et al., 2024) incorporates adversarial concept identification to preserve unrelated generations during fine-tuning. However, these methods have limitations on modern diffusion models, including optimization overhead, reliance on guidance that must be adapted for distilled models, or dependence on LLM agents.

2.2. Activation steering in Large Language Models.

Steering in large language models (LLMs) has emerged as a practical paradigm for controlling behavior. Various approaches have been explored, including specific interventions on weights (Ziegler et al., 2019; Ilharco et al., 2022; Meng et al., 2022), prompt engineering (Zhou et al., 2022), and soft prompting (Khashabi et al., 2022; Lester et al., 2021). At the same time, one of the most promising directions involves modifying activations using specifically calculated steering vectors. These vectors can be estimated using gradient descent (Hernandez et al., 2023; Subramani et al., 2022), PCA decomposition (Zou et al., 2023), or simply as the mean difference between activations from contrastive prompt pairs (Turner et al., 2023; Li et al., 2023). While these methods operate primarily on transformer-based LLMs, they provide foundational insights for extending activation steering to other architectures, such as diffusion models.

2.3. Latent space navigation and Diffusion Steering.

Diffusion models enable controllable synthesis through guidance signals and latent-space manipulations, allowing users to influence the generation process toward desired outcomes. Early works focused on semantic navigation in latent representations, such as interpolating between latent codes to blend concepts, and guidance-based conditioning, including classifier guidance (Dhariwal & Nichol, 2021) where an external classifier steers the denoising process, classifier-free guidance (Ho & Salimans, 2022) that amplifies conditioning signals without additional models or more complicated approaches introducing semantic guidance (Brack et al., 2022).

More recent methods have adapted steering techniques from large language models by intervening directly in diffusion model activations, offering finer-grained control without retraining. In UNet-based latent diffusion models, a widespread control mechanism is cross-attention manipulation: for instance, CASteer (Gaintseva et al., 2025) constructs steering vectors for cross-attention layers using contrastive prompt pairs (e.g., prompts with and without the target concept) to enable concept erasure, style transfer. Other examples include Activation Transport (AcT) (Rodriguez et al., 2024), which applies optimal transport theory to transport activations between source and target distributions for precise steering in both language and diffusion models.

3. Preliminaries

Early text-to-image diffusion models were mostly U-Net based, with explicit self-attention over image latents and cross-attention for text conditioning. Modern text-to-image systems increasingly use diffusion transformers (DiTs), where image latents and text conditioning are represented as token sequences and processed by transformer blocks. In this work, we focus on Flux.1. At each denoising step, the Flux transformer receives image latent tokens together with two text-conditioning inputs: token-level text embeddings, and a pooled text embedding. In the standard Flux.1 pipeline, the token-level embeddings are produced by T5 (Raffel et al., 2020), while the pooled embedding is produced by the CLIP text branch (Radford et al., 2021). The pooled embedding is combined with the diffusion timestep and used to modulate the transformer computation.

Let X_i^ℓ and X_t^ℓ denote the image and text streams entering block ℓ, and let p denote the pooled text embedding. The pooled route is first combined with the timestep τ into a modulation vector

$$e_\tau = \phi(\tau, p), \tag{1}$$

which parametrizes adaptive normalization and gating inside transformer blocks. Thus, Flux uses token-level text embeddings for sequence-level prompt semantics and a pooled text route for global conditioning.

Flux contains two types of transformer blocks: **double-stream** and **single-stream**. In double-stream blocks, text and image tokens are kept as separate residual streams that interact through joint attention. A double-stream block first applies a gated joint-attention residual update and then applies a modality-specific feed-forward residual update. These internal operations could be abstracted as a total residual update for each stream:

$$\begin{aligned} X_i^{\ell+1} &= X_i^\ell + R_{i,\ell}(X_i^\ell, X_t^\ell; e_\tau), \\ X_t^{\ell+1} &= X_t^\ell + R_{t,\ell}(X_t^\ell, X_i^\ell; e_\tau), \end{aligned} \tag{2}$$

where $R_{i,\ell}$ and $R_{t,\ell}$ denote the total residual updates produced by the block for the image and text streams, respectively. Each residual update contains the block's joint-attention and feed-forward computations, including the adaptive modulation determined by e_τ. The outputs $X_t^{\ell+1}$ and $X_i^{\ell+1}$ are the post-block states propagated to subsequent blocks. An illustration of the block structure is provided in Appendix A Figure 8.

By contrast, single-stream blocks concatenate text and image tokens and process them with shared weights. In this work, we primarily steer double-stream blocks, where text and image streams remain explicitly separated.

4. Method

In this work, we investigate activation steering for transformer-based diffusion models, with a primary focus on Flux.1[schnell] and Flux.1[dev] for concept erasure. We also demonstrate the method's applicability to other tasks, like adding and removing objects. We begin by motivating the choice of internal representations to steer in DiT-style architectures, then describe the construction of steering vectors, and finally explain their application during inference.

4.1. Where to apply steering

The residual structure in Eq. 2 gives several possible diffusion-transformer intervention sites. For the text stream of a double-stream block, one may steer the internal residual update $R_{t,\ell}$ or steer the post-block state $X_t^{\ell+1}$. These interventions are not equivalent under difference-based steering.

