TRACE: TRAJECTORY-AWARE CONCEPT ERASURE FOR FLOW MATCHING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept unlearning is essential for suppressing harmful and copyrighted content in text-to-image (T2I) models while preserving their generation utility. While flow matching (FM) models are increasingly widespread, current unlearning techniques are predominantly diffusion-oriented, redirecting model predictions toward safe anchor-conditioned targets or unconditional generations at isolated states. However, directly transferring these diffusion-unlearning objectives to FM models can overlook their inherent continuous flow structure. We theoretically analyze how such adaptation can damage the underlying flow geometry. Flow-aware unlearning is essential for FM models because mismatching the flow paths can introduce conflicting velocity supervision and impose a positive regression floor, which causes a risk of utility degradation during unlearning. To overcome these structural disruptions, we propose TRACE, a flow-aware unlearning framework that erases target concepts via safe flow trajectory replacements. TRACE first jointly selects anchors for different generation seeds using a flow conflict proxy. It then fine-tunes the model along the selected teacher trajectories to replace harmful or undesirable generations while preserving a safe flow structure. Across the eight main-table settings, experiments on Stable Diffusion 3.5 Medium and FLUX.2 4B demonstrate that TRACE outperforms state-of-the-art baselines, reducing target-concept CLIP similarity by up to 7.10%, increasing retained-prompt CLIP alignment by up to 7.78%, and reducing FID by up to 16.55%. Furthermore, under sequential unlearning of 15 concepts on SD 3.5 Medium, TRACE maintains a 74.93% anchor-win rate and 15.29 FID, retaining a clear advantage over the compared baselines.

1 INTRODUCTION

Text-to-image (T2I) generation has achieved remarkable success in generating high-fidelity images conditioned on input text prompts. Flow Matching (FM) models provide a prominent foundation for these models by learning velocity fields that transport noise seeds into images Lipman et al. (2023); Liu et al. (2023). However, their ability to reproduce harmful, copyrighted, or undesirable concepts can negatively impact on models' compliance and legality. To mitigate these risks, concept unlearning techniques have been proposed with the goal of effectively removing specified concepts while preserving generation on unrelated prompts Gandikota et al. (2023); Kumari et al. (2023).

Nevertheless, current concept unlearning techniques in T2I tasks are predominantly diffusion-oriented. To construct unlearning supervision, these methods redirect the model's prediction under the target concept toward that of an anchor concept at latent states sampled from the entire latent space Gandikota et al. (2023); Kumari et al. (2023); Kim et al. (2026). Although effective for diffusion settings, they typically perform unsatisfactorily on FM models. This is because the unlearning latents should be carefully sampled to remain flow-coherent with the original model. Otherwise, the misaligned flow structure can induce conflicting velocity supervisions, which makes convergence difficult and harms non-target utility preservation. While some very recent related works focus on FM models' architectural shift from U-Net to DiT Gao et al. (2025); Grebe et al. (2026), they still lack flow-aware guidance to maintain a coherent flow intentionally.

Achieving effective erasure while preserving utility is particularly challenging for FM models. Unlike traditional diffusion models that predict noise, FM models formulate the generative process through continuous trajectories driven by learned velocity fields Lipman et al. (2023); Liu et al. (2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Therefore, a well-structured underlying flow geometry is crucial when training and editing FM models Tong et al. (2023); Geng et al. (2025). This motivates our first question: *how can we construct flow-coherent supervision for concept unlearning in FM models?*

To theoretically answer this question, we start from transport theory and flow geometry. Our analysis identifies four theory-motivated design principles for FM unlearning. **P1:** source correspondence. Concept unlearning should operate along the model-induced trajectories rather than sampling the entire latent space, which accounts for the failure of most diffusion-oriented methods. **P2 & P3:** trajectory support and instantaneous tangents. Even along the model-induced paths, a flow-coherent unlearning still needs careful endpoint (anchor images) choices and regression velocity choices to avoid flow mismatches. **P4**: joint velocity compatibility. Even when each trajectory is individually coherent, the joint paths selection can still be incompatible. This motivates the second question: *how can we translate these geometric insights into a practical unlearning framework that simultaneously achieves effective target removal and robust non-target utility preservation?*

To this end, we propose TRACE , a two-stage flow-aware concept unlearning framework. In a *pre-unlearning stage*, we optimize a compatibility proxy loss to assign the anchor trajectory for each seed (e.g., Gaussian noise) jointly, aiming to reduce the global velocity conflicts. In the following *unlearning-updating stage*, we fine-tune the model along these selected paths to replace the target predictions (e.g., velocity towards harmful or undesirable images) with anchor predictions (e.g., velocity towards anchor images with safe content). The first stage targets **P4**, whereas the second stage addresses individually flow-coherent unlearning (achieving **P1**-**P3**).

Extensive experiments on Stable Diffusion 3.5 Medium and FLUX.2 4B demonstrate that TRACE simultaneously achieves superior target removal and non-target utility preservation compared to SOTA baselines. Across eight concept–model settings, it consistently achieves the best CLIP(U), Anchor-win, CLIP(R), and FID results among the baselines. Furthermore, TRACE is robust under multiple-concept unlearning. Its removal and utility advantage persists under sequential unlearning of up to 15 concepts. The ablation study also empirically shows the effects of our proposed theoretical principles and TRACE design.

Our main contributions are summarized as follows:

- **Flow-aware geometric analysis.** We theoretically analyze structural vulnerabilities of unlearning supervision in FM models, identifying four design principles regarding source correspondence, trajectory support, instantaneous tangents, and joint velocity compatibility.
- **Geometry-guided concept unlearning.** We propose TRACE , a two-stage framework that combines model-induced tangent supervision with joint anchor assignment, translating the geometric insights into a practical optimization objective and effective FM unlearning algorithm.
- **Removal, utility, and scalability.** We conduct comprehensive evaluations across multiple concept categories on two representative modern FM backbone models. TRACE improves single-concept removal and utility preservation over state-of-the-art baselines and retains its advantage under multiple-concept unlearning. Our ablation study also shows the effects of the TRACE design.

2 Related Work

Concept unlearning in diffusion models. Diffusion concept-erasure methods suppress a target concept while attempting to preserve unrelated generation. Teacher-guided methods replace the target prompt's noise or score prediction with an unconditional or anchor-conditioned target Gandikota et al. (2023); Kumari et al. (2023); complementary approaches edit selected parameters or attention modules, identify salient weights, or add retention and adversarial objectives Fan et al. (2023); Wu et al. (2025b). These methods are well motivated in their native diffusion settings. Our concern is their direct translation to FM. A pointwise diffusion target specifies the prediction to be changed at a sampled noisy state. When this target is imported as FM supervision, one must additionally decide which states support the edit and which source-conditioned path, if any, gives the velocity its transport interpretation. Replacing only the output parameterization leaves these choices implicit.

Concept unlearning in Flow Matching models. Recent FM-specific methods adapt concept erasure to flow-based T2I models. EraseAnything uses parameter-efficient editing with attention regularization Gao et al. (2025), while GEM steers generation toward safe alternatives and away

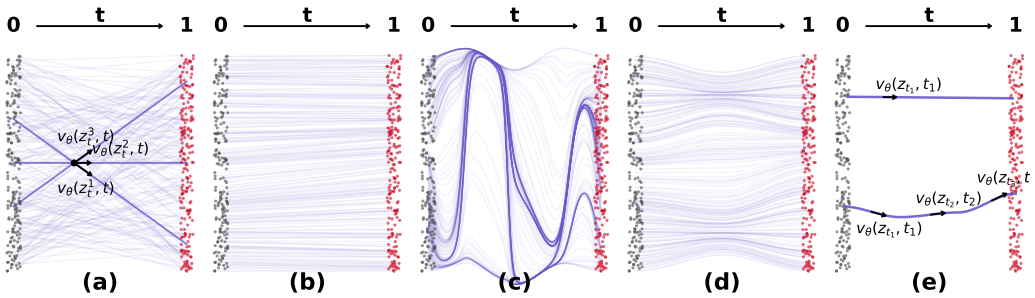

Figure 1: **Illustration of FM training and inference.** Black and red dots represent noise seeds and image endpoints, respectively. The $t = 0 \to 1$ axis denotes the transport process and purple lines denote the trajectories. From left to right: **(a)** Training flow via arbitrary coupling π; **(b)** Training flow improved by OT coupling π_{OT}; **(c)** The learned flow (curved); **(d)** The rectified learned flow (straighter); **(e)** Why curved flow is slow in generation: more sampling steps for approximation.

from the target concept through a contrastive velocity objective Grebe et al. (2026). ContinualFlow studies FM unlearning through energy-based mass subtraction Simone et al. (2025). They still lack flow-aware guidance to intentionally maintain a coherent flow during concept erasure, risking supervision conflicts and utility degradation during unlearning. The FM literature has shown that the coupling and interpolation used to construct conditional paths affect the induced regression problem and the geometry of the learned flow Lipman et al. (2023); Liu et al. (2023); Villani (2009). We bring this perspective together with our theoretical insights to concept unlearning, proposing TRACE to improve flow coherence during unlearning and realize better removal–utility performance.

3 Preliminary: Flow Matching and Concept Unlearning

FM Training and Inference. FM models Lipman et al. (2023); Liu et al. (2023) learn a velocity field that transports a source distribution (Gaussian noise μ_0) to the data distribution p_{data}. We use a coupling π to pair seeds (noise latents z_0) and endpoints (image data latents z_1), and define a conditional path $[z_t]_{t=0}^{1}$ between them. Modern FM models commonly use linear training interpolants, $z_t = (1-t)z_0 + tz_1$ for $t \in [0,1]$, with constant pairwise target velocity $v_t = z_1 - z_0$. The FM model is trained by conditional velocity regression $\mathcal{L}_{\mathrm{CFM}}(\theta) \triangleq \mathbb{E}_{t,(z_0,z_1)\sim\pi}\|v_\theta(z_t,t) - v_t\|_2^2$, where $v_\theta(z_t,t)$ is the velocity predicted at latent z_t and time t Labs et al. (2025). As illustrated in Figure 1(a–c), straight training paths need not yield straight learned trajectories Geng et al. (2025): the regression averages the pairwise velocity targets conditional on the observed state and time. For inference, the model-induced flow solves $\Phi^c_{\theta,s\to t}(x) = x + \int_s^t v_\theta(\Phi^c_{\theta,s\to\tau}(x),\tau,c)\,d\tau$, where x is the state at time s and $c \in \mathcal{C}$ is the condition Esser et al. (2024). We write $\Phi^c_{\theta,t}(z_0)$ when $s = 0$ for brevity.

FM Optimization and Geometric Constraints. Learning an efficient velocity field depends on the supervision geometry. First, arbitrary coupling can assign different pairwise velocity targets to nearby interpolation states, as in Figure 1(a). OT coupling improves source–endpoint ordering, reducing the conflicting directions illustrated in Figure 1(b) Villani (2009); Tong et al. (2023); Cheng & Schwing (2025). Second, reflow straightens learned trajectories, as in Figure 1(c–d), to facilitate numerical generation Labs et al. (2025); Liu et al. (2023); Figure 1(e) illustrates the repeated tangent approximations needed along a curved path. These training interpolants should not be confused with solutions of a fixed conditional ODE: under forward/backward uniqueness, two distinct solution trajectories of the same field cannot meet at the same time Villani (2009). Together, these observations motivate controlling the state–velocity relationships introduced by unlearning.

Diffusion-Style Concept Unlearning. Most T2I concept-erasure methods were developed for diffusion models rather than FM models. Given the target concept c_f, they update the original model θ_* using an anchor prediction for *unlearning direction* such as $\Delta_\epsilon = \epsilon_{\theta_*}(z_t,t,c_a) - \epsilon_{\theta_*}(z_t,t,c_f)$, where c_a is an anchor concept or the null condition $\varnothing$, and ϵ is the predicted noise. A literal transfer to FM replaces ϵ with a velocity prediction v at sampled states. This substitution alone can induce a supervision conflict with the original model's flow structure during unlearning as we analyze in §4.

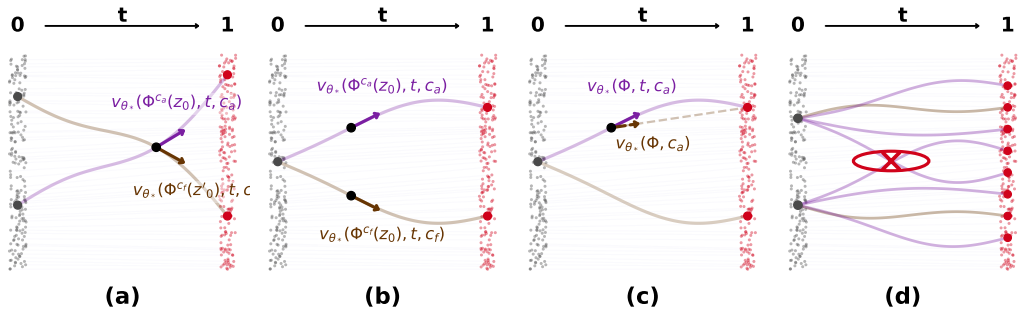

Figure 2: **Four geometric pitfalls in FM unlearning supervision.** **(a)** Cross-source replacement discards the source correspondence (**P1**). **(b)** Querying the anchor field on the forget path fails to follow the intended same-source anchor path (**P2**). **(c)** An endpoint displacement need not match the instantaneous teacher tangent (**P3**). **(d)** Individually coherent paths can request conflicting velocities at nearby states at a common time (**P4**).

4 Theoretical Analysis: Flow-aware FM Unlearning

In this section, we theoretically revisit concept unlearning from a flow-aware perspective. We first introduce the base assumptions. Next, in § 4.1–4.3, we analyze geometric mismatches that can arise when constructing velocity supervision for FM unlearning and derive four design principles, **P1–P4**. Due to space limits, the justifications, assumption details, and proofs are provided in Appendix A.

Assumption 1 (Finite energy) *All source, endpoint, and supervision-velocity random variables used below have finite second moments. For full-time path bounds, teacher trajectories are absolutely continuous and their tangents satisfy* $\mathbb{E}\int_0^1 \|v_{\theta_*}(\Phi^{c_a}_{\theta_*,t}(Z_0), t, c_a)\|_2^2\, dt < \infty$.

Assumption 2 (Local regularity and well-posedness of flow fields) *The original and edited fields are measurable in time and locally Lipschitz in state on a common region containing each compared path, with* $\|v(x,t,c) - v(y,t,c)\|_2 \le L_c\|x-y\|_2$. *The required solutions exist uniquely on* $[0,1]$, *including forward/backward anchor continuations from the queried states.*

Assumption 3 (Independent cross-source replacement) *In § 4.1, a fraction* $p \in [0,1]$ *of edited examples pairs a source* $Z \sim \mu_0$ *with a safe endpoint* $Y_a \sim \nu_a$ *drawn independently of* Z, *where* ν_a *denotes the safe endpoint distribution.*

Assumption 4 (Finite-resolution field) *In § 4.3, a finite nondecreasing* ψ *with* $\psi(0) = 0$ *bounds the forget-conditioned fields by* $\|v(z,t,c_f) - v(z',t,c_f)\|_2 \le \psi(\|z-z'\|_2)$ *on the relevant support.*

4.1 How Cross-Source Replacement Disrupts Source Correspondence

A direct diffusion-to-FM unlearning transfer, via velocity substitution, can disrupt source correspondence when supervision is constructed without preserving the source–endpoint pairing. Specifically, a direct replacement construction can independently pair a source z_0 with a safe endpoint z_1^a. Then we analyze how such cross-seed replacement disrupts source–endpoint ordering as illustrated in Figure 2(a). According to Assumption 3, the new coupling for unlearning π_u is a mixture between the original model-induced coupling π_* and an independent coupling between μ_0 and ν_a, *i.e.*, $\pi_u = (1-p)\pi_* + p(\mu_0 \otimes \nu_a)$. Then we have the following theorem:

Theorem 1 (Cross-source replacement induces two-cycle ordering defects) *For two independent draws* $(Z,Y),(Z',Y') \sim \pi_u$, *let* $S = \langle Z - Z', Y - Y'\rangle$. *Separately draw* $\bar{Z}, \bar{Z}' \sim \mu_0$ *and* $B, B' \sim \nu_a$, *all mutually independent, and define* $S_a = \langle \bar{Z} - \bar{Z}', B - B'\rangle$, $\omega_a = \mathbb{P}(S_a = 0)$, *and* $M_a = \mathbb{E}|S_a|$. *Under Assumptions 1 and 3,*

$$\mathbb{P}(S < 0) \ge \frac{p^2}{2}(1-\omega_a), \qquad \mathbb{E}[\max\{-S, 0\}] \ge \frac{p^2}{2} M_a. \tag{1}$$

Well-structured flows, *e.g.*, quadratic-cost OT flows Villani (2009); Tong et al. (2023), hold two-cycle ordering condition $S \ge 0$ almost surely. However, cross-seed replacement produces ordering reversals

($S < 0$) with probability at least $p^2(1-\omega_a)/2$, so at least this fraction of independently sampled source-endpoint pairs exhibits a two-cycle ordering defect; moreover, the expected ordering-defect magnitude $\mathbb{E}[\max\{-S, 0\}]$ is at least $p^2 M_a/2$, quantifying the average severity of the disrupted source–endpoint ordering (Appendix A.2). To avoid this cross-seed ordering disruption, the direct solution is **P1**: preserve source correspondence by generating each source's safe endpoint with the same seed.

4.2 Why Same-Seed Regression Can Still Cause Velocity-Supervision Shift

Same-seed unlearning alone still does not guarantee preserving the intended teacher trajectory. There can still be support-induced and tangent-induced supervision mismatches, where the supervised states or velocities differ from those of the intended anchor trajectory. Supposing **P1** holds, when unlearning c_f, we first sample $z_0 \sim \mu_0$ and then construct the same-seed trajectories: the safe anchor trajectory $\Phi^{c_a}_{\theta_*,t}(z_0)$, abbreviated as $A_t(z_0)$, and the target trajectory $\Phi^{c_f}_{\theta_*,t}(z_0)$, abbreviated as $F_t(z_0)$. We define three quantities to characterize this support mismatch (suppressing the common source argument z_0 below): the bridge width $g_t := \|F_t - A_t\|_2$ between the two trajectories; the *implicit anchor source* $\widetilde{z}_t := \Phi^{c_a}_{\theta_*,t\to 0}(F_t)$; and the *implicit anchor endpoint* $\widetilde{A}_{1|t} := \Phi^{c_a}_{\theta_*,t\to 1}(F_t)$.

Proposition 1 (Target-support supervision induces source and endpoint shifts) *Under Assumption 2, let L_a bound the anchor field's spatial Lipschitz constant on the comparison region. Sampling z_t along F_t gives:*

$$e^{-L_a t} g_t \le \|\widetilde{z}_t - z_0\|_2 \le e^{L_a t} g_t, \quad e^{-L_a(1-t)} g_t \le \|\widetilde{A}_{1|t} - A_1\|_2 \le e^{L_a(1-t)} g_t. \tag{2}$$

Whenever the bridge width $g_t > 0$, this proposition shows that, under the anchor field, F_t traces back to a different source, $\widetilde{z}_t \neq z_0$, and continues to a shifted endpoint, $\widetilde{A}_{1|t} \neq A_1$. This motivates **P2**: sample the unlearning latents along the intended anchor trajectory.

Proposition 2 (Chord-induced tangent mismatch) *Let $\bar{A}_t := z_0 + t(A_1 - z_0)$ denote the same-source endpoint chord. Under Assumptions 1 and 2, for μ_0-a.e. source z_0 and every $t \in [0,1]$,*

$$A_t - \bar{A}_t = -\int_0^t r_\tau^{\tan}\, d\tau, \quad \sup_{t\in[0,1]} \|A_t - \bar{A}_t\|_2^2 \le \int_0^1 \|r_t^{\tan}\|_2^2\, dt, \tag{3}$$

where $r_t^{\tan} := A_1 - z_0 - v_{\theta_*}(A_t, t, c_a)$ denotes the tangent–chord velocity residual along A_t. The residual vanishes almost everywhere in time iff the anchor trajectory is straight with constant velocity.

Unless the model-induced anchor trajectory is straight with constant velocity ($A_t = \bar{A}_t$ for all t), the velocity residual $r_t^{\tan}$ is nonzero along part of the trajectory, meaning that chord velocity supervision differs from teacher-tangent supervision. Therefore, although we train θ_* using $v_t = z_1 - z_0$, chord velocity $A_1 - z_0$ is not guaranteed to be compatible with the learned anchor trajectory. Appendix A.4 further shows that the reconstructed chord need not be an integral curve of the learned flow, motivating **P3**: regress toward the model-induced tangent velocity.

4.3 Why Individually Safe Paths Can Still Be Jointly Incompatible

P1–P3 make each anchor trajectory individually coherent. However, the forget-conditioned field $v_{\theta_u}(\cdot,\cdot,c_f)$ is trained with multiple anchor paths without receiving their anchor indices, so two such trajectories can still request locally incompatible velocities. At a sampled common time T, we draw two latent–velocity supervision pairs (Z, V) and (Z', V') independently from the same conditional law. Let $\mathcal{V}_\psi$ denote the nonempty class of forget-conditioned velocity fields satisfying Assumption 4. We define $q_{\mathrm{conf}}(r,\delta) := \mathbb{P}(\|Z - Z'\|_2 \le r, \|V - V'\|_2 \ge \delta)$ to measure local conflict, which is the probability mass of nearby supervision states with separated velocity targets. Then we have:

Theorem 2 (Velocity conflict imposes a regression floor) *Under Assumptions 1 and 4, we have*

$$\inf_{v\in\mathcal{V}_\psi} \mathbb{E}_{T,(Z,V)}\big[\|v(Z,T,c_f) - V\|_2^2\big] \ge \frac{q_{\mathrm{conf}}(r,\delta)}{4} \max\{\delta - \psi(r), 0\}^2, \quad \forall r \ge 0,\ \delta > 0. \tag{4}$$

If the requested velocity separation exceeds the field's finite-scale variation budget, $\delta > \psi(r)$, and $q_{\text{conf}}(r, \delta) > 0$, then no field in $\mathcal{V}_\psi$ can fit all supervision with zero regression error. This theorem shows that the union of individually coherent anchor velocities can become jointly unrealizable within $\mathcal{V}_\psi$. Appendix A further analyzes the regression floor and its connection to SGD noise. To this end, we propose TRACE, which jointly assigns anchors across seeds following **P4**: per-source anchor choices should be jointly velocity-compatible.

5 TRACE : FLOW-AWARE CONCEPT UNLEARNING

Given a target concept c_f, TRACE follows **P4** to jointly select anchor trajectories and then applies **P1–P3** to construct their supervision. We seek an anchor assignment $\boldsymbol{q} = (q_1, \ldots, q_N)$ for N seeds from candidates c_a^q, $q = 0, 1, \ldots, Q$. We set $c_a^0 = \varnothing$ as the null condition used in common null-anchor unlearning baselines (Gandikota et al., 2023; Han et al., 2024). In the *pre-unlearning stage*, we sample seeds $z_{0,i}$, generate $\mathcal{O}(N(Q+1))$ teacher trajectories, and determine $\boldsymbol{q}$. In the *unlearning-updating stage*, we update the model along the assigned paths. Write $A_t^{(q)}(z_{0,i}) = \Phi_{\theta_*,t}^{c_a^q}(z_{0,i})$ and $a_t^{(q)}(z_{0,i}) = v_{\theta_*}(A_t^{(q)}(z_{0,i}), t, c_a^q)$ for each candidate path and its tangent.

