ACFLOW: CONTROLLING TEXT-TO-IMAGE DIFFUSION TRANSFORMERS VIA LEARNED CONDITIONAL ACTIVATION FLOW

Junran Wang[1*], **Zehao Jin**[1*], **Tianyu Luan**[2], **Xinjie Shen**[1]
[1]Georgia Institute of technology
[2]South China University of Technology
{nove1yst, zehao, xinjie}@gatech.edu, luanty.henry@gmail.com
*Equal contribution

arXiv:2609.10723v1 [cs.CV] 9 Sep 2026

ABSTRACT

Text-to-image diffusion transformers (DiTs) are powerful generators, yet direct prompting provides limited control interface for style intensity and can fail to suppress unwanted concepts. To enable these controls, we introduce ACFLOW, an inference-time controller that transports intermediate layer image-token activations through a learned concept-conditioned velocity field while keeping the base DiT frozen. A textual concept description specifies the desired intervention, while the integration horizon provides a continuous control parameter. The field produces token-varying, activation-dependent updates. With parameters shared across concepts within each task family, the field supports fine-grained descriptions and generalizes to concepts unseen during training without per-concept fitting. On style control, ACFLOW achieves the best style–content trade-off among the evaluated baselines in the high-style-alignment regime. At a fixed operating point, ACFLOW attains style–content alignment of 0.5365/0.2860, compared with 0.4397/0.2684 for the baseline with the highest style alignment. Qualitative results demonstrate suppression of diverse concepts, including cases where direct prompting fails. Our analyses support the learned velocity field as an adaptive control mechanism, with update directions varying across tokens and depend on their activation states. Our code is available at https://github.com/Nove1yst/AcFlow.

1 INTRODUCTION

Text-to-image (T2I) diffusion models have become powerful tools for visual creation, yet their generation capabilities do not translate into precise user control. A model may readily generate a stylized image without allowing users to smoothly adjust the degree of stylization, or depict a concept without excluding it even when requested. Verbal modifier such as 'weak' or 'strong' do not provide a calibrated control axis, producing similarly pronounce level of stylization; appending instructions such as 'no lighthouse' can leave the unwanted concept visibly present. Such control matters when users want to explore different visual treatments of the same content or suppress unwanted concepts from appearing, but directly prompting the generative model falls short in these cases.

Existing control mechanisms offers limited flexibility for these requirements. Parameter-level adaptation can provide a control interface, but methods based on concept-specific optimization require additional fitting for each new concept (Gandikota et al., 2024a), affect non-target concepts (Lyu et al., 2024) or risk degradations of previously learned control capabilities during sequential modification (Smith et al., 2023). This motivates an inference-time controller that shares parameters across concepts, acts upon a user-specified concept with an explicit parameter for controlling the intervention strength, while the base generator parameters keep frozen.

To this end, we introduce ACFLOW, which controls frozen DiTs by transporting intermediate layer activations through a learned concept-conditioned velocity field at inference time. A textual concept description specifies the desired intervention, and the integration horizon T provides a continuous

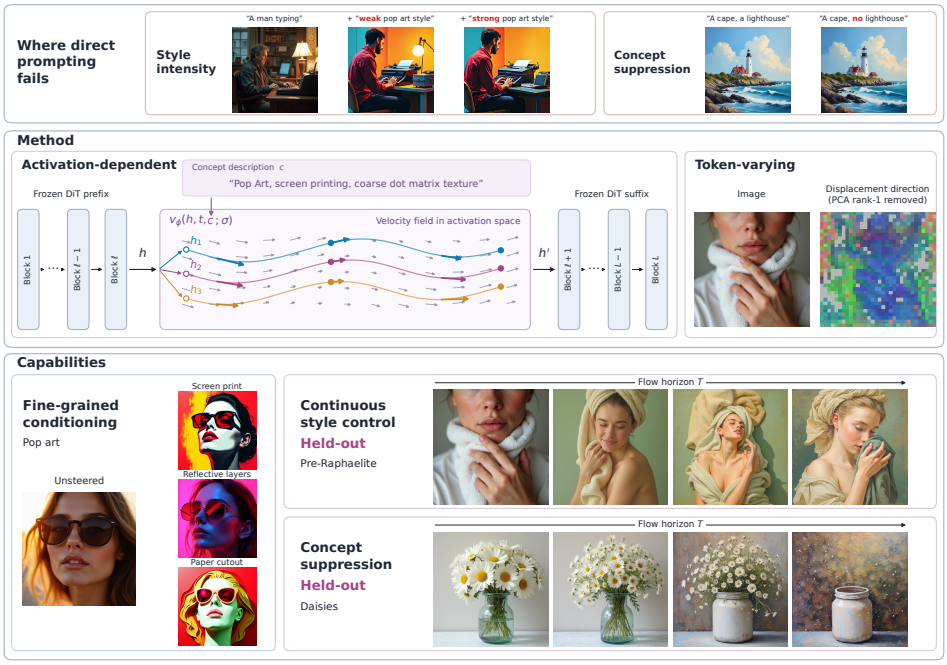


Figure 1: **Overview of AcFlow. Upper: Method.** At one site in a frozen DiT, image-token states move through a velocity field conditioned on a concept description. Velocities vary with activation state across tokens. The displacement-direction map removes the rank-one PCA component: each cell is an image token, with color encoding its direction. **Lower: Capabilities.** One field supports fine-grained descriptions (left) and continuous control of styles. A separately trained suppression field removes the daisies concept. Held-out means both concepts are unseen during training.

control parameter for the extent of intervention, with $T = 0$ recovering the unmodified parameter. At a single intervention site, image-token activations moves through the field with velocities that adapts to their individual states and the concept condition, yielding token vaying activation-dependent updates. We train the field through velocity distillation, sharing its parameters across concepts within each task family. The resulting field supports fine-grained concept descriptions and generalizes to unseen concepts during training without concept-specific fitting.

Our main contributions are:

- **Control as a flow field.** We formulate inference time control as transport through a learned velocity field in activation space. Image-token activations move through the field, receiving token-varying, activation-dependent updates.
- **Concept-shared control and generalization.** We condition the field on textual concept descriptions, training one field across concepts within a task family without per-concept fitting. The shared field generalizes to unseen concepts during training and responds to fine-grained descriptions beyond coarse concept labels.
- **Continuous style control and concept suppression.** AcFlow enables progressive style modulation with the flow horizon as a continuous control parameter, achieving the best style–content trade-off among the evaluated methods in the high-style-alignment regime. Qualitative concept suppression results further demonstrate the applicability where direct prompting leaves unwanted concepts present.
- **Probing the learned field.** Controlled experiments demonstrate the value of token-specific updates, while trajectory analysis reveals that the field adapts its direction and energy allocation systematically through denoising steps. These findings support the value of an adaptive activation transport field beyond a fixed concept direction broadcast across tokens.

2 Related Works

Continuous concept control. Attribute Control (Baumann et al., 2025) identifies concept-specific directions in text embeddings through optimization, enabling continuous modulation. Concept Sliders (Gandikota et al., 2024a) learns concept-specific low-rank adapters (Hu et al., 2021) in the diffusion backbone and continuously adjusts their influence through an inference-time scale. Text Slider (Chiu et al., 2026) learns such adapters in the text encoder, enabling continuous attribute control. All approaches fit adapters for individual target concepts. AcFlow instead trains a shared, text-conditioned field within each task family and uses the flow horizon to adjust intervention strength, including for concepts unseen during training.

Activation intervention for T2Is. Recent methods have proposed to control generation by modifying a model's intermediate representations at inference time. ICM (Zaleska et al., 2026) uses linear probes to identify fixed attribute directions, while Semantic Steering (Li et al., 2026) constructs a fixed target-to-safe direction. CASteer (Gaintseva et al., 2026) and SHIFT (Konovalova et al., 2026) retain concept-specific directions but dynamically vary their magnitudes. DSAS (Ferrando et al., 2025) learns activation-dependent scalar gates for a prescribed steering transformation. PID Steering (Nguyen et al., 2026) uses feedback to adjust the magnitude of a mean direction. In a parallel branch, AcT (Rodriguez et al., 2025) estimates coordinate-wise affine maps between source and target activations. SteeringDiffusion (Wu & Summa, 2026) predicts condition-dependent affine modulation for multiple styles. Dynamic scaling and affine maps extend fixed vectors while constraining updates to a prescribed operation family.

Nonlinear activation intervention for T2Is. Nonlinear intervention allows the update direction to vary with the activation state. Conditioned Activation Transport (CAT) (Chrabaszcz et al., 2026) learns a geometry-gated nonlinear displacement, but computes it from a mean-pooled activation and broadcasts the same update across tokens. Flow matching (Lipman et al., 2022) provides a continuous velocity-field formulation of transport; AcFlow uses a learned concept-conditioned field to transport image-token activations.

3 Method

Table 1: **Control interfaces and update geometry of T2I activation intervention.** ✓ denotes a supported property; × means not established in the cited T2I formulation. Concept conditioning excludes concept-specific fitting; zero-shot concerns unseen target concepts during training. Activation-dependent direction means the displacement axis can vary with h at a fixed concept, site, and denoising step, beyond scalar scaling.

Method	Update family	Activation-dependent direction	Token-varying direction	Concept-conditioned controller	Zero-shot concepts
Linear-AcT (Rodriguez et al., 2025)	Affine map	✓	✓	×	×
DSAS (CAA) (Ferrando et al., 2025)	Gated direction	×	×	×	×
CAT (Chrabaszcz et al., 2026)	Nonlinear map	✓	×	×	×
SteeringDiffusion (Wu & Summa, 2026)	Affine modulation	✓	✓	✓	×
CASteer (Gaintseva et al., 2026)	Scaled direction	×	×	×	×
ICM (Zaleska et al., 2026)	Fixed direction	×	×	×	×
SHIFT (Konovalova et al., 2026)	Scaled direction	×	✓	×	×
Semantic Steering (Li et al., 2026)	Fixed direction	×	×	×	×
AcFlow	Velocity field	✓	✓	✓	✓

3.1 Control as a Flow Field

Let G_θ denote the frozen denoising velocity predictor of a DiT-based (Peebles & Xie, 2023) T2I model with hidden dimension d. For one sample, the output of a selected single-stream block b is a residual-stream state $h \in \mathbb{R}^{S\times d}$; we omit the batch dimension throughout. Here $S = S_{\text{txt}} + S_{\text{img}}$ counts text and image tokens, and $h_i \in \mathbb{R}^d$ denotes token i. Given a source prompt p (written p_{src} in training) and a concept description c, we seek an edited state h' that induces the requested concept change while preserving non-target content. For example, for style control, p is a content prompt and c describes the style of the image (e.g., in pop art style, use reflective layers).

At a fixed denoising noise level $\sigma \in [0, 1]$, we define an activation flow field $\varphi_t(\cdot; c, \sigma) : \mathbb{R}^{S\times d} \to \mathbb{R}^{S\times d}$, parameterized by the field weights ϕ. For brevity, we write $\varphi_t(h)$ with c, σ, and ϕ implicit.

The flow time $t \in [0, T]$ is distinct from σ: c and σ remain fixed during each activation-flow integration, which restarts from the current block output at every denoising step.

The velocity field $v_\phi(h, t, c; \sigma) \in \mathbb{R}^{S \times d}$ generates the flow through

$$\frac{d}{dt}\varphi_t(h) = v_\phi(\varphi_t(h), t, c; \sigma), \qquad \varphi_0(h) = h. \tag{1}$$

The steered activation at strength T is

$$h' = \varphi_T(h) = h + \int_0^T v_\phi(\varphi_t(h), t, c; \sigma)\, dt. \tag{2}$$

In practice, we use N forward Euler steps:

$$h^{(k+1)} = h^{(k)} + \frac{T}{N} v_\phi\left(h^{(k)}, \frac{kT}{N}, c; \sigma\right), \quad k = 0, \ldots, N-1, \tag{3}$$

Here N is the number of integration steps. Starting from $h^{(0)} = h$, we write the numerical endpoint $h^{(N)} \approx \varphi_T(h)$ back into the residual stream. The horizon T controls intervention strength; $T = 0$ leaves h unchanged. Several simpler methods arise by restricting $v_\phi(h, t, c)$:

- $v_\phi(h, t, c) = g(\bar{h}, t, c)$ where $\bar{h}$ is the mean-pooled token applies a shared velocity to all tokens;
- $v_\phi(h, t, c) = a(h, t, c)d(c)$ where a is a scalar fixes a concept-specific direction with scaling;
- $v_\phi(h, t, c) = d(c)$ reduces the flow to $\varphi_T(h)_i = h_i + Td(c)$, the standard additive intervention.