Consider a single contrastive pair and omit ℓ, τ from the notation. Let X_t^- and X_t^+ denote the source and target text states entering the block, and let R_t^- and R_t^+ denote the corresponding total residual updates produced by the block. If we steer only the residual contribution by its exact source-to-target difference, then

$$\widetilde{R}_t = R_t^- + (R_t^+ - R_t^-) = R_t^+. \tag{3}$$

The resulting block output is

$$\widetilde{X}_t^{\ell+1} = X_t^- + \widetilde{R}_t = X_t^- + R_t^+. \tag{4}$$

However, the target block output is

$$X_t^{\ell+1,+} = X_t^+ + R_t^+. \tag{5}$$

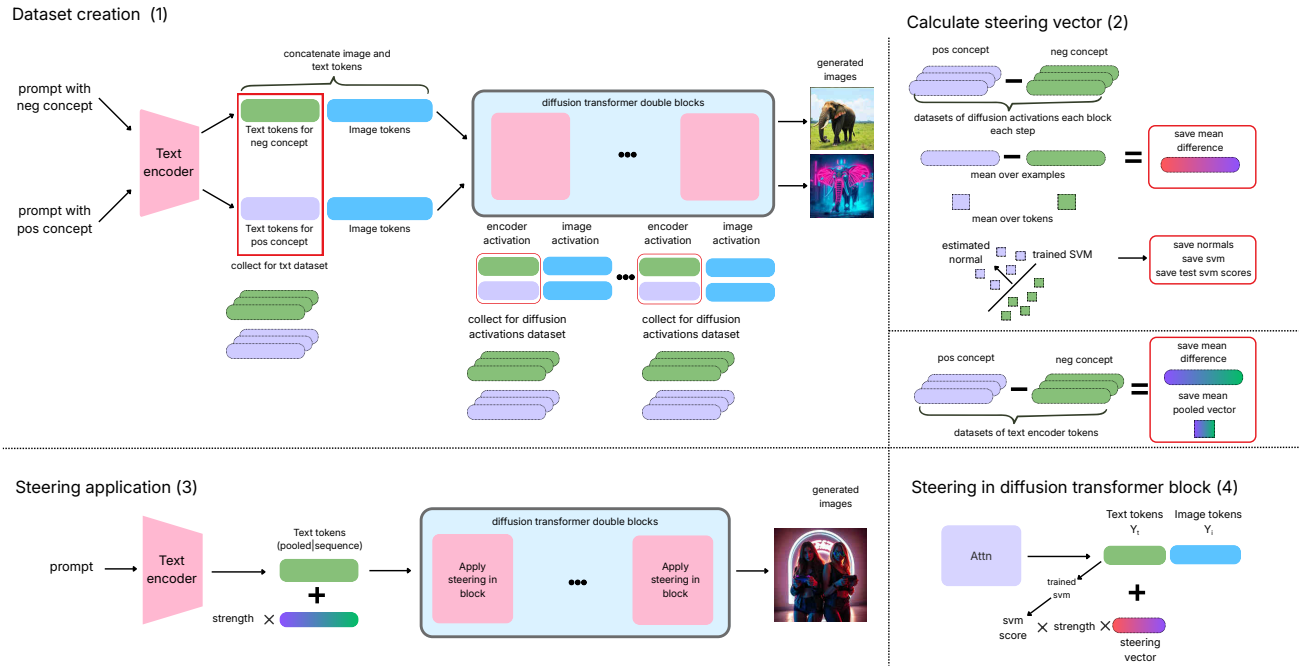

Figure 1. Overview of the steering pipeline: (1) dataset construction from contrastive prompt pairs, (2) steering vector computation based on mean difference and separation plane, (3) application of the vector during inference, and (4) steering inside the diffusion transformer.

Thus, even exact residual-branch steering leaves the mismatch

$$X_t^{\ell+1,+} - \widetilde{X}_t^{\ell+1} = X_t^{+} - X_t^{-}. \tag{6}$$

This mismatch is caused by the source-conditioned stream that remains in the skip connection.

Post-block steering avoids this particular issue by acting directly on the state propagated to downstream blocks. In the single-pair case,

$$\widetilde{X}_t^{\ell+1} = X_t^{\ell+1,-} + \left(X_t^{\ell+1,+} - X_t^{\ell+1,-}\right) = X_t^{\ell+1,+}. \tag{7}$$

Therefore, post-block steering exactly reconstructs the target post-block state in the single-pair setting. In practice, we use a mean direction estimated from multiple contrastive prompt pairs, so Eq. 7 becomes an approximate transport toward the target activation distribution.

This choice is also more architecture-agnostic. Attention and feed-forward internals vary across DiT implementations, whereas block outputs are a common interface for residual transformers. We use post-block text states $X_t^{\ell+1}$ in selected double-stream blocks as our primary diffusion-transformer steering site. We additionally steer the pooled text-conditioning route as a complementary pathway. The pooled embedding is combined with the timestep to form e_τ, which modulates adaptive normalization and gating throughout the transformer. Thus, pooled steering changes the global modulation signal, while post-block text steering changes the propagated token-level text state.

4.2. Steering vector construction

We construct a separate steering vector for each target concept X and for each steering location (text encoder pooled embedding and diffusion transformer tokens). We collect a small paired dataset of n prompt pairs (p_k^-, p_k^+), where p_k^+ adds the target concept to an otherwise neutral prompt, "an elephant" vs. "an elephant in cyberpunk style". Running the model on these prompts, we record activations at the chosen locations and obtain paired samples (X_k^-, X_k^+). The process of dataset construction is illustrated in Fig. 1 (1). After dataset collection, we estimate the steering vectors (Fig. 1 (2)).

Text encoder steering vector. For the pooled text-encoder representation, we use the raw activation-difference vector as the steering direction.