Pre-unlearning Stage. We first introduce two constraints for anchor assignment. **(I) Safe Anchors:** We filter out anchors that fail to suppress the target concept c_f. We encode prompts containing each anchor concept c_a^q into text embeddings $\boldsymbol{e}_a^q$ by model θ_*'s encoder (usually certain CLIP model for modern FM). If the cosine similarity satisfies $\cos(\boldsymbol{e}_a^q, \boldsymbol{e}_f) > \cos(\boldsymbol{e}_a^0, \boldsymbol{e}_f)$, this anchor c_a^q is filtered out. $\mathcal{Q}_i$ denotes the resulting anchor bank available to seed $z_{0,i}$. **(II) Utility Constraint:** We measure how well each anchor endpoint preserves non-target content. For same-seed target and anchor images, we extract CLIP image features and project them out of the target-text subspace. Let d_i^q be the L2-norm of the projected feature difference, measuring anchor c_a^q's context distortion. For an assignment $\boldsymbol{q}$, we define $d(\boldsymbol{q}) = N^{-1} \sum_{i=1}^N d_i^{q_i}$. A smaller $d(\boldsymbol{q})$ implies better utility preservation.

Then we define the local path conflict $C(\boldsymbol{q})$ as a finite-bank proxy for the q_{conf} in Eq.(4) as follows:

$$C(\boldsymbol{q}) = \mathbb{E}_{i,j,t}\left[\frac{\|a_t^{(q_i)}(z_{0,i}) - a_t^{(q_j)}(z_{0,j})\|_2}{\|A_t^{(q_i)}(z_{0,i}) - A_t^{(q_j)}(z_{0,j})\|_2 + \epsilon}\right], \tag{5}$$

where ϵ is a small constant to avoid division by zero. $C(\boldsymbol{q})$ measures how sharply the teacher velocity changes between nearby selected trajectory states, serving as a finite-bank proxy for whether safe paths can be jointly represented by one forget-conditioned velocity field. We optimize $\boldsymbol{q}$ as follows:

$$\arg\min_{\boldsymbol{q}} C(\boldsymbol{q}) \quad \text{s.t.} \quad q_i \in \mathcal{Q}_i \ \ \forall i, \quad \text{and} \quad d(\boldsymbol{q}) \le d(\mathbf{0}), \tag{6}$$

where $\mathbf{0}$ refers to a baseline anchor assignment where all $q_i = 0$. This formulation seamlessly decouples *which safe trajectory to assign* from *how to supervise a source*. The approximately optimized $\boldsymbol{q}^*$ targets **P4**, providing a trajectory-coherent anchor assignment from the perspective of flow geometry. The anchor assignment is frozen during the subsequent parameter-updating stage.

Unlearning-updating Stage. With $\boldsymbol{q}^*$ fixed, we initialize θ_u from the frozen checkpoint θ_*. For each seed $z_{0,i}$, we sample a time t on its assigned teacher trajectory under $c_a^{q_i^*}$ and train the student under c_f to match the corresponding anchor tangent:

$$\mathcal{L}_{\text{removal}}(\theta_u) = \mathbb{E}_{i,t}\left[\left\| v_{\theta_u}\left(A_t^{(q_i^*)}(z_{0,i}), t, c_f\right) - a_t^{(q_i^*)}(z_{0,i})\right\|_2^2\right]. \tag{7}$$

This applies **P1–P3** to the selected safe trajectories. Following prior unlearning methods (Gao et al., 2025; Grebe et al., 2026; Simone et al., 2025), the total objective $\mathcal{L}_{\text{removal}} + \lambda \mathcal{L}_{\text{r}}$ contains the retention loss on non-target prompts $\mathcal{L}_{\text{r}} = \|v_{\theta_u}(z_t, t, c) - v_{\theta_*}(z_t, t, c)\|$, where λ controls its weight.[1]

6 EVALUATION

Experiment Setup. Following current literature, we evaluate TRACE on two modern FM models, SD3.5 Medium and FLUX.2 4B. The baselines include state-of-the-art FM unlearning methods

[1]The main text focuses on algorithm design. Implementation details, including candidate construction, feature encoders, time sampling, LoRA updating, and the assignment solver, are provided in Appendix B.1.

(GEM Grebe et al. (2026) and EA Gao et al. (2025)) and representative T2I unlearning methods (ESD Gandikota et al. (2023), ConAbl Kumari et al. (2023), and FMN Zhang et al. (2024a)). The evaluation metrics assess removal effectiveness (CLIP(U) and Anchor-win) and utility preservation (CLIP(R) and FID). A lower CLIP(U) score indicates weaker alignment with the target concept, while a higher CLIP(R) score indicates better alignment on non-target prompts. Anchor-win measures the percentage of generated images that align more with safe anchor prompts than with the target concept; higher is better. FID Heusel et al. (2017) measures the distributional difference from the COCO30k real-image reference (COCO 2017 training split); lower is better. Each experiment is repeated over 5 random seeds, and we report the mean and relative standard deviation $\mu_{\pm\sigma/|\mu|}$. Table annotations express the relative standard deviation as a percentage. More details are in Appendix B.2.

We divide unlearning concepts into objectives (object/identity concepts) and adjectives (attribute/style). In this section, for objectives, we use Mickey Mouse and Albert Einstein as representatives; for adjectives, we use Aggressive and Van Gogh style. Results for more concepts are provided in Appendix B.5. We also evaluate TRACE's sensitivity to hyper-parameters and the results are reported in Appendix B.6, confirming its robustness under the number of source seeds N and candidate anchors Q.

6.1 Concept Unlearning: Effectiveness and Utility

Table 1 reports the external baselines and progressive ablations of TRACE.

Enhanced Concept Removal. TRACE outperforms all external baselines on both CLIP(U) and Anchor-win across the eight concept–backbone settings. Relative to the best external baseline for each metric, it reduces CLIP(U) by 0.67–7.10% and improves Anchor-win by 4.74–16.13 percentage points. For Mickey Mouse, it achieves Anchor-win rates of 86.17% on SD3.5M and 91.95% on FLUX, compared with GEM's 72.64% and 77.73%.

Improved Model Utility. TRACE also achieves the highest CLIP(R) and lowest FID relative to the external baselines in every setting. It reduces the baseline-induced CLIP(R) drop by 42.91–83.33% and FID degradation by 31.76–79.66%. These percentages measure the fraction of the baseline's degradation relative to the original model that TRACE recovers. For example, on Albert Einstein with SD3.5M, the CLIP(R) drop reduction is $(29.82 - 28.17)/(30.47 - 28.17) = 71.74\%$. Appendix B.2 gives more details for all reported improvements. The advantage over external baselines mostly holds in all 22 additional concept–backbone settings (Appendix B.5). Appendix B.6 reports small changes around $N = 100$–300 and modest utility gains from increasing Q beyond 10.

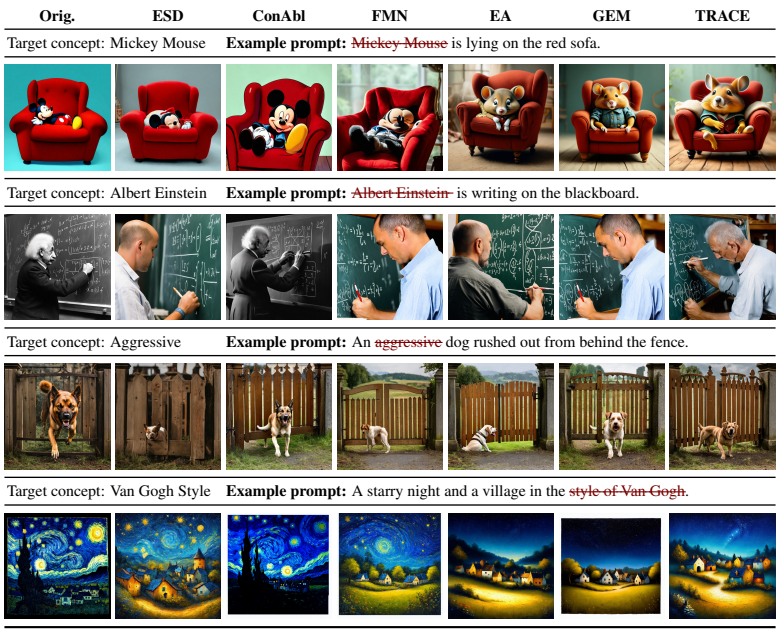

Figure 3: Qualitative comparison on SD3.5M for four representative concepts.

Table 1: Quantitative evaluation. We also report the results of the original unedited model (Orig) for reference. Except for Orig, we highlight the **best** and the second-best results via bold and underline.

Method	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
Object/identity Unlearning: Mickey Mouse								
Orig	$30.13_{\pm 0.7\%}$	$3.05_{\pm 4.8\%}$	$30.47_{\pm 1.2\%}$	$12.72_{\pm 1.1\%}$	$28.21_{\pm 0.8\%}$	$5.68_{\pm 3.7\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$19.72_{\pm 0.6\%}$	$42.60_{\pm 1.9\%}$	$24.34_{\pm 0.9\%}$	$22.38_{\pm 1.3\%}$	$19.35_{\pm 0.9\%}$	$47.48_{\pm 2.2\%}$	$22.61_{\pm 1.0\%}$	$21.08_{\pm 1.6\%}$
ConAbl	$22.59_{\pm 0.8\%}$	$36.25_{\pm 2.8\%}$	$27.09_{\pm 1.1\%}$	$18.84_{\pm 1.4\%}$	$22.75_{\pm 0.5\%}$	$39.38_{\pm 1.7\%}$	$24.29_{\pm 1.7\%}$	$18.90_{\pm 0.9\%}$
FMN	$21.14_{\pm 0.9\%}$	$39.89_{\pm 2.1\%}$	$26.31_{\pm 1.5\%}$	$20.36_{\pm 1.2\%}$	$20.47_{\pm 0.7\%}$	$41.59_{\pm 2.9\%}$	$23.44_{\pm 0.8\%}$	$19.20_{\pm 1.0\%}$
EA	$21.38_{\pm 0.6\%}$	$68.63_{\pm 1.6\%}$	$27.81_{\pm 1.3\%}$	$17.59_{\pm 1.8\%}$	$20.51_{\pm 1.0\%}$	$63.53_{\pm 2.5\%}$	$25.90_{\pm 1.6\%}$	$16.22_{\pm 1.1\%}$
GEM	$19.21_{\pm 0.8\%}$	$72.64_{\pm 2.3\%}$	$28.44_{\pm 0.7\%}$	$16.37_{\pm 1.5\%}$	$\underline{19.17}_{\pm 0.5\%}$	$\underline{77.73}_{\pm 1.8\%}$	$\underline{26.63}_{\pm 1.2\%}$	$14.55_{\pm 1.4\%}$
Native	$22.16_{\pm 0.7\%}$	$37.11_{\pm 2.6\%}$	$26.79_{\pm 1.0\%}$	$20.40_{\pm 1.7\%}$	$22.92_{\pm 0.9\%}$	$40.06_{\pm 2.0\%}$	$25.51_{\pm 1.4\%}$	$19.90_{\pm 1.3\%}$
+P1	$22.56_{\pm 0.5\%}$	$67.55_{\pm 1.5\%}$	$27.63_{\pm 1.8\%}$	$17.43_{\pm 1.2\%}$	$21.14_{\pm 0.8\%}$	$61.14_{\pm 2.7\%}$	$25.49_{\pm 0.9\%}$	$16.47_{\pm 1.6\%}$
+P1P2	$21.73_{\pm 0.8\%}$	$71.10_{\pm 2.1\%}$	$28.27_{\pm 1.4\%}$	$16.98_{\pm 1.4\%}$	$21.85_{\pm 0.7\%}$	$66.33_{\pm 2.1\%}$	$25.61_{\pm 1.6\%}$	$15.99_{\pm 1.3\%}$
+P1P2P3	$\underline{18.76}_{\pm 0.9\%}$	$\underline{80.64}_{\pm 2.2\%}$	$\underline{29.33}_{\pm 1.1\%}$	$\underline{14.76}_{\pm 1.5\%}$	$19.98_{\pm 0.6\%}$	$75.02_{\pm 1.9\%}$	$26.58_{\pm 1.7\%}$	$\underline{14.26}_{\pm 1.0\%}$
TRACE	$\mathbf{18.55}_{\pm 1.0\%}$	$\mathbf{86.17}_{\pm 2.8\%}$	$\mathbf{29.69}_{\pm 1.3\%}$	$\mathbf{14.11}_{\pm 1.4\%}$	$\mathbf{18.42}_{\pm 0.7\%}$	$\mathbf{91.95}_{\pm 2.1\%}$	$\mathbf{28.07}_{\pm 1.5\%}$	$\mathbf{13.35}_{\pm 1.2\%}$
Object/identity Unlearning: Albert Einstein								
Orig	$30.05_{\pm 0.3\%}$	$4.21_{\pm 4.5\%}$	$30.47_{\pm 1.1\%}$	$12.72_{\pm 1.5\%}$	$28.56_{\pm 0.8\%}$	$6.13_{\pm 3.2\%}$	$28.80_{\pm 0.9\%}$	$11.95_{\pm 1.2\%}$
ESD	$19.82_{\pm 0.7\%}$	$43.22_{\pm 2.4\%}$	$24.15_{\pm 1.4\%}$	$22.60_{\pm 1.1\%}$	$19.71_{\pm 0.6\%}$	$46.31_{\pm 2.1\%}$	$22.85_{\pm 1.7\%}$	$20.92_{\pm 1.8\%}$
ConAbl	$22.11_{\pm 0.6\%}$	$34.88_{\pm 1.9\%}$	$27.36_{\pm 0.5\%}$	$19.13_{\pm 1.7\%}$	$23.20_{\pm 1.1\%}$	$40.74_{\pm 4.0\%}$	$24.06_{\pm 1.2\%}$	$18.61_{\pm 1.4\%}$
FMN	$20.83_{\pm 0.8\%}$	$40.75_{\pm 2.8\%}$	$26.14_{\pm 1.6\%}$	$20.08_{\pm 1.0\%}$	$20.12_{\pm 0.9\%}$	$42.13_{\pm 3.8\%}$	$23.79_{\pm 0.6\%}$	$19.47_{\pm 1.3\%}$
EA	$20.92_{\pm 0.9\%}$	$67.19_{\pm 2.1\%}$	$27.56_{\pm 0.8\%}$	$17.86_{\pm 1.6\%}$	$20.06_{\pm 0.7\%}$	$65.28_{\pm 2.5\%}$	$25.61_{\pm 1.5\%}$	$16.54_{\pm 1.1\%}$
GEM	$19.67_{\pm 0.6\%}$	$71.32_{\pm 1.6\%}$	$28.17_{\pm 1.7\%}$	$16.62_{\pm 1.2\%}$	$19.44_{\pm 0.5\%}$	$\underline{76.45}_{\pm 1.9\%}$	$26.34_{\pm 1.0\%}$	$15.33_{\pm 1.5\%}$
Native	$22.48_{\pm 0.8\%}$	$36.06_{\pm 2.3\%}$	$26.92_{\pm 0.6\%}$	$19.71_{\pm 1.8\%}$	$23.37_{\pm 1.0\%}$	$39.24_{\pm 2.9\%}$	$24.18_{\pm 1.3\%}$	$19.23_{\pm 1.7\%}$
+P1	$22.09_{\pm 0.7\%}$	$68.93_{\pm 2.6\%}$	$27.91_{\pm 1.3\%}$	$17.15_{\pm 1.4\%}$	$21.55_{\pm 0.6\%}$	$60.02_{\pm 3.5\%}$	$25.83_{\pm 1.8\%}$	$16.78_{\pm 1.0\%}$
+P1P2	$21.35_{\pm 0.6\%}$	$71.00_{\pm 1.3\%}$	$28.12_{\pm 1.4\%}$	$16.67_{\pm 1.3\%}$	$21.22_{\pm 0.4\%}$	$67.08_{\pm 2.6\%}$	$25.96_{\pm 0.8\%}$	$16.11_{\pm 1.3\%}$
+P1P2P3	$\underline{18.91}_{\pm 0.5\%}$	$\underline{79.37}_{\pm 1.5\%}$	$\underline{29.06}_{\pm 1.5\%}$	$\underline{14.98}_{\pm 1.1\%}$	$\underline{19.31}_{\pm 0.8\%}$	$76.18_{\pm 2.2\%}$	$\underline{26.91}_{\pm 0.7\%}$	$\underline{14.02}_{\pm 1.6\%}$
TRACE	$\mathbf{18.36}_{\pm 0.9\%}$	$\mathbf{87.45}_{\pm 2.2\%}$	$\mathbf{29.82}_{\pm 0.9\%}$	$\mathbf{13.87}_{\pm 1.3\%}$	$\mathbf{18.06}_{\pm 1.1\%}$	$\mathbf{90.66}_{\pm 4.1\%}$	$\mathbf{28.39}_{\pm 1.4\%}$	$\mathbf{13.62}_{\pm 1.8\%}$
Adjective Concepts Unlearning: Aggressive								
Orig	$31.02_{\pm 0.3\%}$	$2.11_{\pm 4.7\%}$	$30.47_{\pm 1.2\%}$	$12.72_{\pm 1.6\%}$	$29.07_{\pm 0.9\%}$	$2.13_{\pm 2.6\%}$	$28.80_{\pm 0.8\%}$	$11.95_{\pm 1.1\%}$
ESD	$20.63_{\pm 0.7\%}$	$45.31_{\pm 2.4\%}$	$23.52_{\pm 0.7\%}$	$23.67_{\pm 1.4\%}$	$20.86_{\pm 0.6\%}$	$44.28_{\pm 3.1\%}$	$21.96_{\pm 1.6\%}$	$22.58_{\pm 1.7\%}$
ConAbl	$22.48_{\pm 0.6\%}$	$38.17_{\pm 1.9\%}$	$26.93_{\pm 1.8\%}$	$20.16_{\pm 1.1\%}$	$22.09_{\pm 0.5\%}$	$39.06_{\pm 3.7\%}$	$23.41_{\pm 1.1\%}$	$19.76_{\pm 1.5\%}$
FMN	$21.35_{\pm 0.8\%}$	$40.89_{\pm 2.8\%}$	$26.41_{\pm 1.0\%}$	$21.72_{\pm 1.7\%}$	$21.16_{\pm 1.0\%}$	$41.23_{\pm 2.4\%}$	$23.08_{\pm 1.2\%}$	$20.19_{\pm 1.2\%}$
EA	$20.88_{\pm 0.9\%}$	$70.62_{\pm 2.1\%}$	$26.07_{\pm 0.5\%}$	$17.83_{\pm 1.0\%}$	$21.43_{\pm 0.8\%}$	$61.87_{\pm 3.0\%}$	$25.82_{\pm 1.7\%}$	$16.75_{\pm 1.4\%}$
GEM	$20.56_{\pm 0.6\%}$	$73.46_{\pm 1.6\%}$	$28.21_{\pm 1.6\%}$	$15.62_{\pm 1.3\%}$	$20.52_{\pm 0.7\%}$	$76.31_{\pm 2.8\%}$	$26.47_{\pm 0.9\%}$	$15.44_{\pm 1.8\%}$
Native	$22.74_{\pm 0.8\%}$	$40.72_{\pm 2.3\%}$	$27.12_{\pm 1.4\%}$	$20.06_{\pm 1.5\%}$	$22.31_{\pm 1.1\%}$	$38.61_{\pm 1.9\%}$	$23.35_{\pm 1.5\%}$	$19.12_{\pm 1.0\%}$
+P1	$21.09_{\pm 0.7\%}$	$68.15_{\pm 2.6\%}$	$26.85_{\pm 1.1\%}$	$17.24_{\pm 1.2\%}$	$21.68_{\pm 0.9\%}$	$68.42_{\pm 3.9\%}$	$25.31_{\pm 1.0\%}$	$16.08_{\pm 1.6\%}$
+P1P2	$20.54_{\pm 0.5\%}$	$73.53_{\pm 1.6\%}$	$27.33_{\pm 0.2\%}$	$16.74_{\pm 1.2\%}$	$20.35_{\pm 0.6\%}$	$72.66_{\pm 3.0\%}$	$25.88_{\pm 1.4\%}$	$15.92_{\pm 1.1\%}$
+P1P2P3	$\underline{19.81}_{\pm 0.5\%}$	$\underline{82.97}_{\pm 1.5\%}$	$\underline{29.14}_{\pm 0.8\%}$	$\underline{13.42}_{\pm 1.8\%}$	$\mathbf{18.97}_{\pm 0.6\%}$	$\mathbf{81.36}_{\pm 3.4\%}$	$\underline{26.79}_{\pm 1.3\%}$	$\underline{14.38}_{\pm 1.1\%}$
TRACE	$\mathbf{19.52}_{\pm 0.9\%}$	$\mathbf{85.73}_{\pm 2.2\%}$	$\mathbf{29.76}_{\pm 1.5\%}$	$\mathbf{13.31}_{\pm 1.4\%}$	$\underline{19.41}_{\pm 0.5\%}$	$\underline{81.05}_{\pm 2.0\%}$	$\mathbf{27.64}_{\pm 1.8\%}$	$\mathbf{13.87}_{\pm 1.7\%}$
Adjective Concepts Unlearning: Van Gogh Style								
Orig	$30.92_{\pm 0.9\%}$	$6.95_{\pm 3.7\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.58_{\pm 1.0\%}$	$7.66_{\pm 3.6\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$19.36_{\pm 0.7\%}$	$44.02_{\pm 2.4\%}$	$23.76_{\pm 0.6\%}$	$23.24_{\pm 1.6\%}$	$19.94_{\pm 0.7\%}$	$45.96_{\pm 2.5\%}$	$21.75_{\pm 1.1\%}$	$22.20_{\pm 1.0\%}$
ConAbl	$21.92_{\pm 0.6\%}$	$39.06_{\pm 1.9\%}$	$27.01_{\pm 1.1\%}$	$19.90_{\pm 1.8\%}$	$22.56_{\pm 0.8\%}$	$37.92_{\pm 3.2\%}$	$23.77_{\pm 0.8\%}$	$19.32_{\pm 1.3\%}$
FMN	$21.89_{\pm 0.8\%}$	$42.24_{\pm 2.8\%}$	$26.00_{\pm 1.4\%}$	$20.95_{\pm 1.1\%}$	$21.78_{\pm 0.6\%}$	$39.77_{\pm 2.1\%}$	$23.30_{\pm 1.6\%}$	$20.57_{\pm 1.7\%}$
EA	$21.35_{\pm 0.9\%}$	$71.89_{\pm 2.1\%}$	$26.26_{\pm 1.5\%}$	$17.58_{\pm 1.4\%}$	$20.92_{\pm 1.1\%}$	$60.51_{\pm 3.8\%}$	$26.16_{\pm 1.0\%}$	$17.04_{\pm 1.8\%}$
GEM	$20.13_{\pm 0.6\%}$	$74.90_{\pm 1.6\%}$	$27.94_{\pm 0.9\%}$	$15.89_{\pm 1.5\%}$	$19.97_{\pm 0.5\%}$	$75.07_{\pm 2.9\%}$	$26.19_{\pm 1.7\%}$	$15.13_{\pm 1.2\%}$
Native	$22.09_{\pm 0.8\%}$	$39.56_{\pm 2.3\%}$	$26.89_{\pm 1.2\%}$	$20.32_{\pm 1.0\%}$	$22.87_{\pm 0.9\%}$	$39.43_{\pm 4.0\%}$	$24.01_{\pm 0.5\%}$	$19.49_{\pm 1.4\%}$
+P1	$21.53_{\pm 0.7\%}$	$69.44_{\pm 2.6\%}$	$27.14_{\pm 1.8\%}$	$16.96_{\pm 1.3\%}$	$21.14_{\pm 0.6\%}$	$67.14_{\pm 1.9\%}$	$25.69_{\pm 1.2\%}$	$16.47_{\pm 1.1\%}$
+P1P2	$20.82_{\pm 0.6\%}$	$71.93_{\pm 1.7\%}$	$27.78_{\pm 1.1\%}$	$15.31_{\pm 1.4\%}$	$21.33_{\pm 1.0\%}$	$69.33_{\pm 2.8\%}$	$25.85_{\pm 0.9\%}$	$15.47_{\pm 1.3\%}$
+P1P2P3	$\underline{19.26}_{\pm 0.5\%}$	$\underline{84.22}_{\pm 1.5\%}$	$\underline{28.87}_{\pm 0.8\%}$	$\mathbf{13.63}_{\pm 1.7\%}$	$\mathbf{19.33}_{\pm 1.0\%}$	$\underline{80.15}_{\pm 3.5\%}$	$\underline{27.06}_{\pm 1.5\%}$	$\underline{14.75}_{\pm 1.6\%}$
TRACE	$\mathbf{19.23}_{\pm 0.9\%}$	$\mathbf{84.41}_{\pm 2.2\%}$	$\mathbf{29.05}_{\pm 1.3\%}$	$\underline{14.09}_{\pm 1.2\%}$	$\underline{19.55}_{\pm 0.8\%}$	$\mathbf{80.28}_{\pm 2.7\%}$	$\mathbf{27.31}_{\pm 0.9\%}$	$\mathbf{14.12}_{\pm 1.5\%}$

Figure 3 illustrates the removal–utility trade-off. For Mickey Mouse, ESD, ConAbl, and FMN retain distinctive round ears and cartoon features. In the aggressive-dog example, EA turns the dog away and GEM changes its forward-running pose, altering non-target content along with the target attribute. TRACE suppresses target-specific visual cues while retaining the red armchair in the first example and the dog-and-fence composition in the second.

