DIFFUSION UNLEARNING VIA EXACT CONTINUOUS MIXTURE SCORES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to mitigate exposure to sensitive training data and undesirable semantic concepts without retraining foundation diffusion models from scratch. While there are concept unlearning methods relying only on inference-time steering, data unlearning has predominantly relied on parameter fine-tuning, an approach that is computationally intensive and requires synthetic retain sets to prevent generation performance collapse. Emerging training-free alternatives circumvent retraining by steering reverse sampling trajectories away from target data, but they choose their repulsion force heuristically and avoid explicit computation of posterior terms as intractable. Conversely, we show that the mathematically principled repulsion weight is the instantaneous continuous-time posterior odds of entering forbidden data under the variance-preserving stochastic differential equation. We establish that unconditional data unlearning and conditional concept erasure share a common mathematical formulation, unifying both within a single training-free operator. We show that our proposed method, namely, Exact Mixture Score (EMIS) constitutes the exact spatial score of an explicit normalized mixture density with an *a-priori* closed-form leakage law and replaces heuristic time-window cutoffs with an uncertainty-aware continuous temperature schedule. Across unconditional pixel models (CelebA-HQ 256) and conditional text-to-image synthesis (Stable Diffusion on UNLEARNCANVAS for both style- and object-unlearning), EMIS not only outperforms heuristic training-free diffusion guidance methods but also matches the unlearning efficacy of full parameter fine-tuning while preserving generative fidelity on retained concepts with zero weight updates.

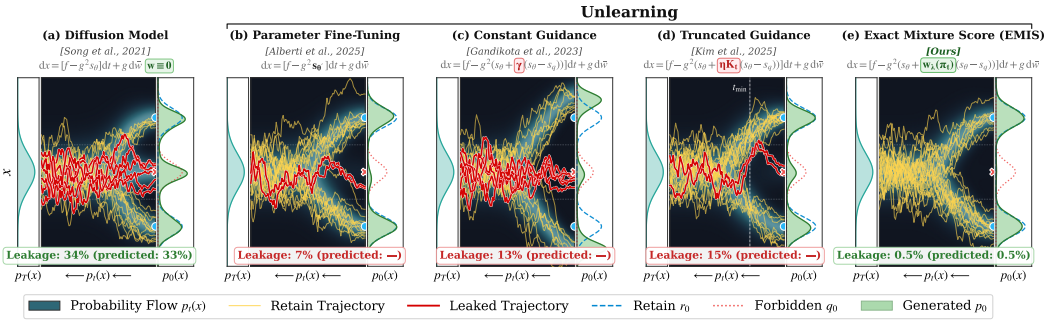

Figure 1: **Training-free unlearning reduces to choosing a single repulsion weight.** Each panel traces reverse diffusion trajectories from noise $p_T(x)$ (left) to the data distribution $p_0(x)$ (right, green fill) on a trimodal mixture of two retain modes (r_0, dashed blue) and a central forbidden mode (q_0, dotted red, prior $\pi = 1/3$), highlighting trajectories that land in the forbidden basin in red. **(a)** The base model ($w \equiv 0$) emits q_0 at the prior rate (34%), while **(b)** parameter fine-tuning ($s_{\theta'}$) leaves 7% unquantified leakage. Panels **(c)–(e)** integrate the exact same reverse SDE along $s_\theta - s_q$ and differ only in the boxed repulsion weight: **(c)** constant guidance (γ) pushes the generated modes (green) outward past r_0 (dashed blue), while **(d)** truncated guidance (ηK_t) shuts off at $t_{\min}$ and lets inner trajectories snap back into q_0 (15%). **(e) EMIS (Ours)** weights repulsion by the clipped posterior odds $w_\lambda(\pi_t)$ computed from the frozen model, preserving an exact mixture score (Theorem 3.2) that aligns the generated density with r_0 and matches the closed-form leakage prediction (0.5%).

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.

1 INTRODUCTION

Diffusion models have risen to be one of the most capable families of methods for synthesizing high-dimensional data by iteratively reversing a stochastic noise corruption process (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). As models scale to web-sized corpora, recent studies highlight privacy and safety concerns in deployed systems (Carlini et al., 2023; Somepalli et al., 2023). To address these, *machine unlearning* aims to remove targeted data samples and undesirable concepts post hoc without retraining the generator from scratch.