Diffusion transformer steering vector. For diffusion-transformer steering, we use the mean-difference direction at the post-block text states defined in Sec. 3. For prompt pair k, denote the recorded post-block text activations at block ℓ and denoising step τ by

$$X^-_{k,\ell,\tau}, X^+_{k,\ell,\tau} \in \mathbb{R}^{T\times C}. \tag{8}$$

Here T is the number of text tokens and C is the channel dimension. We estimate a per-token steering field

$$\Delta X_{\ell,\tau} = \frac{1}{n}\sum_{k=1}^{n}\left(X^{+}_{k,\ell,\tau} - X^{-}_{k,\ell,\tau}\right) \in \mathbb{R}^{T\times C}. \quad (9)$$

To separate direction from strength, we normalize the field over the channel dimension for each token:

$$v_{\ell,\tau,j} = \frac{\Delta X_{\ell,\tau,j}}{\|\Delta X_{\ell,\tau,j}\|_2 + \varepsilon}, \qquad j = 1,\dots,T. \quad (10)$$

4.3. Inference-time Steering

At inference time, we steer the model by adding a concept-specific direction to selected activations (Figure 1 4). Let a denote an activation (either the pooled text embedding or a text-token activation in the diffusion transformer), and let v denote the corresponding steering vector. The intervention is defined as

$$\tilde{X} = X + \alpha\, v, \quad (11)$$

where α is the steering strength. The estimation of α is described in the following paragraph.

Text encoder steering strength. For the text encoder, we steer only the pooled embedding. We set the steering strength based on the cosine similarity between the initial-prompt embedding e_{init} and the target-concept embedding e_{target}, scaled by a user-defined coefficient γ:

$$\alpha_{\text{pool}} = \gamma\, \cos\big(e_{\text{init}}, e_{\text{target}}\big). \quad (12)$$

This cosine-based scaling helps preserve non-target concepts while suppressing the target concept.

Diffusion transformer steering strength. For diffusion-transformer steering, we regularize the steering strength using a lightweight classifier (SVM) signal that estimates whether the target concept remains present in the current hidden activations. For each selected site (ℓ,τ), we pool text-token activations over the token dimension:

$$\bar{X}^{+}_{k,\ell,\tau} = \frac{1}{T}\sum_{j=1}^{T} X^{+}_{k,\ell,\tau,j}, \qquad \bar{X}^{-}_{k,\ell,\tau} = \frac{1}{T}\sum_{j=1}^{T} X^{-}_{k,\ell,\tau,j}. \quad (13)$$

We train a linear classifier on

$$\mathcal{D}_{\ell,\tau} = \{(\bar{X}^{+}_{k,\ell,\tau}, 1)\}_{k=1}^{n} \cup \{(\bar{X}^{-}_{k,\ell,\tau}, 0)\}_{k=1}^{n}. \quad (14)$$

At inference time, the classifier produces a target-concept confidence $p_{\text{cls}} \in [0,1]$ for the current activation. We convert this score into a bounded scaling factor

$$\eta_{\text{cls}} = \text{clip}\left(\frac{1}{(1-p_{\text{cls}}) + \varepsilon} - 1, 0, \eta_{\max}\right), \quad (15)$$

and modulate the base steering strength γ as

$$\alpha_{\text{diff}} = \gamma\, \eta_{\text{cls}}. \quad (16)$$

This mechanism increases steering when the classifier detects residual target-concept evidence and suppresses steering when the concept is already weak or poorly separated at the selected site.

5. Experiments

5.1. Overview

We evaluate SHIFT on a diverse set of tasks covering both *abstract* and *concrete* concepts. In particular, we focus on (i) erasing safety-critical abstract concepts (e.g., nudity), (ii) concrete defined concepts. We additionally include small local objects removal examples (e.g., hats and glasses) to provide visually interpretable results.

5.2. Implementation Details

5.2.1. SHIFT

We compute steering vectors as activation differences between prompts with and without the target concept. For concrete object and concept erasure we use 20 prompt pairs; for nudity erasure we use 135 pairs. The examples of prompts are presented in Appendix B.2. We also train a linear SVM and use its score as a regularizer during generation. Unless stated otherwise, we fix the steering strength to 6 for text pooled text-encoder activations. We use block-specific steering vectors, and apply a single vector across all timesteps starting from step 0. We evaluate on Flux.1[dev] and Flux.1[schnell]. For Flux.1[dev] we use 28 steps with guidance scale 3.5; for Flux.1[schnell] we use 4 steps with guidance scale 0.0. Our main experiments use 1024 resolution, for nudity erasure we use 512 following EraseAnything (Gao et al., 2025) recommendation.

5.2.2. Baselines

We compare SHIFT against several concept-erasure baselines. Several of these methods were originally introduced

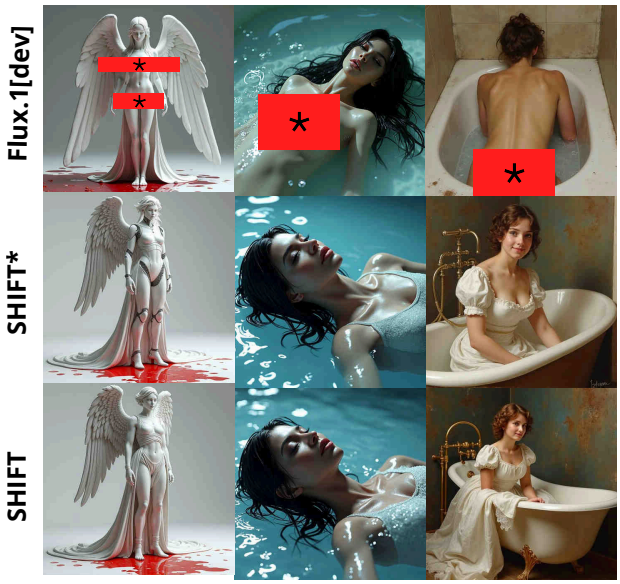

Figure 2. Flux.1[dev] original and steered generation. * denotes that we use steering vector from Flux.1[schnell] activations to steer Flux.1[dev].