6.2 Ablation Study

We progressively incorporate the four design principles using the variants in Table 1.

Native adaptation denotes direct diffusion-to-FM transfer with cross-source supervision, rather than the P1 design. Its Anchor-win rates for Mickey Mouse are 37.11% on SD3.5M and 40.06% on FLUX, with substantial utility degradation.

Flow-aware supervision. Introducing source correspondence (+P1) raises Anchor-win rates to 67.55% and 61.14%, respectively. Adding safe-path support (+P1P2) further increases Anchor-win

Table 2: Multiple concepts unlearning on SD 3.5M. The results on FLUX are in Appendix B.3.

Method	Unlearning 5 Concepts				Unlearning 10 Concepts				Unlearning 15 Concepts			
	CLIP(U)↓	A.W.(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	A.W.(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	A.W.(%)↑	CLIP(R)↑	FID↓
Orig	$30.35_{\pm 0.6\%}$	$3.09_{\pm 4.2\%}$	$30.47_{\pm 1.2\%}$	$12.72_{\pm 1.1\%}$	$30.53_{\pm 0.4\%}$	$3.44_{\pm 4.1\%}$	$30.47_{\pm 1.2\%}$	$12.72_{\pm 1.1\%}$	$30.50_{\pm 0.9\%}$	$3.18_{\pm 4.8\%}$	$30.47_{\pm 1.2\%}$	$12.72_{\pm 1.1\%}$
ESD	$22.48_{\pm 0.6\%}$	$33.11_{\pm 1.9\%}$	$24.24_{\pm 0.9\%}$	$22.31_{\pm 1.3\%}$	$22.66_{\pm 0.9\%}$	$31.50_{\pm 2.1\%}$	$24.17_{\pm 0.8\%}$	$22.85_{\pm 1.5\%}$	$23.15_{\pm 1.4\%}$	$29.50_{\pm 2.5\%}$	$24.12_{\pm 1.3\%}$	$23.01_{\pm 1.8\%}$
ConAbl	$23.99_{\pm 0.8\%}$	$29.94_{\pm 2.3\%}$	$26.35_{\pm 1.1\%}$	$20.25_{\pm 1.4\%}$	$23.75_{\pm 0.7\%}$	$27.37_{\pm 2.9\%}$	$26.02_{\pm 1.3\%}$	$21.00_{\pm 1.6\%}$	$24.05_{\pm 1.1\%}$	$25.37_{\pm 3.3\%}$	$25.88_{\pm 1.9\%}$	$21.45_{\pm 1.7\%}$
FMN	$23.17_{\pm 0.9\%}$	$32.15_{\pm 2.1\%}$	$25.64_{\pm 1.5\%}$	$21.27_{\pm 1.2\%}$	$23.26_{\pm 0.7\%}$	$26.38_{\pm 2.0\%}$	$24.92_{\pm 1.6\%}$	$21.66_{\pm 1.1\%}$	$24.46_{\pm 0.9\%}$	$24.03_{\pm 2.7\%}$	$24.23_{\pm 1.8\%}$	$22.27_{\pm 1.9\%}$
EA	$22.82_{\pm 0.6\%}$	$59.45_{\pm 1.6\%}$	$27.62_{\pm 1.3\%}$	$17.88_{\pm 1.8\%}$	$23.57_{\pm 0.7\%}$	$52.68_{\pm 1.4\%}$	$26.34_{\pm 1.6\%}$	$18.29_{\pm 1.9\%}$	$23.89_{\pm 1.3\%}$	$49.80_{\pm 1.8\%}$	$25.94_{\pm 2.0\%}$	$18.74_{\pm 2.2\%}$
GEM	$20.44_{\pm 0.8\%}$	$64.29_{\pm 2.0\%}$	$28.46_{\pm 0.7\%}$	$16.66_{\pm 1.4\%}$	$21.03_{\pm 0.8\%}$	$60.00_{\pm 2.4\%}$	$28.29_{\pm 0.6\%}$	$16.90_{\pm 1.3\%}$	$21.23_{\pm 0.9\%}$	$58.12_{\pm 2.3\%}$	$27.99_{\pm 0.7\%}$	$16.98_{\pm 1.6\%}$
TRACE	$\mathbf{19.49}_{\pm 1.0\%}$	$\mathbf{81.18}_{\pm 2.3\%}$	$\mathbf{29.67}_{\pm 1.1\%}$	$\mathbf{14.17}_{\pm 1.2\%}$	$\mathbf{19.98}_{\pm 1.3\%}$	$\mathbf{77.93}_{\pm 2.8\%}$	$\mathbf{29.11}_{\pm 1.3\%}$	$\mathbf{15.09}_{\pm 1.2\%}$	$\mathbf{20.25}_{\pm 1.6\%}$	$\mathbf{74.93}_{\pm 3.1\%}$	$\mathbf{28.73}_{\pm 1.7\%}$	$\mathbf{15.29}_{\pm 1.3\%}$

and CLIP(R) and reduces FID in all eight settings. Combining source correspondence, safe-path support, and instantaneous tangents (+P1P2P3) further improves the removal–utility trade-off. On SD3.5M with Mickey Mouse, this variant achieves 80.64% Anchor-win and 14.76 FID, compared with GEM's 72.64% and 16.37. These support the combined geometry-guided unlearning design.

Complete TRACE. In Table 1, joint anchor assignment (P4) yields its largest Anchor-win gains on the object/identity examples. On FLUX, Anchor-win increases from 75.02% to 91.95% for Mickey Mouse and from 76.18% to 90.66% for Albert Einstein. Attribute/style results exhibit metric-specific trade-offs: on FLUX with Aggressive, +P1P2P3 has better CLIP(U) and Anchor-win, whereas TRACE improves CLIP(R) and FID. On SD3.5M with Van Gogh style, joint assignment improves removal and CLIP(R), but increases FID from 13.63 to 14.09.

6.3 Multiple Concepts Unlearning: Scalability and Robustness

In real-world deployment, generative models often receive successive concept-removal requests rather than isolated single edits. To evaluate our framework under repeated editing, we sequentially unlearn 5, 10, and 15 concepts on SD3.5M, using a mixture of object/identity and attribute/style concepts. Table 2 reports the results; Appendix B.3 provides the corresponding FLUX evaluation.

Sustained Effectiveness and Utility. TRACE maintains the best removal and utility scores among the evaluated methods at each sequence length. After removing 15 concepts, it achieves 74.93% Anchor-win, 28.73 CLIP(R), and 15.29 FID, compared with GEM's 58.12%, 27.99, and 16.98. Increasing the sequence from 5 to 15 concepts lowers TRACE 's Anchor-win by 6.25 percentage points and increases FID by 1.12. GEM changes by 6.17 percentage points and 0.32 FID over the same range, so the advantage is in the attained performance rather than a uniformly smaller degradation rate. Notably, TRACE 's reported Anchor-win and FID after 15 removals remain better than GEM's after only 5 (74.93% versus 64.29%; 15.29 versus 16.66). EA's Anchor-win falls from 59.45% to 49.80%, illustrating a larger loss in removal effectiveness over this sequence.

Comparison with Diffusion-Derived Baselines. ESD, ConAbl, and FMN retain lower removal effectiveness and utility throughout the evaluated sequence. At 15 concepts, their Anchor-win rates range from 24.03% to 29.50%, CLIP(R) from 24.12 to 25.88, and FID from 21.45 to 23.01. In comparison, TRACE retains both a stronger preference for safe alternatives and better non-target generation. These results support the practical benefit of flow-aware supervision under repeated edits, without attributing all performance differences to a single geometric mechanism.

7 Conclusion

In this paper, we study concept unlearning in FM models through the geometry of velocity supervision. Our analysis identifies four design principles: source correspondence, safe-path support, instantaneous tangents, and joint velocity compatibility. Guided by these insights, TRACE jointly selects anchor trajectories using a compatibility proxy under a context constraint, then updates the model along the selected teacher paths. Experiments on two flow-based backbones show a stronger removal–utility trade-off than the compared baselines across the evaluated concepts. Progressive ablations support the supervision principles and two-stage design, while additional-concept and sensitivity experiments characterize the robustness and trade-offs of the framework. Its advantage also persists under sequential concept removal. These results demonstrate the value of constructing and coordinating safe trajectory supervision for FM unlearning.

AI USE STATEMENT

In this work, we used generative AI tools to polish academic writing, check for potentially missing citations, and confirm the implementation code.

ETHICS STATEMENT

TRACE is intended to support authorized concept removal and responsible maintenance of text-to-image models. Decisions about which concepts to remove should account for legitimate uses and the risk of suppressing unrelated content. Our evaluation considers non-target utility alongside removal effectiveness to make this trade-off visible. Neither geometric consistency nor benchmark removal scores establish legal compliance or guarantee that a concept cannot be recovered. Deployment should therefore combine concept erasure with appropriate access controls, evaluation of unintended effects, and ongoing safety assessment rather than treating it as a complete safeguard.

REPRODUCIBILITY STATEMENT

Section 5 specifies the anchor-assignment and teacher-path updating objectives, and Appendix A provides the theoretical assumptions and proofs. Section 6 and Appendix B.2 define the reported metrics and improvement calculations. Appendix B.2 also records the default source-bank size, trajectory-sampling budget, anchor count, and COCO30k FID reference; Appendices B.3–B.6 provide the additional sequential, concept, and sensitivity results. Experiments are repeated over five random seeds, with means and relative standard deviations reported. Code and scripts required to reproduce the experiments will be published upon acceptance.

REFERENCES

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, et al. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16420–16430, 2025.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Black Forest Labs. FLUX.2 [klein] Base 4B: Model card. Hugging Face model card, 2026. URL `https://huggingface.co/black-forest-labs/FLUX.2-klein-base-4B`.

Anh Bui, Thuy-Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *International Conference on Learning Representations*, volume 2025, pp. 64032–64074, 2025.

Yuze Cai, Jiahao Lu, Hongxiang Shi, Yichao Zhou, and Hong Lu. Prototype-guided concept erasure in diffusion models. *arXiv preprint arXiv:2603.08271*, 2026.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–10. IEEE, 2025.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15875–15884, 2025.

Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

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.

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.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2534. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6d13e085b79d454da5910e4ca82a3d9d-Abstract-Conference.html.

Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13393–13402, 2025.

Jonas Henry Grebe, Tobias Braun, Anna Rohrbach, and Marcus Rohrbach. Geometric erasure by contrastive velocity matching in rectified flows. *arXiv preprint arXiv:2606.00140*, 2026.

Tingxu Han, Weisong Sun, Yanrong Hu, Chunrong Fang, Yonglong Zhang, Shiqing Ma, Tao Zheng, Zhenyu Chen, and Zhenting Wang. Continuous concepts removal in text-to-image diffusion models. *arXiv preprint arXiv:2412.00580*, 2024.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21143–21151, 2024.

Miso Kim, Georu Lee, Yunji Kim, Hoki Kim, Jinseong Park, and Woojin Lee. Co-occurring associated retained concepts in diffusion unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ryc7jKP6H9.

Myeongseob Ko, Henry Li, Zhun Wang, Jonathan Patsenker, Jiachen T Wang, Qinbin Li, Ming Jin, Dawn Song, and Ruoxi Jia. Boosting alignment for post-unlearning text-to-image generative models. *Advances in Neural Information Processing Systems*, 37:85131–85154, 2024.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. *Advances in Neural Information Processing Systems*, 38:70876–70911, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18596–18606, 2025a.

Byung Hyun Lee, Sungjin Lim, Seunggyu Lee, Dong Un Kang, and Se Young Chun. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *International Conference on Learning Representations*, volume 2025, pp. 99119–99158, 2025b.

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. In *International Conference on Learning Representations*, volume 2026, pp. 30660–30682, 2026a.

Uichan Lee, Jeonghyeon Kim, and Sangheum Hwang. Localized concept erasure in text-to-image diffusion models via high-level representation misdirection. *arXiv preprint arXiv:2602.19631*, 2026b.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. *arXiv preprint arXiv:2602.06771*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? *Advances in Neural Information Processing Systems*, 38:14751–14775, 2026.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Fengyuan Miao, Shancheng Fang, Lingyun Yu, Yadong Qu, Yuhao Sun, Xiaorui Wang, and Hongtao Xie. Sderasure: Concept-specific trajectory shifting for concept erasure via adaptive diffusion classifier. In *The Fourteenth International Conference on Learning Representations*, 2026.

Saemi Moon, Minjong Lee, Sangdon Park, and Dongwoo Kim. Holistic unlearning benchmark: A multi-faceted evaluation for text-to-image diffusion model unlearning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16356–16366. IEEE, 2025.

Saemi Moon, Suhyeon Jun, Seoyeon Lee, and Dongwoo Kim. Concept unlearning via cross-attention activation projection for diffusion models. *arXiv preprint arXiv:2605.25765*, 2026.

Jiadong Pan, Hongcheng Gao, Zongyu Wu, Taihang Hu, Li Su, Qingming Huang, and Liang Li. Leveraging catastrophic forgetting to develop safe diffusion models against malicious finetuning. *Advances in Neural Information Processing Systems*, 37:115208–115232, 2024.

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.

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.

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2966–2975. IEEE, 2025.

Hoigi Seo, Byung Hyun Lee, Jaehyun Cho, Sungjin Lim, and Se Young Chun. Erasing thousands of concepts: Towards scalable and practical concept erasure for text-to-image diffusion models. *arXiv preprint arXiv:2604.16481*, 2026.

Reza Shirkavand, Peiran Yu, Shangqian Gao, Gowthami Somepalli, Tom Goldstein, and Heng Huang. Efficient fine-tuning and concept suppression for pruned diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18619–18629, 2025.

Lorenzo Simone, Davide Bacciu, and Shuangge Ma. Continualflow: Learning and unlearning with neural flow matching. *arXiv preprint arXiv:2506.18747*, 2025.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-8328-0. URL `https://bookstore.ams.org/gsm-140/`.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9121–9130, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Jiahang Tu, Ye Li, Yiming Wu, Hanbin Zhao, Chao Zhang, and Hui Qian. Mass concept erasure in diffusion models with concept hierarchy. *arXiv preprint arXiv:2601.03305*, 2026.

Korada Sri Vardhana and Soma Biswas. Generase: Generalizable and semantically-aware concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2327–2335, 2026.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768. IEEE, 2025.

Siyuan Wen, Jiahao Zeng, and Ningning Ding. AutoAnchor: Stable diffusion unlearning using cross-attention as a manifold surrogate. *arXiv preprint arXiv:2607.08337*, 2026.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28263–28273, 2025a.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025b.

Lexiang Xiong, Chengyu Liu, Jingwen Ye, Yan Liu, and Yuecong Xu. Semantic surgery: Zero-shot concept erasure in diffusion models. *arXiv preprint arXiv:2510.22851*, 2025.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1755–1764, 2024a.

Lingyun Zhang, Yu Xie, Yanwei Fu, and Ping Chen. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8172–8181, 2025a.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025b.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

A Rigorous Theory and Proofs

This appendix develops the four geometric design principles of Section 4. We distinguish properties of a supervision construction, approximation bounds for a specified field class, and stability bounds for an edited ODE. None alone certifies semantic removal or generated image quality. Restatements retain the numbers of their main-text counterparts; additional results use an independent appendix counter. In particular, the branch-resolved form of Theorem 1 below makes explicit which distribution each expectation uses.

A.1 Notation and technical conditions

States and flow maps. Let $Z_0 \sim \mu_0$ be the source in $\mathbb{R}^d$, c_f the forget condition, and c_a a fixed anchor unless an index is displayed. The frozen and edited fields are $v_{\theta_\star}$ and v_{θ_u}. Define the two-time flow by

$$\Phi^c_{\theta,s\to t}(x) = x + \int_s^t v_\theta\big(\Phi^c_{\theta,s\to r}(x), r, c\big)\, dr, \qquad \Phi^c_{\theta,s\to s}(x) = x. \tag{8}$$

Here x is the state at time s, not necessarily a time-zero source. We write $\Phi^c_{\theta,t} = \Phi^c_{\theta,0\to t}$. A flow-map evaluation is one vector; the function of t is a trajectory. On their domains, uniqueness gives

$$\Phi^c_{\theta,r\to t} \circ \Phi^c_{\theta,s\to r} = \Phi^c_{\theta,s\to t}, \qquad (\Phi^c_{\theta,s\to t})^{-1} = \Phi^c_{\theta,t\to s}. \tag{9}$$

These identities describe the exact ODE, not a finite Euler step.

The teacher trajectories and tangents are

$$\begin{aligned} F_t(z_0) &= \Phi^{c_f}_{\theta_\star,t}(z_0), \quad f_t(z_0) = v_{\theta_\star}(F_t(z_0), t, c_f), \\ A_t(z_0) &= \Phi^{c_a}_{\theta_\star,t}(z_0), \quad a_t(z_0) = v_{\theta_\star}(A_t(z_0), t, c_a). \end{aligned} \tag{10}$$

Their derivatives equal the displayed tangents almost everywhere. For multiple anchors, superscript (q) denotes condition $c_a^{(q)}$. The checkpoint coupling is $\pi_\star = (\mathrm{Id}, F_1)_{\#}\mu_0$, where pushforward denotes the law obtained by applying the map to a source draw. The generic safe endpoint law is ν_a; for an anchor teacher it is $(A_1)_{\#}\mu_0$, not necessarily the true conditional data distribution. The checkpoint coupling is neither assumed to be its pretraining coupling nor assumed to be an optimal transport plan.

Technical forms of the standing assumptions. The following extensions supply the domains and integrability used in the proofs. They are modular; for example, the source-pairing result does not require a regular edited ODE.

Assumption 1 (technical form) 1 *Sources and endpoints have finite second moments. For a supervision target V at a sampled time T, assume $\mathbb{E}\|V\|_2^2 < \infty$. For full-time path bounds assume almost-sure absolute continuity of $A_\cdot(Z_0)$ and $\mathbb{E}\int_0^1 \|a_t(Z_0)\|_2^2 dt < \infty$. Endpoint Wasserstein bounds require both endpoint laws to have finite second moments. The velocity moment condition concerns the sampled supervision vector.*

Justification. These conditions make quadratic risks and transport costs integrable. Cauchy–Schwarz makes the endpoint inner products integrable. Full-time path energy also bounds intermediate states through $\sup_t \|A_t\|_2^2 \le 2\|Z_0\|_2^2 + 2\int_0^1 \|a_t\|_2^2 dt$. Finite energy under a training-time law is not automatically finite energy under Lebesgue time; converting between them needs the coverage condition stated below.

Assumption 2 (technical form) 2 *Each field used in an ODE comparison is measurable in time, continuous in state for almost every time, and bounded on compact state regions by an integrable temporal envelope. On the common region containing the compared paths it satisfies*

$$\|v(x,t,c) - v(y,t,c)\|_2 \le L_c(t)\|x-y\|_2, \qquad L_c \ge 0, \quad L_c \in L^1([0,1]). \tag{11}$$

All required solutions exist on the full comparison interval. For off-path anchor queries this includes backward and forward anchor continuations from $F_t(z_0)$. Chord comparisons include the chord in the common region; student tracking includes both the selected teacher path and student rollout.

Justification. These hypotheses give a Carathéodory existence and uniqueness setting. Existence on the full interval is a separate condition; local Lipschitz continuity alone does not exclude finite-time escape (Teschl, 2012). Linear growth is one sufficient alternative, but is not assumed when the required continuations are specified directly. The main-text bounds use the special case $L_c(t) \le L_c$. No useful global Transformer Lipschitz constant is claimed. Population bounds with one constant require that constant to hold uniformly over the compared sources. Uniqueness constrains trajectories of one fixed conditional field, not intersections between different conditional fields.

Assumption 3 (technical form) 3 *The analyzed coupling is*

$$\pi_u = (1-p)\pi_\star + p(\mu_0 \otimes \nu_a), \qquad p \in [0,1]. \tag{12}$$

Each independent draw has a branch indicator $H \sim \mathrm{Bernoulli}(p)$, and conditional on $H = 1$ the source and safe endpoint are independent.

Justification. This is a construction-specific assumption, not a consequence of generic velocity substitution. A fixed permutation of a finite bank need not give the population product law. If loss weighting is absorbed into the law, independence must still hold after weighting: source/endpoint-dependent weights can destroy it. Here p is branch probability, not a semantic erasure-strength coefficient.

Assumption 4 (technical form) 4 *Let $\psi : [0,\infty) \to [0,\infty)$ be finite and nondecreasing with $\psi(0) = 0$. The nonempty class $\mathcal{V}_\psi$ contains jointly measurable, finite-risk fields satisfying, at almost every training time,*

$$\|v(z,t,c_f) - v(z',t,c_f)\|_2 \le \psi(\|z - z'\|_2) \tag{13}$$

for all relevant state pairs. The same ψ is used for the entire class.

Justification. This bound is the only regularity ingredient in the regression obstruction. Lipschitz and Hölder cases have $\psi(r) = Lr$ and $\psi(r) = Lr^\alpha$, respectively. For a continuity interpretation one also needs $\lim_{r\downarrow 0} \psi(r) = 0$, but the finite-scale inequality itself uses only the displayed budget. LoRA does not automatically impose this budget. A sampled upper quantile of secant changes is a diagnostic, not a bound on every population pair.

Sampling and conditioning. Draw $T \sim \lambda$ and then $(Z,V) \sim P_T$, where P_t is a measurable probability kernel of state–target pairs. To measure conflict, draw two pairs independently *conditional on the same T*; they need not be independent after marginalizing this common time. The full law and risk are

$$\mathsf{Q}(dt\,dz\,dv) = \lambda(dt)P_t(dz\,dv), \qquad R_{\mathsf{Q}}(v) = \mathbb{E}\|v(Z,T,c_f) - V\|_2^2. \tag{14}$$

Time-only weights can be absorbed into λ after normalization; general sample weights also change P_t. A source-indexed path Γ is teacher-consistent under c when $\Gamma_0(z_0) = z_0$ and $\dot\Gamma_t = v_{\theta_\star}(\Gamma_t, t, c)$ almost everywhere. This defines consistency of the supervision, not semantic safety or finite-data student generalization.

A.2 P1: Independent replacement and source correspondence

Lemma A.1 (Quadratic two-cycle monotonicity) *For a quadratic-cost optimal coupling π, two independent pairs satisfy $\langle Z - Z', Y - Y'\rangle \ge 0$ almost surely.*

Proof 1 *An optimal quadratic plan is concentrated on a cyclically monotone set (Villani, 2009). Its two-point inequality is $c(z,y) + c(z',y') \le c(z,y') + c(z',y)$, where $c(z,y) = \|z-y\|_2^2/2$. Expanding the difference gives $-\langle z - z', y - y'\rangle \le 0$.*

This is a necessary OT condition, not a requirement for every valid FM flow. In dimensions greater than one it is not a complete OT characterization.

Separating the mixture and replacement moments. For two mixture draws let $S = \langle Z - Z', Y - Y' \rangle$, $\rho(\pi_u) = \mathbb{P}(S < 0)$, and $\Delta(\pi_u) = \mathbb{E}[(-S)_+]$. Separately draw $\bar{Z}, \bar{Z}' \overset{\text{iid}}{\sim} \mu_0$ and $B, B' \overset{\text{iid}}{\sim} \nu_a$, all mutually independent, and define

$$S_a = \langle \bar{Z} - \bar{Z}', B - B' \rangle, \qquad \omega_a = \mathbb{P}(S_a = 0), \qquad M_a = \mathbb{E}|S_a|. \tag{15}$$

These moments describe the replacement–replacement branch only.