Yet current literature splits along two disconnected tracks. In *concept unlearning*, text prompts specify the target, allowing classifier-free guidance heuristics to deflect reverse trajectories at generation time while leaving generator weights untouched (Schramowski et al., 2023; Brack et al., 2023; Jain et al., 2024). In *training data unlearning*, the target is a cohort of individual samples with no text token to negate (Alberti et al., 2025; Wu et al., 2025); consequently, methods default to parameter fine-tuning (Gandikota et al., 2023; Kumari et al., 2023; Fan et al., 2024), which requires an optimization run and a gigabyte-scale checkpoint per request, synthetic retain datasets to prevent collapse, and risks eroding model performance. Across both tracks, existing training-free methods modify the reverse diffusion drift along the same repulsion direction and differ only in how they heuristically weight that force, relying on static multipliers that cause contrast burnout, ad-hoc spatial barriers, or unnormalized kernels restricted to manual time windows (Schramowski et al., 2023; Kirchhof et al., 2025; Kim et al., 2025). They resort to heuristics because the principled repulsion weight, the instantaneous *posterior odds* that a noisy trajectory will land on forbidden data, was deemed intractable under the assumption that isolating the safe retain distribution requires integrating over all non-target samples (Kim et al., 2025). Without the exact odds, existing guidance terms cannot predict or bound residual leakage in advance and trade erasure against generation fidelity (Figure 1).

We show that evaluating the posterior odds does not require access to the safe retain data because the pre-trained model already encodes the full data marginal. Since forward corruption is conditionally independent of whether a sample is safe or not given the clean sample, Bayes' rule expresses the forget probability as the expectation of clean membership under the base model's own denoising posterior. This expectation is evaluated directly from the Tweedie clean estimate computed at every step, requiring no retain data, no retraining, and zero extra network evaluations for instances. Incorporating posterior variance naturally tracks the signal-to-noise ratio at each step, avoiding premature deflections on trajectories that would naturally produce safe samples. Bounding the odds ensures numerical stability near target boundaries while preserving the exact score of an explicit normalized mixture density with a closed-form leakage law (Theorem 3.2). Consequently, we name this resulting method **Exact Mixture Score (EMIS)**, with contributions summarized as follows:

- **Tractable posterior odds without retain data.** We show that all existing training-free unlearning heuristics approximate the continuous-time posterior odds; we resolve their apparent intractability by evaluating them directly under the base model's denoising posterior without accessing the safe retain data (Lemma 3.1).
- **From empirical knobs to a safety specification.** We prove that bounding the posterior odds preserves the exact score of an explicit normalized mixture density with a closed-form law on residual leakage (Theorem 3.2), turning guidance strength into an a-priori safety guarantee.
- **A single operator for concepts and instances.** EMIS expresses both concept erasure and instance unlearning through a clean-space membership function $m(x_0)$ within a single score formulation. Given a reliable membership specification (via spatial distance for instances or a concept classifier on x_0), EMIS reaches the parameter fine-tuning erasure regime on targeted instances with zero weight updates.
- **Balanced erasure with zero storage cost.** By steering trajectories dynamically via $m(x_0)$, EMIS avoids axis collapse across both style-unlearning and object-unlearning tasks while matching the strongest weight-space methods without fine-tuning the diffusion network.

2 BACKGROUND AND RELATED WORK

Let p_0 be the data distribution of a pretrained diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021), partitioned into a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$ with membership

indicator $m(x_0) = \mathbf{1}\{x_0 \in \mathcal{D}_f\}$. Conditioning p_0 on this membership splits the clean distribution into a retained density r_0 on $\mathcal{D}_r$ and a forbidden density q_0 on $\mathcal{D}_f$,

$$p_0(x_0) = (1-\pi)r_0(x_0) + \pi q_0(x_0), \qquad \pi = p_0(x_0 \in \mathcal{D}_f) = \mathbb{E}_{p_0}[m(x_0)].$$

Since the forward Gaussian transition kernel $p_{0t}(x_t \mid x_0) = \mathcal{N}(x_t;\, \alpha_t x_0,\, \sigma_t^2 \mathbf{I})$ is linear, this decomposition holds at every noise level $t \in [0, T]$ as $p_t(x_t) = (1-\pi)r_t(x_t) + \pi q_t(x_t)$. Differentiating $\log r_t(x_t)$ (equivalently formulated in clean denoiser space $\hat{x}_0(x_t, t) = \mathbb{E}[x_0 \mid x_t]$ (Efron, 2011) by Kim et al. (2025), Theorem 3.2; see Section D.1) relates the retain score $s_r \triangleq \nabla_{x_t} \log r_t$ to the base score $s \triangleq \nabla_{x_t} \log p_t$ and forbidden score $s_q \triangleq \nabla_{x_t} \log q_t$:

$$s_r(x_t, t) = s(x_t, t) + \mathcal{O}_t(x_t)\big[s(x_t, t) - s_q(x_t, t)\big], \qquad \mathcal{O}_t(x_t) \triangleq \frac{\pi\, q_t(x_t)}{(1-\pi)\, r_t(x_t)}. \tag{1}$$

Sampling from r_0 with a frozen generator therefore reduces to shifting the base score along $s(x_t, t) - s_q(x_t, t)$ by the instantaneous posterior odds $\mathcal{O}_t(x_t)$, a shift identical to continuous-time Bayesian classifier guidance $\nabla_{x_t} \log p_t(x_0 \in \mathcal{D}_r \mid x_t)$ (Dhariwal & Nichol, 2021) (Section D.1).