for SD1.5 and later adapted to Flux without explicitly recommended hyperparameters, which makes reproduction more challenging. Therefore, in each corresponding experimental section and Appendix B.1, we explicitly report all training details used for these baselines.

5.3. Results

5.3.1. Abstract Concept Erasure

In this section, we study **abstract concept erasure** on the I2P benchmark (Schramowski et al., 2023). I2P contains 4,703 prompts and corresponding seeds designed to generate inappropriate generations, including nudity. Following prior work, we use NudeNet with a threshold of 0.6 to detect nude body parts for the nudity-erasure task. We evaluate both Flux.1[dev] and Flux.1[schnell]. For Flux.1[schnell], we compare against EAP (Bui et al., 2024), CA (Kumari et al., 2023), and UCE (Gandikota et al., 2024) using the recommended public configurations. All models are evaluated at 512×512 resolution. For additional evaluation of our method impact on overall image quality generation we validate our method on selected 5,000 captions from MS-COCO with corresponding seed following EraseAnything (Gao et al., 2025) recommendation.

The results are presented in Table 1. SHIFT clearly outperforms all baselines on safety-related metrics, achieving more than $3\times$ and $4\times$ stronger suppression at different steering strengths. At the same time, CLIP and FID remain nearly unchanged, indicating good prompt alignment. The table also shows that increasing steering strength improves erasure but reduces CLIP and increases FID metrics, showing a trade-off between erasure and quality preservation.

Figure 3. **Qualitative Comparison with Baselines:** the first row – original generation of Flux.1[schnell], from top to bottom – UCE, CA and SHIFT (Ours)

For Flux.1[dev], we report the baseline metrics from EraseAnything (Gao et al., 2025). We further evaluate Flux.1[dev] using steering vectors computed from both its own activations and Flux.1[schnell] activations. As shown in Table 2, our method remains effective for the non-distilled model even when the steering vector is transferred from the distilled model.

Table 1. (Left) Quantity of explicit content detected using the NudeNet detector on the I2P benchmark for Flux.1[schnell]. (Right) Comparison of FID and CLIP on MS-COCO 5K sampled images.

Method	**Detected Nudity (Quantity)**				**MS-COCO 5K**	
	Common	**Female**	**Male**	**Total↓**	**FID↓**	**CLIP↑**
ESD	280	121	25	426	31.88	30.86
EAP	373	175	10	558	32.02	31.51
CA	234	98	10	342	31.81	31.25
UCE	232	127	5	364	**31.65**	**31.55**
SHIFT ($\gamma = 25$)	76	30	6	112	32.1	31.42
SHIFT ($\gamma = 30$)	**61**	**20**	**4**	**85**	32.3	31.18
Flux.1[schnell]	412	190	10	612	31.63	31.59

We additionally provide a qualitative comparison in Fig. 3 for Flux.1[schnell] and Fig. 2 for Flux.1[dev], showing that our method can erase undesirable concepts while preserving overall image quality.

Table 2. Quantity of explicit content detected using the NudeNet detector on the I2P benchmark for Flux.1[dev].

Method	Detected Nudity (Quantity)			
	Common	Female	Male	Total↓
CA	253	65	26	344
ESD	329	145	32	506
UCE	122	39	12	173
MACE	173	55	28	256
EAP	287	86	13	386
Meta-Unlearning	355	140	26	521
EraseAnything	129	48	22	199
SHIFT (dev acts)	**62**	**11**	**6**	**79**
SHIFT (schnell acts)	**62**	14	7	83
Flux.1[dev]	406	161	38	605

5.3.2. Concrete Concept Erasure

We additionally evaluate our method on concrete concept erasure. Following the protocol established by SPM (Lyu et al., 2024), we assess concrete concept removal using 80 fixed prompts and 9 different seeds. Specifically, we evaluate erasure of the Snoopy concept while preserving five related concepts: Mickey, SpongeBob, Pikachu, dog, and legislator. For quality evaluation, we compute CLIP scores between the target prompt and the generated image, and FID between original and erased generations. For non-target concepts, we expect lower FID and higher CLIP, while for the erased target concept we expect a decrease in CLIP. We conduct these experiments on Flux.1[schnell]. Results for Flux.1[dev] are provided in the Appendix C.2 Table 9 and qualitative results are presented in Figure 4.

Figure 4. Qualitative Comparison with Baselines for Flux.1[dev] (left) and Flux.1[schnell] (right). * denotes that we use steering vector from Flux.1[schnell] activations to steer Flux.1[dev]

Table 3. Quantitative evaluation of concrete object erasure for Flux.1[schnell] model.