Theorem 1 (branch-resolved formulation) 5 *Under finite endpoint second moments and independent replacement,*

$$\rho(\pi_u) \geq \frac{p^2}{2}(1 - \omega_a), \qquad \Delta(\pi_u) \geq \frac{p^2}{2} M_a. \tag{16}$$

Both inequalities are equalities when $p = 1$. A positive first lower bound excludes quadratic-cost optimality of π_u.

Proof idea. Both draws use the replacement branch with probability p^2. On that event, swapping the two independent endpoints reverses the inner product without changing its distribution.

Proof 2 *Let H, H' be the independent branch indicators. Conditional on $H = H' = 1$, the four variables have the product law in Eq. 15. Given $\bar{Z} - \bar{Z}'$, swapping B, B' shows symmetry of S_a. Thus*

$$\mathbb{P}(S_a < 0) = \tfrac{1}{2}(1 - \omega_a), \qquad \mathbb{E}[(-S_a)_+] = \tfrac{1}{2} M_a. \tag{17}$$

The latter moment is finite by Cauchy–Schwarz. Restricting the nonnegative sign probability and negative-part expectation to $\{H = H' = 1\}$ proves Eq. 16. At $p = 1$ the restriction is the whole law. The OT conclusion follows from Lemma A.1.

The replacement moments in Eq. 15 are the same ω_a, M_a used in the main-text bound. They isolate the p^2 contribution; the other branch contributions are characterized next.

Proposition A.2 (Exact branch decomposition) *Define $\rho_{hh'} = \mathbb{P}(S < 0 \mid H = h, H' = h')$ and $\Delta_{hh'} = \mathbb{E}[(-S)_+ \mid H = h, H' = h']$ through the branch kernels. Then*

$$\begin{aligned} \rho(\pi_u) &= (1-p)^2 \rho_{00} + 2p(1-p)\rho_{01} + p^2 \rho_{11}, \\ \Delta(\pi_u) &= (1-p)^2 \Delta_{00} + 2p(1-p)\Delta_{01} + p^2 \Delta_{11}, \end{aligned} \tag{18}$$

where $\rho_{11} = (1 - \omega_a)/2$ and $\Delta_{11} = M_a/2$.

Proof 3 *Condition on both indicators. Swapping the complete pair draws leaves S unchanged, so the 01 and 10 terms agree. Total probability and expectation give the result. Branch kernels define the terms even at zero mixture weights.*

The p^2 term is a guaranteed contribution, not an upper bound on the total defect; mixed terms can contribute at first order in p.

Corollary A.3 (Interior contraction of endpoint chords) *For fixed endpoint pairs put $a = z - z'$, $b = y - y'$ and $h(t) = (1-t)a + tb$. If $\langle a, b \rangle < 0$, the minimum separation occurs at*

$$t_* = \frac{\|a\|_2^2 - \langle a, b \rangle}{\|a - b\|_2^2} \in (0, 1), \quad \|h(t_*)\|_2^2 = \frac{\|a\|_2^2 \|b\|_2^2 - \langle a, b \rangle^2}{\|a - b\|_2^2} < \min\{\|a\|_2^2, \|b\|_2^2\}. \tag{19}$$

Proof 4 *Expand $\|a + t(b - a)\|_2^2$ and set its derivative to zero. The numerator of t_* is positive; the denominator minus that numerator equals $\|b\|_2^2 - \langle a, b \rangle > 0$. Substitution gives the minimum. A strictly convex quadratic with an interior minimizer is strictly smaller there than at either endpoint.*

The conclusion concerns straight interpolants, not necessarily teacher ODE paths. Contraction is not exact intersection: $a = (1, 0)$ and $b = (-1, 1)$ have negative inner product but squared minimum separation $1/5$. No practically large conflict mass or visual failure follows from this alone.

Lemma A.4 (Exact bridge to common-time conflict) *For the straight-pair law $X_t = (1-t)Z_0 + tY$, $U = Y - Z_0$, define*

$$J_r(z, y, z', y') = \int_0^1 \mathbf{1}\{\|(1-t)(z-z') + t(y-y')\|_2 \le r\}\, \lambda(dt). \tag{20}$$

Then

$$q_{\mathrm{conf}}(r, \delta) = \mathbb{E}_{\pi\otimes\pi}\left[J_r(Z_0, Y, Z_0', Y')\mathbf{1}\{\|(Y - Z_0) - (Y' - Z_0')\|_2 \ge \delta\}\right]. \tag{21}$$

Proof 5 *Condition on the two endpoint pairs, integrate their common-time proximity indicator, and use their constant displacement labels. Tonelli applies to the nonnegative integrand.*

This identity is the exact straight-pair connection to P4. High-dimensional small-ball probabilities can be very small, even when strictly positive.

A.3 P2: Safe support and anchor-characteristic shifts

Lemma A.5 (Two-sided stability on a common flow domain) *Let $\Lambda_c(s,t) = \int_{\min(s,t)}^{\max(s,t)} L_c(r)dr$. For the solutions covered by the ODE conditions above,*

$$e^{-\Lambda_c(s,t)}\|x - y\|_2 \le \|\Phi^c_{s\to t}(x) - \Phi^c_{s\to t}(y)\|_2 \le e^{\Lambda_c(s,t)}\|x - y\|_2. \tag{22}$$

Proof 6 *For $s < t$, subtract the two integral equations and use the state Lipschitz bound to obtain $\|e_r\|_2 \le \|x-y\|_2 + \int_s^r L_c(u)\|e_u\|_2 du$. Grönwall gives the upper bound. Apply it to the inverse map along the same solution segments to obtain $\|x-y\|_2 \le e^{\Lambda_c(s,t)}\|e_t\|_2$, giving the lower bound. Time reversal handles $t < s$. The lower bound concerns the flow map, not the instantaneous velocity field.*

At a source z_0 and time t define

$$\begin{aligned} g_t &= \|F_t(z_0) - A_t(z_0)\|_2, \\ \widetilde{z}_t &= (\Phi^{c_a}_{\theta_\star,t})^{-1}(F_t(z_0)), \\ \widetilde{A}_{1|t} &= \Phi^{c_a}_{\theta_\star,t\to 1}(F_t(z_0)) = A_1(\widetilde{z}_t). \end{aligned} \tag{23}$$

The implicit endpoint $\widetilde{A}_{1|t}$ depends on the switching time t. It is an endpoint of a frozen-anchor continuation, not necessarily the edited student's endpoint. The inverse requires $F_t(z_0)$ in the relevant anchor-flow image.

Proposition 1 (extension) 6 *Under the common-domain conditions,*

$$\begin{aligned} e^{-\Lambda_a(0,t)}g_t \le \|\widetilde{z}_t - z_0\|_2 \le e^{\Lambda_a(0,t)}g_t, \\ e^{-\Lambda_a(t,1)}g_t \le \|\widetilde{A}_{1|t} - A_1(z_0)\|_2 \le e^{\Lambda_a(t,1)}g_t. \end{aligned} \tag{24}$$

Thus $g_t = 0$, $\widetilde{z}_t = z_0$, and $\widetilde{A}_{1|t} = A_1(z_0)$ are equivalent. The constant-L_a bounds in the main text are a special case.

Proof 7 *Apply Lemma A.5 to the two time-t states under the same anchor field, first backward to 0, then forward to 1. Composition maps $A_t(z_0)$ to z_0 and $A_1(z_0)$, respectively. The factors are finite and strictly positive, proving the equivalences.*

Meaning and limits. When $g_t > 0$, the target-path state has a different source and endpoint *when interpreted through the anchor field*. Its actual forget-conditioned source remains z_0. The continuation is still anchor-conditioned, not an unknown concept. This proves exact same-source inconsistency, not bad image quality or a nonzero student endpoint error.

Corollary A.6 (Continuity of off-path predictions) *At the compared states,*

$$\|v_{\theta_\star}(F_t, t, c_a) - a_t\|_2 \le L_a(t)g_t. \tag{25}$$

The difference between the norms of the anchor-minus-forget contrasts at F_t and A_t is at most $(L_a(t) + L_f(t))g_t$.

Proof 8 *The first claim is the Lipschitz inequality. For the second, use the reverse triangle inequality on the contrast vectors and apply the two field Lipschitz bounds.*

These are upper bounds only. A state-independent anchor field gives equal velocities at distinct states, so a nonzero support gap does not imply a nonzero velocity-vector difference.

Example A.7 (Conditional fields may intersect) *On $\mathbb{R}$, let $v_f(x,t) = 0$ and $v_a(x,t) = 1 - 4t$. From $z_0 = 0$, the unique solutions are $F_t = 0$ and $A_t = t - 2t^2$. At $t = 1/2$ their states coincide but velocities are 0 and -1. There is no uniqueness contradiction because the ODEs differ. Consequently, a difference between conditional predictions at one state does not prove different original sources, much less independent source–endpoint pairing.*

A.4 P3: Instantaneous tangents and teacher-flow consistency

Two residuals, two evaluation locations. Set $\Delta(z_0) = A_1(z_0) - z_0$ and $\bar{A}_t(z_0) = z_0 + t\Delta(z_0)$. Define

$$\begin{aligned} r_t^{\tan}(z_0) &= \Delta(z_0) - v_{\theta_\star}(A_t(z_0), t, c_a), \\ r_t^{\mathrm{com}}(z_0) &= \Delta(z_0) - v_{\theta_\star}(\bar{A}_t(z_0), t, c_a). \end{aligned} \tag{26}$$

The tangent residual is measured on the actual teacher path. The commutation residual tests the teacher ODE on the reconstructed chord. They need not have the same numerical value.

Proposition 2 (precise restatement) 7 *For a source with an absolutely continuous teacher path and square-integrable tangent,*

$$\begin{aligned} A_t(z_0) - \bar{A}_t(z_0) &= -\int_0^t r_s^{\tan}(z_0) ds, \\ \sup_{t\in[0,1]} \|A_t(z_0) - \bar{A}_t(z_0)\|_2^2 &\le \int_0^1 \|r_s^{\tan}(z_0)\|_2^2 ds. \end{aligned} \tag{27}$$

Furthermore, $r_t^{\tan}(z_0) = 0$ for Lebesgue-almost every t if and only if $A_t(z_0) = z_0 + t\Delta(z_0)$ for every t. The required path is its constant-speed endpoint chord*, not merely a trajectory whose geometric trace lies on a line.*

Proof 9 *Absolute continuity gives $A_t - z_0 = \int_0^t a_s ds$ and $\Delta = \int_0^1 a_s ds$. Subtract $t\Delta$ to obtain the identity. Cauchy–Schwarz yields $\| \int_0^t r_s ds\|_2^2 \le t \int_0^t \|r_s\|_2^2 ds$, which is bounded by the full-time energy. Take the supremum. A zero residual gives path equality by integration. Conversely, equal absolutely continuous paths have equal derivatives almost everywhere.*

Corollary A.8 (A sharper endpoint-pinned bound) *Under the same conditions,*

$$\|A_t - \bar{A}_t\|_2^2 \le t(1-t) \int_0^1 \|r_s^{\tan}\|_2^2 ds, \qquad \int_0^1 \|r_s^{\tan}\|_2^2 ds \ge 4 \sup_t \|A_t - \bar{A}_t\|_2^2. \tag{28}$$

Proof 10 *The residual integrates to zero, since both paths have the same endpoints. Hence $\int_0^t r_s ds = \int_0^1 (\mathbf{1}_{[0,t]}(s) - t) r_s ds$. The squared L^2 norm of this scalar coefficient is $t(1-t)$. Apply Cauchy–Schwarz and then $t(1-t) \le 1/4$.*

A measured large path–chord gap therefore forces a large integrated tangent error. The converse does not follow because residuals may cancel over time. A non-chord trajectory has a nonzero residual on a positive-measure set of times, not necessarily at every time.

Example A.9 (Nonconstant speed and temporal cancellation) *For $A_t = z_0 + t^2 b$ with $b \neq 0$, the trajectory lies on a straight segment, but $a_t = 2tb$ and $r_t^{\tan} = (1-2t)b$ is not zero a.e. For a cancellation example, in one dimension let $A_t^{[n]} = z_0 + t + n^{-1/2}\sin(2\pi n t)$, for integer $n \ge 1$. Its endpoint chord is $z_0 + t$, its maximum squared deviation is $1/n$, and its tangent-residual energy is $2\pi^2 n$. It is generated by the smooth state-independent field $1 + 2\pi\sqrt{n}\cos(2\pi n t)$. Thus arbitrarily large residual energy need not produce a large path gap.*

Proposition A.10 (Chord–teacher commutation and risk transfer) *Assume a common spatial Lipschitz constant L_a on a region containing both the teacher path and chord. Define the commutation energy*

$$\mathcal{C}_{\mathrm{com}} = \mathbb{E}_{Z_0} \int_0^1 \|r_t^{\mathrm{com}}(Z_0)\|_2^2 dt.$$

Then

$$\mathbb{E} \sup_t \|A_t(Z_0) - \bar{A}_t(Z_0)\|_2^2 \le e^{2L_a} \mathcal{C}_{\mathrm{com}}. \tag{29}$$

Zero commutation energy is equivalent to equality of the paths for μ_0-almost every source. If $e_\theta(z,t) = v_{\theta_u}(z,t,c_f) - v_{\theta_\star}(z,t,c_a)$ is K_e-Lipschitz on this region, define $R_{\mathrm{path}} = \mathbb{E}\|e_\theta(A_T(Z_0),T)\|_2^2$ and $R_{\mathrm{chord}} = \mathbb{E}\|e_\theta(\bar{A}_T(Z_0),T)\|_2^2$ under the same source and time laws. Then

$$|\sqrt{R_{\mathrm{path}}} - \sqrt{R_{\mathrm{chord}}}| \le K_e e^{L_a} \sqrt{\mathcal{C}_{\mathrm{com}}}. \tag{30}$$

Proof 11 *For $e_t = A_t - \bar{A}_t$, subtract the dynamics to obtain*

$$\dot{e}_t = v_{\theta_\star}(A_t,t,c_a) - v_{\theta_\star}(\bar{A}_t,t,c_a) - r_t^{\mathrm{com}}, \qquad e_0 = 0. \tag{31}$$

The inhomogeneous Grönwall inequality bounds $\|e_t\|_2$ by $\int_0^t e^{L_a(t-s)}\|r_s^{\mathrm{com}}\|_2 ds$. Cauchy–Schwarz and $t \le 1$ give Eq. 29 after supremum and expectation. With zero residual, the chord solves the same ODE from the same source, hence equals the teacher path by uniqueness. Conversely, path equality gives equality of derivatives a.e. For risk transfer, the reverse triangle inequality in L^2 gives

$$\begin{aligned} |\sqrt{R_{\mathrm{path}}} - \sqrt{R_{\mathrm{chord}}}| &\le \big(\mathbb{E}\|e_\theta(A_T,T) - e_\theta(\bar{A}_T,T)\|_2^2\big)^{1/2} \\ &\le K_e\big(\mathbb{E}\|A_T - \bar{A}_T\|_2^2\big)^{1/2} \le K_e e^{L_a}\sqrt{\mathcal{C}_{\mathrm{com}}}. \end{aligned} \tag{32}$$

Which loss is being compared. R_{chord} matches the teacher field evaluated at chord states; it is not automatically the training loss against the displacement Δ. For $R_\Delta = \mathbb{E}\|v_{\theta_u}(\bar{A}_T,T,c_f) - \Delta\|_2^2$,

$$\sqrt{R_{\mathrm{chord}}} \le \sqrt{R_\Delta} + \big(\mathbb{E}\|r_T^{\mathrm{com}}(Z_0)\|_2^2\big)^{1/2}. \tag{33}$$

The last expectation uses the actual time law; it equals $\mathcal{C}_{\mathrm{com}}$ only under uniform time. Under uniqueness, $r^{\tan} = 0$ a.e., $r^{\mathrm{com}} = 0$ a.e., and path equality are equivalent. The two residuals quantify different nonzero discrepancies, not different zero-consistency criteria.

Geometric interpretation. Straight conditional interpolation is legitimate FM training (Lipman et al., 2023; Liu et al., 2023); the marginal ODE need not follow these individual interpolants (Geng et al., 2025). Reconstructing a chord from a teacher endpoint need not recover the teacher's dynamics. Displacement labels on *actual teacher states* change labels but do not prove that the resulting student rollout is a chord. Reconstructing both chord states and labels defines a different supervision path. Neither case alone proves image degradation or greater optimization cost.

A.4.1 The constructive counterpart: safe-path tracking

Proposition A.11 (Teacher-path error controls the edited rollout) *For a fixed measurable assignment $q(z_0)$, write $\Gamma_t(z_0) = A_t^{(q(z_0))}(z_0)$, $\widehat{\Gamma}_t(z_0) = \Phi_{\theta_u,t}^{c_f}(z_0)$, and $\varepsilon_t(z_0) = v_{\theta_u}(\Gamma_t(z_0),t,c_f) - \dot{\Gamma}_t(z_0)$. Assume full-interval student existence, a common spatial Lipschitz constant L_u on the comparison region, and finite energies and endpoint moments. Then*

$$\begin{aligned} \|\widehat{\Gamma}_t - \Gamma_t\|_2 &\le \int_0^t e^{L_u(t-s)}\|\varepsilon_s\|_2 ds, \\ \mathbb{E}\sup_t \|\widehat{\Gamma}_t - \Gamma_t\|_2^2 &\le K(L_u)\mathbb{E}\int_0^1 \|\varepsilon_t\|_2^2 dt, \end{aligned} \tag{34}$$

where $K(L) = (e^{2L}-1)/(2L)$ for $L > 0$ and $K(0) = 1$. Also,

$$W_2^2\big((\widehat{\Gamma}_1)_{\#}\mu_0, (\Gamma_1)_{\#}\mu_0\big) \le K(L_u)\mathbb{E}\int_0^1 \|\varepsilon_t\|_2^2 dt. \tag{35}$$

Zero full-time residual is equivalent to equality of the trajectories for almost every source and all times.

Proof 12 *Subtract the two dynamics and add and subtract $v_{\theta_u}(\Gamma_t, t, c_f)$. The integral difference obeys*

$$\|\widehat{\Gamma}_t - \Gamma_t\|_2 \le L_u \int_0^t \|\widehat{\Gamma}_s - \Gamma_s\|_2 ds + \int_0^t \|\varepsilon_s\|_2 ds. \tag{36}$$

Grönwall gives the first estimate. Cauchy–Schwarz yields

$$\left(\int_0^t e^{L_u(t-s)}\|\varepsilon_s\|_2 ds\right)^2 \le \left(\int_0^t e^{2L_u(t-s)} ds\right) \int_0^t \|\varepsilon_s\|_2^2 ds \le K(L_u) \int_0^1 \|\varepsilon_s\|_2^2 ds. \tag{37}$$

Taking supremum and expectation gives the second estimate. The same-source joint law $(\widehat{\Gamma}_1(Z_0), \Gamma_1(Z_0))$ is an admissible coupling of the two endpoint laws, so its expected squared distance upper-bounds W_2^2. Zero residual gives zero gap for every time a.s.; the converse follows by differentiation of equal absolutely continuous paths.

The selected routes may come from different teacher conditions. The result is conditional on a small residual; P4 explains why this residual may be unattainable for an incompatible selection. It does not turn latent-space closeness into a semantic-removal certificate.

Finite-source interpretation. For a fixed bank of distinct sources $\{z_{0,i}\}_{i=1}^N$ and a fixed assignment q_i, the same proof applies with $\mu_N = N^{-1}\sum_i \delta_{z_{0,i}}$ in place of μ_0, provided the stated ODE and regularity conditions hold on those paths. In particular,

$$W_2^2\left(\frac{1}{N}\sum_i \delta_{\widehat{\Gamma}_1(z_{0,i})}, \frac{1}{N}\sum_i \delta_{\Gamma_1(z_{0,i})}\right) \le \frac{K(L_u)}{N} \sum_i \int_0^1 \|\varepsilon_t(z_{0,i})\|_2^2 \, dt.$$

This bankwise statement requires no assignment outside the observed sources. The population version instead requires a measurable assignment on the source distribution and population residual control. Held-out-seed performance is evaluated separately as described in Appendix B.1.

Corollary A.12 (Time-law coverage and finite time grids) *If $\lambda(dt) = w(t)dt$ with $w(t) \ge w_{\min} > 0$ a.e., then*

$$\mathbb{E}\int_0^1 \|\varepsilon_t\|_2^2 dt \le w_{\min}^{-1}\mathbb{E}_{Z_0,T\sim\lambda}\|\varepsilon_T\|_2^2. \tag{38}$$

Alternatively, partition $[0,1]$ into cells of lengths $h_m \le h$, with one representative t_m per cell. If the pathwise residual is B-Lipschitz in time with a common constant, then

$$\mathbb{E}\int_0^1 \|\varepsilon_t\|_2^2 dt \le 2\mathbb{E}\sum_m h_m \|\varepsilon_{t_m}\|_2^2 + 2B^2h^2. \tag{39}$$

Proof 13 *The first estimate uses $1 \le w/w_{\min}$. On cell m, $\|\varepsilon_t\|_2^2 \le 2\|\varepsilon_{t_m}\|_2^2 + 2B^2h^2$. Integrate over cells and sources for the second estimate.*

If $w > 0$ a.e. but lacks a positive essential lower bound, zero weighted loss still gives zero full-time residual, but not this uniform quantitative constant. Finite time samples or a finite source bank give no population guarantee without temporal or source generalization control.

Logit-normal time sampling. For the logit-normal density p_{LN} in Appendix B.1, with finite μ_{LN} and $\sigma_{\mathrm{LN}} > 0$, the density is positive on $(0,1)$ but tends to zero at both endpoints. Thus its essential infimum on $[0,1]$ is zero. For any $0 < \eta < 1/2$, let $m_\eta = \min_{t\in[\eta,1-\eta]} p_{\mathrm{LN}}(t) > 0$. Splitting the time integral gives

$$\mathbb{E}\int_0^1 \|\varepsilon_t\|_2^2 \, dt \le \frac{\mathbb{E}_{Z_0,T\sim p_{\mathrm{LN}}}\|\varepsilon_T\|_2^2}{m_\eta} + \mathbb{E}\int_{[0,\eta)\cup(1-\eta,1]} \|\varepsilon_t\|_2^2 \, dt.$$

The interior estimate follows from $p_{\mathrm{LN}} \ge m_\eta$ there; substitution into Proposition A.11 gives a tracking bound with an explicit boundary-residual term. The finite-grid estimate in Corollary A.12 remains applicable under its separate temporal-regularity condition.

Guidance and modified supervision. The path, tangent, and deployment ODE must use consistent fields. At CFG scale w, let $v^{\mathrm{cfg}}_{\theta,w}(z,t,c) = wv_\theta(z,t,c) + (1-w)v_\theta(z,t,\varnothing)$. The guided teacher–student residual is exactly $we_c + (1-w)e_\varnothing$, where e_c compares the student's forget-conditioned prediction with the teacher's anchor-conditioned prediction at the same state, and $e_\varnothing$ is the null-branch residual. Thus raw conditional matching alone does not control a changed null branch or a different guided path. The tracking theorem applies when the same CFG-effective field is used throughout. If practical labels are $\dot{\Gamma}_t + b_t$, a training residual $\widehat{\varepsilon}_t$ gives $\varepsilon_t = \widehat{\varepsilon}_t + b_t$ and $\mathbb{E}\int \|\varepsilon_t\|_2^2 dt \le 2\mathbb{E}\int \|\widehat{\varepsilon}_t\|_2^2 dt + 2\mathbb{E}\int \|b_t\|_2^2 dt$. Negative guidance is therefore not exact tangent distillation unless its additional term vanishes. Numerical ODE paths require separate solver-error control; the continuous result does not certify arbitrary finite-step sampling.