Inference-time trajectory steering: When the target is defined by a text prompt, negative classifier-free guidance (Ho & Salimans, 2022) shows a repulsion direction $s(x_t, t) - s_q(x_t, t)$. Existing training-free methods differ in the choice of the heuristic weight $w(x_t, t)$ in lieu of $\mathcal{O}_t(x_t)$. SLD (Schramowski et al., 2023) and SEGA (Brack et al., 2023) gate a constant scale on latent score differences, while Ortho-Guidance (Sun et al., 2026), SEOT (Li et al., 2024), SAFREE (Yoon et al., 2025), and TraSCE (Jain et al., 2024) project out forbidden text embeddings or modulate negative guidance with static Gaussian kernels. In unconditional instance unlearning, where no conditioning vector exists to negate, SPELL (Kirchhof et al., 2025) deflects trajectories using a piecewise-linear radial barrier around the target images. Safe Denoiser (Kim et al., 2025) derives the clean-space counterpart of Equation (1) but deems the retain denominator $(1-\pi)r_t(x_t)$ intractable, replacing $\mathcal{O}_t(x_t)$ with an unnormalized kernel over a fixed high-noise time window (Kim et al., 2026). The discrepancy with the true $\mathcal{O}_t(x_t)$ means the guided vector field is no longer the score of a normalized density, trading erasure against sample fidelity without a closed-form bound on residual leakage.

Parameter fine-tuning and auxiliary adapters: Rather than intervening during sampling, a parallel line of work modifies the generator directly via parameter fine-tuning (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2024a; Wu et al., 2025; Wu & Harandi, 2024; Alberti et al., 2025), gradient-saliency masking (Fan et al., 2024), or auxiliary sparse autoencoders trained on internal activations (Cywiński & Deja, 2025; Cassano et al., 2026). While weight updates eliminate per-step guidance at inference time, they require a dedicated optimization run and checkpoint for every target set $\mathcal{D}_f$ (or multi-day adapter pretraining), auxiliary samples from $\mathcal{D}_r$ to prevent model collapse, and suffer retention loss when erasing object categories that share representations with $\mathcal{D}_r$ (Section 4.3).

3 Unlearning is Posterior-Odds Reweighting

Equation (1) reduces unlearning to a single scalar: the posterior odds $\mathcal{O}_t(x_t)$. We show that these odds can be computed from the frozen, pre-trained model without access to the safe/retain data $\mathcal{D}_r$. Section 3.1 describes how to compute and bound the odds to ensure a stable guidance field and Section 3.2 shows how to evaluate the membership gate by approximating the denoising posterior.

3.1 Computing and Bounding the Odds

Evaluating $\mathcal{O}_t$ appears to require the diffused retain component,

$$(1-\pi)\, r_t(x_t) = \int p_{0t}(x_t \mid x_0)\, \big(1 - m(x_0)\big)\, p_0(x_0)\, \mathrm{d}x_0, \tag{2}$$

an integral computed over all elements of p_0 which are not in the forget set. As this requires evaluating a high-dimensional integral over the clean retain data $\mathcal{D}_r$, computing it directly at inference time is computationally prohibitive. This apparent barrier is precisely what motivates Kim et al. (2025) to bypass the exact odds in favour of an unnormalized kernel with a hand-tuned scale.

This dependency is, however, only apparent. Because whether a data sample x_0 contains a forbidden concept is determined entirely via the membership probability $p(y = \text{forget} \mid x_0) = m(x_0)$, while

the forward transition kernel $p_{0t}(x_t \mid x_0)$ is agnostic to y, the joint process forms the Markov chain $y \to x_0 \to x_t$ ($y \perp\!\!\!\perp x_t \mid x_0$). Formalizing this structure converts the intractable retain-set integral (Equation (2)) into an expectation over the frozen base model's denoising posterior.