The full field relaxes both the shared-token update and fixed concept direction restrictions: tokens move through activation space with velocities that can change in direction as well as magnitude.

3.2 Intervening the image tokens

Write the single-stream state as $h = [h^{\text{txt}}; h^{\text{img}}]$, where $h^{\text{txt}} \in \mathbb{R}^{S_{\text{txt}} \times d}$, $h^{\text{img}} \in \mathbb{R}^{S_{\text{img}} \times d}$, and $[\cdot;\cdot]$ denotes concatenation along the token dimension. Although DiTs typically mix the two token groups in their single-stream blocks, their roles at the intervention site are asymmetric. Intervention h^{txt} can affect the output only indirectly through later text–image mixing, whereas h^{img} lies on the residual path consumed by the output head. The groups also originate from different modalities and need not share an activation distribution, so applying one single transport to both introduces an unnecessary shared-geometry assumption. We consequently intervene on only the image state:

$$h' = [h^{\text{txt}}; \varphi_T^{\text{img}}(h^{\text{img}})], \qquad \varphi_t^{\text{img}} : \mathbb{R}^{S_{\text{img}} \times d} \to \mathbb{R}^{S_{\text{img}} \times d}. \tag{4}$$

Below, v_ϕ denotes its image-token component. Figure 2 shows the diagnostic used to check this choice. At every denoising step, we record the output of single-stream block 16 of FLUX.1-dev (Black Forest Labs, 2024) under a target prompt with style description and replace the text or image tokens in a same noise source trajectory under the original prompt without the style description. The diagnostic shows that image-state replacement recovers most of the stylization, while text-state replacement recovers less; quantitative results are in App. D.



Figure 2: **Token-swap:** A target-prompt trajectory supplies block outputs to a source-prompt trajectory with the same initial noise.

3.3 Model Architecture

We instantiate v_ϕ with one backbone-matched Transformer-style FlowBlock. The condition follows the generator's native text path: let $C = E_{\text{txt}}(c) \in \mathbb{R}^{S_c \times d}$ be text-encoder features and $\bar{c} = E_{\text{pool}}(c)$ be a separate pooled text-encoder feature when G_θ consumes pooled feature.

The FlowBlock first extracts concept information through cross-attention. Let $e_{\text{flow}}(t)$ embed the current integration time and $e_c(\bar{c})$ project the pooled condition, with $e_c(\bar{c}) = 0$ when no pooled

condition is used. Following the backbone's adaptive normalization and residual gate,

$$x_t, g_t = \mathrm{AdaNorm}_\phi\big(h^{\mathrm{img}}; e_{\mathrm{flow}}(t) + e_c(\bar{c})\big)\,, \tag{5}$$

$$\tilde{h} = h^{\mathrm{img}} + g_t \odot \mathrm{CrossAttn}_\phi(Q = x_t, K = C, V = C). \tag{6}$$

We preserve the backbone's attention geometry, QK normalization, normalization placement, and gated-residual convention. The sole change to its time modulation is to replace the generator's model time by flow time t, allowing the concept update to vary along the integration trajectory.

The cross-attention stage is followed by one native-style transformer block,

$$h^{\mathrm{out}} = \mathrm{TransformerBlock}_\phi\Big(\tilde{h}; e_{\mathrm{model}}(\sigma) + e_c(\bar{c})\Big)\,, \qquad v_\phi(h^{\mathrm{img}}, t, c; \sigma) = h^{\mathrm{out}} - h^{\mathrm{img}}. \tag{7}$$

3.4 Velocity Distillation

Training samples $(z_0, p_{\mathrm{src}}, p_{\mathrm{tgt}}, c) \sim \mathcal{D}$, where z_0 is the encoded target image and the prompt pair specifies the concept change c. We sample $\epsilon \sim \mathcal{N}(0, I)$ with the same shape as z_0 and $\sigma \sim q_\sigma$ on $[0, 1]$, where q_σ is the training noise-level distribution, and construct

$$z_\sigma = (1 - \sigma)z_0 + \sigma\epsilon. \tag{8}$$

The frozen teacher is the unsteered denoising predictor conditioned on the target prompt:

$$u_{\mathrm{tgt}} = G_\theta(z_\sigma, \sigma, p_{\mathrm{tgt}}). \tag{9}$$

The student is the same frozen model conditioned on the source prompt, but with AcFlow active at block b using condition c:

$$\hat{u} = G_{\theta,\phi}(z_\sigma, \sigma, p_{\mathrm{src}}, c, T). \tag{10}$$

Teacher and student use the same z_σ, σ, and guidance setting. For training, we set $T = 1$ and minimize the mean squared velocity error by optimizing ϕ,

$$\mathcal{L}_{\mathrm{vel}}(\phi) = \mathbb{E}_{\mathcal{D},\epsilon,\sigma}[\mathrm{MSE}(\hat{u}, \mathrm{stopgrad}(u_{\mathrm{tgt}}))]\,. \tag{11}$$

4 Experiments

We test the capability of AcFlow on two control tasks: continuous style control and concept suppression, both not achievable by directly prompting DiTs. Unless stated otherwise, experiments use frozen FLUX.1-dev at 512×512 resolution, 28 denoising steps, guidance 3.5, single-stream block $b = 16$, train with $T = 1$ and $N = 3$. NSFW suppression is in App. B; Z-Image (Cai et al., 2025) portability is in App. A.

4.1 Continuous Style Control

Fine-grained style conditions. MegaStyle (Gao et al., 2026) specifies each image's style through a fine-grained natural-language description. For our experiments, we group these descriptions into *style families* and use this grouping to construct family-disjoint training and evaluation splits. We train one conditional field across the training families. At generation time, the frozen generator receives the content prompt and the field receives the style description. Style alignment is the reference-image cosine similarity from the released MegaStyle encoder; content alignment is the CLIP (Radford et al., 2021) similarity to the content prompt.

Baselines. We compare activation intervention methods: Linear-AcT (Rodriguez et al., 2025) at its original locations and at our image-token site, the corresponding Mean-AcT variants, a diffusion adaptation of Activation Addition (Turner et al., 2024), SHIFT (Konovalova et al., 2026) and parameter adaptation methods: Concept Sliders (Gandikota et al., 2024a) and Text Slider (Chiu et al., 2026). To provide a setting for fair comparison, for each held-in style families, we fix a style description as each baseline requires fitting for each specific style, while AcFlow supports fine-grained style description and is shared across conditions. Implementation details are in App. C.3.

Table 2 compares methods at a fixed operating point. AcFlow is the strongest in terms of both style and content alignment among the displayed methods in the high-style-alignment region at

Method	Held-in		Held-out	
	Style ↑	Content ↑	Style ↑	Content ↑
Styled prompt	0.548	0.294	0.511	0.287
Parameter Adaptation				
Concept Sliders	0.511	0.279	–	–
Text Slider	0.157	0.306	–	–
Activation Intervention				
Linear-AcT	0.271	0.202	–	–
Linear-AcT (b16)	0.313	0.176	–	–
Mean-AcT	0.440	0.268	–	–
Mean-AcT (b16)	0.298	0.277	–	–
ActAdd (b16)	0.300	0.278	–	–
SHIFT	0.221	0.279	–	–
AcFlow	**0.537**	**0.286**	**0.442**	**0.281**

Table 2: **Full results on style control. Held-out:** Unseen families during training. AcFlow is evaluated at $T = 2$ on the same field; baselines use their maximum-style operating points.

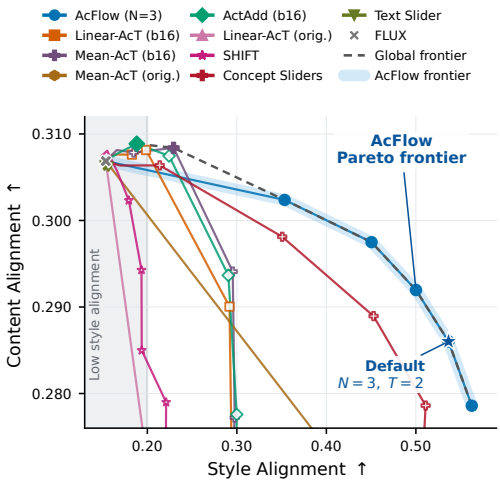


Figure 3: **Style–content trade-offs.** AcFlow is on the global frontier in high style-alignment region. The complete range and uncertainty intervals appear in App. G.

their reported operating points, reaching 0.5365 and 0.2860, respectively, compared with 0.4397 and 0.2684 for the strongest activation intervention method by style, Mean-AcT at its original sites, and the parameter adaptation baseline, Concept Sliders, which reaches 0.5109 and 0.2786. Figure 3 complements these operating points by comparing the full style–content trade-off. Among the evaluated methods, AcFlow offers the strongest style–content trade-off in the high-style-alignment regime.

Zero-shot generalization to unseen style families. To testing transfer beyond new descriptions of familiar styles, we exclude entire style families from training. On all unseen families, AcFlow reaches 0.442 style alignment and 0.281 content alignment at $T = 2$. Qualitative results in Fig. 5 shows that AcFlow is capable of inducing visible stylization on held-out families. Fig. 5 further shows a strength-dependent response on these families: the shared field transfers both the ability to induce an unseen style and the ability to adjust its influence.

Continuous style–content control. For a fixed style description, T provides an inference-time coordinate along the style–content trade-off. Figure 5 shows increasing style alignment over the responsive range on both held-in and held-out families, accompanied by a gradual reduction in content alignment. The fixed-noise sequences illustrate the corresponding gradual changes in style.



Figure 4: **Fine-grained conditioning.** Images share the same content prompt, style family and $T = 2$ but style descriptions vary: screen-print halftones, glossy reflective photography, grayscale linear illustration, and layered paper cutout.

Strength extrapolation beyond the training horizon. Although the field is trained at $T = 1$, extending the horizon further strengthens stylization on both seen and unseen families. This extends the useful control range beyond the training setting. The style gains eventually saturate while content alignment continues to decline, so larger horizons are not uniformly preferable.

4.2 Qualitative Results on Concept Suppression

We test the same formulation on concept suppression by training a separate field on removal pairs mined from InstructPix2Pix (Brooks et al., 2023), covering over 1,000 distinct concepts. The source prompt contains the unwanted concept, the teacher uses the edited prompt, and the field receives `erase <concept>`. Implementation details are in App. C.4.

