SHIFT: Steering Hidden Intermediates in Flow Transformers

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

Introduction

Modern DiT-based diffusion models create highly prompt-faithful images, but they can also generate harmful or copyrighted content, making concept removal a critical safety challenge. Existing methods often rely on gradient-based fine-tuning, which becomes prohibitively expensive for 12B-parameter DiTs and often fails on guidance-distilled models.

We propose a radically different approach, inspired by activation steering in LLMs. Instead of retraining the model, SHIFT simply nudges the internal activations along a learned direction at inference time. **No weight updates. No extra optimization.**

SHIFT — a lightweight, training-free steering framework tailored for Multimodal Diffusion Transformers.

Where to steer

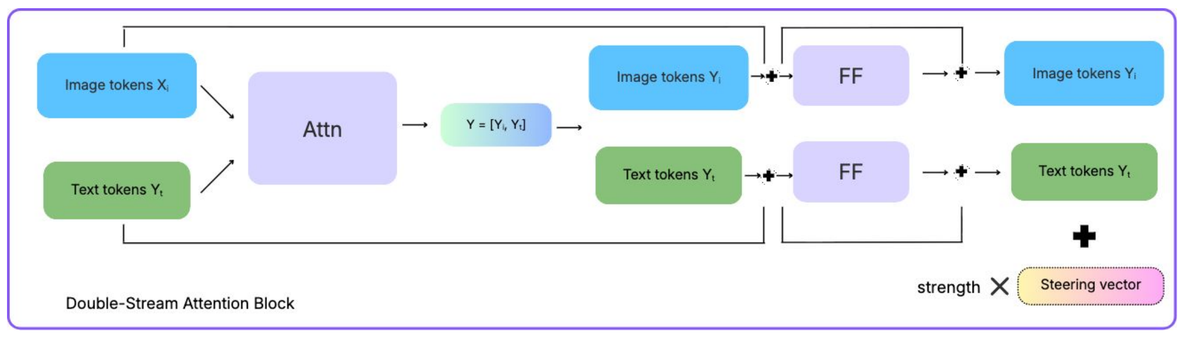

Fig.2. Double Stream transformer block illustration and applied steering for Flux.1.

In double-stream blocks, text and image tokens interact through joint attention. Residual-only steering leaves a source-state mismatch, while post-block steering acts directly on the propagated text representation. We steer the post-block text state, with pooled steering as a complementary global route.

$$X_t^{\ell+1} = X_t^{\ell} + R_{t,\ell}(X_t^{\ell}, X_i^{\ell}; e_\tau)$$
$$X_i^{\ell+1} = X_i^{\ell} + R_{i,\ell}(X_i^{\ell}, X_t^{\ell}; e_\tau)$$

Learning a concept direction

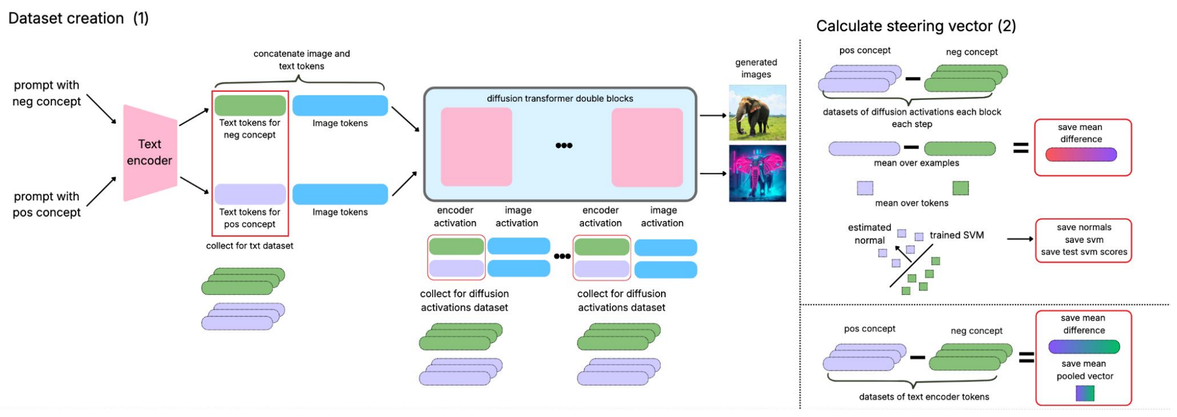

Fig.1. Overview of the steering pipeline: (1) dataset construction from contrastive prompt pairs, (2) steering vector computation based on mean difference and separation plane.

From n contrastive prompt pairs (neutral prompt vs the same prompt with the target concept) — we extract activations and compute their normalized mean difference.