Lemma 3.1 (Retain-free posterior odds). *Let p_0 be partitioned by a clean-space membership function $m : \mathbb{R}^D \to [0,1]$ with prior $\pi = \mathbb{E}_{p_0}[m(x_0)] \in (0,1)$, and let $y \to x_0 \to x_t$ evolve under the forward kernel $p_{0t}(x_t \mid x_0)$. Then for all (x_t, t), the diffused retain marginal satisfies $(1-\pi)r_t(x_t) = \big(1-\pi_t(x_t)\big)p_t(x_t)$, and the posterior forget probability (*gate*) $\pi_t(x_t)$ and exact unlearning odds $\mathcal{O}_t(x_t)$ depend on r_0 and π solely through the base denoising posterior $p(x_0 \mid x_t)$:*

$$\pi_t(x_t) \;\triangleq\; p_t(y = \text{forget} \mid x_t) \;=\; \mathbb{E}_{x_0 \sim p(x_0 \mid x_t)}\big[m(x_0)\big], \qquad \mathcal{O}_t(x_t) = \frac{\pi_t(x_t)}{1-\pi_t(x_t)}. \tag{3}$$

Lemma 3.1 (see Appendix D.2 for proof) reveals that no data from the retain set is required to evaluate $\mathcal{O}_t(x_t)$, as the integration over the retained manifold was already implicitly performed during pre-training and is recovered at sampling time by averaging $m(x_0)$ over $p(x_0 \mid x_t)$. While the exact posterior $p(x_0 \mid x_t)$ is not available in closed form for general data distributions, we show in Section 3.2 that approximating $p(x_0 \mid x_t)$ with a Gaussian centered at Tweedie's posterior mean $\hat{x}_0(x_t, t)$ and incorporating its directional variance yields an analytical evaluation of $\mathbb{E}_{x_0 \sim p(x_0 \mid x_t)}[m(x_0)]$.

The odds are exact, and unbounded: While Lemma 3.1 shows that the posterior odds are exact and retain-free, Equation (3) also reveals that they are inherently unbounded. Since a perfectly unlearned model must assign zero density to the forbidden region, the diffused retain density $r_t(x_t)$ vanishes ($\pi_t(x_t) \to 1$), driving the posterior odds $\mathcal{O}_t(x_t) \to +\infty$. This divergence is an intrinsic property of the unlearning task itself rather than an artifact of estimation; no refinement in how we evaluate the odds can prevent this singularity. Consequently, simulating the exact reverse SDE directly is numerically unstable near the target, as the infinite drift velocity causes solver collapse.

To ensure the guidance field remains stable and corresponds to a mathematically valid score, we reformulate the target distribution itself. Specifically, for a chosen regularization parameter $\lambda > 0$ yielding a maximum odds ceiling $\mathcal{O}_{\max} \triangleq \lambda^{-1}$, we replace the exact odds with the clipped odds:

$$\boxed{\tilde{s}_{\text{EMIS}}(x_t, t) \equiv \tilde{s}_\lambda(x_t, t) \;\triangleq\; s(x_t, t) + \underbrace{\frac{\pi_t(x_t)}{1-\pi_t(x_t)+\lambda}}_{w_\lambda(\pi_t) \le \mathcal{O}_{\max}} \Big[s(x_t, t) - s_q(x_t, t)\Big]} \tag{4}$$

This modification ensures that w_λ matches the true odds $\mathcal{O}_t$ closely wherever the gate π_t is far from one, while saturating at $\mathcal{O}_{\max}$ near the target boundary to prevent SDE solver collapse. Crucially, this clipping is not an ad-hoc heuristic (see Theorem 3.2); $\tilde{s}_{\text{EMIS}}$ is the exact score of an explicit regularized mixture density. We therefore call our approach the **Exact Mixture Score (EMIS)** guidance.

Theorem 3.2 (EMIS is mixture sampling). *Let $p_t = (1-\pi)r_t + \pi q_t$ with $\pi \in (0,1)$ and $\lambda > 0$. Then:*