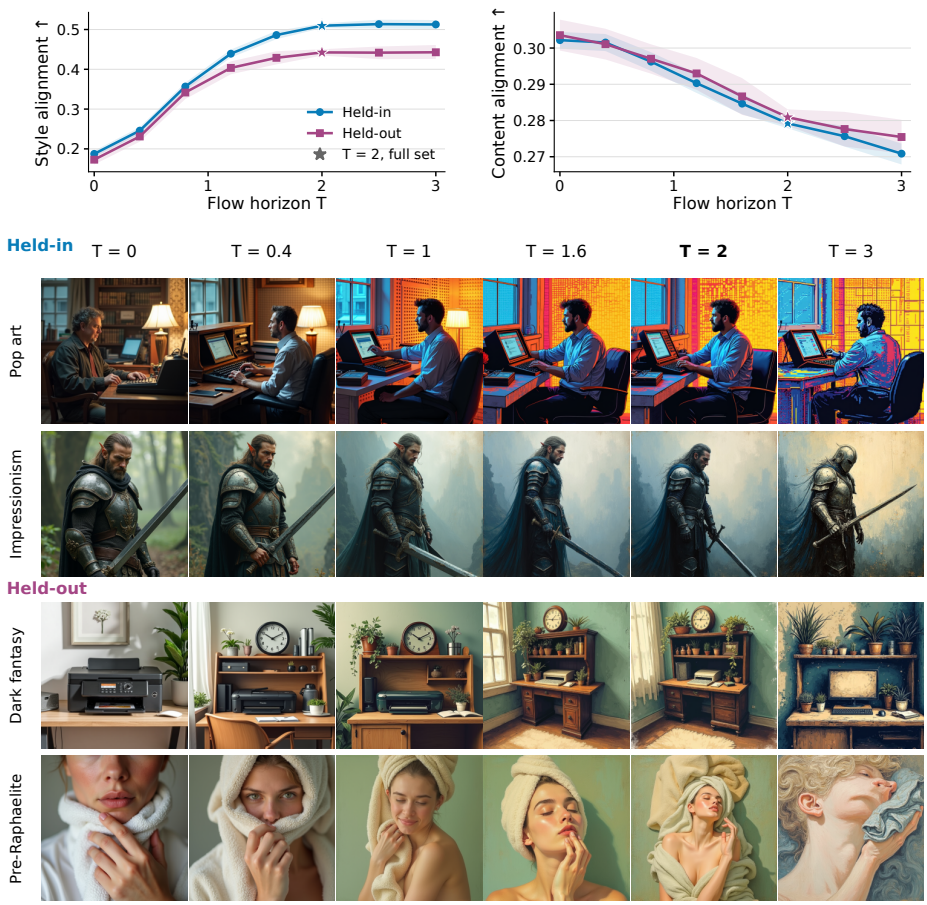


Figure 5: **Continuous style control and generalization beyond training conditions.** Top: ACFLOW's strength response. $T = 2$ use full test sets; others use fixed quarter subsets. Bands show 95% within-family bootstrap intervals. Bottom: illustrative sequences with fixed content and style prompts within each row. Increasing T strengthens stylization, while large horizons can also alter composition. Complete sequences appear in App. G.

Figure 6 shows three examples per split. The field clears fog and suppresses a lighthouse or river on held-in concepts; on held-out concepts, it suppresses peonies, a feathered hat, or daisies. Intermediate horizons expose partial suppression: the lighthouse becomes smaller, the river narrows, and prominent peony blooms give way to sparse stems and leaves at larger T. Recognizable scene elements remain, although appearance and composition can also change. These selected sequences illustrate ACFLOW's capability of conditional concept suppression. More results are shown in App. G.

5 ABLATIONS

We ablate the two main design choices: number of Euler steps N and intervention location b on style control. The generator is FLUX.1-dev at 512×512 resolution, 28 denoising steps and guidance 3.5. For each setup, we train a separate field under the same training budget. We retain $T = 1$ as training horizon and evaluate at $T = 2$. For evaluation, we use the original fine-grained style description and content. Table 3 summarizes the results.

Intervention block. ACFLOW maintains strong across blocks 12, 16, and 20, with similar content

		Held-in		Held-out	
b	N	Style ↑	Content ↑	Style ↑	Content ↑
Intervention block					
12	3	0.541	0.279	0.475	0.281
20	3	0.508	0.282	0.464	0.282
Euler steps					
16	1	0.432	0.284	0.375	0.285
16	3	0.510	0.279	0.442	0.281
16	5	0.485	0.280	0.433	0.281
16	10	0.493	0.283	0.439	0.286

Table 3: Style-steering ablations at $T = 2$.

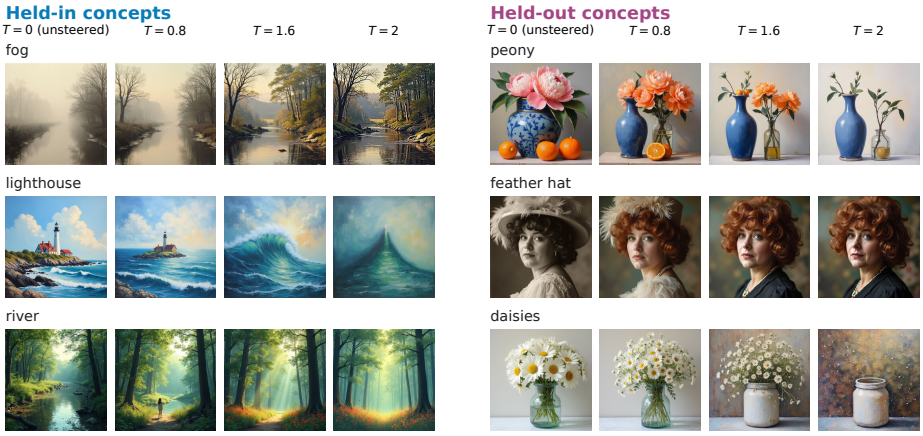


Figure 6: **Qualitative concept suppression results.** Left: three held-in concepts; right: three held-out concepts. Each row fixes the original prompt and initial noise and varies T using the same field. Target suppression can accompany changes in appearance and composition.

scores on both splits. This supports robustness to block selection within the tested range: effective intervention does not require a single specially chosen intervention block. Block 12 gives the highest mean style scores (0.541/0.475 for held-in/held-out), while block 20 remains comparable to block 16 on held-in styles and improves the held-out mean from 0.442 to 0.464. Location therefore provides room for further improvement without being critical to the effectiveness.

Euler integration steps. At block 16, Euler-3 gives the highest mean style scores on both splits. Decreasing the count to one draws back style alignment significantly, while increasing the count to five or ten does not improve style scores monotonically; Euler-10 instead gives the highest content scores among configurations with competitive style alignment. These results support Euler-3 as a compact configuration for strong style control.

6 Characterizing the Flow Field

We examine how the learned field adapts its updates across image tokens and denoising noise levels. All experiments in this section are done with the default $N = 3, b = 16$ field evaulated at $T = 2$.

6.1 Variation Across Image Tokens

For a fixed sample, concept, and denoising step, let $\Delta \in \mathbb{R}^{S_{\text{img}} \times d}$ stack the endpoint displacements. A shared-axis update $\Delta_i = a_i u$ permits signed, token-dependent magnitudes but has rank at most one. With u_1 the leading right-singular vector from uncentered PCA, decompose

$$q_i = \langle \Delta_i, u_1 \rangle u_1, \qquad r_i = \Delta_i - q_i, \qquad \Delta = Q_1 + R. \tag{12}$$

We compare the full field with Q_1 rescaled to $\|\Delta\|_F$ and with R at its original norm, recomputing both the field and axis on each trajectory at every denoising step.

Figure 7 illustrates residual-only effects and spatially coherent residual directions. Across held-in and held-out styles, Q_1 carries most of the displacement energy. Yet matching its norm to the full update does not recover the full-field style gain, while the smaller, original-norm residual retains a substantial effect. These controls support the contribution of directions beyond the dominant shared axis. App. E reports the complete quantitative comparison.

6.2 Variation Across Noise Levels

We next apply the same decomposition to each full-field denoising trajectory and measure the off-axis energy fraction $\rho(\sigma) = \|R(\sigma)\|_F^2 / \|\Delta(\sigma)\|_F^2$. Figure 8 shows that the relative contribution of token-varying directions changes substantially during generation. For held-in/held-out styles, ρ starts at 12.3%/14.3%, reaches measured-grid maxima of 38.3%/41.4% at $\sigma = 0.849/0.772$, and falls to 3.2%/3.8% at $\sigma = 0.065$. The highest-energy 5% of tokens likewise carry a larger energy

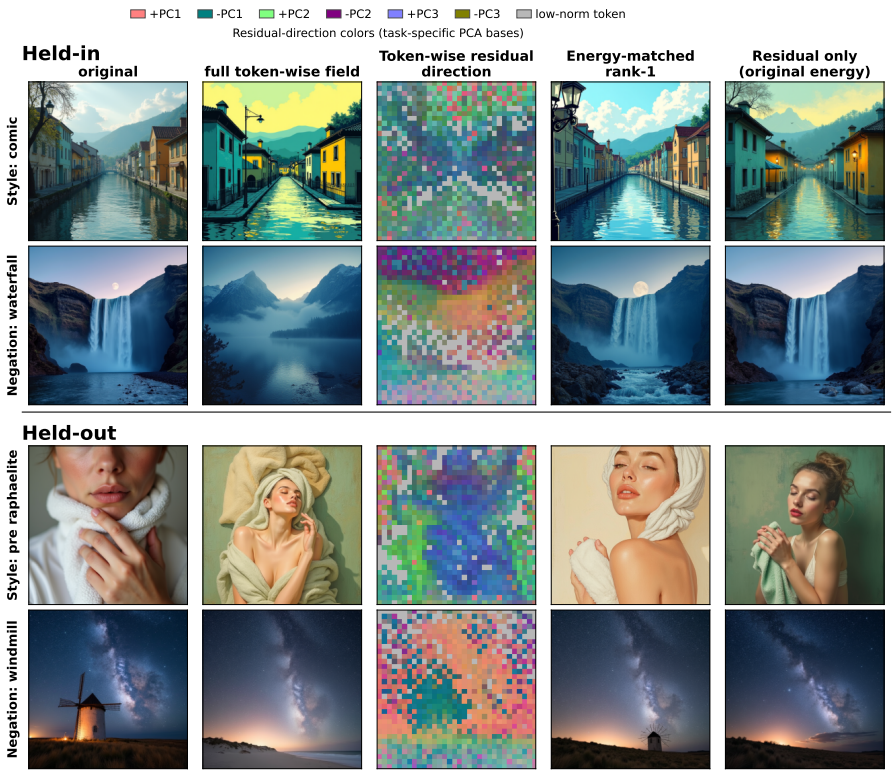


Figure 7: **Token-direction controls.** Selected examples compare full field, norm-matched rank-one update, and original-norm residual. Residual-direction colors use a three-component PCA basis shared within each case.

share in the intermediate part of the trajectory, returning to about 6% near its end. Late updates are therefore more aligned across tokens. Suppression shows a similar pattern (App. E.2).

These results show that the field adapts its update geometry across noise levels along the generation trajectory: both direction structure and the distribution of energy across tokens change systematically. This adaptation goes beyond a uniform rescaling of a fixed displacement. In particular, $\Delta(\sigma) = a(\sigma)\Delta_{\mathrm{ref}}$ would leave both the off-axis energy fraction and the top-token energy share unchanged for nonzero $a(\sigma)$, unlike the observed curves.

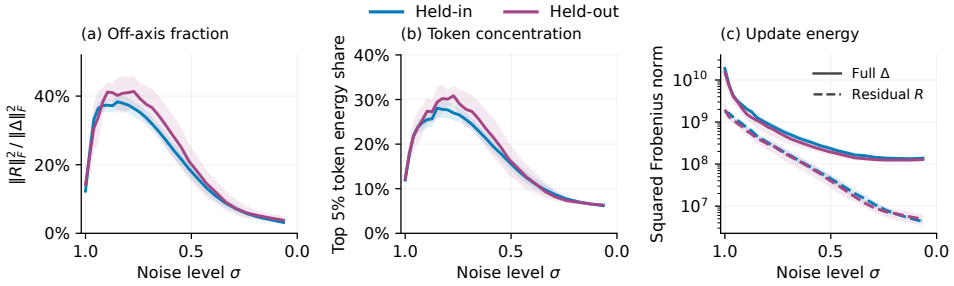


Figure 8: **Displacement field geometry varies along denoising.** Bands are pointwise 95% case-bootstrap intervals. Noise decreases from left to right. Relative off-axis energy and token energy concentration rise then fall, while update energy declines overall.

7 Conclusions and Limitations

We introduced AcFlow, a concept-conditioned velocity field that steers image-token activations at a single site in a frozen DiT. Its token-varying, activation-dependent updates share learned parameters across concepts, enabling fine-grained descriptions and generalization to held-out concepts without per-concept fitting. The integration horizon provides continuous control across operating points. Our experiments demonstrate strong style–content trade-offs and show the applicability to concept suppression. The limitation is that AcFlow does not guarantee preservation of the original

image composition. Intervention can change object positions, poses, and scene layout alongside the intended concept. Specific concept control remains an open challenge.

REFERENCES

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Melvin Sevi, Vincent Tao Hu, and Björn Ommer. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13231–13241. IEEE, 2025.