A.5 P4: Joint compatibility of the supervision distribution

Use the sampling experiment of Eq. 14. For the two conditionally independent pairs at the same T, define

$$q_{\mathrm{conf}}(r,\delta) = \mathbb{P}\{\|Z - Z'\|_2 \le r,\ \|V - V'\|_2 \ge \delta\}. \tag{40}$$

This is a probability over the common time and both pairs, not a pointwise assertion about every possible pair.

Theorem 2 (restatement) 8 *For every $r \ge 0$, $\delta > 0$, and $v \in \mathcal{V}_\psi$,*

$$R_{\mathsf{Q}}(v) \ge \frac{q_{\mathrm{conf}}(r,\delta)}{4} \max\{\delta - \psi(r), 0\}^2. \tag{41}$$

The same bound holds for the infimum over $\mathcal{V}_\psi$. It is strictly positive if both $q_{\mathrm{conf}}(r,\delta) > 0$ and $\delta > \psi(r)$ hold.

Proof idea. At a conflict pair, local field variation accounts for at most $\psi(r)$ of the target separation. The two prediction errors absorb the rest.

Proof 14 *Let $e = \|v(Z,T,c_f) - V\|_2$, $e' = \|v(Z',T,c_f) - V'\|_2$, and let $\mathcal{E}$ be the event in Eq. 40. On this event, the triangle inequality gives*

$$\delta \le \|V - V'\|_2 \le e + \|v(Z,T,c_f) - v(Z',T,c_f)\|_2 + e' \le e + \psi(r) + e'. \tag{42}$$

Hence $e + e' \ge b = \max\{\delta - \psi(r), 0\}$ there, and $e^2 + e'^2 \ge b^2 \mathbf{1}_{\mathcal{E}}/2$ everywhere. The pairs have identical marginals, so $\mathbb{E}(e^2 + e'^2) = 2R_{\mathsf{Q}}(v)$. Taking expectations gives the factor $1/4$. The bound is independent of v, allowing the infimum.

This is a sufficient certificate of positive optimal risk within the specified class, not an equivalence characterizing all nonzero risks. Its practical magnitude depends on conflict mass and the over-budget gap. A particular neural parameterization may incur additional error even when the larger function class is compatible.

Proposition A.13 (Risk decomposition and mean-field obstruction) *Let $\Xi = (Z,T)$ and $m_{\mathsf{Q}}(\Xi) = \mathbb{E}[V \mid \Xi]$. Define*

$$\mathfrak{V}(\mathsf{Q}) = \mathbb{E}\|V - m_{\mathsf{Q}}(\Xi)\|_2^2, \qquad \mathcal{E}_\psi(\mathsf{Q}) = \inf_{v \in \mathcal{V}_\psi} \mathbb{E}\|v(Z,T,c_f) - m_{\mathsf{Q}}(\Xi)\|_2^2. \tag{43}$$

Then

$$\begin{aligned} R_{\mathsf{Q}}(v) &= \mathbb{E}\|v(Z,T,c_f) - m_{\mathsf{Q}}(\Xi)\|_2^2 + \mathfrak{V}(\mathsf{Q}), \\ \inf_{v \in \mathcal{V}_\psi} R_{\mathsf{Q}}(v) &= \mathfrak{V}(\mathsf{Q}) + \mathcal{E}_\psi(\mathsf{Q}). \end{aligned} \tag{44}$$

Define $\widetilde{q}_{\mathrm{conf}}$ by replacing V, V' in Eq. 40 with their conditional means. Then

$$\mathcal{E}_\psi(\mathsf{Q}) \ge \frac{\widetilde{q}_{\mathrm{conf}}(r,\delta)}{4} \max\{\delta - \psi(r), 0\}^2. \tag{45}$$

Also $\max\{\mathfrak{V}, \mathcal{E}_\psi\} \ge q_{\mathrm{conf}}(r,\delta) \max\{\delta - \psi(r), 0\}^2/8$.

Proof 15 *Expand $v - V = (v - m_{\mathsf{Q}}) + (m_{\mathsf{Q}} - V)$. The cross term has zero conditional expectation given Ξ, proving the decomposition. Only its first term depends on v, so taking the infimum gives the second identity. Repeat the paired proof with conditional-mean labels for the mean-field bound. Finally $a + b \geq B$ for nonnegative a, b implies $\max(a, b) \geq B/2$.*

The two components are conditional target variance and restricted-class mean-field approximation error. They are not "anchor error plus SGD error." Gradient noise is a separate consequence. A positive Bayes variance alone does not make the mean field unrepresentable or its endpoint law inaccurate.

Corollary A.14 (Hidden multiple anchors and the single-anchor control) *For a fixed measurable assignment $q(Z_0)$, sample $Z = A_T^{(q(Z_0))}(Z_0)$ and $V = a_T^{(q(Z_0))}(Z_0)$, with every student example conditioned on c_f. Positive conflict mass exceeding the budget implies a positive risk floor. For a single fixed anchor, $\mathfrak{V}(\mathsf{Q}) = 0$; if its teacher field on the sampled support satisfies ψ, the optimal risk over $\mathcal{V}_\psi$ is zero.*

Proof 16 *The first statement substitutes the selected law into the main bound. For a fixed anchor, $V = v_{\theta_\star}(Z, T, c_a)$ is a deterministic function of the observed state and time. Its conditional variance is zero, and that same function, used under c_f, is an exact-fit candidate when it belongs to the class.*

Multiple anchors need not conflict: they may agree on overlapping regions or occupy separate regions. A single anchor can also exceed a more restrictive student regularity budget. The conclusion does not guarantee realizability by finite-rank LoRA. Supplying anchor indices to the student would change the regression problem; the present result concerns a common observed forget condition.

A.5.1 Condition-aware extension

Let $\mathcal{C}$ be a standard Borel condition space with a measurable metric $d_{\mathcal{C}}$. Extend the law to $\lambda(dt)$ $P_t(dz\, dc\, dv)$. For nondecreasing moduli vanishing at zero, impose

$$\|v(z, t, c) - v(z', t, c')\|_2 \leq \psi_z(\|z - z'\|_2) + \psi_c(d_{\mathcal{C}}(c, c')). \tag{46}$$

The metric may be applied to the actual conditioning representation; it is not presumed to be an exact semantic metric. Define $q_{\mathrm{cond}}(r_z, r_c, \delta)$ as the probability that two pairs at the same time have state distance at most r_z, condition distance at most r_c, and target separation at least δ.

Corollary A.15 (Finite-resolution condition interference) *Every finite-risk field satisfying Eq. 46 obeys*

$$\mathbb{E}\|v(Z, T, C) - V\|_2^2 \geq \frac{q_{\mathrm{cond}}(r_z, r_c, \delta)}{4} \max\{\delta - \psi_z(r_z) - \psi_c(r_c), 0\}^2. \tag{47}$$

The conditional-mean counterpart holds when the means condition on (Z, T, C).

Proof 17 *On the defining event, insert the two model predictions between V, V'. Their difference is at most $\psi_z(r_z) + \psi_c(r_c)$. The two error norms must cover the positive remainder. Squaring and averaging as above yields the result; substitute conditional means for the mean-field version.*

This is a conditional mechanism for interference between neighboring forget and retain prompts. It does not identify which task absorbs the error. For identical observed conditions, the condition budget vanishes on the compared pairs; it does not require the model to ignore conditions elsewhere.

A.5.2 Conditional variance and stochastic-gradient noise

Proposition A.16 (Conditional gradient covariance) *Let Ξ include all observed inputs and $\ell_\theta = \|v_\theta(\Xi) - V\|_2^2/2$. Assume differentiability and finite gradient second moments. Write $J_\theta = \partial v_\theta / \partial \theta \in \mathbb{R}^{d \times p_\theta}$ and $\Sigma(\Xi) = \mathrm{Cov}(V \mid \Xi)$. Then*

$$\begin{aligned} \mathbb{E}[\nabla_\theta \ell_\theta \mid \Xi] &= J_\theta^\top (v_\theta - m_{\mathsf{Q}}), \\ \mathrm{Cov}(\nabla_\theta \ell_\theta \mid \Xi) &= J_\theta^\top \Sigma J_\theta. \end{aligned} \tag{48}$$

At $v_\theta = m_{\mathsf{Q}}$ a.s.,

$$\mathbb{E}[\|\nabla_\theta \ell_\theta\|_2^2 \mid \Xi] = \mathrm{tr}(J_\theta J_\theta^\top \Sigma). \tag{49}$$

If $h^\top J_\theta J_\theta^\top h \geq \gamma_J \|h\|_2^2$ for every h in the range of $\Sigma(\Xi)$ with a common $\gamma_J > 0$, then $\mathbb{E}\|\nabla_\theta \ell_\theta\|_2^2 \geq \gamma_J \mathfrak{V}(\mathsf{Q})$ at this conditional-mean fit.

Proof 18 *The chain rule gives $\nabla_\theta \ell_\theta = J_\theta^\top (v_\theta - V)$. Conditioning on Ξ leaves only V random, proving the mean and covariance. At the conditional mean, the gradient mean vanishes, so squared norm expectation is the trace of the covariance. Diagonalize $\Sigma = \sum_k \lambda_k h_k h_k^\top$ and apply the lower bound to each eigenvector with $\lambda_k > 0$, then average over Ξ.*

For B independent full training examples, mean-gradient covariance is divided by B. If a minibatch shares one random time, this scaling is conditional on that time; variation of the conditional gradient mean across time is an additional term. Disagreement in a Jacobian-null direction need not create parameter-gradient noise. No convergence-speed theorem follows.

A.6 Additional mechanisms and scope

Proposition A.17 (Mixture ambiguity and its first-order scale) *Let $\mathsf{Q}_p = (1-p)\mathsf{Q}_0 + p\mathsf{Q}_1$, $p \in [0,1]$, with branch H. Let s_0, s_1 be densities of the observed input Ξ relative to a common measure ζ, and $m_h(\xi) = \mathbb{E}_{\mathsf{Q}_h}[V \mid \Xi = \xi]$. The between-branch part of conditional target variance is*

$$\mathfrak{V}_{\text{between}}(\mathsf{Q}_p) = \int \frac{p(1-p)s_0(\xi)s_1(\xi)}{(1-p)s_0(\xi) + ps_1(\xi)} \|m_1(\xi) - m_0(\xi)\|_2^2 d\zeta(\xi), \tag{50}$$

with the integrand defined as zero whenever $p(1-p)s_0(\xi)s_1(\xi) = 0$, including every point where the mixture density vanishes. Suppose on a measurable $B \subset \{s_0 > 0\}$ that $\int_B s_0 d\zeta \geq \alpha > 0$, $s_1/s_0 \geq \kappa > 0$, and $\|m_1 - m_0\|_2 \geq \delta_0 > 0$. Then

$$\mathfrak{V}_{\text{between}}(\mathsf{Q}_p) \geq \frac{p(1-p)\kappa}{1-p+p\kappa} \alpha \delta_0^2. \tag{51}$$

This is an $\Omega(p)$ lower bound as $p \downarrow 0$. If additionally $\int_{\{s_0>0\}} s_1 \|m_1 - m_0\|_2^2 d\zeta < \infty$, the between-branch term is $O(p)$ and hence $\Theta(p)$ under the lower-bound conditions.

Proof 19 *For $0 < p < 1$, Bayes' rule on the positive-mixture-density support gives $\eta(\xi) = \mathbb{P}(H = 1 \mid \Xi = \xi) = ps_1/((1-p)s_0 + ps_1)$. Total conditional variance contributes $\eta(1-\eta)\|m_1 - m_0\|_2^2$. Multiplying by the mixture input density and integrating proves Eq. 50. On B, put $u = s_1/s_0$. The function $u/(1-p+pu)$ is increasing, so its value is at least $\kappa/(1-p+p\kappa)$. Integrate the resulting bound against s_0. For the upper bound, the density factor is at most ps_1 for $p < 1$, and is zero outside the overlap. This proves the asserted asymptotic orders. At $p = 0$ or $p = 1$, the branch is deterministic, so the between-branch variance and the displayed lower bound are both zero under the stated integrand convention.*

The observation Ξ must include all actual conditions. Erase and retain examples with different observed prompts are not a hidden same-input mixture; their proximity is instead treated by Corollary A.15.

Proposition A.18 (Covariance stress and marginal acceleration) *Consider straight interpolation $X_t = Z_0 + tU$ with constant sample velocity U, and define $\rho_t(x)$, $m_t(x) = \mathbb{E}[U \mid X_t = x]$, and $\Sigma_t(x) = \mathrm{Cov}(U \mid X_t = x)$. Assume sufficient smoothness, finite moments, and decay to justify the moment equations, and work where $\rho_t > 0$. Then*

$$(\partial_t + m_t \cdot \nabla) m_t = -\frac{1}{\rho_t} \nabla \cdot (\rho_t \Sigma_t). \tag{52}$$

Proof 20 *The joint phase-space law obeys free transport $\partial_t f_t + u \cdot \nabla_x f_t = 0$, classically or weakly. Integrating against 1 and u yields*

$$\partial_t \rho_t + \nabla \cdot (\rho_t m_t) = 0, \qquad \partial_t (\rho_t m_t) + \nabla \cdot \big(\rho_t (m_t \otimes m_t + \Sigma_t)\big) = 0. \tag{53}$$

Expand the second equation componentwise, subtract m_t times the first, and divide by the positive density. This gives the identity.

The spatial divergence of covariance stress, not its trace alone, controls material acceleration. Acceleration includes speed changes as well as turning. Neither conditional variance nor raw pair conflict alone establishes a monotone curvature or few-step sampling claim.

Example A.19 (Positive defect and Bayes floor with an exact endpoint map) *Let* $Z_0, E \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ *and* $Y = b + E$ *with* $b \neq 0$. *Compare this independent coupling with the deterministic translation* $Y = Z_0 + b$. *For* $D_t = (1-t)^2 + t^2$, *Gaussian conditioning gives*

$$m_t(x) = b + \frac{2t-1}{D_t}(x - tb), \qquad \mathrm{Var}(Y - Z_0 \mid X_t) = \frac{1}{D_t}. \tag{54}$$

Its two-cycle sign-defect probability is $1/2$ *and, under uniform time,* $\mathfrak{V} = \int_0^1 D_t^{-1} dt = \pi/2 > 0$. *Nevertheless* m_t *is spatially* 1*-Lipschitz and its exact ODE flow is*

$$x_t = tb + \sqrt{D_t}\, x_0, \qquad x_1 = b + x_0. \tag{55}$$

Thus it has the same endpoint map as the deterministic translation.

Proof 21 *The centered state is* $(1-t)Z_0 + tE$, *with variance* D_t. *The centered label is* $E - Z_0$, *with variance* 2 *and covariance* $2t-1$ *with the state. The Gaussian conditional mean and variance are therefore as in Eq. 54, using* $2D_t - (2t-1)^2 = 1$. *Set* $u = 2t - 1$; *then* $D_t = (1+u^2)/2$ *and the variance integral is* $\int_{-1}^{1}(1+u^2)^{-1}du = \pi/2$. *The spatial derivative is* $2u/(1+u^2)$, *whose magnitude is at most one. Differentiating Eq. 55 reproduces* $m_t(x_t)$ *and the source initial condition; uniqueness identifies the flow. For the independent coupling, source and endpoint differences are independent centered nondegenerate Gaussians, so their product is negative with probability* $1/2$. *The mean field itself belongs to the* 1*-Lipschitz class, and consequently its restricted approximation gap is zero.*

This example separates pair-regression ambiguity from endpoint failure. Theorem 2 lower-bounds the total regression risk; it does not by itself imply inaccurate samples or failed semantic removal.

A.7 From the regression theorem to a computable assignment score

The theorem uses common-time supervision probabilities and a prescribed field budget. The method uses finite candidate paths and a sampled pair score. The following statements identify the relationship without equating the two objects.

Proposition A.20 (Secant-score control of conflict mass) *Under any fixed pair-and-time sampling law* H, *define*

$$R_\epsilon = \frac{\|V - V'\|_2}{\|Z - Z'\|_2 + \epsilon}, \qquad C_\rho = \mathbb{E}_H \rho(R_\epsilon), \qquad \epsilon > 0, \tag{56}$$

where ρ *is nonnegative and nondecreasing. Write* $q_H(r,\delta)$ *for the conflict probability under this same law. If* $\rho(\delta/(r+\epsilon)) > 0$, *then*

$$q_H(r,\delta) \le \min\left\{1, \frac{C_\rho}{\rho(\delta/(r+\epsilon))}\right\}. \tag{57}$$

Proof 22 *On the conflict event,* $R_\epsilon \ge \delta/(r+\epsilon)$. *Monotonicity gives* $\rho(R_\epsilon) \ge \rho(\delta/(r+\epsilon))\mathbf{1}_{\mathcal{E}}$. *Take expectations and divide; the probability is also at most one.*

For the raw-ratio main-text score, take $\rho(s) = s$. For a logarithmically robust score, take $\rho(s) = \log(1+s^2)$. This controls a tail probability for the *same* pair sampling law. It does not give an upper bound on the optimal regression risk: reducing an upper bound on the mass in a risk lower bound is not a risk upper-bound argument. Nor does a smaller average score order every realized tail probability across assignments. State/velocity normalization and nonlinear penalties define different operational scores and must be reported as such.

Proposition A.21 (What a sparse comparison graph certifies) *Let a fixed nonempty undirected graph* $\mathcal{E}$ *compare* N *sources. Choose an edge uniformly, give its endpoints a uniformly random*

orientation, and sample the same time for both paths. Let d_i be source i's degree, $d_{\max} = \max_i d_i$, and let $R_i(v)$ be its time-averaged regression risk. For the graph conflict probability $q_{\mathcal{E}}$,

$$R_{\mathcal{E}}(v) := \sum_i \frac{d_i}{2|\mathcal{E}|} R_i(v) \geq \frac{q_{\mathcal{E}}(r,\delta)}{4} \max\{\delta - \psi(r), 0\}^2. \tag{58}$$

The uniform-source risk satisfies

$$\frac{1}{N}\sum_i R_i(v) \geq \frac{2|\mathcal{E}|}{N d_{\max}} \frac{q_{\mathcal{E}}(r,\delta)}{4} \max\{\delta - \psi(r), 0\}^2. \tag{59}$$

Proof 23 *The triangle-inequality proof only needs identical pair marginals, not independence. Random orientation gives source marginal $d_i/(2|\mathcal{E}|)$, proving the first inequality. Nonnegativity gives $R_{\mathcal{E}} \leq N d_{\max}(\sum_i R_i/N)/(2|\mathcal{E}|)$; rearrange and substitute.*

This is an empirical-bank statement, not a population certificate. A source-neighbor graph can miss paths that approach later in time. If a different graph law, time law, pooled state representation, or assignment-dependent graph is used, its relation to the training law must be checked separately. A mixed graph containing frozen forget branches also need not describe the labels actually regressed on by a safe-path loss; those edges are checkpoint-relative diagnostics, not automatically samples from the P4 training distribution.

Proposition A.22 (Finite assignment feasibility and reference fallback) *Suppose each finite admissible set $\mathcal{Q}_i$ contains 0, and all costs are finite. Put $d(\boldsymbol{q}) = N^{-1}\sum_i d_i^{(q_i)}$. The program minimizing $C(\boldsymbol{q})$ subject to $q_i \in \mathcal{Q}_i$ and $d(\boldsymbol{q}) \leq d(\mathbf{0})$ has a minimizer, with objective at most $C(\mathbf{0})$. A heuristic returning the lowest-scoring feasible assignment among a set of candidates that includes $\mathbf{0}$ has the same baseline inequality, although it need not attain the global minimum.*

Proof 24 *The feasible set is a nonempty finite set because it contains $\mathbf{0}$. A finite function attains a minimum on such a set. The heuristic claim is the corresponding minimum over its visited feasible candidates.*

These are guarantees for the specified finite scores only. They do not imply safety of all generated outputs, zero regression floor, improved student FID, or use of more than one anchor. The condition $0 \in \mathcal{Q}_i$ is ensured by the strict screening inequality in Section 5. Projecting CLIP features away from a target-text subspace removes that linear component, not every semantic trace of the concept. Both anchor screening and context costs remain proxies whose evaluation requires a separate experimental protocol.

A.8 Result map and scope

Main-text question	Formal statement and scope
P1: source correspondence	Section A.2: branch-resolved ordering defects; contraction concerns endpoint chords, not universal ODE crossings.
P2: safe support	Section A.3: two-sided implicit-source and frozen-anchor-continuation bounds on a common flow domain.
P3: teacher tangents	Section A.4: constant-speed consistency, commutation, risk transfer, and conditional student tracking.
P4: joint compatibility	Section A.5: a class-explicit risk floor, its decomposition, and condition/optimization extensions.
Assignment implementation	Section A.7: finite-score and graph-law statements, not population or semantic certificates.

The theory separates three questions: whether supervision is consistent with a particular teacher flow, whether a field in a stated class can fit it, and whether a resulting model performs well on a semantic task. Controlled experiments must establish the connection to target suppression, utility, and sequential editing. In particular, no theorem here establishes irreversible concept removal, universal superiority over diffusion-derived methods, or error-free accumulation of repeated edits.

B Additional Evaluation Details and Results

B.1 Implementation Settings

Unless otherwise stated, TRACE uses $N = 100$ noise seeds in the source bank and samples 15 points along each seed's teacher trajectory for supervision. The candidate anchor count is $Q = 10$; following the indexing in Section 5, the null condition $q = 0$ serves as an additional reference. Candidate trajectories and their teacher velocities are constructed during pre-unlearning, and the selected anchor assignment is fixed during parameter updating. The 15-point budget describes sampled trajectory states, rather than an inference-step count. All reported evaluations use held-out generation seeds that are disjoint from the trajectory source bank.

Backbones and generation settings. We evaluate at 1024×1024 resolution using Stable Diffusion 3.5 Medium and the undistilled `FLUX.2-klein-base-4B` checkpoint. The latter is abbreviated as FLUX.2 [klein] 4B or FLUX in the tables. FLUX inference uses 50 steps and a guidance scale of 4.0, matching the checkpoint's official example (Black Forest Labs, 2026).

Candidate anchors and feature encoders. Following LLM-assisted anchor construction in prior unlearning work (Kim et al., 2026; Wen et al., 2026), we first generate prompts containing the forget concept c_f, mask the target-concept span, and ask an LLM (ChatGPT 5.6 Thinking) to complete the masked span with its most likely alternative in the remaining context. Repeating this procedure supplies the Q candidate anchor concepts, which are then screened by the criterion in Section 5. For this text-only screening, SD3.5M uses its CLIP ViT-L component, whereas FLUX uses Qwen3-4B. The utility constraint uses the image and text representations of an external OpenAI CLIP ViT-L model on both backbones, so that the target-text projection and image-feature comparison share an aligned feature space. Each backbone retains its native text-conditioning pipeline during teacher-trajectory construction and parameter updating; CLIP utility scoring does not replace FLUX's Qwen3-4B conditioning encoder.

Trajectory-time sampling. We use logit-normal time sampling following rectified-flow training (Esser et al., 2024). In the noise-to-image convention of this paper, write

$$\xi \sim \mathcal{N}(\mu_{\mathrm{LN}}, \sigma_{\mathrm{LN}}^2), \qquad t = \mathrm{sigmoid}(\xi) = \frac{1}{1 + e^{-\xi}}, \qquad 0 < t < 1. \tag{60}$$

The corresponding density is

$$p_{\mathrm{LN}}(t) = \frac{1}{\sigma_{\mathrm{LN}}\sqrt{2\pi}\, t(1-t)} \exp\left[-\frac{\left(\log \frac{t}{1-t} - \mu_{\mathrm{LN}}\right)^2}{2\sigma_{\mathrm{LN}}^2} \right]. \tag{61}$$

This choice is motivated by concentrating a finite supervision budget on intermediate portions of the teacher paths rather than sampling uniformly over time. When converting a data-to-noise time coordinate τ to our convention, we use $t = 1 - \tau$; a logit-normal location parameter therefore changes sign under this conversion.

Parameter-efficient updating. Following parameter-efficient FM unlearning (Grebe et al., 2026), we train LoRA adapters with rank $r = 32$ on the attention query, key, and value projections and the MLP layers of the Transformer blocks, while keeping the pretrained weights frozen. We use AdamW with learning rate 10^{-5}, no learning-rate warmup, and batch size 4.