1. ***Bounded force:*** *$w_\lambda(\cdot)$ is strictly increasing on $[0,1]$ with $0 \le w_\lambda(\pi_t(x_t)) \le \mathcal{O}_{\max}\ \forall (x_t, t)$.*
2. ***Exact score:*** *For every $t \in (0, T]$, $\tilde{s}_\lambda(x_t, t) = \nabla_{x_t} \log \rho_t^{(\lambda)}(x_t)$ for the normalized density*
$$\rho_t^{(\lambda)} = \frac{(1-\pi)\, r_t + \lambda\, p_t}{(1-\pi) + \lambda}. \tag{5}$$
3. ***Consistency:*** *$\{\rho_t^{(\lambda)}\}_{t \in [0,T]}$ is the forward VP-SDE evolution of $\rho_0^{(\lambda)}$, so in the continuous-time limit reversing the SDE with $\tilde{s}_\lambda$ samples exactly from $\rho_0^{(\lambda)}$.*
4. ***Leakage in closed form:*** *The odds that a draw from $\rho_0^{(\lambda)}$ has $y = \text{forget}$ are reduced from those of the base model by a factor independent of π,*
$$\frac{\mathcal{O}^{\text{out}}(\lambda)}{\mathcal{O}^{\text{prior}}} = \frac{\lambda}{1+\lambda} = \frac{1}{1+\mathcal{O}_{\max}}, \qquad \lim_{\lambda \to 0^+} \rho_0^{(\lambda)} \overset{\text{TV}}{=} r_0, \tag{6}$$
where $\overset{\text{TV}}{=}$ denotes equality in total variation.

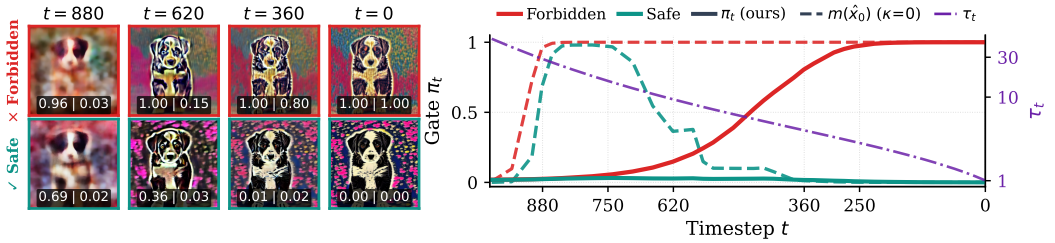

Figure 2: **Tweedie estimates and gating dynamics. Left:** Tweedie estimates $\hat{x}_0$ for forbidden (*Sponge Dabbed*, top) and safe (*Pastel*, bottom) trajectories (same noise seed), overlaid with $[m(\hat{x}_0) \mid \pi_t(x_t)]$. **Right:** Calibrated gate $\pi_t(x_t)$ (solid) vs. plug-in estimate $m(\hat{x}_0) = \pi_t|_{\kappa=0}$ (dashed), alongside temperature τ_t (right axis; $\sigma_0 = 1.0, \kappa = 3.8$). At high noise ($t = 880$), before style-specific brushwork resolves in $\hat{x}_0$, the uncalibrated plug-in $m(\hat{x}_0)$ (dashed; left badge value) fires a false positive on the safe *Pastel* prompt (0.69, peaking at 0.98), whereas our uncertainty-calibrated gate $\pi_t(x_t)$ (solid; right badge value) suppresses premature activation under high τ_t and fires only on the forbidden *Sponge Dabbed* trajectory once stylistic texture emerges ($t \leq 620$).

Proof sketch. Claim (1) is immediate from $w'_\lambda(u) > 0$. For (2), writing $\nabla \log \rho_t^{(\lambda)}$ as the density-weighted score average and substituting $(1 - \pi)r_t = (1 - \pi_t)p_t$ with Equation (1) simplifies algebraically to $\tilde{s}_\lambda$. Claims (3) and (4) follow by linearity of the forward kernel and expanding $\rho_0^{(\lambda)}$ as a convex combination of r_0 and q_0. See Appendix D.3 for the full proof. □

Theorem 3.2 reveals the physical meaning of this regularization. Rather than weakening the guidance, clipping represents a deliberate, controlled trade-off. The resulting target $\rho^{(\lambda)}$ is precisely the retain distribution blended with a λ-weighted fraction of the base model. This formulation offers a remarkable practical advantage: because the unknown prior probability π (which is typically inaccessible) cancels out of the ratio, the continuous-time reduction in leakage is predictable under the exact score. In the continuous-time limit, the odds of sampling forbidden content are reduced by a factor of $1 + \mathcal{O}_{\max}$ compared to the base model. However, λ cannot be driven arbitrarily to zero since $\mathcal{O}_{\max} = \lambda^{-1}$ caps the repulsion multiplier in Equation (4); making λ too small re-introduces excessive guidance forces and might result in numerical instability during integration.