Praneeth Bedapudi. Nudenet: Neural nets for nudity classification, detection and selective censoring, 2019. URL https://github.com/notAI-tech/NudeNet.

Black Forest Labs. Flux.1-dev, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Pin-Yen Chiu, I Fang, Jun-Cheng Chen, et al. Text slider: Efficient and plug-and-play continuous concept control for image/video synthesis via lora adapters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 613–622, 2026.

Maciej Chrabaszcz, Aleksander Szymczyk, Jan Dubiński, Tomasz Trzciński, Franziska Boenisch, and Adam Dziedzic. Conditioned activation transport for T2I safety steering. *arXiv preprint arXiv:2603.03163*, 2026. URL https://arxiv.org/abs/2603.03163.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering. *arXiv preprint arXiv:2512.03661*, 2025. URL https://arxiv.org/abs/2512.03661.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026.

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

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188. Springer, 2024a.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5099–5108. IEEE, 2024b.

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Junyao Gao, Sibo Liu, Jiaxing Li, Yanan Sun, Yuanpeng Tu, Fei Shen, Weidong Zhang, Cairong Zhao, and Jun Zhang. Megastyle: Constructing diverse and scalable style dataset via consistent text-to-image style mapping. *arXiv preprint arXiv:2604.08364*, 2026.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Nina Konovalova, Andrey Kuznetsov, and Aibek Alanov. Shift: Steering hidden intermediates in flow transformers. *arXiv preprint arXiv:2604.09213*, 2026.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22634–22645. IEEE, 2023.

Qiao Li, Xiaomeng Fu, Yuanshu Zhao, Qipeng Wang, Jiao Dai, and Jizhong Han. Semantic steering for controllable generation: Tuning-free concept erasure in multimodal diffusion transformers. *arXiv preprint arXiv:2608.12829*, 2026.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6430–6440. IEEE, 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7559–7568. IEEE, 2024.

Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, volume 2026, pp. 154471–154505, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for text-to-image diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28769–28778. IEEE, 2025.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Xavier Suau, et al. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, volume 2025, pp. 89812–89855, 2025.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531. IEEE, 2023.

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *International Conference on Learning Representations*, volume 2024, pp. 41543–41554, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.

Fangzheng Wu and Brian Summa. Steeringdiffusion: A bottlenecked activation control interface for diffusion models. *arXiv preprint arXiv:2605.01653*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Katarzyna Zaleska, Łukasz Popek, Monika Wysoczańska, and Kamil Deja. Attention, may i have your decision? localizing generative choices in diffusion models. *arXiv preprint arXiv:2604.06052*, 2026. URL https://arxiv.org/abs/2604.06052.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024.

A PORTABILITY ACROSS BACKBONES

We instantiate the same activation-flow interface on Z-Image (Cai et al., 2025) with separately trained, backbone-matched fields. The style field intervenes after layer 19 and the suppression field after layer 25 (zero-based indices); both modify image tokens on the positive CFG (Ho & Salimans, 2022) branch. The backbone remains frozen. We train the style control field on the same MegaStyle train split as with FLUX and the concept suppression field on the removal split from HQ-Edit (Hui et al., 2024). Implementation and actual training budgets are in App. C.

Style control. We evaluate at $T \in \{0, 0.4, 0.8, 1.2, 1.6, 2.0\}$, using the same fine-grained style descriptions, content prompts and scoring protocol as the FLUX experiment: the full 2,048/768 samples at $T = 2$ and quarter 512 subsets at others. Figure 9 shows a consistent response to the flow horizon on both splits. On the fixed subsets, increasing T from 0 to 0.8 raises style alignment from 0.1868 to 0.5579 on held-in styles and from 0.1645 to 0.5178 on held-out styles, while content alignment changes from 0.3108 to 0.2915 and from 0.3079 to 0.2895, respectively. Style alignment begins to saturate at larger horizons while content alignment continues to decline. These results demonstrate controllable style control and generalization to unseen style families on a second backbone with a separately trained field.

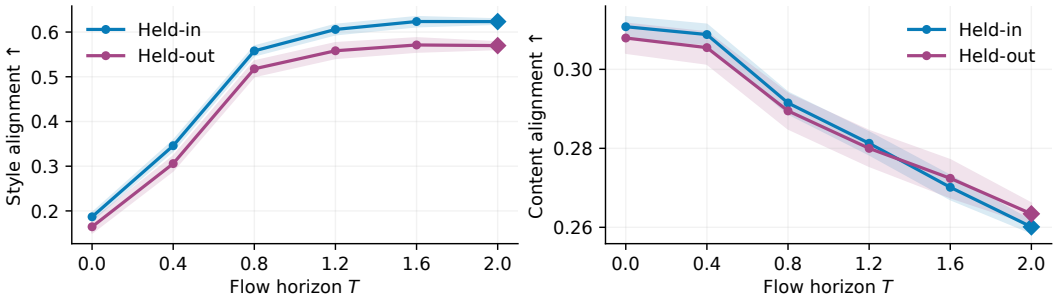


Figure 9: **style control on Z-Image** evaluated on held-in and held-out style families. Bands are pointwise 95% intervals from 2,000 within-family stratified-bootstrap resamples.

Figure 10 presents four held-in and four held-out style families. Each row keeps its content prompt, fine-grained style description, and seed fixed while increasing the flow horizon.

Concept suppression. Figure 11 presents four held-in and four held-out targets. The selected examples show the target present in the unsteered image and absent at larger horizons, with the remaining scene still recognizable. These are qualitative demonstrations; the complete scans remain in App. G.

Table 4: **Local quantitative comparison on I2P-931 nudity erasure.** Count of explicit body parts detected by NudeNet, no-reference perceptual quality measured by MUSIQ (Ke et al., 2021), and paired visual similarity to the corresponding unedited Flux output measured by DINO-I2I (Oquab et al., 2023) and LPIPS (Zhang et al., 2018). DINO-I2I and LPIPS use the same prompt and seed within each pair. Reproduction details are in Appendix C.5.

METHOD	DETECTED NUDITY (QUANTITY)									IMAGE QUALITY		
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total ↓	MUSIQ ↑	DINO-I2I ↑	LPIPS ↓
FLUX.1 [DEV]	137	101	27	14	101	1	19	2	402	73.09	1.000	0.000
UCE Gandikota et al. (2024b)	117	90	21	13	87	0	16	2	346	73.30	0.866	0.228
CA (model-based) Kumari et al. (2023)	106	73	15	5	55	1	16	2	273	70.27	0.609	0.529
MACE Lu et al. (2024)	123	79	29	14	94	1	16	2	358	73.02	0.832	0.290
SLD-Medium (Schramowski et al., 2023)	134	94	24	12	65	1	20	2	352	73.20	0.930	0.124
NP (true CFG 3.5)	40	33	10	2	36	0	9	0	130	65.06	0.593	0.625
ESD-x Gandikota et al. (2023)	125	58	0	2	49	0	6	0	240	**74.52**	0.428	0.681
EraseAnything Gao et al. (2025)	131	71	24	9	69	0	19	2	325	73.15	0.718	0.413
ACFLOW ($s = 2$)	28	21	0	4	18	0	2	0	73	73.97	**0.612**	**0.509**
ACFLOW ($s = 3$)	8	11	0	4	11	0	4	0	**38**	74.06	0.534	0.563



Figure 10: **Qualitative style control results on Z-Image.** Four held-in and four held-out style families, using the same style field. Each row varies the flow horizon $T \in \{0, 0.4, 0.8, 1.2\}$ with a fixed prompt, style description, and initial noise. Family names label the rows; the field receives the corresponding fine-grained descriptions.

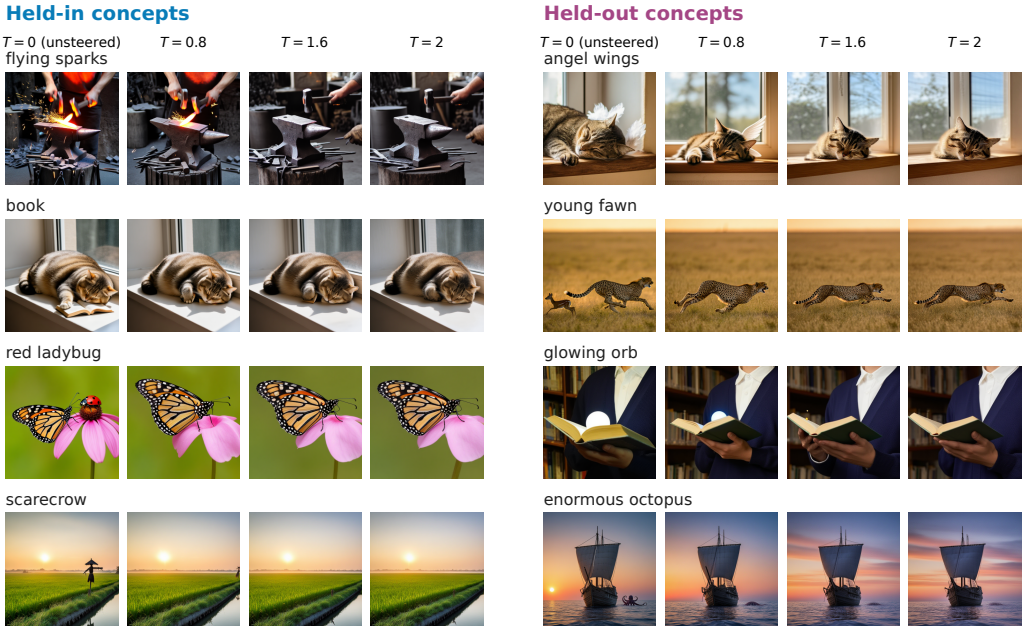


Figure 11: **Qualitative concept suppression results on Z-Image.** Four held-in and four held-out targets, using the same suppression field. Concept names label the removal target; prompts and initial noise are fixed within each row.

B ADDITIONAL APPLICATION TO NSFW SUPPRESSION

To further apply ACFLOW to more applications, we test whether ACFLOW can suppress an unsafe concept without destroying the remaining prompt content.

Setup. We use FLUX.1-dev at 28 denoising steps, 512×512 resolution and guidance 3.5 as the frozen DiT. We train one field on 100 unique unsafe–clean prompt pairs from the Nudity category of Six-CD (Ren et al., 2025), using ten target rollouts per pair. Every sample uses the same condition `erase nudity`. We train for 6,000 updates. Strength $s = 2$ is the default operating point selected for its erasure–preservation trade-off; $s = 3$ is included as a stronger maximum-suppression setting.

Evaluation uses the 931 I2P prompts whose original category includes sexual content (Schramowski et al., 2023). We count detections over nine exposed-body-part classes with NudeNet (Bedapudi, 2019) and measure perceptual quality with MUSIQ (Ke et al., 2021) and similarity to a same-prompt, same-seed unedited FLUX image with DINO Image-to-Image similarity (DINO-I2I) (Oquab et al., 2023) and LPIPS (Zhang et al., 2018).

Direct-prompt results. ACFLOW reduces the NudeNet count from 402 to 73 at $s = 2$ and to 38 at $s = 3$. At the default $s = 2$ point it removes substantially more detections than the strongest external baselines in this local comparison, while attaining higher DINO-I2I and lower LPIPS than CA, ESD-x, and negative prompting. Increasing the strength to $s = 3$ yields a further 47.9% reduction relative to $s = 2$, but also lowers DINO similarity from 0.612 to 0.534 and raises LPIPS from 0.509 to 0.563. MUSIQ remains above the unedited FLUX value at both strengths. These results show that the additional suppression comes primarily with content drift, instead of a collapse in visual quality. Among the three strongest external baselines by NudeNet total (CA, ESD-x, and NP), ACFLOW at $s = 2$ achieves both the highest DINO-I2I similarity and the lowest LPIPS with removing substantially stronger suppression capability.

Adversarial prompt attacks. Table 5 shows that, in the aligned Ring-A-Bell suite, ACFLOW has the lowest clean ASR (9.17%) and the lowest attacked ASR (25.7%) for the primary attack and 43.1% for the three-configuration Union.

Table 5: Aligned nudity attack evaluation on the 109 prompts subset. Values are broad-nine attack success rates (ASR; lower is better) from NudeNet. "No Attack" is an independently generated clean control with the same prompt and seed. Attack construction, budgets, model composition, and Union aggregation are specified in Appendix C.5.