Method	Snoopy		Mickey		Spongebob		Pikachu		Dog		Legislator	
	CS↓	FID	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
CA	22.49	168.37	24.25	129.91	**28.25**	131.78	**27.44**	129.54	22.86	96.94	21.18	83.83
SHIFT	**18.74**	151.35	**26.48**	**51.71**	27.76	**46.56**	27.21	**40.10**	**24.63**	**37.15**	**21.54**	**40.40**
Flux.1[schnell]	28.01	–	26.72	–	27.94	–	27.15	–	24.62	–	21.89	–

5.3.3. Small Object Erasure

Additionally, we evaluate our method for domain shifting, where we suppress small objects in generated images. For example, we prevent generations of people with glasses, hats, or remove smiles. These experiments are conducted on Flux.1[schnell] with 4 inference steps. Qualitative results are shown in Figures 5–7. Our method can erase not only global concepts but also small local objects. Table 4 presents quantitative evaluation, with '%' denoting the percentage of manually identified concepts in the set of generated images. However, it is not an image-editing method, and background consistency is not preserved.

Table 4. Remove-task steering quantitative results for the Flux.1[schnell] model.

Method	Glasses			Hat			Smile		
	% ↓	FID ↓	CLIP ↑	% ↓	FID ↓	CLIP ↑	% ↓	FID ↓	CLIP ↑
Flux.1[schnell]	93.2	–	30.74	90.0	–	31.52	71.8	–	31.57
SHIFT ($\gamma = 10$)	87.7	52.95	30.98	83.4	46.25	31.81	70.1	56.01	31.66
SHIFT ($\gamma = 20$)	79.8	69.43	31.40	72.7	63.82	31.87	65.8	74.32	31.86
SHIFT ($\gamma = 30$)	51.4	104.08	31.53	44.1	89.64	32.17	53.5	99.06	31.81

Figure 5. Steering to remove small concepts: glasses

Figure 6. Steering to remove small concepts: hat

Figure 7. Steering to remove small concepts: smile

5.4. Ablations

In this section, we provide a comprehensive ablation analysis of the proposed steering process.

5.4.1. Where to apply steering

In the main part of the paper, we apply steering to the outputs of transformer blocks. Here, we investigate an alternative location: the outputs of the joint attention operation.

Specifically, the joint attention in a diffusion transformer produces representations for both text and image tokens:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right), \quad [Y_t, Y_i] = Y = AV, \tag{17}$$

where Y_t and Y_i denote the text and image token representations. We apply the steering vector to the text token representations Y_t (Figure 9):

$$\tilde{Y} = Y_t + \alpha \cdot v. \tag{18}$$

Table 5. (Left) Quantity of explicit content detected using the NudeNet detector on the I2P benchmark for Flux.1[schnell] attention-based and main method. (Right) Comparison of FID and CLIP on MS-COCO 5K sampled images.

Method	**Detected Nudity (Quantity)**				**MS-COCO 5K**	
	Common	**Female**	**Male**	**Total↓**	**FID↓**	**CLIP↑**
SHIFT-attn ($\gamma = 250$)	87	32	3	122	33.9	31.22
SHIFT-attn ($\gamma = 500$)	73	23	**1**	97	34.5	31.09
SHIFT ($\gamma = 25$)	76	30	6	112	**32.1**	**31.42**
SHIFT ($\gamma = 30$)	**61**	**20**	4	**85**	32.3	31.18
Flux.1[schnell]	412	190	10	612	31.63	31.59

We evaluate the SHIFT-attn variant on both concrete object erasure (Table 6) and abstract concept erasure (nudity, Table 5, Figure 11). Although attention-level steering proves effective, it requires significantly higher steering strength γ and leads to slightly greater degradation of non-target concepts than block-output steering.

Table 6. Quantitative evaluation of concrete object erasure for Flux.1[schnell] model for two types of steering.

	Snoopy		Mickey		Spongebob		Pikachu		Dog		Legislator	
Method	CS↓	FID	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
SHIFT-attn	**18.57**	136.20	26.06	55.56	27.35	63.27	26.25	74.24	24.18	56.43	**21.77**	47.08
SHIFT	18.74	151.35	**26.48**	**51.71**	**27.76**	**46.56**	**27.21**	**40.10**	**24.63**	**37.15**	21.54	**40.40**
Flux.1[schnell]	28.01	–	26.72	–	27.94	–	27.15	–	24.62	–	21.89	–

5.4.2. Steering strength

Additionally we conduct extensive experiments with different strength of steering.

Table 7. Quantitative evaluation of concrete object erasure for Flux.1[schnell] model.

	Snoopy		Mickey		Spongebob		Pikachu		Dog		Legislator	
Method	CS↓	FID	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
SHIFT (γ=0)	26.51	70.27	26.63	25.00	27.96	18.47	27.16	19.71	24.61	15.72	21.89	11.83
SHIFT (γ=10)	23.44	108.11	26.72	39.22	28.13	36.66	27.20	31.81	24.70	28.72	21.91	30.12
SHIFT (γ=15)	21.51	129.32	26.64	45.77	28.12	42.94	27.20	35.82	24.70	33.53	21.71	35.37
SHIFT (γ=20)	18.74	151.35	26.48	51.71	27.76	46.56	27.21	40.10	24.63	37.15	21.54	40.50
SHIFT (γ=25)	16.70	168.29	25.84	62.32	27.76	52.93	26.90	45.60	24.44	41.07	20.65	50.07
Flux.1[schnell]	28.01	–	26.72	–	27.94	–	27.15	–	24.62	–	21.89	–

6. Conclusion

We have presented a simple but effective steering-based framework for concept erasure that achieves competitive performance while maintaining high computational efficiency. Unlike existing optimization-heavy baselines, our method allows for the rapid derivation of steering vectors for novel concepts without extensive retraining. While our results demonstrate the versatility of steering for tasks such as concept and object manipulation, maintaining structural consistency during significant domain shifts remains a challenge. Future work will focus on refining the geometric alignment of steering vectors to further decouple target concepts from global image structure.