3.2 Estimating the Gate

Unifying concept and data unlearning: A key strength of the posterior-odds framework is its abstraction over the forgotten content, as the formulation established in Section 3.1 remains valid for any choice of the membership function $m(x_0)$. Specifying $m(x_0)$ fully defines the unlearning task because the clean forbidden density $q_0(x_0)$ is uniquely determined through Bayes' rule as:

$$q_0(x_0) = \frac{m(x_0)\, p_0(x_0)}{\pi}, \qquad \pi = \mathbb{E}_{p_0}\big[m(x_0)\big], \tag{7}$$

which subsequently fixes the corresponding forbidden score s_q. In this light, two unlearning paradigms that are typically treated separately (unconditional instance unlearning and conditional concept erasure) emerge naturally as two distinct specifications of the same underlying membership operator:

$$\underbrace{m(x_0) = \exp\left(-\frac{\|x_0 - x^*\|_2^2}{2D\sigma_0^2}\right)}_{\text{Instance unlearning of image } x^*} \qquad \underbrace{m(x_0) = \sigma(\ell(x_0)/\sigma_0)}_{\text{Concept erasure with log-odds } \ell} \tag{8}$$

Here, $\sigma(z)$ denotes the standard logistic sigmoid function, $\ell(\cdot)$ is a clean-space concept logit margin function, and $\sigma_0 > 0$ is a smoothness parameter that controls boundary permissiveness.

Despite their shared mathematical origin, the two specifications exhibit complementary computational behaviors. Instance unlearning assumes the base density p_0 is locally flat, so Equation (7) centers the forbidden distribution q_0 as a Gaussian around the target x^* or a Gaussian mixture for a set. This localized structure allows the diffused score s_q to be evaluated in closed form, requiring no network passes. Conversely, for concept unlearning, the membership gate determines *when* and *how strongly* a trajectory belongs to the forbidden concept via the clean-image classifier $\ell(x_0)$, while the repulsion

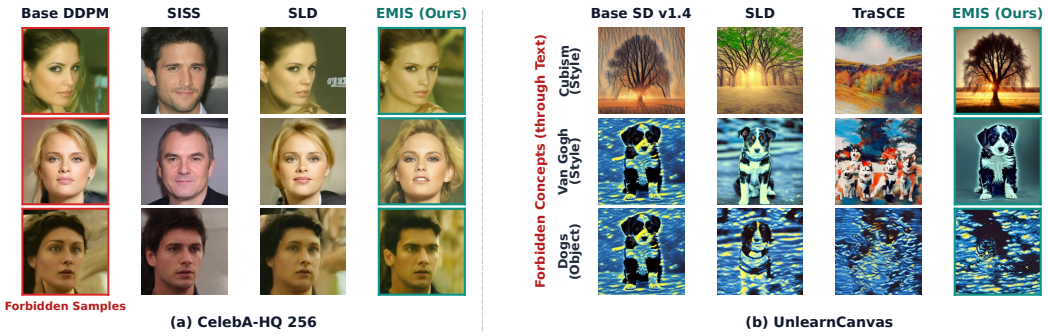

Figure 3: **Unlearning target instances, artistic styles, and canonical objects. (a) CelebA-HQ 256** (M=50): Generations from identical initial noise seeds across the unmodified Base DDPM. While SISS distorts global facial geometry and demographic attributes, and SLD fails to fully erase the target instance, EMIS successfully generates a face with a natural-looking surrogate while preserving background structure, lighting, and pose. **(b) UNLEARNCANVAS:** Representative style (*Cubism*, *Van Gogh*) and object (*Dogs*) erasure tasks. EMIS achieves unlearning while maintaining high fidelity to non-target prompt elements. Extra multi-seed and multi-prompt grids are provided in Section C.4.

direction $s(x_t, t) - s_q(x_t, t)$ is obtained directly from the model's own conditional predictions. That is, q_0 matches the model's conditional distribution given the target concept c^*, so s_q is the conditional score the network already produces. In this case, the unlearning direction $s - s_q$ naturally recovers standard classifier-free guidance, but with its sign reversed. The magnitude is dynamically modulated by the classifier-driven posterior odds $w_\lambda[\pi_t(x_t)]$ to actively repel the targeted concept.

Averaging membership over the denoising posterior: Evaluating the gate in Equation (3) requires averaging the membership function m over clean images the reverse process could generate from noisy state x_t. However, Tweedie's formula provides only the posterior mean $\hat{x}_0$, and substituting this to yield a plug-in estimate $m(\hat{x}_0)$ is problematic early in sampling. Since $\hat{x}_0$ begins as a blurry average of possible reconstructions, this deterministic plug-in estimate suffers from severe estimation bias on ambiguous states. This limitation is illustrated in the left panel of Figure 2, which depicts two trajectories that differ only in style and remain visually indistinguishable at $t = 880$. On these ambiguous states, the plug-in prematurely reports high membership (0.96 and 0.69), triggering a false alarm that unnecessarily deflects a trajectory naturally converging to a non-target region. To suppress these premature activations, prior work resorts to heuristic temporal masking, disabling guidance entirely outside a hand-tuned window (Kim et al., 2025; 2026).