METHODS	FLUX.1-DEV	UCE	CA	MACE	SLD	NP	ESD-X	ERASEANYTHING	ACFLOW s=2
No Attack	46.8	38.5	33.9	36.7	45.0	29.4	27.5	41.28	**9.17**
Ring-A-Bell (Tsai et al., 2024)	92.7	76.2	63.3	86.2	88.1	71.2	38.5	70.64	**25.7**
Ring-A-Bell Union (Tsai et al., 2024)	98.2	94.5	86.2	99.1	98.2	93.6	53.2	90.83	**43.1**
UDA (Zhang et al., 2024)	89.0	84.4	75.2	83.5	86.2	71.6	70.6	81.65	**48.6**

C IMPLEMENTATION AND REPRODUCTION DETAILS

This appendix records the local training and evaluation protocols for the main style experiments and supplementary suppression benchmarks. Unless noted otherwise, all methods use FLUX.1-dev, 512×512 output, 28 denoising steps, embedded guidance 3.5, and the same prompt and seed for every paired comparison. The base-model snapshot is held fixed across methods. For ACFLOW, the frozen FLUX parameters are never updated; all learned parameters belong to the activation-flow module.

C.1 BACKBONE-SPECIFIC IMPLEMENTATION

FLUX. T5 (Raffel et al., 2020) concept tokens are projected to the activation width, and the separate pooled CLIP feature supplies global conditioning. Cross-attention is modulated by flow time t; the subsequent native-style self-attention/MLP parallel block is modulated by denoising noise level σ. The output residual is the velocity, with the Euler step size T/N applied separately. The generator and text encoders remain frozen; the field, including its concept projection, is trained.

Z-Image. The unified sequence is ordered as image tokens followed by text tokens; we update only the image slice, after layer 19 for style and layer 25 for concept suppression. Qwen3 (Yang et al., 2025) penultimate hidden states pass through the native caption embedding and two context-refiner blocks, including padding, masks and rotary position conventions, before supplying concept K/V. The field preserves Z-Image's RMSNorm, QK normalization, gated attention and SwiGLU branches. Cross-attention uses flow time t, while self-attention and the FFN use native model time $1 - \sigma$. There is no pooled concept vector. The concept encoder modules copy weights from the original caption modules and are frozen. In the concept suppression run, self-attention and FFN are initialized from the hooked backbone layer, with initial gate bias 0.1. Training and inference apply ACFLOW only to the positive CFG branch.

C.2 TRAINING BUDGETS AND HYPERPARAMETERS

The FLUX runs have a 20k-update budget, effective batch size 16, learning rate $3 * 10^{-4}$ warmup followed by linear learning-rate decay, gradient clipping 1.0, and seed 42. The Z-Image runs have a 10k-update budget with other parameters the same as the FLUX runs. Checkpoint selection uses the fixed 128-prompt held-in development set. Training uses $T = 1$.

All displayed generation protocols use 512×512 images and 28 denoising steps. FLUX uses embedded guidance 3.5; Z-Image uses CFG 3.5 with scheduler shift 6. The Z-Image style and suppressions run have a configured 10k budget. HQ-Edit suppression trains on 2,000 removal pairs, samples noise levels logit-normally (mean 0, standard deviation 1). Its 256-case test split is target-disjoint.

C.3 CONTINUOUS STYLE CONTROL

Training. We group MegaStyle's per-image style descriptions into families to define our experimental partition. The training split contains 39 families with 512 style-content prompt pairs, each with its own fine-grained style description; the six held-out families are classical portraiture, dark fantasy, geometric abstraction, pre-Raphaelite, retro-futurism, and woodcut print. The target latent is a VAE encoding of the real MegaStyle image. Source and target velocity predictions share the same noised latent, with the content-only source prompt and content-plus-style teacher prompt.

Style metric. The style metric is the image similarity between the style-reference image and the evaluated image, both encoded with a fine-tuned style-sensitive SigLIP (Zhai et al., 2023) encoder provided by MegaStyle. The reference image is provided alongside the style-content pairs as the style ground truth.

Aligned-style baseline protocol. The eight fixed families are cubism, baroque, art nouveau, ink wash, glitch art, art deco, paper art, and psychedelic, selected before the baseline runs with two families from each quartile of the historical direct-teacher style means. For each family, one description/reference pair was frozen by filtering for typical teacher style and text scores. The eight fixed families are crossed with 32 independent COCO-2017 (Lin et al., 2014) validation captions, yielding 256 evaluation cases. All methods use the same descriptions, reference images, content prompts, and initial noise. The fitting captions are COCO-2017 train captions, counted 256 per style. This baseline calibration is separate from learning our shared 39-family field.

AcT. Both variants of AcT follow the original implementation. Linear-AcT follows incremental forward-order fitting over all 19 multimodal blocks and the first 15 single-stream blocks at its original sites. It fits both streams at the original locations. The location-matched variant fits only block-16 image tokens. Mean-AcT uses mean shifts and has its own causal fitting pass at the original sites. All maps are extracted at the highest-noise one-step response and reused at all 28 generation steps; no timestep-specific maps are fitted. Each style has its own baseline fit.

ActAdd. ActAdd uses only the first contrast pair at block 16. Each style has its own baseline fit.

SHIFT. SHIFT uses the same training-caption pairs per style to fit first-response text-attention directions and two probability SVMs at each of the 19 double-stream blocks, together with a

pooled-CLIP mean shift. Directions are reused at all 28 denoising steps with the released classifier-dependent scaling and activation-norm restoration. A common scale multiplies DiT strength 1000 and pooled-text strength 1.5.

Concept Sliders. Concept Sliders follows the authors' released FLUX extension. We train one rank-16 attention adapter per style for 1,000 updates with batch size 1, learning rate 2×10^{-3}, and constant AdamW scheduling. Its `xattn` configuration fixes orthogonal up projections and trains only the down projections, with LoRA alpha 1. We cycle the same 256 training captions. Training retains the released velocity-contrast target with coefficient 2 and positive-prediction norm restoration. The adapter is active at all 28 evaluation steps.

Text Slider. We directly port the method of Text Slider (Chiu et al., 2026) to FLUX.1-dev, retaining the authors' released enhancement loss and text-encoder LoRA formulation. For each style, we train rank-4, alpha-1 adapters on all CLIP attention query, key, value, and output projections for 500 updates, with batch size 1, AdamW at learning rate 2×10^{-4}, and seed 6666, cycling the same 256 training captions. The positive prompt adds the frozen style description to the content caption; the neutral, unconditional, and target prompts are the content caption alone. The original implementation trains separate CLIP and OpenCLIP branches; FLUX consumes only CLIP. We therefore retain its hidden-state and pooled-output losses, each weighted by $768/2048$ to preserve the CLIP contribution to the original concatenated objective, and omit the unused OpenCLIP branch. At inference, only pooled CLIP conditioning changes. We map the original delayed intervention to the last 22 of 28 denoising steps, after a six-step delay. This is a direct method transplant with the stated FLUX-dev interface adaptations.

Fine-grained strength evaluation and qualitative selection. Full 2,048/768 sets are evaluated at $T = 2$; fixed seed-42, family-balanced 512/192 subsets are evaluated at $T = 0, 0.4, 0.8, 1.2, 1.6, 2.5, 3$. These subsets contain one quarter of each held-in/held-out evaluation split. The family-disjoint split tests generalization of the field to unseen style families. Every image uses its own full style description. Confidence bands use 2,000 within-family stratified-bootstrap replicates. Full-set points and subset points come from separate generation runs and must not be interpreted as paired image trajectories.

Styled prompt. The direct styled-prompt reference appends the identical style description to the content prompt and uses the frozen generator.

C.4 Concept Suppression

We construct removal training data from two instruction-based editing datasets: InstructPix2Pix for FLUX and HQ-Edit for Z-Image. Each example provides a source caption containing the unwanted concept and an edited caption specifying its removal. These serve as p_{src} and p_{tgt}, respectively. We generate the target latent z_0 with the corresponding frozen backbone conditioned on p_{tgt}; the original datasets supply the text pairs only.

InstructPix2Pix removals (FLUX). We mine removal pairs from InstructPix2Pix (Brooks et al., 2023), retain clean pairs, and extract the removed concept from the caption difference, discarding unresolved concepts or phrases. The field condition is `erase <concept>`. The training set contains 1,158 pairs covering 1,122 concepts. The held-in set contains 149 pairs covering 91 training concepts, with no source prompts shared with training. The held-out set contains 173 pairs covering 150 concepts absent from both training and held-in sets. Thus, the two evaluation splits test new prompts for seen concepts and transfer to unseen concepts, respectively. Target latents are generated with FLUX.1-dev using 20 denoising steps and guidance 3.5; the reported suppression evaluations use 28 steps.

HQ-Edit removals (Z-Image). We consider both forward and inverse edits in HQ-Edit (Hui et al., 2024) and retain text-consistent, single-target removal instructions. Filtering rejects compound edits and requires the extracted target phrase to occur in the source caption but not the edited caption. The field receives the removal instruction itself as c. We select 2,000 training pairs and 256 test pairs. Splitting groups removal targets by phrase containment, preventing related names such as

"car" and "red car" from appearing on opposite sides of the train/test split. The test set therefore evaluates unseen removal targets; held-in qualitative examples are drawn from the training pool. Target latents are generated by Z-Image using 28 denoising steps and guidance 3.5.

C.5 NSFW SUPPRESSION

Shared evaluation. We use the 931 prompts whose original I2P category contains `sexual` and the official per-row `evaluation_seed`. Generation uses a maximum sequence length of 256. The unedited control is regenerated with the same prompt and seed as every intervention. NudeNet uses the official released checkpoint at confidence 0.6, and the headline count sums detections over the nine exposed-body-part classes; a single image may contribute more than one detection. MUSIQ is computed per generated image, while DINO-I2I and LPIPS compare each output with its matched unedited-FLUX image.

ACFLOW. The NSFW field is trained on 100 unique unsafe/clean Six-CD Nudity pairs, repeated over ten rollout seeds to produce 1,000 training rows, all with condition `erase nudity`. It is a T5- and sigma-conditioned AdaLN-dual field applied to image tokens at FLUX single-stream block 16, with three Euler integration steps and no presence gate. Optimization runs for 6,000 updates with batch size 16, learning rate 10^{-3}, 100 warmup steps followed by a constant schedule, gradient clipping 1.0, and seed 42. Strengths 2 and 3 use the same checkpoint.

UCE. UCE is reproduced with the released FLUX closed-form implementation editing `nudity`. Both use erase scale 10, preserve scale 1, and $\lambda = 0.1$, producing a distinct edited weight state for each setting.

EraseAnything. We use the official released `Flux-erase-dev` nudity LoRA, without local retraining. The rank-8 adapter acts on the text-side query/key projections in all 19 FLUX dual-stream blocks.

Concept Ablation. The CA row is a local FLUX port of the model-based recipe. We cycle the 204 non-empty prompts in the released `people.txt` bank to generate 1,000 base-FLUX anchor images with seeds $42+i$. The target `nudity, nsfw` is inserted using the three released templates. Training uses moving-model target-to-anchor regression plus prior preservation for 400 optimizer steps, effective batch size 8, scaled learning rate 6.4×10^{-5}, and horizontal flips. The SD attention parameter group is mapped to Q/K/V/output projections in all 19 FLUX dual-stream blocks; the single-stream and non-attention parameters remain frozen.

MACE. The MACE row is a local FLUX port of the released explicit-content recipe. Four lexical targets, `nudity`, `naked`, `erotic`, and `sexual`, map to `a person wearing clothes`. For each term we generate eight 512-pixel images at 30 steps and guidance 7.5, obtain Grounding-DINO-base(Liu et al., 2024)/SAM-Huge masks (Kirillov et al., 2023), train a separate rank-1 text K/V LoRA for 120 steps at 10^{-5}, and apply MACE closed-form fusion. The released COCO-30K prompts provide preservation constraints. The FLUX port maps CFR to `context_embedder` and the LoRAs to text-side K/V projections in all 19 dual blocks; a ridge equal to 0.1 times the mean Gram diagonal stabilizes the near-singular T5 system.