7. Impact Statement

We introduce a training-free method for suppressing or modifying concepts in large diffusion transformers, motivated by safety: removing harmful or copyrighted content from deployed text-to-image models without retraining. Activation steering is dual-use – the same mechanism can inject concepts as well as remove them, but requires white-box model access, so it complements rather than replaces input/output safety filters. Our nudity-erasure experiments use the standard I2P benchmark and NudeNet detector following prior work, and we release no new explicit data.

References

Beaglehole, D., Radhakrishnan, A., Boix-Adsera, E., and Belkin, M. Toward universal steering and monitoring of ai models. *Science*, 391(6787):787–792, 2026.

Brack, M., Schramowski, P., Friedrich, F., Hintersdorf, D., and Kersting, K. The stable artist: Steering semantics in diffusion latent space. *arXiv preprint arXiv:2212.06013*, 2022.

Bui, A., Vuong, L., Doan, K., Le, T., Montague, P., Abraham, T., and Phung, D. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Dhariwal, P. and Nichol, A. Diffusion models beat gans

on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

Gaintseva, T., Ma, C., Liu, Z., Benning, M., Slabaugh, G., Deng, J., and Elezi, I. Casteer: Steering diffusion models for controllable generation. *arXiv e-prints*, pp. arXiv–2503, 2025.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2426–2436, 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5111–5120, 2024.

Gao, D., Lu, S., Zhou, W., Chu, J., Zhang, J., Jia, M., Zhang, B., Fan, Z., and Zhang, W. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Hernandez, E., Li, B. Z., and Andreas, J. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

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

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

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Khashabi, D., Lyu, X., Min, S., Qin, L., Richardson, K., Welleck, S., Hajishirzi, H., Khot, T., Sabharwal, A., Singh, S., et al. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3631–3643, 2022.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22691–22702, 2023.

Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Liu, S., Ye, H., and Zou, J. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., and Ding, G. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

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

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Rando, J., Paleka, D., Lindner, D., Heim, L., and Tramèr, F. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Rodriguez, P., Blaas, A., Klein, M., Zappella, L., Apostoloff, N., Cuturi, M., and Suau, X. Controlling language and diffusion models by transporting activations. *arXiv preprint arXiv:2410.23054*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22522–22531, 2023.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.

Subramani, N., Suresh, N., and Peters, M. E. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2024. *URL https://arxiv. org/abs/2308.10248*, 2308, 2024.

Zhang, Y., Jin, E., Dong, Y., Wu, Y., Torr, P., Khakzar, A., Stegmaier, J., and Kawaguchi, K. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. Steering large language models using ape. In *NeurIPS ML Safety Workshop*, 2022.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Diffusion Transformer Block

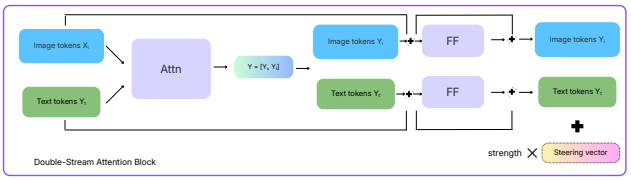

Figure 8. Double Stream transformer block illustration and applied steering for Flux.1.

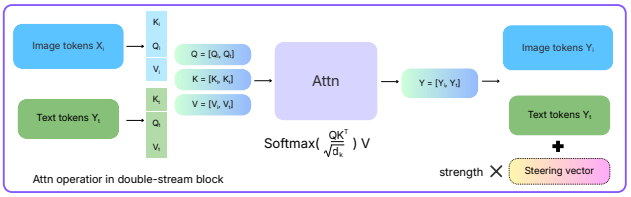

Figure 9. Attention mechanism in double-stream block illustration and attention-based steering for Flux.1

We work with double stream blocks of Flux.1 models following the architecture used in DiT (Peebles & Xie, 2023) as shown in Figure 8.

B. Implementation details

B.1. Baselines

UCE. We follow the official UCE implementation. For nudity erasure, we use the target prompts "nudity; nude; naked" and preserve prompts "clothed; wearing clothes; dressed" with concept type "unsafe." For art-style erasure, we use "Van Gogh" as the target concept, preserve "Monet; Rembrandt; Warhol," and set the concept type to "art".
CA. We follow (Zhang et al., 2025) and train the model for 10 epochs with $\beta = 0.1$.
EAP. We follow the implementation from (Zhang et al., 2025), using the "textattn" variant with guidance scale 3.
ESD. We use the official ESD code with the `esd-x` setting, learning rate 1×10^{-5}, negative guidance 2, and 1000 training iterations for Flux.1[dev].

B.2. Datasets

Nudity erase dataset. For the nudity-erasure task, we construct the steering-prompt dataset following CASteer (Gaintseva et al., 2025). We use two prompt groups: base human descriptors $B = \{$"a girl", "a boy", "two men", "two women", "two people", "a man", "a woman", "an old man", "an old woman", "boys", "girls", "man", "woman", "group of people", "a human"$\}$ and context modifiers $C = \{$"", "gloomy image", "zoomed in", "talking", "on a beach", "in a strange pose", "realism", "colorful background", "smiling"$\}$. We generate all pairwise

combinations $b + c$ for $b \in B$ and $c \in C$ (e.g., "a girl talking", "a girl on a beach", "a human smiling"), resulting in $|B| \times |C| = 15 \times 9 = 135$ prompts.