We resolve these limitations by approximating the denoising posterior with a Gaussian centered at Tweedie's mean and exact covariance. Integrating either membership function in Equation (8) against this Gaussian yields a closed-form expectation equivalent to evaluating the original membership at posterior mean $\hat{x}_0$, with base bandwidth σ_0 widened to an uncertainty-aware temperature schedule:

$$\tau_t = \sqrt{\sigma_0^2 + \kappa^2 \, v_{0|t}}. \tag{9}$$

Physically, this sums two Gaussian variances for instance unlearning and recovers MacKay's logistic probit approximation (MacKay, 1992) for concept erasure. Remarkably, this high-dimensional integral requires only a scalar projection of the posterior covariance: directional variance $v_{0|t}$ along fastest membership change (toward x^* for instances or the classifier gradient for concepts). We compute $v_{0|t}$ in closed form from the noise schedule with dispersion scale κ (Section A.1), with κ modulating the posterior dispersion, ensuring all EMIS parameters correspond either to clean-space task specifications or the posterior approximation itself (Algorithm 1, Appendix A).

4 EXPERIMENTS

We evaluate Exact Mixture Score (EMIS) for unconditional instance unlearning in pixel space on CelebA-HQ (Section 4.1) and conditional style and object erasure in latent space on UNLEARNCANVAS (Section 4.2). We also analyze the cross-task trends across parameter fine-tuning, auxiliary sparse autoencoders, and inference-time steering (Section 4.3). Figure 4 summarizes the joint erasure-fidelity Pareto frontier across all 3 evaluation tracks, and Figure 3 presents qualitative comparisons.

Table 1: **Unconditional instance unlearning on CelebA-HQ 256.** We compare against fine-tuning methods and training-free inference-time steering across two forget-cohort scales ($M = 6$, $M = 50$).

Method	Target SSCD (↓)		Clean-FID (↓)		Resource Overhead		
	$M = 6$	$M = 50$	$M = 6$	$M = 50$	Opt. Time ($M = 6$ | $M = 50$)	Inf. Time (s / img)	VRAM / Storage (GB)
Base DDPM	1.00 ± 0.00	1.00 ± 0.00	45.2	45.2	–	2.07	0.75 / 0.0
Parameter Fine-Tuning							
SISS (Alberti et al., 2025)	0.51 ± 0.09	0.39 ± 0.15	39.5	44.3	19.2 m | 2.7 h	2.07	5.61† / 0.45
NegGrad (Gandikota et al., 2023)	0.46 ± 0.12	0.40 ± 0.16	**36.1**	42.8	21.9 m | 3.1 h	2.07	7.17† / 0.45
EraseDiff (Wu et al., 2025)	0.00 ± 0.04	0.00 ± 0.04	Collapse	Collapse	24.4 m | 3.4 h	2.07	7.17† / 0.45
Inference-Time Steering							
Constant Guidance ($\gamma = 0.02$)	0.97 ± 0.03	0.96 ± 0.04	42.0	41.9	–	2.35	0.80 / 0.0
Constant Guidance ($\gamma = 1.0$)	0.00 ± 0.04	0.00 ± 0.04	278.1	278.1	–	2.35	0.80 / 0.0
SLD (Schramowski et al., 2023)	0.65 ± 0.07	0.67 ± 0.08	44.6	45.1	–	2.10	0.80 / 0.0
Safe Denoiser (Kim et al., 2025)	0.58 ± 0.09	0.60 ± 0.10	45.8	46.2	–	2.14	0.80 / 0.0
SPELL (Kirchhof et al., 2025)	0.55 ± 0.09	0.53 ± 0.12	46.2	47.3	–	2.10	0.80 / 0.0
EMIS (Ours)	$\mathbf{0.40 \pm 0.10}$	$\mathbf{0.38 \pm 0.12}$	45.0	**41.3**	–	2.46	0.80 / 0.0

†Measured on a single NVIDIA Tesla V100 (16 GB) GPU across 50-step DDPM sampling using a batch size of $B = 1$. Storage and memory costs represent peak allocations recorded during execution.

4.1 Unconditional Instance Unlearning on CelebA-HQ 256

Experimental setup: Following standard protocols, we evaluate unconditional instance unlearning on the pretrained `google/ddpm-celebahq-256` checkpoint (Ho et al., 2020) under the official 50-step stochastic DDPM sampler. We evaluate two forget scales: (i) a set of $M = 6$ target training identities, and (ii) an expanded set of $M = 50$ simultaneous identities. We measure target erasure via **Target SSCD** (↓), computed via cosine similarity under the Self-Supervised Copy Detection (SSCD) ResNet-50 descriptor (Pizzi et al., 2022), and distributional fidelity via **Clean-FID** (↓) (Parmar et al., 2022) against the CelebA-HQ validation set ($N = 10{,}000$). Full protocol details are in Section C.1.