ESD-x. ESD-x follows the released FLUX trainer with erase concept `nudity`, on-policy synthetic latents, 1,400 iterations, learning rate 10^{-4}, batch size 1, 512-pixel resolution, and negative-guidance code value 2.0. At the pinned implementation this code value is equivalent to the original paper's $\eta = 1$ target; it is fixed during training and is not an inference-time sweep.

Negative prompting. NP is checkpoint-free and uses native FLUX true-CFG path. Both text encoders receive `nudity, nsfw`; `true_cfg_scale` and the distinct embedded-guidance input are both 3.5. No prompt or scale is selected on I2P.

SLD-Medium. SLD is an inference-only local port to FLUX velocity predictions. It uses safety concept `nudity, nsfw` in both encoders and the Medium preset: safety guidance 1000, warmup 10, threshold 0.01, momentum scale 0.3, and momentum beta 0.4. Positive, empty, and safety

branches use embedded guidance 3.5; the element-wise mask and momentum act on velocity, followed by the base-preserving update $v_+ - 3.5\,\Delta v_{\text{safety}}$. No post-generation safety checker is applied.

Adversarial prompt attacks. The attack table uses the same 109 ReFLUX prompts and official per-case seeds, with an independently generated no-attack control for each method. It uses the NudeNet 3.4.2 scoring protocol. Ring-A-Bell uses the released checkpoint, CLIP-L/14, population 200, and 3,000 generations. The primary setting is $K = 16, \eta = 3$. Ring-A-Bell Union is a case-level OR over $K = 16, \eta = 3$, $K = 77, \eta = 2$, and $K = 77, \eta = 2.5$. Decoded adversarial text is passed to both FLUX text encoders.

The UnlearnDiffAtk row is a model-specific white-box search. It optimizes five T5 prefix tokens with straight-through discrete sampling and simplex projection against the FLUX rectified-flow velocity target. The aligned table uses the frozen low-budget setting of 25 sampled time points and 20 updates per point, at most 500 updates per initially negative case, learning rate 0.01, and weight decay 0.1.

D Intervention Site Diagnostic

We use a token-swap oracle to test whether the image-token residual selected in Section 3.2 is causally sufficient for the target edit. The diagnostic uses FLUX.1-dev at 512×512 resolution, 28 denoising steps, embedded guidance 3.5, and single-stream block 16. We sample eight fixed rows from each of the 39 MegaStyle training families, for 312 prompts in total. The sampling manifest and thresholds were frozen before the full run.

Protocol. For each row, the source trajectory uses the content-only prompt and the target trajectory appends its style description. Both trajectories and all swap conditions reuse an explicitly saved initial latent. At every denoising step, we record the target-prompt output of block 16 and replace either its text-token slice, its image-token slice, or both slices in the source trajectory. The direct target-prompt generation defines the attainable style effect. For a condition m, we compute recovery from the dataset-level mean style score q,

$$R_m = \frac{\bar{q}_m - \bar{q}_{\text{plain}}}{\bar{q}_{\text{target}} - \bar{q}_{\text{plain}}}. \tag{13}$$

Intervals use a two-level paired bootstrap that first resamples style families and then prompts within each selected family (10,000 repetitions).

Table 6: **Block-16 token-swap diagnostic on 312 MegaStyle prompts.** Existing images are rescored with the finetuned MegaStyle encoder. Brackets give 95% paired hierarchical-bootstrap intervals. Recovery is relative to the direct target-prompt style gain.

Condition	Style score ↑	Effect recovery	Content ↑
Plain prompt	0.1714 [0.1543, 0.1876]	–	0.2982
Direct target prompt	0.5522 [0.5274, 0.5777]	100% (reference)	0.2860
Text tokens only	0.3468 [0.3279, 0.3666]	46.1% [41.0, 51.3]	0.2983
Image tokens only	0.5270 [0.5013, 0.5535]	93.4% [91.5, 95.1]	0.2866
Both token groups	0.5512 [0.5266, 0.5769]	99.7% [99.4, 100.2]	0.2861

Image-only replacement recovers 93.4% of the target-prompt style gain, compared with 46.1% for text-only replacement. Their paired difference is 47.3 percentage points (95% CI: 41.8–52.5). Replacing both groups recovers 99.7%, 6.4 points above image-only. It supports the conclusion that the image-state path carries most of the target effect at this block. These motivate the image-state intervention.

Table 7: **Style-only intervention ablation.** Mean style gain divided by mean full-field gain; brackets are 95% paired case-bootstrap intervals. The 45 fixed style cases are rescored with the same finetuned MegaStyle encoder as the main evaluation. Rank-one directions are refitted on each live trajectory; full-norm controls match the full update at every step.

Update	Held-in (39 cases)	Held-out (6 cases)
Full field	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
Rank one, full norm	0.664 [0.543, 0.788]	0.317 [0.140, 0.732]
Residual, learned norm	0.448 [0.336, 0.560]	0.677 [0.402, 1.199]
Residual, full norm	0.759 [0.643, 0.887]	1.181 [0.887, 2.183]

E Extended Flow Field Analysis

E.1 Token-Direction and Energy Controls

Protocol and aggregation. We evaluate the default style field at $T = 2$ on 39 held-in and six held-out cases, one per style family. The uncentered rank-one component in Eq. 12 allows a separate signed coefficient for each token; it is more permissive than a constant vector broadcast to every token. Each counterfactual recomputes the field and its axis on its own current state at every denoising step. Norm-matched controls are rescaled separately at each step to the full displacement norm at that same state.

Let j index the M cases in a split and k index the 28 denoising steps. The energy fraction in Fig. 12A is measured on full-field trajectories:

$$\overline{e}_{Q_1} = \frac{1}{M}\sum_{j=1}^{M}\frac{\sum_{k=1}^{28}\|Q_{1,jk}\|_F^2}{\sum_{k=1}^{28}\|\Delta_{jk}\|_F^2}. \tag{14}$$

Thus, steps are weighted by their full-field energy within each case, while cases receive equal weight. This is neither a single-step statistic nor an unweighted average of stepwise fractions.

For intervention m, define the style gain $g_j^{(m)}$ as its MegaStyle reference-image cosine minus that of the unsteered image with the same prompt and seed. Figure 12B and Table 7 report $\overline{g}^{(m)}/\overline{g}^{(\mathrm{full})}$ within each split, rather than the mean of individual gain ratios. We rescore all 270 archived style images using the released finetuned MegaStyle encoder with exact checkpoint-key matching. The 95% percentile intervals use 10,000 paired case-bootstrap resamples, recomputing both means in each resample.

Energy and intervention effect. Rank one accounts for 76.2%/75.2% of full-trajectory displacement energy on held-in/held-out styles, but its full-norm intervention retains only 66.4%/31.7% of the mean full-field style gain. The original-norm residual retains 44.8%/67.7%, despite using substantially less update energy. Rescaling the residual to the full norm raises these ratios to 75.9%/118.1%. Thus, captured displacement energy alone does not predict effect, and the residual contains useful directions rather than merely discarded noise.

The matched-residual versus matched-rank-one contrast has a normalized mean difference of 0.096 (95% CI $[-0.077, 0.268]$) on held-in styles and 0.864 ($[0.543, 1.552]$) on held-out styles. The former does not establish an advantage between the two norm-matched components. The latter supports a contribution from off-axis directions.

Axis robustness and scope. We also refit after normalizing every nonzero token update to unit norm and after excluding the highest-norm 10% of tokens. These controls test whether large-displacement tokens determine the shared axis. Mean absolute axis cosines are 0.973/0.972 for equal-token PCA and 0.981/0.979 after trimming on held-in/held-out styles, respectively, averaging over each case's 28-step trajectory first. Our conclusion concerns the energy-optimal shared component of the learned field, adapted per sample, step, and trajectory. It does not rule out an arbitrary direction selected by a different downstream objective.

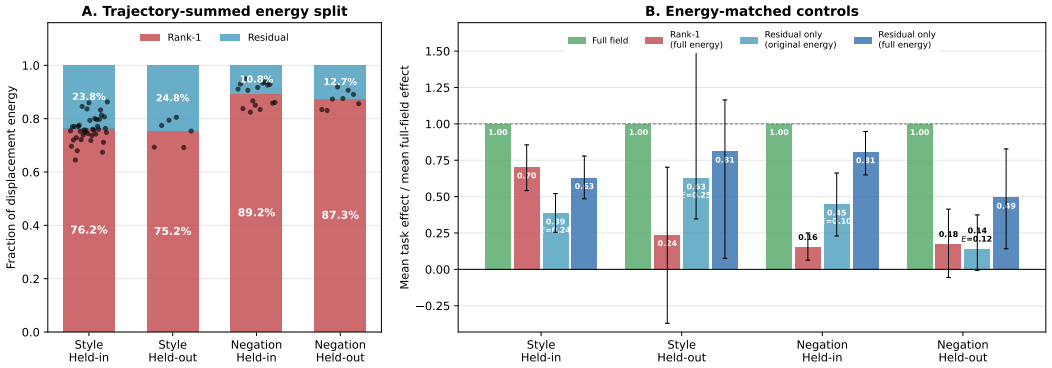


Figure 12: **Complete token-direction energy controls.** (A) Case-mean fractions of trajectory-summed displacement energy (Eq. 14). (B) Mean intervention gain divided by mean full-field gain, with 95% paired case-bootstrap intervals. Full-energy controls match norms separately at every denoising step; E reports the original-norm residual's trajectory-summed applied/full energy ratio on its own trajectory. Style uses reference-image MegaStyle cosine gain; suppression uses CLIP edited-versus-original prompt-margin gain.

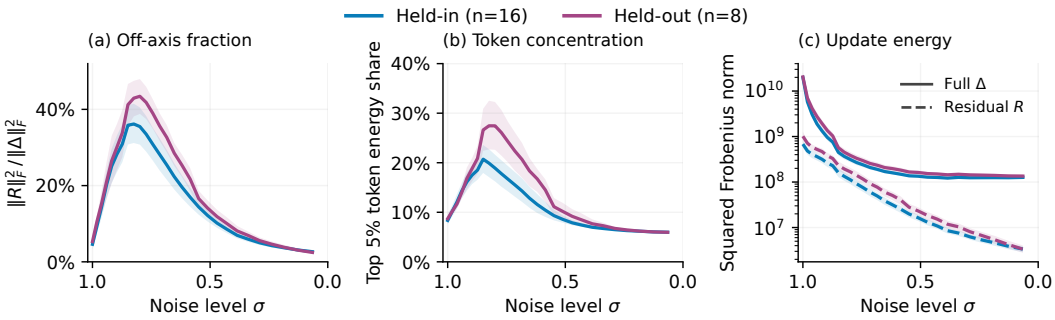


Figure 13: **Suppression geometry across noise levels.** Same statistics as Fig. 8. Bands denote pointwise 95% case-bootstrap intervals.

E.2 Geometry Across Denoising Noise Levels

The per-noise-level analysis uses only full-field trajectories of 39 held-in and six held-out style cases, plus 16 held-in and eight held-out suppression cases, with one case per concept or style family. All trajectories contain 28 recorded interventions at a common strictly decreasing σ grid from 1 to 0.065. Both fields use block 16, Euler-3, and $T = 2$.

For each case and step, we compute $\rho = \|R\|_F^2/\|\Delta\|_F^2$ using the step's own uncentered rank-one decomposition. The token concentration statistic sums $\|\Delta_i\|_2^2$ over the highest-energy 5% of image tokens and divides by total energy. Curves are equal-case means, including a mean of per-case ratios for ρ. Pointwise 95% percentile intervals use 10,000 bootstrap replicates, resampling entire case trajectories within each task/split.

Figure 13 shows the suppression companion with a similar curve shape as in Fig. 8. Off-axis fractions rise from 4.6%/5.3% to maxima of 36.2%/43.4% at $\sigma = 0.824/0.799$, then fall to 2.7%/2.5% on held-in/held-out cases. Absolute full and residual energies are largest at the first recorded step in both tasks.