For FID, we use COCO30k from the COCO 2017 training split as the real-image reference. This is a real-image-reference evaluation, not a distance to the unedited model's generated images. The N and Q sensitivity experiments vary the corresponding bank size around the default setting and report the same four removal and utility metrics as the main evaluation.

Baseline scope. ContinualFlow (Simone et al., 2025) reports low-dimensional benchmarks and latent-space experiments on MNIST and CIFAR-10, rather than the pretrained high-resolution T2I setting evaluated here. We therefore do not include it as a directly comparable baseline for our 1024×1024 experiments.

B.1.1 CONSTRAINED ANCHOR-ASSIGNMENT SEARCH

We approximately solve Eq. equation 6 using constrained iterated conditional modes (ICM), which updates one seed's anchor while holding the other assignments fixed.

Cached costs. Before search, we cache the context costs $D_{i,q} = d_i^q$ and the time-averaged pairwise scores

$$K_{ij}(q,r) = \frac{1}{M}\sum_{m=1}^{M} \frac{\|a_{t_m}^{(q)}(z_{0,i}) - a_{t_m}^{(r)}(z_{0,j})\|_2}{\|A_{t_m}^{(q)}(z_{0,i}) - A_{t_m}^{(r)}(z_{0,j})\|_2 + \epsilon}, \qquad i \neq j, \tag{62}$$

where the paired states use a common time t_m, and the sampled times and costs are fixed throughout search. With uniform averaging over distinct source pairs,

$$\widehat{C}(\boldsymbol{q}) = \frac{1}{\binom{N}{2}}\sum_{i<j} K_{ij}(q_i,q_j), \qquad d(\boldsymbol{q}) = \frac{1}{N}\sum_i D_{i,q_i}. \tag{63}$$

The score uses the raw ratio in Eq. equation 5; the displayed averages specify its time and pair normalization. The cost table has the symmetry $K_{ij}(q,r) = K_{ji}(r,q)$.

Connection to the finite-bank theory. Condition on the cached bank, the common time grid, and a fixed assignment. The uniform-pair score in Eq. equation 63 uses the complete comparison graph, with $|\mathcal{E}| = N(N-1)/2$ and $d_i = d_{\max} = N-1$, and the empirical time law $\lambda_M = M^{-1}\sum_m \delta_{t_m}$. Consequently, the graph-source marginals in Proposition A.21 are uniform and

$$R_{\mathcal{E}}(v) = \frac{1}{NM}\sum_{i,m}\left\|v(A_{t_m}^{(q_i)}(z_{0,i}), t_m, c_f) - a_{t_m}^{(q_i)}(z_{0,i})\right\|_2^2.$$

The factor $2|\mathcal{E}|/(Nd_{\max})$ in Eq. 59 is therefore one. Proposition A.20 applies to $\widehat{C}$ with $\rho(s) = s$ under this same pair-and-time law. These substitutions identify the empirical quantities controlled by the cached score; they do not replace the score–risk distinction in Appendix A.7.

Feasible coordinate updates. Initialize $\boldsymbol{q} = \mathbf{0}$, so that $d(\boldsymbol{q}) = d(\mathbf{0})$. For a proposed replacement $q_i \leftarrow v$, update the cached totals by

$$\begin{aligned} d_i(v) &= d(\boldsymbol{q}) + \frac{D_{i,v} - D_{i,q_i}}{N}, \\ \Delta C_i(v) &= \frac{1}{\binom{N}{2}}\sum_{j\neq i}\big[K_{ij}(v,q_j) - K_{ij}(q_i,q_j)\big]. \end{aligned} \tag{64}$$

Each sweep visits the seeds in a randomly permuted order. For seed i, we evaluate candidates $v \in \mathcal{Q}_i$ satisfying $d_i(v) \leq d(\mathbf{0})$, including its current assignment, and choose one minimizing $\Delta C_i(v)$. We accept a change only if it strictly decreases $\widehat{C}$; otherwise, we keep the current assignment. Accepted changes update both cached totals. Search stops after a full sweep without an accepted change or when the prescribed sweep budget is reached, and the resulting assignment is frozen for parameter updating.

The all-null initialization is feasible, every accepted iterate remains feasible, and the score never increases. Thus the returned assignment satisfies $d(\boldsymbol{q}^*) \leq d(\mathbf{0})$ and $\widehat{C}(\boldsymbol{q}^*) \leq \widehat{C}(\mathbf{0})$, as in Proposition A.22. Termination without an improving move gives a one-coordinate local optimum; the heuristic does not require a global solution.

Search cost. The cached tables require $O(N(Q+1))$ context costs and $O(N^2(Q+1)^2)$ pair scores. At $N = 100$ and $Q = 10$, the full pair table contains 1,210,000 scalars, or approximately 4.84 MB in float32, excluding cached trajectory states and velocities. Once the tables have been constructed, evaluating all candidates for one coordinate costs $O(N(Q+1))$, and a full sweep costs $O(N^2(Q+1))$. Teacher-trajectory generation and pair-score construction are separate precomputation costs.

B.2 Metric Reporting and Relative Improvements

The main evaluation uses CLIP(U) and Anchor-win to assess target removal, and CLIP(R) and FID to assess non-target utility. Lower CLIP(U) means weaker target-concept alignment; higher Anchor-win means a greater fraction of generations prefer the evaluated safe alternatives. CLIP(R) measures alignment between non-target prompts and their generated images. All reported means are accompanied by the relative standard deviation over five random seeds, expressed as $100\sigma/|\mu|$ percent. Thus an annotation $\mu \pm r\%$ corresponds to an absolute standard deviation $\sigma = |\mu|r/100$, where r is the displayed numerical percentage. Bold and underlined entries rank the reported means; these annotations are not confidence intervals or significance-test results.

Let O, B, and M denote the original model, an external baseline, and TRACE , respectively. For each concept–backbone setting, the strongest external baseline is chosen separately for each metric from ESD, ConAbl, FMN, EA, and GEM. Internal ablations are not included in this baseline comparison. Relative improvements in the reported scores are

$$\begin{aligned} \text{CLIP(U) reduction} &= \frac{\text{CLIP(U)}_B - \text{CLIP(U)}_M}{\text{CLIP(U)}_B}, \\ \text{CLIP(R) increase} &= \frac{\text{CLIP(R)}_M - \text{CLIP(R)}_B}{\text{CLIP(R)}_B}, \\ \text{FID reduction} &= \frac{\text{FID}_B - \text{FID}_M}{\text{FID}_B}. \end{aligned} \tag{65}$$

Multiplying by 100% gives relative percentage changes. The abstract reports the main-table comparison in Table 1, whose maximum CLIP(U) reduction, CLIP(R) increase, and FID reduction are 7.10%, 7.78%, and 16.55%, respectively. The full single-concept evaluation combines the four main-table concepts with the eleven additional concepts in Appendix B.5, for 30 concept–backbone settings. Across this full evaluation, the corresponding maxima are 7.10%, 8.26%, and 16.55%. The 8.26% CLIP(R) increase occurs for Naruto on FLUX, relative to EA: $(28.84 - 26.64)/26.64 = 8.26\%$.

To quantify recovery of utility lost during unlearning, we instead normalize by the baseline's degradation relative to the original model:

$$\begin{aligned} \text{CLIP(R) drop reduction} &= \frac{\text{CLIP(R)}_M - \text{CLIP(R)}_B}{\text{CLIP(R)}_O - \text{CLIP(R)}_B}, \\ \text{FID degradation reduction} &= \frac{\text{FID}_B - \text{FID}_M}{\text{FID}_B - \text{FID}_O}. \end{aligned} \tag{66}$$

The denominators are positive in the evaluated settings. For Albert Einstein, TRACE reduces GEM's CLIP(R) drop by 71.74% on SD3.5M and 83.33% on FLUX. For Mickey Mouse, the corresponding FID degradation reductions are 61.92% and 46.15%. These are fractions of recovered degradation, not relative changes in the final scores. Changes in Anchor-win are reported in percentage points.

B.3 Complete Sequential Results on FLUX.2 4B

Table 3 gives all four metrics at 5, 10, and 15 sequentially removed concepts on FLUX.2 [klein] 4B. TRACE achieves the best edited-model mean for every reported metric at each sequence length, extending the SD3.5M results in Table 2 to the second backbone.

Removal across sequential edits. Relative to GEM, TRACE improves Anchor-win by 15.85, 16.70, and 16.75 percentage points after 5, 10, and 15 concepts, respectively. At 15 concepts, it retains 76.85% Anchor-win and 19.65 CLIP(U), compared with GEM's 60.10% and 20.65. EA attains 50.90% Anchor-win, while the diffusion-derived baselines range from 25.80% to 31.40%.

Non-target utility. TRACE also maintains higher CLIP(R) and lower FID at each evaluated length. At 15 concepts, its CLIP(R) is 26.90 and FID is 14.50, compared with GEM's 26.15 and 16.10. Its own utility nevertheless declines as edits accumulate: from 5 to 15 concepts, CLIP(R) decreases by 0.95 and FID increases by 0.95. Thus, the results show a persistent performance advantage, not the elimination of cumulative editing effects.

Table 3: Complete sequential unlearning results on FLUX.2. Best edited-model means are bold; uncertainty annotations are relative standard deviations in percent.

Method	CLIP(U)↓	Anchor-win (%)↑	CLIP(R)↑	FID↓
		5 concepts		
Orig	$28.53_{\pm 0.6\%}$	$4.12_{\pm 3.8\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.1\%}$
ESD	$21.65_{\pm 0.8\%}$	$36.21_{\pm 2.2\%}$	$22.84_{\pm 1.1\%}$	$21.05_{\pm 1.5\%}$
ConAbl	$23.10_{\pm 0.9\%}$	$31.85_{\pm 2.6\%}$	$24.62_{\pm 0.9\%}$	$19.10_{\pm 1.3\%}$
FMN	$22.45_{\pm 1.0\%}$	$35.12_{\pm 2.1\%}$	$23.95_{\pm 1.6\%}$	$20.05_{\pm 1.2\%}$
EA	$21.90_{\pm 0.5\%}$	$61.20_{\pm 1.9\%}$	$25.80_{\pm 1.5\%}$	$16.70_{\pm 1.7\%}$
GEM	$19.80_{\pm 0.7\%}$	$68.45_{\pm 2.4\%}$	$26.65_{\pm 0.8\%}$	$15.40_{\pm 1.4\%}$
TRACE	$\mathbf{18.75}_{\pm 0.9\%}$	$\mathbf{84.30}_{\pm 2.5\%}$	$\mathbf{27.85}_{\pm 1.2\%}$	$\mathbf{13.55}_{\pm 1.6\%}$
		10 concepts		
Orig	$28.61_{\pm 0.7\%}$	$4.38_{\pm 3.5\%}$	$28.80_{\pm 1.2\%}$	$11.95_{\pm 1.3\%}$
ESD	$21.85_{\pm 1.1\%}$	$33.50_{\pm 2.8\%}$	$22.65_{\pm 0.7\%}$	$21.65_{\pm 1.8\%}$
ConAbl	$23.00_{\pm 0.6\%}$	$29.40_{\pm 3.1\%}$	$24.30_{\pm 1.7\%}$	$19.90_{\pm 1.1\%}$
FMN	$22.50_{\pm 0.8\%}$	$28.95_{\pm 2.0\%}$	$23.40_{\pm 1.4\%}$	$20.45_{\pm 1.5\%}$
EA	$22.45_{\pm 0.9\%}$	$54.10_{\pm 2.3\%}$	$24.70_{\pm 1.0\%}$	$17.20_{\pm 1.6\%}$
GEM	$20.40_{\pm 1.0\%}$	$63.80_{\pm 2.9\%}$	$26.50_{\pm 1.3\%}$	$15.80_{\pm 1.4\%}$
TRACE	$\mathbf{19.30}_{\pm 0.8\%}$	$\mathbf{80.50}_{\pm 3.2\%}$	$\mathbf{27.35}_{\pm 1.5\%}$	$\mathbf{14.20}_{\pm 1.2\%}$
		15 concepts		
Orig	$28.58_{\pm 0.5\%}$	$4.25_{\pm 4.0\%}$	$28.80_{\pm 1.6\%}$	$11.95_{\pm 1.4\%}$
ESD	$22.30_{\pm 0.9\%}$	$31.40_{\pm 2.7\%}$	$22.45_{\pm 1.5\%}$	$21.90_{\pm 1.2\%}$
ConAbl	$23.35_{\pm 1.0\%}$	$27.60_{\pm 3.9\%}$	$24.05_{\pm 1.1\%}$	$20.50_{\pm 1.7\%}$
FMN	$23.60_{\pm 0.6\%}$	$25.80_{\pm 3.4\%}$	$22.95_{\pm 1.8\%}$	$21.00_{\pm 1.5\%}$
EA	$22.95_{\pm 0.7\%}$	$50.90_{\pm 2.2\%}$	$24.35_{\pm 0.9\%}$	$17.75_{\pm 1.3\%}$
GEM	$20.65_{\pm 1.1\%}$	$60.10_{\pm 2.6\%}$	$26.15_{\pm 1.2\%}$	$16.10_{\pm 1.0\%}$
TRACE	$\mathbf{19.65}_{\pm 0.8\%}$	$\mathbf{76.85}_{\pm 3.6\%}$	$\mathbf{26.90}_{\pm 1.7\%}$	$\mathbf{14.50}_{\pm 1.6\%}$

B.4 Unlearning More Concepts

We extend the evaluation to eleven concepts beyond the four main-text representatives: five object/identity concepts (Snoopy, Goku, Naruto, Alaskan Malamute, and Porsche) and six attribute/style concepts (happy, sad, nude, Monet-style, Simple sketch, and bright). Tables 4–6 report the external baselines and the same progressive variants as Table 1.

B.5 Unlearning More Concepts

We extend the evaluation to eleven concepts beyond the four main-text representatives: five object/identity concepts (Snoopy, Goku, Naruto, Alaskan Malamute, and Porsche) and six attribute/style concepts (happy, sad, nude, Monet-style, Simple sketch, and bright). Tables 4–6 report the external baselines and the same progressive variants as Table 1.

Removal and utility across additional concepts. TRACE achieves higher mean Anchor-win and CLIP(R) and lower FID than the majority of evaluated external baselines across the 22 additional concept–backbone settings. Relative to the strongest external baseline for each metric, Anchor-win increases by 2.51–15.92 percentage points, CLIP(R) increases by up to 9.2%, and FID decreases by up to 19.5%. CLIP(U) is lower in the majority of the settings. The exceptions (e.g., Snoopy and Goku on SD3.5M, where GEM attains 18.80 and 18.68 versus TRACE 's 19.08 and 18.96, or Alaskan Malamute on FLUX) retain the Anchor-win and non-target utility advantages.

Object and identity concepts. For Snoopy, TRACE achieves 85.39% and 92.80% Anchor-win on SD3.5M and FLUX, respectively, compared with GEM's 73.44% and 77.02%. For Naruto, it reduces FID from GEM's 16.58 to 13.35 on SD3.5M and from 14.58 to 13.25 on FLUX, while increasing CLIP(R) from 28.40 to 31.34 and from 27.65 to 28.55. The gains over external baselines are not restricted to character concepts: the additional tables also show better removal and utility for Alaskan Malamute and Porsche.

Attribute and style concepts. The Anchor-win and non-target utility advantages over external baselines persist across all six additional attribute/style tasks. On SD3.5M, TRACE achieves Anchor-win rates of 85.44%, 85.61%, and 88.61% for happy, sad, and nude, respectively, compared with GEM's 73.60%, 73.73%, and 76.74%. The corresponding FIDs decrease from 14.89, 16.11, and 16.47 to 13.64, 12.95, and 13.41. For Monet-style, Simple sketch, and bright, Anchor-win ranges from 82.95% to 85.95% on SD3.5M and from 78.87% to 81.75% on FLUX, alongside the CLIP(U) exceptions noted above.

Progressive supervision and joint assignment. Adding instantaneous-tangent supervision (+P1P2P3) improves the vast majority of reported metric means over +P1P2. Joint assignment then yields concept-dependent gains and trade-offs. For happy on FLUX, +P1P2P3 has better removal scores, whereas TRACE increases CLIP(R) by 2.72 and lowers FID by 1.32. For sad and nude on FLUX, the trade-offs vary, with +P1P2P3 offering better CLIP(R) and FID than TRACE . For Monet-style and Simple sketch on FLUX, joint assignment improves CLIP(R) and lowers FID by 1.95 and 2.03, respectively. For bright, joint assignment reduces CLIP(R) by 1.86 and increases FID by 1.33 on SD3.5M, but on FLUX it improves CLIP(R) by 0.59 and lowers FID by 0.44. Thus, P4 provides substantial gains in several settings without uniformly improving every metric over tangent supervision alone.

B.6 Sensitivity to Source and Anchor Bank Sizes

We evaluate sensitivity to the number of source seeds N and candidate anchors Q using Van Gogh style on both backbones. Tables 7 and 8 report the results around the default $N = 100$, $Q = 10$ setting.

Source-bank size N. A small bank of $N = 10$ gives relatively weak removal, with Anchor-win of 38.50% on SD3.5M and 37.50% on FLUX. Increasing N to 100 raises these scores to 84.41% and 80.28%. Changes from $N = 100$ to 300 are small: Anchor-win increases by only 0.09 and 0.10 percentage points, and FID decreases by 0.06 and 0.07. At $N = 500$, further improvements remain visible: compared with the default, Anchor-win increases by 2.14 and 2.57 percentage points and FID decreases by 0.27 and 0.32. Thus, the results show a local plateau around $N = 100$–300 rather than a strict saturation at $N = 100$. The default uses fewer source trajectories while retaining most of the removal benefit. The $N = 50$ setting has the best utility scores in this sweep but lower Anchor-win than the default by 9.91 and 5.08 percentage points, illustrating that source-bank size also changes the removal–utility trade-off.

Anchor-bank size Q. Increasing Q from 3 to 10 improves removal, but additional candidates beyond 10 yield little further change. From $Q = 10$ to 50, CLIP(U) decreases by only 0.03 on SD3.5M and 0.04 on FLUX, and Anchor-win rises by 0.19 and 0.12 percentage points. Utility improves modestly over the same range: CLIP(R) increases by 0.45 and 0.34, while FID decreases by 0.24 and 0.27. One possible explanation is that a larger candidate set allows source trajectories to be distributed among more anchor conditions, so fewer paths need to be assigned to any one anchor and there is more flexibility to find low-conflict, context-preserving trajectories. All selected paths still supervise the same forget-conditioned model. This interpretation is consistent with the utility trend, but the sweep does not directly measure per-anchor assignment load or conflict reduction. The default $Q = 10$ therefore captures most of the observed removal gain, while larger banks offer a modest utility benefit at the cost of additional candidate trajectories.

B.7 Wall-Clock Runtime

Table 9 reports the wall-clock times of the evaluated methods and the two stages of TRACE . Total runtimes are 11 min 36 s on SD3.5M and 10 min 24 s on FLUX, compared with 6 min 46 s and 5 min 17 s for GEM. Thus, the total cost is approximately $1.71\times$ and $1.97\times$ that of GEM, respectively, while remaining below ESD, ConAbl, FMN, and EA on both backbones.

The stage breakdown shows that most of TRACE 's runtime is spent in pre-unlearning: trajectory construction and anchor assignment take 8 min 45 s on SD3.5M and 8 min 16 s on FLUX. Parameter updating along the selected teacher paths takes 2 min 50 s and 2 min 08 s, respectively. The two-stage design therefore combines an up-front supervision-construction cost with a short parameter-updating stage.

Table 4: Additional object/identity concepts: Snoopy, Goku, Naruto, Alaskan Malamute, and Porsche. Best edited-model means are **bold**; second-best are underlined.