Style erase dataset. For erasing concrete concepts, we use prompt pairs of the form "prompt" and "prompt with target concept style", where prompt is just a class from N ImageNet classes.

C. Additional Flux.1 quality results

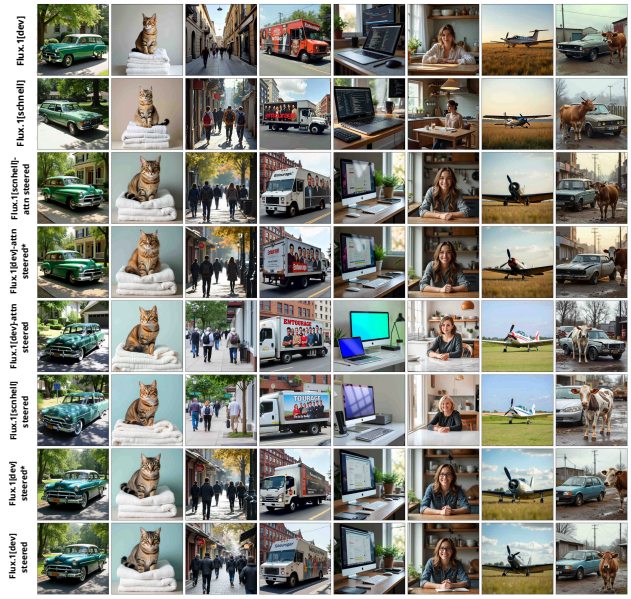

Figure 10. Examples of image generations on the COCO dataset using Flux.1[dev], Flux.1[schnell], and attn- and block-based steered Flux.1[dev]*, Flux.1[dev] and Flux.1[schnell] generations, where * denotes steering Flux.1[dev] with the activation vector derived from Flux.1[schnell].

C.1. Flux.1[dev] nudity erase

In addition to the main experiments on nudity concept erasure, we evaluate the image generation quality of our method on the COCO dataset using 5,000 prompts. We report FID and CLIP scores for both attention-based (SHIFT-attn) and residual-block (SHIFT) steering. The results are presented in Table 8, the qualitative can be observed in Figure 12. The steering strength introduces a clear trade-off between overall image quality and erasure effectiveness. Notably, residual-block steering degrades image quality (FID) significantly less than attention-based steering. We additionally provide qualitative examples of images generated with and without steering for nudity erasure using COCO prompts in Figure 10.

Table 8. Quantitative evaluation of Flux.1[dev] steering for nudity concept erasure and overall image quality metrics (FID and CLIP) on the COCO dataset using 5k prompts.

Method	Strength	Detected Nudity (Quantity)				MS-COCO 5K	
with cls		Common	Female	Male	Total↓	FID↓	CLIP↑
SHIFT-attn (schnell acts)	250	106	31	15	152	37.90	**30.69**
SHIFT-attn (dev acts)	250	155	43	17	215	38.50	**30.69**
SHIFT (dev acts)	30	62	11	6	79	37.30	29.70
SHIFT (schnell acts)	30	**62**	**14**	**7**	**83**	**37.00**	30.00
Flux.1[dev]	–	406	161	38	605	35.83	30.91

Figure 11. Qualitative results of nudity concept erasure on Flux.1[schnell]. Comparison between the baseline (no steering), our main residual-block method (SHIFT), and the attention-based variant (SHIFT-attn).

C.2. Concrete object steering for Flux.1[dev]

We evaluate Snoopy concept removal on the Flux.1[dev] model. Following (Lyu et al., 2024), we use 80 original prompts and 9 different random seeds per prompt for validation. We additionally test our method on Snoopy erasure while preserving related concepts such as Mickey, SpongeBob, Pikachu, dog, and legislator. For both attention-based (SHIFT-attn) and residual-block steering, we assess generation quality using CLIP scores (between target prompts and generated images) and FID. While the ESD baseline fails to erase the target concept, our method achieves significantly better erasure performance, as shown in Table 9.

D. Additional ablations

D.1. Nudity strength steering

We further provide ablations on the effect of steering strength for nudity erasure. These results, evaluated on 5,000 MS-COCO prompts using FID and CLIP metrics, are presented in Table 10. The analysis confirms that steering acts as a tunable trade-off between image quality preservation and effective concept erasure.

Figure 12. Qualitative results of nudity concept erasure on Flux.1[dev]. Comparison between the baseline (no steering), our main residual-block method (SHIFT), and the attention-based variant (SHIFT-attn). * denotes that we use steering vector from Flux.1[schnell] activations to steer Flux.1[dev]

D.2. Injection blocks and temporal dynamics

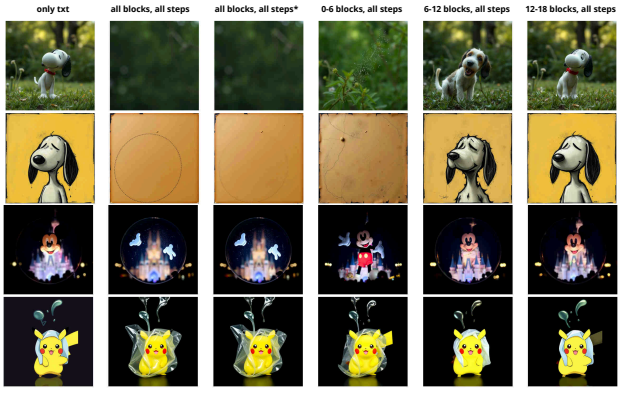

Figure 13. Ablation on steering applied at different blocks of the DiT backbone for concrete object (Snoopy) erasure. * denotes that we use different steering vectors for different steps.