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

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

Inference time steering

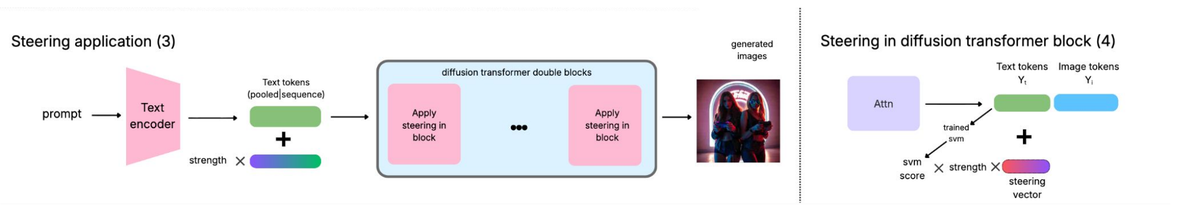

Fig.3. (3) Application of the vector during inference, and (4) Steering inside the diffusion.

The SVM confidence score estimates whether the target concept remains present in the current activations. We additionally steer the pooled CLIP embedding using cosine similarity to the target concept.

$$\tilde{X} = X + \alpha\, v,$$
$$\alpha_{\text{diff}} = \gamma\, \eta_{\text{cls}}$$
$$\eta_{\text{cls}} = \text{clip}\left(\frac{1}{(1-p_{\text{cls}})+\varepsilon} - 1, 0, \eta_{\max}\right)$$

Results

Nudity erasure

Fig. . Qualitative Comparison with Baselines: the first row – original generation of Flux.1[schnell], from top to bottom – UCE, CA and SHIFT.

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

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.

Snoopy erasure

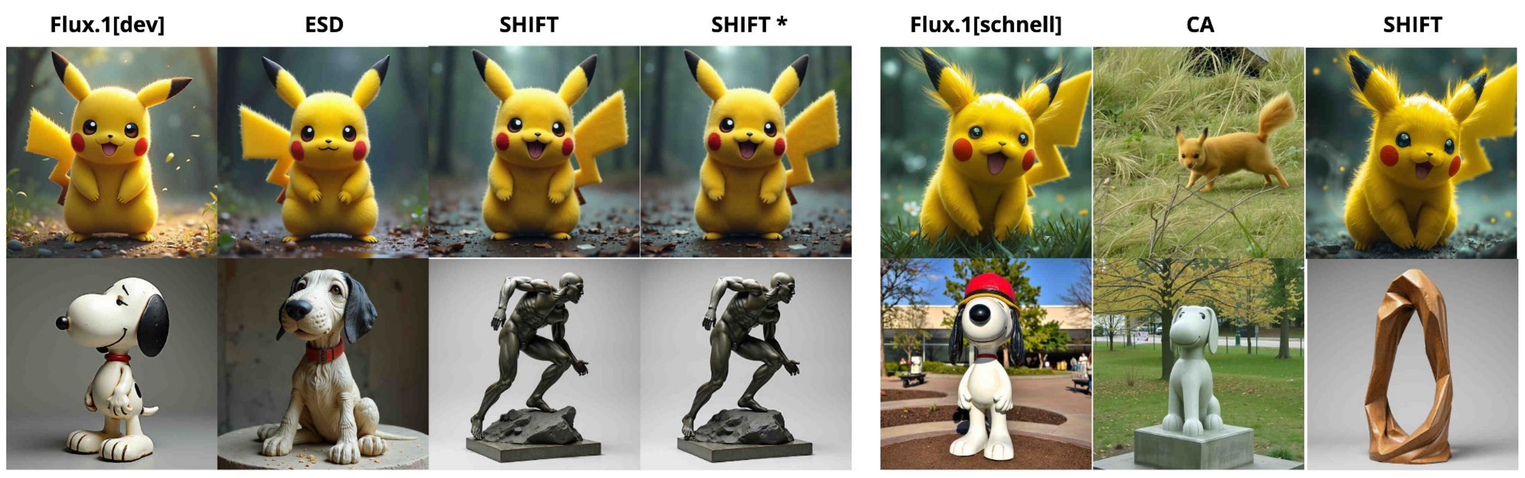

Fig.. Qualitative Comparison with Baselines for Flux.1[dev] (left) and Flux.1[schnell] (right). * denotes that we use steering vector from Flux.1[schnell] activations to steer Flux.1[dev]

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	–

Table 2. Quantitative evaluation of concrete object erasure for Flux.1[schnell] model.

Remove glasses

Old → Young

Fig.. 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]

Limitation: 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.

1. FusionBrain Lab 2. HSE University 3. Lomonosov Moscow State University

Mechanistic Interpretability Workshop, ICML 2026

FUSIONBRAIN LAB