Method	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
Object/Identity Concepts Unlearning: Snoopy								
Orig	$30.46_{\pm 0.7\%}$	$3.08_{\pm 4.4\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$28.52_{\pm 1.0\%}$	$5.78_{\pm 3.5\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$20.22_{\pm 0.7\%}$	$42.11_{\pm 2.2\%}$	$23.88_{\pm 0.6\%}$	$23.66_{\pm 1.6\%}$	$18.93_{\pm 0.8\%}$	$48.08_{\pm 2.3\%}$	$21.42_{\pm 1.0\%}$	$21.44_{\pm 1.6\%}$
ConAbl	$23.57_{\pm 0.8\%}$	$35.79_{\pm 2.5\%}$	$26.84_{\pm 1.3\%}$	$17.59_{\pm 1.3\%}$	$23.93_{\pm 0.6\%}$	$39.81_{\pm 1.8\%}$	$24.01_{\pm 1.7\%}$	$18.57_{\pm 0.9\%}$
FMN	$21.29_{\pm 0.8\%}$	$40.35_{\pm 2.1\%}$	$25.83_{\pm 1.5\%}$	$21.74_{\pm 1.1\%}$	$19.90_{\pm 0.7\%}$	$41.10_{\pm 2.9\%}$	$24.34_{\pm 1.0\%}$	$18.30_{\pm 1.2\%}$
EA	$21.84_{\pm 0.7\%}$	$67.83_{\pm 1.6\%}$	$28.69_{\pm 1.2\%}$	$16.85_{\pm 1.9\%}$	$21.51_{\pm 1.0\%}$	$64.25_{\pm 2.3\%}$	$25.16_{\pm 1.8\%}$	$17.16_{\pm 0.9\%}$
GEM	$18.80_{\pm 0.8\%}$	$73.44_{\pm 2.1\%}$	$28.97_{\pm 0.9\%}$	$15.82_{\pm 1.2\%}$	$19.33_{\pm 0.5\%}$	$77.02_{\pm 2.0\%}$	$25.51_{\pm 0.9\%}$	$14.50_{\pm 1.6\%}$
Native	$23.10_{\pm 0.7\%}$	$36.68_{\pm 2.8\%}$	$26.54_{\pm 0.8\%}$	$21.06_{\pm 1.8\%}$	$22.42_{\pm 1.0\%}$	$40.51_{\pm 2.0\%}$	$26.57_{\pm 1.5\%}$	$19.45_{\pm 1.1\%}$
+P1	$21.83_{\pm 0.6\%}$	$68.29_{\pm 1.5\%}$	$27.57_{\pm 1.8\%}$	$18.18_{\pm 1.2\%}$	$20.46_{\pm 0.8\%}$	$60.60_{\pm 2.7\%}$	$26.21_{\pm 0.9\%}$	$17.32_{\pm 1.9\%}$
+P1P2	$19.91_{\pm 0.8\%}$	$78.71_{\pm 1.9\%}$	$29.10_{\pm 1.3\%}$	$14.45_{\pm 1.3\%}$	$20.35_{\pm 0.9\%}$	$74.59_{\pm 2.0\%}$	$25.57_{\pm 1.6\%}$	$15.32_{\pm 1.2\%}$
+P1P2P3	$\mathbf{17.97}_{\pm 0.7\%}$	$81.36_{\pm 2.2\%}$	$29.35_{\pm 0.9\%}$	$15.45_{\pm 1.8\%}$	$19.11_{\pm 0.6\%}$	$74.35_{\pm 1.7\%}$	$27.69_{\pm 1.8\%}$	$13.80_{\pm 1.0\%}$
TRACE	$19.08_{\pm 1.0\%}$	$\mathbf{85.39}_{\pm 2.6\%}$	$\mathbf{29.71}_{\pm 1.3\%}$	$\mathbf{14.61}_{\pm 1.7\%}$	$\mathbf{17.48}_{\pm 0.8\%}$	$\mathbf{92.80}_{\pm 1.9\%}$	$\mathbf{28.10}_{\pm 1.5\%}$	$\mathbf{13.04}_{\pm 1.2\%}$
Object/Identity Concepts Unlearning: Goku								
Orig	$30.45_{\pm 0.7\%}$	$2.99_{\pm 4.6\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$28.52_{\pm 1.0\%}$	$5.71_{\pm 3.4\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$19.35_{\pm 0.8\%}$	$42.98_{\pm 1.7\%}$	$25.05_{\pm 0.9\%}$	$21.41_{\pm 1.0\%}$	$19.77_{\pm 0.7\%}$	$46.94_{\pm 2.4\%}$	$22.88_{\pm 0.8\%}$	$20.95_{\pm 1.9\%}$
ConAbl	$23.25_{\pm 1.0\%}$	$36.67_{\pm 2.6\%}$	$26.51_{\pm 1.1\%}$	$19.83_{\pm 1.5\%}$	$21.81_{\pm 0.7\%}$	$38.92_{\pm 1.9\%}$	$24.81_{\pm 1.7\%}$	$18.30_{\pm 0.9\%}$
FMN	$20.06_{\pm 0.7\%}$	$40.21_{\pm 2.3\%}$	$27.91_{\pm 1.5\%}$	$19.42_{\pm 1.0\%}$	$21.05_{\pm 0.8\%}$	$41.23_{\pm 2.9\%}$	$22.52_{\pm 0.7\%}$	$19.06_{\pm 1.0\%}$
EA	$22.21_{\pm 0.6\%}$	$68.01_{\pm 1.8\%}$	$27.50_{\pm 1.3\%}$	$18.31_{\pm 2.0\%}$	$20.13_{\pm 1.1\%}$	$64.11_{\pm 2.5\%}$	$26.64_{\pm 1.4\%}$	$15.92_{\pm 1.1\%}$
GEM	$18.68_{\pm 0.8\%}$	$73.31_{\pm 2.5\%}$	$28.91_{\pm 0.7\%}$	$15.36_{\pm 1.4\%}$	$19.97_{\pm 0.6\%}$	$76.90_{\pm 1.6\%}$	$27.15_{\pm 1.2\%}$	$14.16_{\pm 1.2\%}$
Native	$23.30_{\pm 0.7\%}$	$36.78_{\pm 2.4\%}$	$26.21_{\pm 1.0\%}$	$20.44_{\pm 1.8\%}$	$21.58_{\pm 0.7\%}$	$40.41_{\pm 2.2\%}$	$25.73_{\pm 1.4\%}$	$19.94_{\pm 1.3\%}$
+P1	$22.14_{\pm 0.6\%}$	$68.15_{\pm 1.3\%}$	$29.26_{\pm 2.1\%}$	$16.74_{\pm 1.2\%}$	$22.02_{\pm 0.7\%}$	$60.47_{\pm 2.8\%}$	$25.19_{\pm 0.9\%}$	$15.95_{\pm 1.4\%}$
+P1P2	$22.36_{\pm 0.7\%}$	$70.47_{\pm 2.1\%}$	$28.23_{\pm 1.4\%}$	$17.52_{\pm 1.5\%}$	$21.17_{\pm 0.6\%}$	$67.07_{\pm 2.3\%}$	$25.57_{\pm 1.6\%}$	$15.70_{\pm 1.1\%}$
+P1P2P3	$\mathbf{18.04}_{\pm 0.7\%}$	$81.39_{\pm 2.4\%}$	$30.14_{\pm 1.1\%}$	$14.24_{\pm 1.5\%}$	$21.16_{\pm 0.7\%}$	$74.20_{\pm 1.9\%}$	$26.81_{\pm 1.4\%}$	$14.15_{\pm 1.0\%}$
TRACE	$18.96_{\pm 1.0\%}$	$\mathbf{85.25}_{\pm 2.8\%}$	$\mathbf{28.76}_{\pm 1.3\%}$	$\mathbf{14.43}_{\pm 1.2\%}$	$\mathbf{18.09}_{\pm 0.9\%}$	$\mathbf{92.82}_{\pm 2.3\%}$	$\mathbf{28.33}_{\pm 1.5\%}$	$\mathbf{12.83}_{\pm 1.2\%}$
Object/Identity Concepts Unlearning: Naruto								
Orig	$29.82_{\pm 0.6\%}$	$3.14_{\pm 4.7\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$27.92_{\pm 0.9\%}$	$5.81_{\pm 3.8\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$20.29_{\pm 0.7\%}$	$42.22_{\pm 1.8\%}$	$23.81_{\pm 0.8\%}$	$22.42_{\pm 1.4\%}$	$19.32_{\pm 0.8\%}$	$47.05_{\pm 2.1\%}$	$23.72_{\pm 1.1\%}$	$20.49_{\pm 1.5\%}$
ConAbl	$21.93_{\pm 0.9\%}$	$36.57_{\pm 2.7\%}$	$27.54_{\pm 1.2\%}$	$17.79_{\pm 1.2\%}$	$23.22_{\pm 0.4\%}$	$39.05_{\pm 1.8\%}$	$24.74_{\pm 1.6\%}$	$19.14_{\pm 0.9\%}$
FMN	$22.44_{\pm 0.9\%}$	$39.53_{\pm 2.2\%}$	$26.01_{\pm 1.4\%}$	$20.80_{\pm 1.3\%}$	$19.87_{\pm 0.7\%}$	$41.97_{\pm 2.8\%}$	$23.41_{\pm 0.9\%}$	$18.84_{\pm 1.1\%}$
EA	$20.94_{\pm 0.7\%}$	$69.30_{\pm 1.5\%}$	$29.12_{\pm 1.3\%}$	$16.93_{\pm 1.6\%}$	$20.49_{\pm 1.1\%}$	$62.94_{\pm 2.6\%}$	$25.31_{\pm 1.4\%}$	$15.90_{\pm 1.2\%}$
GEM	$19.99_{\pm 0.9\%}$	$71.96_{\pm 2.2\%}$	$28.40_{\pm 0.8\%}$	$16.58_{\pm 1.6\%}$	$18.43_{\pm 0.6\%}$	$78.56_{\pm 1.7\%}$	$27.65_{\pm 1.4\%}$	$14.58_{\pm 1.3\%}$
Native	$21.04_{\pm 0.6\%}$	$37.45_{\pm 2.5\%}$	$27.52_{\pm 1.1\%}$	$19.89_{\pm 1.6\%}$	$24.06_{\pm 0.8\%}$	$39.71_{\pm 1.9\%}$	$25.72_{\pm 1.5\%}$	$19.81_{\pm 1.2\%}$
+P1	$23.00_{\pm 0.6\%}$	$66.91_{\pm 1.6\%}$	$26.49_{\pm 1.6\%}$	$17.64_{\pm 1.3\%}$	$20.49_{\pm 0.7\%}$	$61.80_{\pm 2.6\%}$	$25.46_{\pm 1.0\%}$	$16.00_{\pm 1.4\%}$
+P1P2	$20.88_{\pm 0.7\%}$	$71.76_{\pm 2.2\%}$	$28.74_{\pm 1.3\%}$	$15.82_{\pm 1.4\%}$	$22.75_{\pm 0.8\%}$	$65.62_{\pm 2.0\%}$	$26.09_{\pm 1.5\%}$	$15.96_{\pm 1.4\%}$
+P1P2P3	$19.32_{\pm 0.8\%}$	$79.87_{\pm 2.3\%}$	$28.41_{\pm 1.1\%}$	$14.79_{\pm 1.6\%}$	$19.59_{\pm 0.7\%}$	$75.83_{\pm 1.8\%}$	$27.07_{\pm 1.6\%}$	$14.14_{\pm 1.1\%}$
TRACE	$\mathbf{18.17}_{\pm 1.1\%}$	$\mathbf{87.07}_{\pm 2.7\%}$	$\mathbf{31.34}_{\pm 1.3\%}$	$\mathbf{13.35}_{\pm 1.3\%}$	$\mathbf{18.95}_{\pm 0.6\%}$	$\mathbf{91.09}_{\pm 2.2\%}$	$\mathbf{28.55}_{\pm 1.4\%}$	$\mathbf{13.25}_{\pm 1.3\%}$
Object/Identity Concepts Unlearning: Alaskan Malamute								
Orig	$30.46_{\pm 0.7\%}$	$2.95_{\pm 4.5\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$28.51_{\pm 0.8\%}$	$5.75_{\pm 3.6\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$18.75_{\pm 0.8\%}$	$47.98_{\pm 1.8\%}$	$23.64_{\pm 0.9\%}$	$23.25_{\pm 1.4\%}$	$17.82_{\pm 0.7\%}$	$51.89_{\pm 2.2\%}$	$23.32_{\pm 1.0\%}$	$20.63_{\pm 1.6\%}$
ConAbl	$22.12_{\pm 0.7\%}$	$36.65_{\pm 2.8\%}$	$27.11_{\pm 1.0\%}$	$18.06_{\pm 1.3\%}$	$23.40_{\pm 0.7\%}$	$38.96_{\pm 1.9\%}$	$24.06_{\pm 1.9\%}$	$19.67_{\pm 1.0\%}$
FMN	$21.33_{\pm 0.8\%}$	$40.25_{\pm 2.2\%}$	$26.32_{\pm 1.4\%}$	$20.95_{\pm 1.2\%}$	$20.03_{\pm 0.8\%}$	$41.19_{\pm 3.0\%}$	$23.46_{\pm 0.7\%}$	$18.60_{\pm 1.0\%}$
EA	$20.40_{\pm 0.8\%}$	$72.81_{\pm 1.6\%}$	$29.52_{\pm 1.4\%}$	$17.04_{\pm 1.9\%}$	$19.53_{\pm 0.9\%}$	$69.17_{\pm 2.5\%}$	$24.38_{\pm 1.5\%}$	$16.85_{\pm 1.3\%}$
GEM	$18.79_{\pm 0.8\%}$	$78.38_{\pm 2.3\%}$	$27.33_{\pm 0.7\%}$	$16.68_{\pm 1.5\%}$	$18.02_{\pm 0.6\%}$	$81.85_{\pm 1.8\%}$	$28.00_{\pm 1.0\%}$	$14.24_{\pm 1.3\%}$
Native	$21.95_{\pm 0.7\%}$	$36.70_{\pm 2.4\%}$	$27.89_{\pm 1.0\%}$	$19.41_{\pm 1.7\%}$	$23.16_{\pm 0.8\%}$	$40.42_{\pm 2.0\%}$	$25.02_{\pm 1.4\%}$	$21.13_{\pm 1.3\%}$
+P1	$21.58_{\pm 0.6\%}$	$73.20_{\pm 1.3\%}$	$26.00_{\pm 1.9\%}$	$17.93_{\pm 1.0\%}$	$20.17_{\pm 0.7\%}$	$65.45_{\pm 3.0\%}$	$26.54_{\pm 0.9\%}$	$15.96_{\pm 1.6\%}$
+P1P2	$17.88_{\pm 0.7\%}$	$82.32_{\pm 2.1\%}$	$29.10_{\pm 1.4\%}$	$16.62_{\pm 1.4\%}$	$20.01_{\pm 0.8\%}$	$79.23_{\pm 2.1\%}$	$24.61_{\pm 1.6\%}$	$16.29_{\pm 1.1\%}$
+P1P2P3	$\mathbf{18.15}_{\pm 0.7\%}$	$\mathbf{86.41}_{\pm 2.4\%}$	$28.73_{\pm 1.1\%}$	$15.53_{\pm 1.6\%}$	$18.83_{\pm 0.5\%}$	$79.15_{\pm 1.9\%}$	$27.10_{\pm 1.7\%}$	$13.43_{\pm 1.1\%}$
TRACE	$17.55_{\pm 1.0\%}$	$85.27_{\pm 2.8\%}$	$\mathbf{30.57}_{\pm 1.3\%}$	$\mathbf{13.57}_{\pm 1.3\%}$	$\mathbf{18.48}_{\pm 0.9\%}$	$\mathbf{92.83}_{\pm 2.1\%}$	$\mathbf{27.25}_{\pm 1.3\%}$	$\mathbf{13.63}_{\pm 1.1\%}$
Object/Identity Concepts Unlearning: Porsche								
Orig	$29.81_{\pm 0.6\%}$	$3.13_{\pm 4.6\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$27.91_{\pm 0.9\%}$	$5.82_{\pm 3.9\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$19.54_{\pm 0.7\%}$	$42.15_{\pm 1.8\%}$	$24.81_{\pm 0.8\%}$	$21.29_{\pm 1.3\%}$	$20.15_{\pm 0.8\%}$	$46.98_{\pm 2.1\%}$	$22.59_{\pm 1.1\%}$	$21.93_{\pm 1.6\%}$
ConAbl	$23.75_{\pm 0.9\%}$	$30.51_{\pm 2.7\%}$	$25.46_{\pm 1.1\%}$	$19.43_{\pm 1.3\%}$	$23.43_{\pm 0.6\%}$	$32.98_{\pm 1.8\%}$	$25.49_{\pm 1.6\%}$	$18.18_{\pm 0.9\%}$
FMN	$21.63_{\pm 0.8\%}$	$33.51_{\pm 2.2\%}$	$27.08_{\pm 1.4\%}$	$19.98_{\pm 1.1\%}$	$22.96_{\pm 0.9\%}$	$35.87_{\pm 2.8\%}$	$22.73_{\pm 0.9\%}$	$19.60_{\pm 0.9\%}$
EA	$21.82_{\pm 0.7\%}$	$69.30_{\pm 1.5\%}$	$27.23_{\pm 1.4\%}$	$18.50_{\pm 1.8\%}$	$19.91_{\pm 0.9\%}$	$62.87_{\pm 2.6\%}$	$26.40_{\pm 1.5\%}$	$15.28_{\pm 1.1\%}$
GEM	$18.99_{\pm 0.9\%}$	$71.93_{\pm 2.2\%}$	$29.85_{\pm 0.6\%}$	$15.74_{\pm 1.5\%}$	$19.15_{\pm 0.6\%}$	$78.49_{\pm 1.7\%}$	$26.08_{\pm 1.3\%}$	$15.16_{\pm 1.3\%}$
Native	$23.32_{\pm 0.8\%}$	$31.46_{\pm 2.5\%}$	$25.97_{\pm 1.1\%}$	$21.65_{\pm 1.7\%}$	$23.35_{\pm 0.8\%}$	$33.68_{\pm 2.1\%}$	$26.25_{\pm 1.3\%}$	$19.52_{\pm 1.4\%}$
+P1	$22.76_{\pm 0.6\%}$	$60.90_{\pm 1.6\%}$	$28.73_{\pm 1.8\%}$	$17.11_{\pm 1.3\%}$	$22.52_{\pm 0.7\%}$	$55.67_{\pm 2.6\%}$	$24.21_{\pm 1.0\%}$	$16.82_{\pm 1.5\%}$
+P1P2	$23.17_{\pm 0.7\%}$	$64.84_{\pm 2.0\%}$	$27.73_{\pm 1.5\%}$	$17.80_{\pm 1.4\%}$	$22.42_{\pm 0.8\%}$	$58.90_{\pm 2.2\%}$	$26.15_{\pm 1.5\%}$	$15.49_{\pm 1.2\%}$
+P1P2P3	$18.78_{\pm 0.8\%}$	$79.81_{\pm 2.3\%}$	$\mathbf{31.12}_{\pm 1.0\%}$	$14.30_{\pm 1.6\%}$	$20.40_{\pm 0.7\%}$	$75.80_{\pm 1.8\%}$	$26.07_{\pm 1.6\%}$	$14.96_{\pm 1.1\%}$
TRACE	$\mathbf{19.15}_{\pm 1.0\%}$	$\mathbf{83.94}_{\pm 2.7\%}$	$28.50_{\pm 1.2\%}$	$\mathbf{14.40}_{\pm 1.3\%}$	$\mathbf{18.39}_{\pm 0.8\%}$	$\mathbf{85.05}_{\pm 2.0\%}$	$\mathbf{29.76}_{\pm 1.6\%}$	$\mathbf{13.10}_{\pm 1.3\%}$

C Additional Related Work

Complementing the diffusion- and FM-specific methods discussed in Section 2, related concept-erasure research develops adaptive targets and prototype-based guidance (Bui et al., 2025; Li et al., 2026; Cai et al., 2026), as well as objectives for balancing erasure with alignment, model preservation, and preference-based supervision (Ko et al., 2024; Hong et al., 2024; Wu et al., 2025a; Park et al.,

Table 5: Additional attribute/style concepts: happy, sad, and nude. Best edited-model means are **bold**; second-best are underlined.

Method	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
Attribute/Style Concepts Unlearning: happy								
Orig	$31.42_{\pm 0.4\%}$	$2.07_{\pm 4.6\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.49_{\pm 0.8\%}$	$2.13_{\pm 3.2\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$21.09_{\pm 0.8\%}$	$45.12_{\pm 2.2\%}$	$23.22_{\pm 0.6\%}$	$24.20_{\pm 1.5\%}$	$20.57_{\pm 0.5\%}$	$43.89_{\pm 2.7\%}$	$22.83_{\pm 1.5\%}$	$22.01_{\pm 1.9\%}$
ConAbl	$22.88_{\pm 0.5\%}$	$38.04_{\pm 2.0\%}$	$26.81_{\pm 1.8\%}$	$19.30_{\pm 1.1\%}$	$22.64_{\pm 0.6\%}$	$38.75_{\pm 3.5\%}$	$23.13_{\pm 1.2\%}$	$20.06_{\pm 1.3\%}$
FMN	$21.08_{\pm 0.9\%}$	$40.81_{\pm 2.6\%}$	$27.10_{\pm 0.9\%}$	$20.63_{\pm 1.7\%}$	$22.10_{\pm 1.1\%}$	$41.00_{\pm 2.3\%}$	$23.92_{\pm 1.1\%}$	$19.20_{\pm 1.3\%}$
EA	$21.33_{\pm 0.8\%}$	$70.51_{\pm 1.9\%}$	$25.58_{\pm 0.6\%}$	$18.18_{\pm 1.0\%}$	$21.26_{\pm 0.7\%}$	$61.54_{\pm 2.8\%}$	$26.68_{\pm 1.9\%}$	$15.83_{\pm 1.5\%}$
GEM	$20.16_{\pm 0.7\%}$	$73.60_{\pm 1.7\%}$	$29.13_{\pm 1.4\%}$	$14.89_{\pm 1.4\%}$	$21.14_{\pm 0.8\%}$	$76.04_{\pm 2.5\%}$	$27.24_{\pm 0.8\%}$	$14.56_{\pm 1.7\%}$
Native	$23.55_{\pm 0.9\%}$	$40.46_{\pm 2.1\%}$	$26.99_{\pm 1.2\%}$	$20.49_{\pm 1.5\%}$	$21.83_{\pm 1.0\%}$	$38.53_{\pm 2.0\%}$	$24.22_{\pm 1.3\%}$	$18.30_{\pm 1.2\%}$
+P1	$20.67_{\pm 0.6\%}$	$67.78_{\pm 2.5\%}$	$27.71_{\pm 1.2\%}$	$16.64_{\pm 1.3\%}$	$22.41_{\pm 0.9\%}$	$68.31_{\pm 4.0\%}$	$24.74_{\pm 1.1\%}$	$16.35_{\pm 1.6\%}$
+P1P2	$21.10_{\pm 0.5\%}$	$72.89_{\pm 1.5\%}$	$26.99_{\pm 0.5\%}$	$17.14_{\pm 1.1\%}$	$19.85_{\pm 0.6\%}$	$72.61_{\pm 2.9\%}$	$26.81_{\pm 1.5\%}$	$15.27_{\pm 1.2\%}$
+P1P2P3	$19.37_{\pm 0.6\%}$	$82.45_{\pm 1.4\%}$	$30.15_{\pm 0.9\%}$	$12.86_{\pm 1.6\%}$	$\mathbf{19.58}_{\pm 0.7\%}$	$\mathbf{81.22}_{\pm 3.2\%}$	$26.05_{\pm 1.2\%}$	$14.61_{\pm 1.1\%}$
TRACE	$\mathbf{20.07}_{\pm 0.8\%}$	$\mathbf{85.44}_{\pm 2.3\%}$	$\mathbf{29.12}_{\pm 1.6\%}$	$\mathbf{13.64}_{\pm 1.5\%}$	$18.91_{\pm 0.5\%}$	$81.02_{\pm 2.1\%}$	$\mathbf{28.77}_{\pm 1.7\%}$	$\mathbf{13.29}_{\pm 1.7\%}$
Attribute/Style Concepts Unlearning: sad								
Orig	$31.45_{\pm 0.5\%}$	$2.10_{\pm 4.5\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.55_{\pm 0.9\%}$	$2.10_{\pm 2.8\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$21.26_{\pm 0.8\%}$	$45.13_{\pm 2.1\%}$	$23.03_{\pm 0.7\%}$	$24.01_{\pm 1.5\%}$	$20.52_{\pm 0.6\%}$	$44.03_{\pm 3.1\%}$	$23.00_{\pm 1.4\%}$	$21.69_{\pm 1.8\%}$
ConAbl	$22.25_{\pm 0.4\%}$	$38.13_{\pm 2.2\%}$	$27.86_{\pm 1.6\%}$	$19.32_{\pm 1.0\%}$	$23.04_{\pm 0.7\%}$	$38.76_{\pm 3.3\%}$	$23.12_{\pm 1.1\%}$	$20.01_{\pm 1.5\%}$
FMN	$22.13_{\pm 0.7\%}$	$40.93_{\pm 2.5\%}$	$25.75_{\pm 0.8\%}$	$22.07_{\pm 1.8\%}$	$20.78_{\pm 1.2\%}$	$40.98_{\pm 2.4\%}$	$24.16_{\pm 1.0\%}$	$19.30_{\pm 1.2\%}$
EA	$20.37_{\pm 0.9\%}$	$70.49_{\pm 1.8\%}$	$26.98_{\pm 0.7\%}$	$17.13_{\pm 1.1\%}$	$22.30_{\pm 0.8\%}$	$61.66_{\pm 2.7\%}$	$25.33_{\pm 1.8\%}$	$16.98_{\pm 1.4\%}$
GEM	$21.22_{\pm 0.8\%}$	$73.73_{\pm 1.6\%}$	$27.38_{\pm 1.4\%}$	$16.11_{\pm 1.3\%}$	$20.05_{\pm 0.9\%}$	$76.02_{\pm 2.6\%}$	$27.60_{\pm 0.9\%}$	$14.79_{\pm 1.9\%}$
Native	$22.15_{\pm 0.7\%}$	$40.49_{\pm 2.3\%}$	$28.08_{\pm 1.2\%}$	$19.31_{\pm 1.5\%}$	$23.11_{\pm 0.9\%}$	$38.53_{\pm 1.9\%}$	$22.85_{\pm 1.5\%}$	$19.42_{\pm 1.0\%}$
+P1	$21.84_{\pm 0.8\%}$	$67.88_{\pm 2.6\%}$	$26.14_{\pm 1.1\%}$	$17.67_{\pm 1.4\%}$	$21.20_{\pm 0.7\%}$	$68.30_{\pm 3.9\%}$	$26.25_{\pm 1.2\%}$	$15.35_{\pm 1.6\%}$
+P1P2	$20.09_{\pm 0.6\%}$	$72.95_{\pm 1.6\%}$	$28.33_{\pm 0.6\%}$	$16.22_{\pm 1.0\%}$	$21.06_{\pm 0.5\%}$	$72.90_{\pm 3.0\%}$	$25.21_{\pm 1.4\%}$	$16.20_{\pm 1.3\%}$
+P1P2P3	$20.49_{\pm 0.7\%}$	$82.60_{\pm 1.5\%}$	$28.43_{\pm 0.8\%}$	$13.78_{\pm 1.6\%}$	$\mathbf{18.48}_{\pm 0.8\%}$	$\mathbf{81.23}_{\pm 3.3\%}$	$27.73_{\pm 1.3\%}$	$13.83_{\pm 1.2\%}$
TRACE	$\mathbf{19.14}_{\pm 0.6\%}$	$\mathbf{85.61}_{\pm 2.2\%}$	$\mathbf{30.89}_{\pm 1.5\%}$	$\mathbf{12.95}_{\pm 1.4\%}$	$19.99_{\pm 0.4\%}$	$81.12_{\pm 2.0\%}$	$\mathbf{26.93}_{\pm 1.6\%}$	$\mathbf{14.15}_{\pm 1.5\%}$
Attribute/Style Concepts Unlearning: nude								
Orig	$31.51_{\pm 0.5\%}$	$2.08_{\pm 4.7\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.56_{\pm 0.7\%}$	$2.11_{\pm 2.8\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$20.82_{\pm 0.9\%}$	$48.15_{\pm 2.2\%}$	$22.57_{\pm 0.8\%}$	$24.52_{\pm 1.4\%}$	$20.09_{\pm 0.6\%}$	$47.03_{\pm 3.1\%}$	$22.57_{\pm 1.6\%}$	$22.15_{\pm 1.6\%}$
ConAbl	$21.53_{\pm 0.5\%}$	$41.11_{\pm 1.9\%}$	$27.44_{\pm 1.5\%}$	$19.80_{\pm 1.3\%}$	$22.62_{\pm 0.5\%}$	$41.75_{\pm 3.7\%}$	$22.63_{\pm 1.1\%}$	$20.62_{\pm 1.5\%}$
FMN	$21.67_{\pm 0.8\%}$	$43.95_{\pm 2.6\%}$	$25.32_{\pm 1.1\%}$	$22.64_{\pm 1.9\%}$	$20.32_{\pm 0.9\%}$	$44.07_{\pm 2.4\%}$	$23.63_{\pm 1.2\%}$	$19.77_{\pm 1.4\%}$
EA	$19.85_{\pm 0.7\%}$	$74.74_{\pm 2.1\%}$	$26.42_{\pm 0.7\%}$	$17.63_{\pm 1.1\%}$	$21.82_{\pm 0.8\%}$	$64.70_{\pm 2.9\%}$	$24.93_{\pm 1.9\%}$	$17.47_{\pm 1.6\%}$
GEM	$20.70_{\pm 0.6\%}$	$76.74_{\pm 1.6\%}$	$26.99_{\pm 1.5\%}$	$16.47_{\pm 1.3\%}$	$19.65_{\pm 0.7\%}$	$79.12_{\pm 2.8\%}$	$27.05_{\pm 0.9\%}$	$15.28_{\pm 1.6\%}$
Native	$21.70_{\pm 0.9\%}$	$43.53_{\pm 2.3\%}$	$27.56_{\pm 1.4\%}$	$19.85_{\pm 1.5\%}$	$22.64_{\pm 1.0\%}$	$41.97_{\pm 1.8\%}$	$22.37_{\pm 1.5\%}$	$19.97_{\pm 1.2\%}$
+P1	$21.27_{\pm 0.7\%}$	$70.97_{\pm 2.4\%}$	$25.66_{\pm 1.3\%}$	$18.23_{\pm 1.2\%}$	$20.66_{\pm 0.8\%}$	$72.31_{\pm 3.8\%}$	$25.78_{\pm 0.9\%}$	$15.86_{\pm 1.6\%}$
+P1P2	$19.65_{\pm 0.6\%}$	$75.97_{\pm 1.5\%}$	$27.90_{\pm 0.7\%}$	$16.65_{\pm 1.3\%}$	$20.59_{\pm 0.5\%}$	$75.71_{\pm 3.1\%}$	$24.81_{\pm 1.3\%}$	$16.76_{\pm 1.1\%}$
+P1P2P3	$20.02_{\pm 0.6\%}$	$85.60_{\pm 1.4\%}$	$28.00_{\pm 0.9\%}$	$14.29_{\pm 1.7\%}$	$\mathbf{18.02}_{\pm 0.8\%}$	$\mathbf{84.50}_{\pm 3.2\%}$	$27.26_{\pm 1.1\%}$	$14.27_{\pm 1.2\%}$
TRACE	$\mathbf{18.67}_{\pm 0.8\%}$	$\mathbf{88.61}_{\pm 2.4\%}$	$\mathbf{30.41}_{\pm 1.4\%}$	$\mathbf{13.41}_{\pm 1.5\%}$	$19.58_{\pm 0.6\%}$	$84.33_{\pm 2.0\%}$	$\mathbf{26.70}_{\pm 1.8\%}$	$\mathbf{14.69}_{\pm 1.7\%}$