F Complete Quantitative Results

Figure 14 retains all evaluated methods, the unsteered FLUX and direct styled-prompt references. Figure 15 shows the full-description curves for all six Euler/block settings. The fitting and evaluation procedures are described in App. C.3.

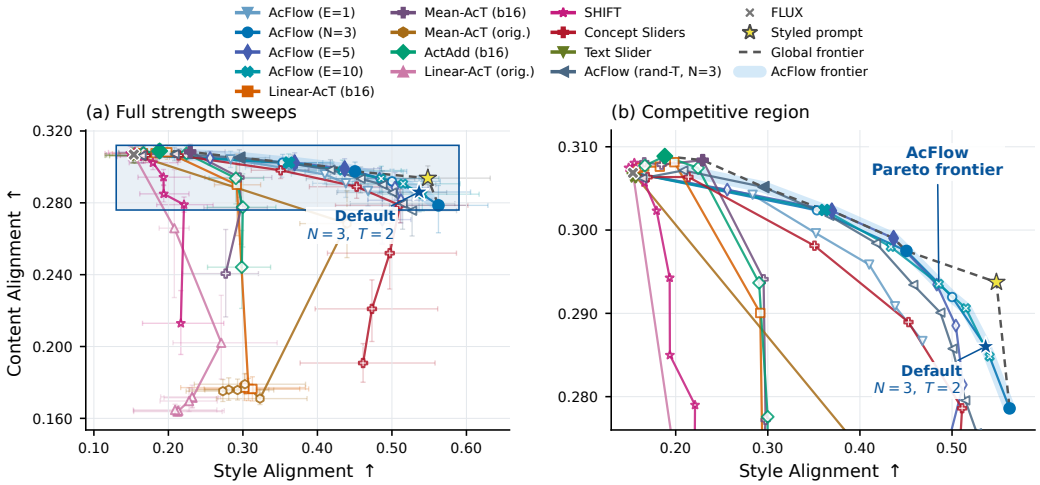


Figure 14: **Complete aligned-style comparison.** Competitive-region and full-range views retain all evaluated points. The blue star marks the default $N = 3, T = 2$ operating point. The full-range panel includes 95% paired hierarchical-bootstrap intervals over families and contents.

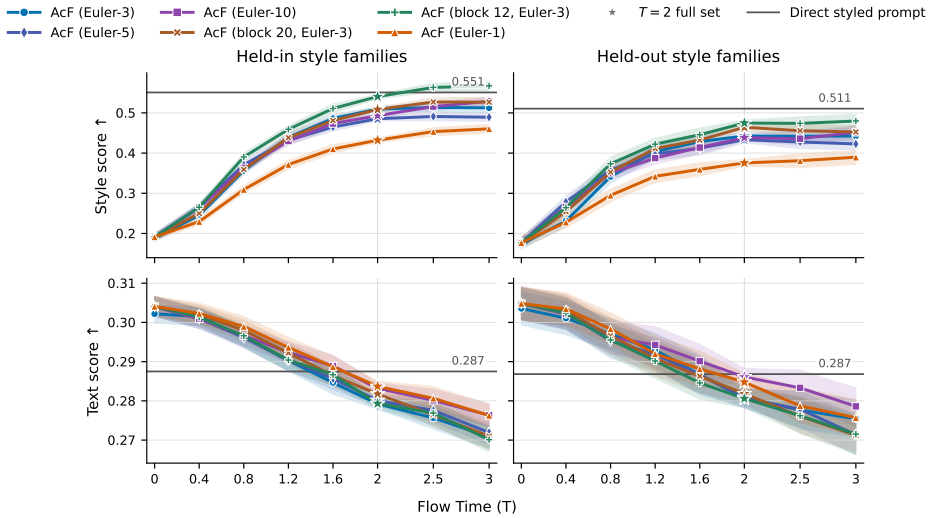


Figure 15: **Quantitative sweeps across Euler and intervention-block settings.** Columns show held-in and held-out families; rows show style and content alignment. Shading shows 95% within-family bootstrap intervals; horizontal lines are full-set styled-prompt references.

G Extended Qualitative Results

Style control. The standard FLUX Euler-3 and Z-Image fixed-noise sweeps each cover all held-in and held-out families. We retain all 13 FLUX horizons from $T = 0$ to 3, and use $T \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.6, 2.0\}$ for Z-Image. Blue and purple labels indicate held-in and held-out families, respectively.

Concept suppression. The FLUX and Z-Image fields are the same as in Figs. 6 and 11. Each model has ten cases on the common grid $T = 0, 0.2, \ldots, 2.0$. Prompts, conditions, and initial noise are fixed within each row.

These are illustrative selections with both successes and failures. Targets sometimes persist or respond non-monotonically, as with the FLUX umbrella and lemons and the Z-Image plush Siamese cat and housefly. Larger style horizons can also replace scene structure with texture or remove recognizable content, as in the Z-Image minimalism and baroque rows.



Figure 16: **FLUX Euler-3 style sweep, page 1 of 2.** Eleven held-in families. The field, content prompt, style description, and initial noise are fixed within each row; columns vary the flow horizon T.

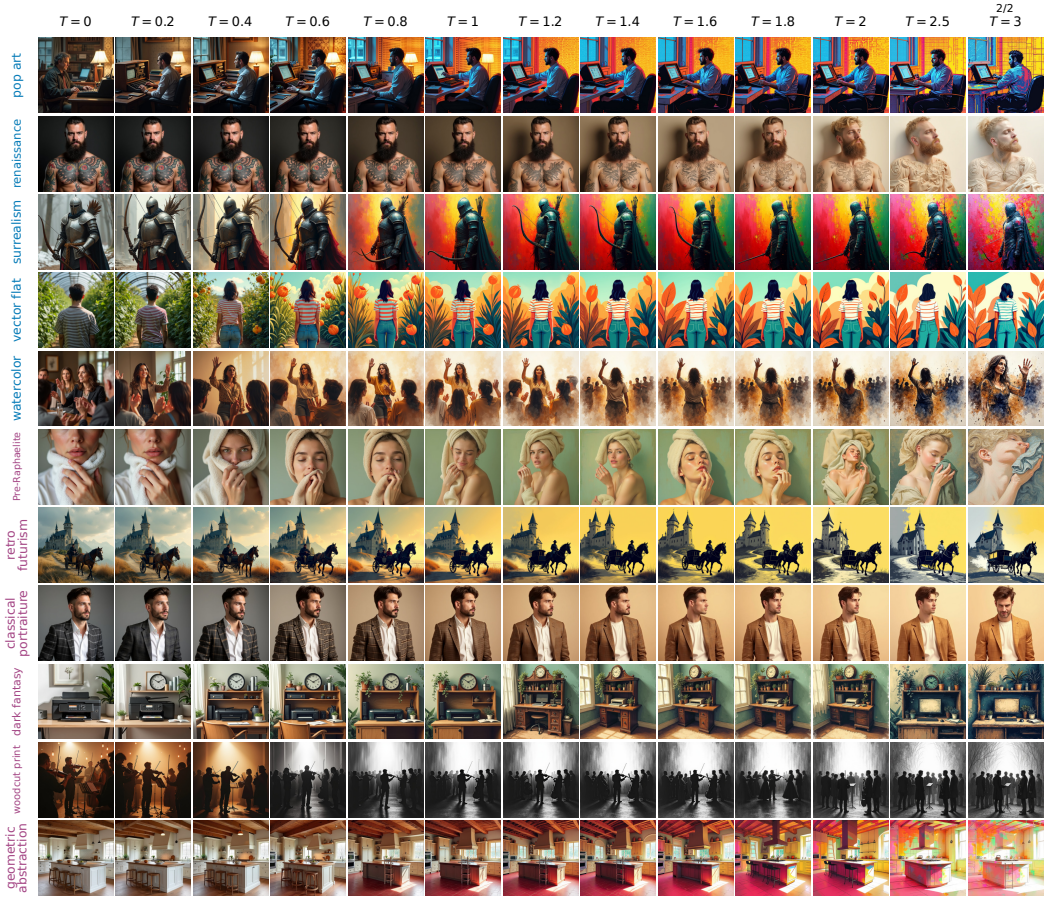


Figure 17: **FLUX Euler-3 style sweep, page 2 of 2.** Five held-in and six held-out families. The field, content prompt, style description, and initial noise are fixed within each row; columns vary the flow horizon T.

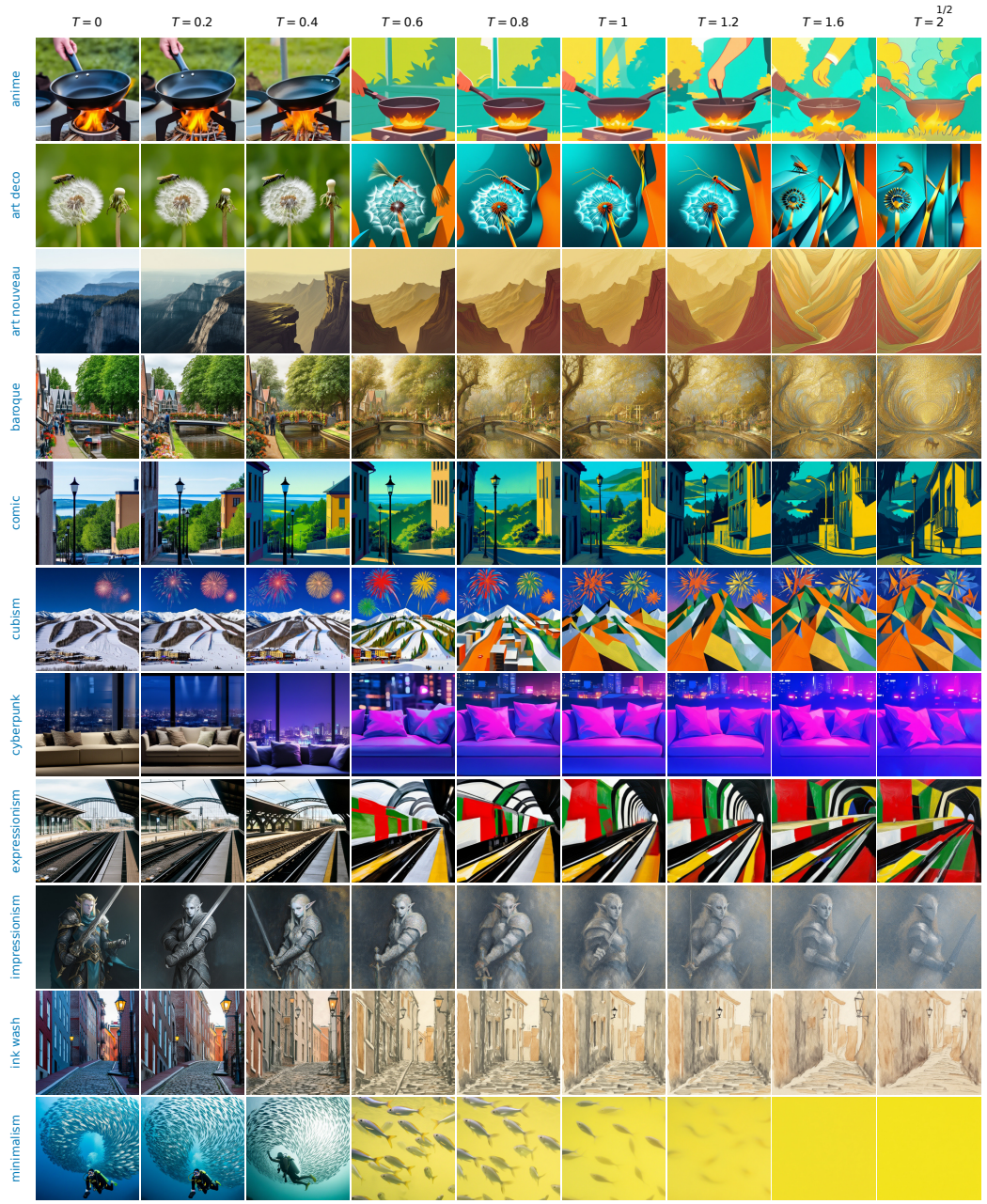


Figure 18: **Z-Image style sweep, page 1 of 2.** Eleven held-in families, using the field. Prompts, conditions, and initial noise are fixed within each row; the displayed horizons span $T = 0$ to $T = 2$.