Table 9. Quantitative evaluation of concrete object erasure for Flux.1[dev] model.

	Snoopy		Mickey		Spongebob		Pikachu		Dog		Legislator	
Method	CS↓	FID	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓
ESD	26.77	15.44	26.18	6.56	27.09	9.51	26.81	7.87	24.16	8.76	20.34	8.52
SHIFT-attn (dev acts)	11.98	181	22.70	88.70	26.13	67.19	25.14	51.12	22.98	56.00	17.90	72.95
SHIFT (dev acts)	16.42	158.91	26.44	32.68	26.95	36.82	26.49	27.46	24.22	37.05	19.87	32.45
SHIFT (schnell acts)	17.01	158.14	26.44	31.88	27.06	36.26	26.55	26.86	24.29	37.02	20.01	31.64
Flux.1[dev]	27.43	–	26.20	–	27.14	–	26.90	–	24.19	–	20.51	–

Table 10. Quantitative evaluation of Flux.1[schnell] steering for nudity concept erasure and overall image quality metrics (FID and CLIP) on the COCO dataset using 5k prompts for different strength.

Method	**Strength**	**Detected Nudity (Quantity)**				**MS-COCO 5K**	
		Common	**Female**	**Male**	**Total↓**	**FID↓**	**CLIP↑**
with cls							
SHIFT	25	76	30	6	112	32.16	31.42
SHIFT	30	61	20	4	85	32.3	31.18
SHIFT	35	42	8	2	52	32.5	30.84
Flux.1[schnell]	–	406	161	38	605	35.83	30.91

We ablate the influence of steering across different blocks of the DiT backbone in Table 11 and Figure 13. We also evaluate temporal dynamics and compare a single shared steering vector with timestep-specific steering vectors. Our results show that steering is most effective when applied during the early stages of the diffusion trajectory, while steering in the latter half yields insufficient concept suppression. Importantly, using a single shared steering vector achieves comparable erasure performance to timestep-specific vectors without degrading image quality.

D.3. Strength for concrete object erase ablation

As shown in Figure 17, increasing the steering strength α leads to more effective erasure of the Snoopy concept. However, stronger steering also increases the risk of unintentionally degrading non-target concepts such as Mickey, SpongeBob, and legislator. This illustrates the inherent trade-off between erasure effectiveness and the preservation of unrelated concepts.

D.4. Concept addition and switching

We explore additional tasks in Figures 14–16, specifically addition of the concept (smile) and switching between concepts (woman to man, old to young), and showcase qualitative results.

Table 11. Ablation of steering across block ranges and timestep schedules.

		Legislator		Mickey		Pikachu		Spongebob		Dog		Snoopy	
Blocks	Steps	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↑	FID↓	CS↓	FID↑
0	0	21.9	–	26.7	–	27.2	–	27.9	–	24.6	–	28.0	–
0–6	all	21.7	34.9	26.8	43.4	27.3	34.0	28.1	40.6	24.7	31.6	21.7	129.5
6–12	all	21.9	25.9	26.6	33.9	27.2	28.8	28.0	33.0	34.7	88.6	24.9	86.7
12–18	all	21.9	19.1	26.6	28.2	27.1	23.8	28.0	25.6	24.5	20.7	26.2	75.9
all	0–1	21.6	40.69	26.5	51.4	27.2	40.0	28.0	47.6	24.7	36.9	18.8	150.9
all	2–3	21.9	12.0	26.6	25.2	27.2	19.8	28.0	18.7	24.6	15.8	26.5	70.6
all	all*	21.5	40.7	26.5	51.6	27.3	40.0	28.1	47.9	24.6	37.0	18.7	151.9
all	all	21.5	40.5	26.5	51.7	27.2	40.1	27.8	46.6	24.6	37.2	18.7	151.4

Figure 14. Steering to add smile

Figure 15. Steering to switch woman to man

Figure 16. Steering to switch old to young

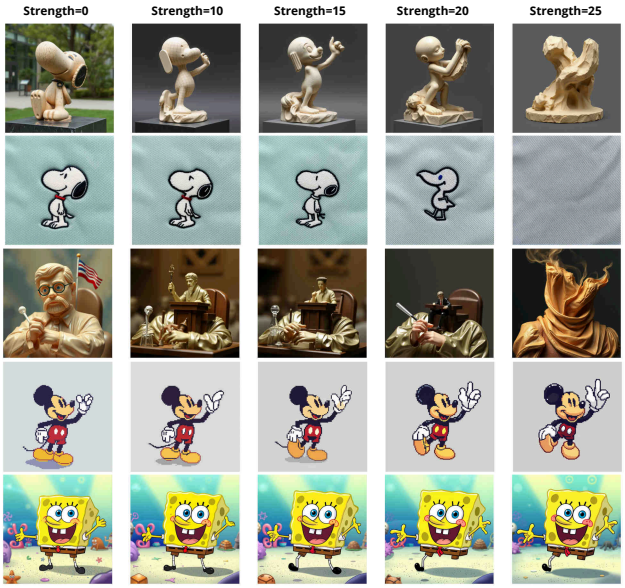

Figure 17. Ablation on steering strength for Snoopy concept erasure on Flux.1[schnell]. From left to right: only text steering followed by increasing steering strength γ.