2024). Robustness-oriented methods address adversarial prompts and subsequent fine-tuning (Zhang et al., 2024b; Li et al., 2025; Pan et al., 2024). Complementary interventions modify text representations (Lee et al., 2026b; Xiong et al., 2025), use attention mappings, steering, or spatial gating (Chen et al., 2025; Lee et al., 2025b; Gaintseva et al., 2026; Zhang et al., 2025a; Lee et al., 2025a), and isolate target information through geometric projections or sparse features (Wang et al., 2025; Vardhana & Biswas, 2026; Moon et al., 2026; Biswas et al., 2025; Cywiński & Deja, 2025). Lightweight adapters and fine-grained erasure further address transferability and preservation of related concepts (Lyu et al., 2024; Thakral et al., 2025). These approaches emphasize complementary choices of semantic targets, update objectives, and intervention locations for selective concept removal.

Another line connects erasure to the generation process through concept-specific trajectory shifting (Miao et al., 2026), GFlowNet-based trajectory-balance alignment (Kusumba et al., 2026), and final-output distributional objectives (Zhang et al., 2025b). Large-scale and continual approaches address interference across deletion requests (Lu et al., 2024; Tu et al., 2026; Seo et al., 2026; Heng & Soh, 2023; Lee et al., 2026a), while related work studies concept suppression in pruned diffusion models (Shirkavand et al., 2025). Benchmarking and probing studies examine non-target degradation, multi-faceted erasure quality, and concept recoverability across prompts, latent inputs, and later model updates (Amara et al., 2025; Ren et al., 2025; Moon et al., 2025; Lu et al., 2026; Rusanovsky et al., 2025; George et al., 2025). Within this landscape, TRACE focuses on the construction and coordination of velocity supervision in FM models: it jointly assigns same-source anchor trajectories across seeds using a finite-bank compatibility proxy under a context-preservation constraint, then matches their instantaneous teacher tangents under the forget condition. In particular, EraseFlow also uses anchor trajectories, but optimizes a trajectory-balance objective; TRACE instead couples cross-seed anchor selection with teacher-tangent regression.

Table 6: Additional attribute/style concepts: Monet-style, Simple sketch, and bright. Best edited-model means are **bold**; second-best are underlined.

Method	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
Attribute/Style Concepts Unlearning: Monet-style								
Orig	$31.22_{\pm 0.8\%}$	$7.08_{\pm 3.7\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.12_{\pm 1.1\%}$	$7.68_{\pm 3.6\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$\underline{18.73}_{\pm 0.7\%}$	$44.56_{\pm 2.4\%}$	$24.12_{\pm 0.6\%}$	$22.68_{\pm 1.5\%}$	$20.07_{\pm 0.7\%}$	$46.58_{\pm 2.7\%}$	$22.15_{\pm 1.1\%}$	$21.36_{\pm 1.2\%}$
ConAbl	$21.90_{\pm 0.6\%}$	$38.64_{\pm 1.9\%}$	$28.21_{\pm 1.1\%}$	$19.53_{\pm 1.7\%}$	$22.68_{\pm 0.8\%}$	$38.57_{\pm 3.2\%}$	$24.02_{\pm 1.0\%}$	$18.52_{\pm 1.3\%}$
FMN	$21.95_{\pm 0.9\%}$	$42.99_{\pm 2.6\%}$	$25.13_{\pm 1.4\%}$	$21.85_{\pm 1.1\%}$	$21.15_{\pm 0.6\%}$	$40.38_{\pm 2.1\%}$	$22.21_{\pm 1.5\%}$	$20.57_{\pm 1.8\%}$
EA	$20.84_{\pm 0.9\%}$	$71.10_{\pm 2.1\%}$	$26.24_{\pm 1.5\%}$	$18.42_{\pm 1.4\%}$	$21.06_{\pm 1.0\%}$	$61.43_{\pm 3.8\%}$	$26.47_{\pm 1.2\%}$	$16.32_{\pm 1.5\%}$
GEM	$20.51_{\pm 0.6\%}$	$75.82_{\pm 1.8\%}$	$28.32_{\pm 0.9\%}$	$15.65_{\pm 1.5\%}$	$19.75_{\pm 0.5\%}$	$73.99_{\pm 2.9\%}$	$25.88_{\pm 1.7\%}$	$\underline{15.36}_{\pm 1.2\%}$
Native	$22.84_{\pm 0.8\%}$	$39.29_{\pm 2.3\%}$	$26.42_{\pm 1.2\%}$	$19.82_{\pm 1.0\%}$	$22.56_{\pm 0.9\%}$	$38.99_{\pm 4.0\%}$	$24.97_{\pm 0.7\%}$	$18.79_{\pm 1.4\%}$
+P1	$20.45_{\pm 0.7\%}$	$70.68_{\pm 2.6\%}$	$27.32_{\pm 1.9\%}$	$17.65_{\pm 1.3\%}$	$21.61_{\pm 0.6\%}$	$68.43_{\pm 1.9\%}$	$24.56_{\pm 1.2\%}$	$16.08_{\pm 1.1\%}$
+P1P2	$21.72_{\pm 0.6\%}$	$70.55_{\pm 1.5\%}$	$27.68_{\pm 1.1\%}$	$15.48_{\pm 1.4\%}$	$21.15_{\pm 1.0\%}$	$68.92_{\pm 2.6\%}$	$26.92_{\pm 0.9\%}$	$14.65_{\pm 1.3\%}$
+P1P2P3	$19.98_{\pm 0.9\%}$	$\underline{83.24}_{\pm 2.2\%}$	$\underline{28.85}_{\pm 1.3\%}$	$\underline{14.26}_{\pm 1.2\%}$	$\underline{19.24}_{\pm 0.8\%}$	$\underline{79.92}_{\pm 2.7\%}$	$\mathbf{28.64}_{\pm 0.9\%}$	$\mathbf{13.51}_{\pm 1.5\%}$
TRACE	$\mathbf{18.42}_{\pm 0.5\%}$	$\mathbf{85.29}_{\pm 1.5\%}$	$\mathbf{29.42}_{\pm 0.8\%}$	$\mathbf{13.36}_{\pm 1.6\%}$	$\mathbf{19.64}_{\pm 0.9\%}$	$\mathbf{81.59}_{\pm 3.5\%}$	$\underline{25.92}_{\pm 1.5\%}$	$15.46_{\pm 1.7\%}$
Attribute/Style Concepts Unlearning: Simple sketch								
Orig	$31.42_{\pm 0.7\%}$	$6.84_{\pm 3.9\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.14_{\pm 1.0\%}$	$7.82_{\pm 3.4\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$\underline{18.52}_{\pm 0.7\%}$	$44.78_{\pm 2.4\%}$	$24.08_{\pm 0.6\%}$	$22.95_{\pm 1.6\%}$	$20.12_{\pm 0.7\%}$	$46.81_{\pm 2.5\%}$	$21.91_{\pm 1.1\%}$	$21.76_{\pm 1.0\%}$
ConAbl	$22.94_{\pm 0.6\%}$	$38.39_{\pm 2.1\%}$	$26.64_{\pm 1.1\%}$	$20.21_{\pm 1.8\%}$	$22.42_{\pm 0.9\%}$	$37.31_{\pm 3.2\%}$	$24.87_{\pm 0.8\%}$	$18.56_{\pm 1.3\%}$
FMN	$21.03_{\pm 0.8\%}$	$42.95_{\pm 2.8\%}$	$26.24_{\pm 1.4\%}$	$20.75_{\pm 1.1\%}$	$21.91_{\pm 0.6\%}$	$40.47_{\pm 2.1\%}$	$22.25_{\pm 1.9\%}$	$21.56_{\pm 1.7\%}$
EA	$22.36_{\pm 0.9\%}$	$70.73_{\pm 2.1\%}$	$25.92_{\pm 1.5\%}$	$17.85_{\pm 1.4\%}$	$20.64_{\pm 1.1\%}$	$59.41_{\pm 3.8\%}$	$\underline{27.42}_{\pm 1.0\%}$	$16.24_{\pm 1.8\%}$
GEM	$19.18_{\pm 0.6\%}$	$76.16_{\pm 1.6\%}$	$28.24_{\pm 0.9\%}$	$15.71_{\pm 1.5\%}$	$\underline{20.19}_{\pm 0.5\%}$	$76.36_{\pm 2.9\%}$	$24.96_{\pm 1.4\%}$	$15.93_{\pm 1.2\%}$
Native	$23.13_{\pm 0.8\%}$	$38.85_{\pm 2.3\%}$	$26.49_{\pm 1.2\%}$	$20.53_{\pm 1.1\%}$	$22.68_{\pm 0.9\%}$	$38.72_{\pm 4.0\%}$	$25.16_{\pm 0.6\%}$	$18.49_{\pm 1.4\%}$
+P1	$20.58_{\pm 0.7\%}$	$70.63_{\pm 2.6\%}$	$27.42_{\pm 1.8\%}$	$16.74_{\pm 1.3\%}$	$21.36_{\pm 0.6\%}$	$68.26_{\pm 2.1\%}$	$24.48_{\pm 1.2\%}$	$17.25_{\pm 1.1\%}$
+P1P2	$21.74_{\pm 0.6\%}$	$70.71_{\pm 1.7\%}$	$27.52_{\pm 1.1\%}$	$15.53_{\pm 1.4\%}$	$21.12_{\pm 0.9\%}$	$68.15_{\pm 2.8\%}$	$27.12_{\pm 0.9\%}$	$14.82_{\pm 1.4\%}$
+P1P2P3	$20.18_{\pm 0.9\%}$	$\underline{82.95}_{\pm 2.2\%}$	$\mathbf{28.71}_{\pm 1.3\%}$	$\mathbf{14.28}_{\pm 1.2\%}$	$19.88_{\pm 0.8\%}$	$\underline{78.87}_{\pm 2.7\%}$	$\mathbf{28.58}_{\pm 0.9\%}$	$\mathbf{13.45}_{\pm 1.5\%}$
TRACE	$\mathbf{18.39}_{\pm 0.5\%}$	$\mathbf{85.66}_{\pm 1.5\%}$	$\underline{29.15}_{\pm 0.8\%}$	$\underline{13.42}_{\pm 1.7\%}$	$\mathbf{19.52}_{\pm 1.0\%}$	$\mathbf{81.51}_{\pm 3.5\%}$	$25.86_{\pm 1.5\%}$	$\underline{15.48}_{\pm 1.6\%}$
Attribute/Style Concepts Unlearning: bright								
Orig	$30.41_{\pm 0.8\%}$	$7.06_{\pm 3.8\%}$	$30.47_{\pm 1.7\%}$	$12.72_{\pm 1.2\%}$	$29.12_{\pm 1.1\%}$	$7.75_{\pm 3.5\%}$	$28.80_{\pm 1.4\%}$	$11.95_{\pm 1.5\%}$
ESD	$19.24_{\pm 0.7\%}$	$43.34_{\pm 2.3\%}$	$24.18_{\pm 0.7\%}$	$22.84_{\pm 1.5\%}$	$20.86_{\pm 0.6\%}$	$45.24_{\pm 2.6\%}$	$20.91_{\pm 1.2\%}$	$23.12_{\pm 1.1\%}$
ConAbl	$22.14_{\pm 0.5\%}$	$39.70_{\pm 2.0\%}$	$25.26_{\pm 1.0\%}$	$21.45_{\pm 1.9\%}$	$21.58_{\pm 0.9\%}$	$38.59_{\pm 3.1\%}$	$23.41_{\pm 0.7\%}$	$19.66_{\pm 1.4\%}$
FMN	$21.65_{\pm 0.9\%}$	$41.54_{\pm 2.9\%}$	$26.54_{\pm 1.5\%}$	$20.61_{\pm 1.0\%}$	$22.75_{\pm 0.7\%}$	$39.06_{\pm 2.2\%}$	$22.52_{\pm 1.5\%}$	$21.36_{\pm 1.6\%}$
EA	$21.51_{\pm 0.8\%}$	$73.13_{\pm 2.2\%}$	$24.58_{\pm 1.6\%}$	$19.04_{\pm 1.5\%}$	$19.98_{\pm 1.0\%}$	$61.51_{\pm 3.7\%}$	$25.85_{\pm 1.1\%}$	$17.42_{\pm 1.9\%}$
GEM	$19.95_{\pm 0.5\%}$	$73.53_{\pm 1.7\%}$	$27.56_{\pm 1.0\%}$	$15.75_{\pm 1.6\%}$	$20.84_{\pm 0.6\%}$	$73.70_{\pm 3.0\%}$	$25.32_{\pm 1.6\%}$	$15.98_{\pm 1.3\%}$
Native	$22.24_{\pm 0.9\%}$	$40.29_{\pm 2.2\%}$	$25.08_{\pm 1.1\%}$	$21.94_{\pm 1.1\%}$	$21.82_{\pm 0.8\%}$	$40.17_{\pm 3.9\%}$	$23.64_{\pm 0.6\%}$	$19.86_{\pm 1.3\%}$
+P1	$21.36_{\pm 0.8\%}$	$68.16_{\pm 2.5\%}$	$27.76_{\pm 1.7\%}$	$16.82_{\pm 1.2\%}$	$22.18_{\pm 0.5\%}$	$65.91_{\pm 2.0\%}$	$24.78_{\pm 1.3\%}$	$17.21_{\pm 1.0\%}$
+P1P2	$21.02_{\pm 0.5\%}$	$73.25_{\pm 1.6\%}$	$25.92_{\pm 1.2\%}$	$16.74_{\pm 1.5\%}$	$20.37_{\pm 0.9\%}$	$70.61_{\pm 2.7\%}$	$25.56_{\pm 0.8\%}$	$15.86_{\pm 1.2\%}$
+P1P2P3	$\underline{19.06}_{\pm 0.6\%}$	$\underline{82.84}_{\pm 1.4\%}$	$\mathbf{29.74}_{\pm 0.9\%}$	$\mathbf{14.05}_{\pm 1.6\%}$	$\underline{20.49}_{\pm 0.9\%}$	$\underline{78.81}_{\pm 3.6\%}$	$\mathbf{26.43}_{\pm 1.4\%}$	$\mathbf{14.92}_{\pm 1.5\%}$
TRACE	$\mathbf{19.41}_{\pm 0.8\%}$	$\mathbf{85.95}_{\pm 2.3\%}$	$\underline{27.88}_{\pm 1.2\%}$	$\underline{15.38}_{\pm 1.3\%}$	$\mathbf{18.66}_{\pm 1.0\%}$	$\mathbf{81.75}_{\pm 2.8\%}$	$\underline{27.02}_{\pm 1.0\%}$	$\underline{14.48}_{\pm 1.4\%}$

Table 7: Sensitivity to the number of source seeds N for Van Gogh style. The default is $N = 100$. Best means in each metric are bold; the default row is shaded.

Method (TRACE)	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
$N = 10$	$21.95_{\pm 0.8\%}$	$38.50_{\pm 2.3\%}$	$26.85_{\pm 1.5\%}$	$20.10_{\pm 1.6\%}$	$22.60_{\pm 0.7\%}$	$37.50_{\pm 3.5\%}$	$23.55_{\pm 1.0\%}$	$19.50_{\pm 1.4\%}$
$N = 50$	$20.10_{\pm 0.6\%}$	$74.50_{\pm 1.8\%}$	$\mathbf{29.80}_{\pm 0.9\%}$	$\mathbf{13.50}_{\pm 1.1\%}$	$19.95_{\pm 0.8\%}$	$75.20_{\pm 2.6\%}$	$\mathbf{28.10}_{\pm 1.2\%}$	$\mathbf{13.40}_{\pm 1.3\%}$
$N = 100$ (Default)	$19.23_{\pm 0.9\%}$	$84.41_{\pm 2.2\%}$	$29.05_{\pm 1.3\%}$	$14.09_{\pm 1.2\%}$	$19.55_{\pm 0.8\%}$	$80.28_{\pm 2.7\%}$	$27.31_{\pm 0.9\%}$	$14.12_{\pm 1.5\%}$
$N = 200$	$19.21_{\pm 0.7\%}$	$84.45_{\pm 2.0\%}$	$29.06_{\pm 1.4\%}$	$14.07_{\pm 1.3\%}$	$19.53_{\pm 0.9\%}$	$80.32_{\pm 2.5\%}$	$27.33_{\pm 1.1\%}$	$14.10_{\pm 1.6\%}$
$N = 300$	$19.20_{\pm 0.8\%}$	$84.50_{\pm 2.1\%}$	$29.10_{\pm 1.2\%}$	$14.03_{\pm 1.4\%}$	$19.50_{\pm 1.0\%}$	$80.38_{\pm 2.8\%}$	$27.36_{\pm 0.8\%}$	$14.05_{\pm 1.4\%}$
$N = 500$	$\mathbf{18.95}_{\pm 0.9\%}$	$\mathbf{86.55}_{\pm 1.9\%}$	$29.52_{\pm 1.1\%}$	$13.82_{\pm 1.2\%}$	$\mathbf{19.25}_{\pm 0.6\%}$	$\mathbf{82.85}_{\pm 2.4\%}$	$27.71_{\pm 1.2\%}$	$13.80_{\pm 1.3\%}$

Table 8: Sensitivity to the candidate anchor count Q for Van Gogh style. The default is $Q = 10$. Best means in each metric are bold; the default row is shaded.

Method (TRACE)	Stable Diffusion 3.5 Medium				FLUX.2 [klein] 4B			
	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓	CLIP(U)↓	Anchor-win(%)↑	CLIP(R)↑	FID↓
$Q = 3$	$20.50_{\pm 0.8\%}$	$78.50_{\pm 2.1\%}$	$29.03_{\pm 1.3\%}$	$14.11_{\pm 1.2\%}$	$20.80_{\pm 0.9\%}$	$74.50_{\pm 2.5\%}$	$27.29_{\pm 1.0\%}$	$14.14_{\pm 1.4\%}$
$Q = 5$	$19.65_{\pm 0.8\%}$	$82.10_{\pm 2.0\%}$	$29.04_{\pm 1.2\%}$	$14.10_{\pm 1.3\%}$	$19.85_{\pm 0.7\%}$	$78.50_{\pm 2.4\%}$	$27.30_{\pm 0.9\%}$	$14.13_{\pm 1.5\%}$
$Q = 10$ (Default)	$19.23_{\pm 0.9\%}$	$84.41_{\pm 2.2\%}$	$29.05_{\pm 1.3\%}$	$14.09_{\pm 1.2\%}$	$19.55_{\pm 0.8\%}$	$80.28_{\pm 2.7\%}$	$27.31_{\pm 0.9\%}$	$14.12_{\pm 1.5\%}$
$Q = 20$	$19.21_{\pm 0.9\%}$	$84.50_{\pm 2.1\%}$	$29.15_{\pm 1.4\%}$	$14.04_{\pm 1.1\%}$	$19.53_{\pm 0.7\%}$	$80.35_{\pm 2.6\%}$	$27.42_{\pm 1.0\%}$	$14.05_{\pm 1.6\%}$
$Q = 50$	$\mathbf{19.20}_{\pm 0.8\%}$	$\mathbf{84.60}_{\pm 2.0\%}$	$\mathbf{29.50}_{\pm 1.3\%}$	$\mathbf{13.85}_{\pm 1.2\%}$	$\mathbf{19.51}_{\pm 0.8\%}$	$\mathbf{80.40}_{\pm 2.5\%}$	$\mathbf{27.65}_{\pm 0.8\%}$	$\mathbf{13.85}_{\pm 1.4\%}$

Table 9: Wall-clock runtime. Durations are in minutes:seconds. Total-time annotations are relative standard deviations in percent; stage times are reported separately without uncertainty estimates.

Method	SD3.5M	FLUX.2 [klein] 4B
ESD	$27{:}18_{\pm 2.2\%}$	$35{:}07_{\pm 1.9\%}$
ConAbl	$29{:}55_{\pm 2.1\%}$	$42{:}13_{\pm 1.8\%}$
FMN	$18{:}20_{\pm 2.8\%}$	$21{:}40_{\pm 2.3\%}$
EA	$24{:}11_{\pm 1.9\%}$	$27{:}05_{\pm 2.5\%}$
GEM	$6{:}46_{\pm 2.3\%}$	$5{:}17_{\pm 2.2\%}$
TRACE (total)	$11{:}36_{\pm 2.5\%}$	$10{:}24_{\pm 2.0\%}$
Stage 1: pre-unlearning	8:45	8:16
Stage 2: parameter updating	2:50	2:08