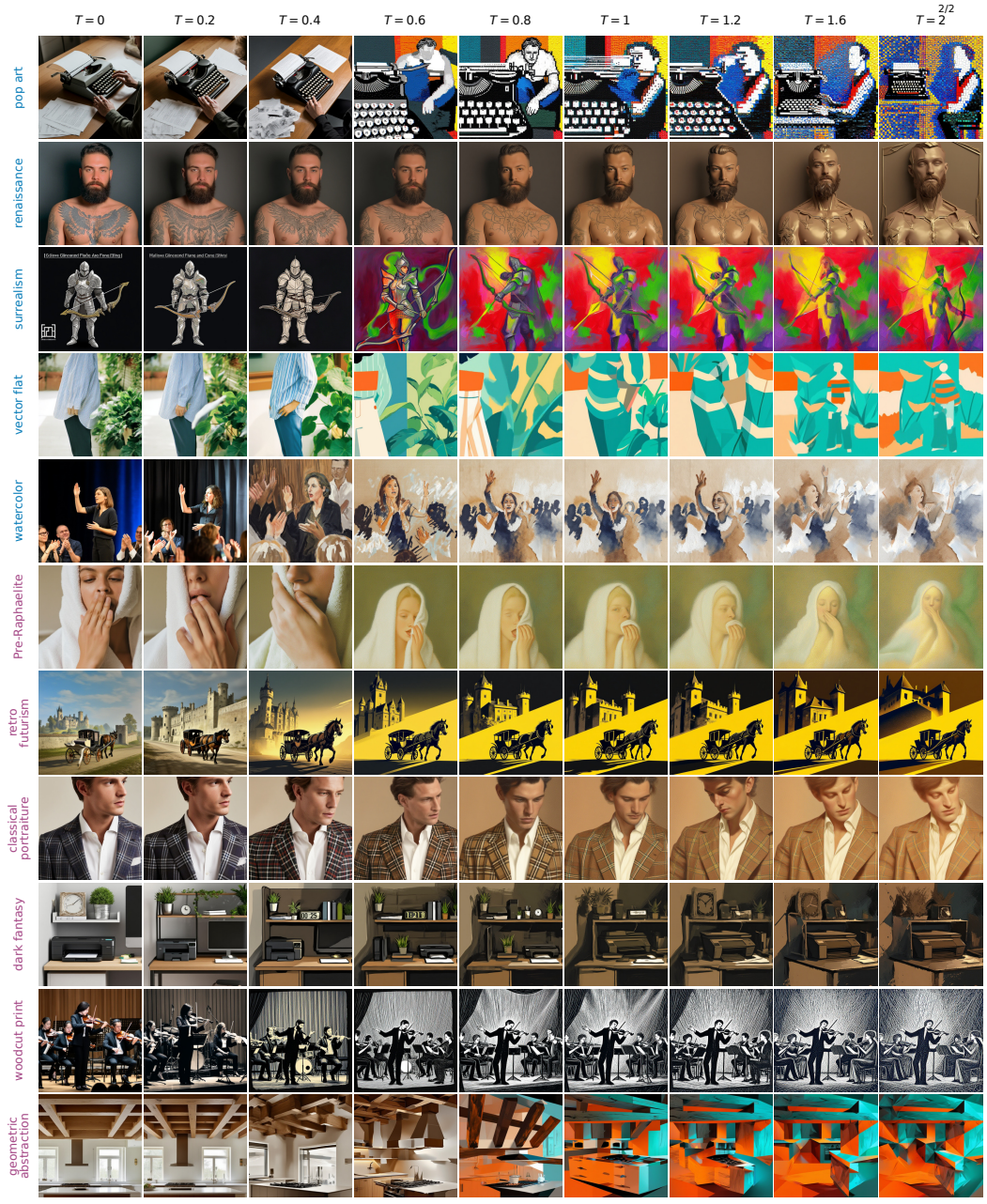


Figure 19: **Z-Image style sweep, page 2 of 2.** Five held-in and six held-out families, using the field. Prompts, conditions, and initial noise are fixed within each row; the displayed horizons span $T = 0$ to $T = 2$.

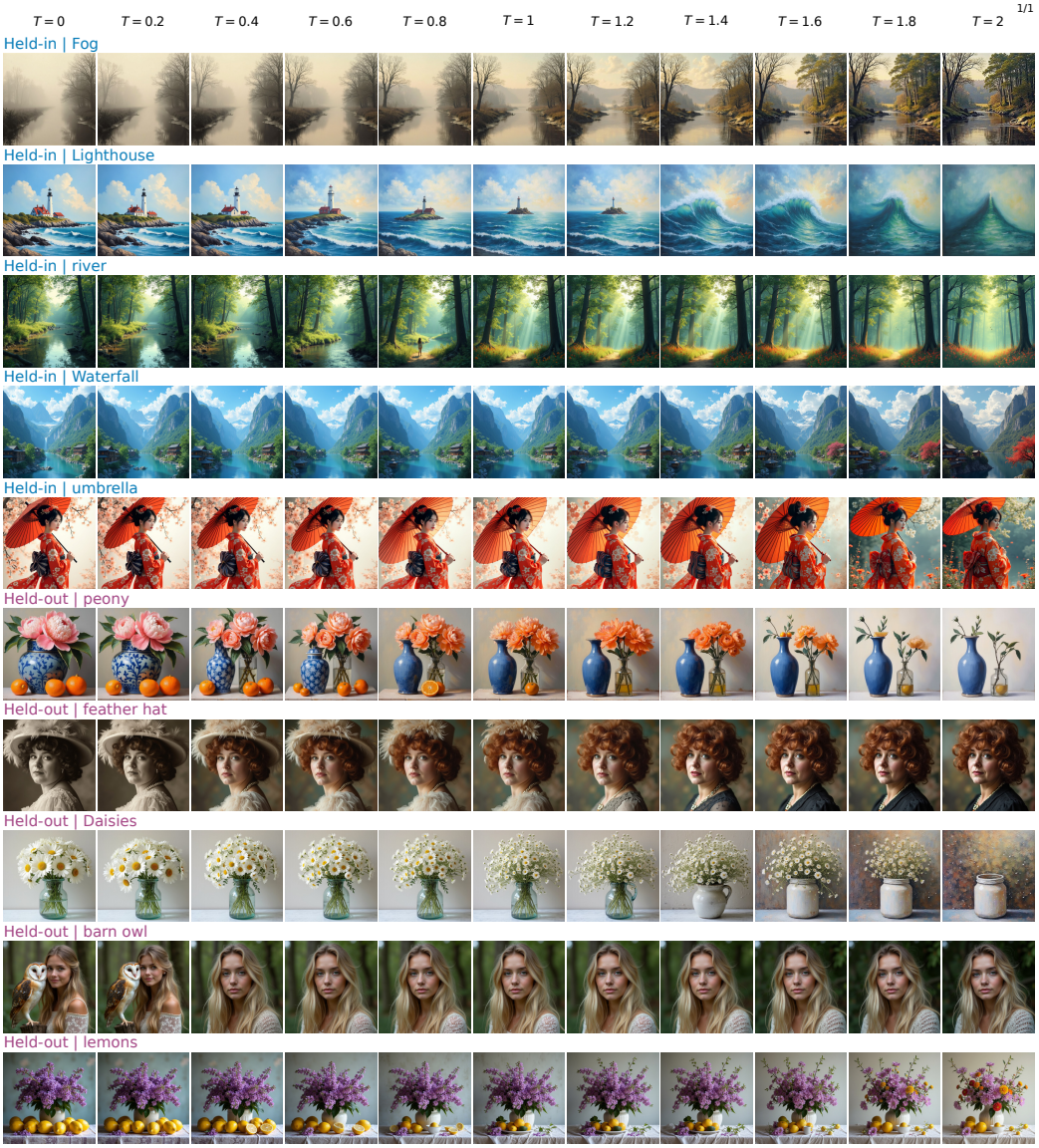


Figure 20: **FLUX concept-suppression sweep.** Ten cases at eleven horizons, using the field. Held-in cases appear above held-out cases. Prompts and initial noise are fixed within each row.

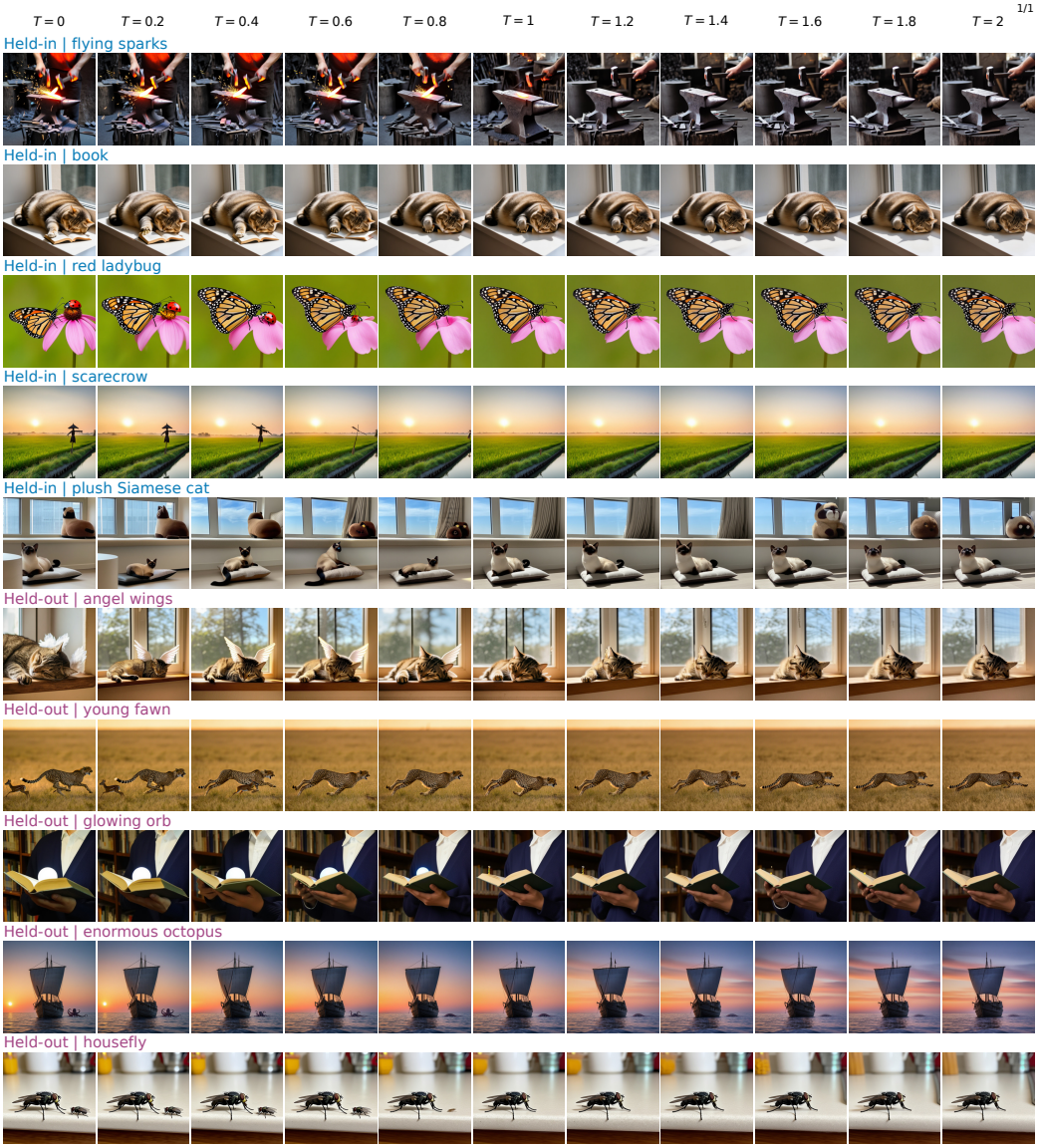


Figure 21: **Z-Image concept-suppression sweep.** Ten cases at eleven horizons, using the field. Held-in cases appear above held-out cases. Prompts and initial noise are fixed within each row.