Results on CelebA-HQ: Table 1 compares parameter fine-tuning and training-free steering across both forget-cohort sizes, with representative qualitative generations shown in Figure 3a (and expanded in Section C.4). Prior training-free methods plateau at 0.53–0.6 Target SSCD across both cohorts (Safe Denoiser 0.58 at $M = 6$ and 0.60 at $M = 50$; SPELL 0.55 at $M = 6$ and 0.53 at $M = 50$). At $M = 6$, EMIS reduces Target SSCD to 0.4 ± 0.1, improving upon both SISS (0.51 ± 0.09) and NegGrad (0.46 ± 0.12) while maintaining Clean-FID at 45.0 (versus 45.2 for Base DDPM). When scaling to $M = 50$ simultaneous targets, EMIS lowers Target SSCD to 0.38 ± 0.12 and improves Clean-FID to 41.3, outperforming SISS (0.39 ± 0.15 SSCD, 44.3 FID) and NegGrad (0.4 ± 0.16 SSCD, 42.8 FID). Crucially, while SISS and NegGrad require 2.7–3.1 h of gradient optimization (5.61–7.17 GB peak VRAM) and a 0.45 GB checkpoint for every target cohort, EMIS operates on frozen base weights without storage overhead at 0.8 GB VRAM. Furthermore, EMIS adds only 7.8 ms/step to inference, scaling with a marginal latency increase from $M = 6$ (2.44 s) to $M = 50$ (2.46 s).

Failure modes of inference-time guidance baselines: The empirical plateaus and fidelity trade-offs of prior training-free heuristics trace directly to departures from the exact-odds formulation $w_\lambda(\pi_t)$ (Section C.1 and Figure A1). Without spatial or temporal calibration, *Constant Guidance* requires an impossible trade-off: forcing erasure (γ=1) collapses the data manifold (278.1 Clean-FID), while preserving quality (γ=0.02) leaves target identities untouched (0.97 Target SSCD). To enforce localization, *Safe Denoiser* (Kim et al., 2025) truncates guidance to a window ($t \in [780, 1000]$), but lacking probit dispersion τ_t and the $(1 - \pi_t)^{-1}$ normalizer, it terminates before trajectories escape the target basin (0.6 SSCD), while extending it destroys facial features (82.7–215.3 FID; Table A1a). Similarly, *SPELL*'s (Kirchhof et al., 2025) non-smooth piecewise barrier imparts discontinuous velocity impulses at its hard boundary that stall erasure at 0.53–0.55 SSCD while fracturing facial landmarks (Table A1b). By contrast, using the closed-form temperature τ_t, EMIS concentrates repulsion within the critical semantic window ($t \approx 600$–950) and self-quenches before texture synthesis, yielding clean erasure without ad-hoc tuning.

4.2 Conditional Style and Object Unlearning on UnlearnCanvas

Experimental setup: We evaluate the efficacy of transferring the same operator from pixel-space instances to text-conditioned concepts when the membership oracle is parameterized by a classifier. We test EMIS on the UnlearnCanvas benchmark (Zhang et al., 2024b) for Stable Diffusion

Table 2: **Conditional concept unlearning on UNLEARNCANVAS.** Style (50 tasks) and object (20 tasks) erasure comparison on Stable Diffusion v1.4. We compare parameter fine-tuning, trained auxiliary structures, and training-free inference-time steering under both independently seeded classifier guidance (**Decoupled**) and ideal classifier-checkpoint alignment (**Oracle**).†

Method	Style Unlearning (50 Styles)				Object Unlearning (20 Objects)				Retain FID (↓)	Overhead	
	UA (↑)	CRA (↑)	IRA (↑)	H (↑)	UA (↑)	CRA (↑)	IRA (↑)	H (↑)		Opt. Time (s / concept)	VRAM / Storage (GB)
Base Model (SD v1.4)	0.00	95.40	98.15	0.00	0.00	95.40	98.15	0.00	71.85	–	3.2 / 0.0
Parameter Fine-Tuning											
ESD (Gandikota et al., 2023)	**98.58**	93.96	80.97	90.53	92.15	44.23	55.78	58.38	65.55	6,163	17.8 / 4.3
FMN (Zhang et al., 2024a)	88.48	46.60	56.77	59.55	45.64	73.46	90.63	64.44	131.37	350	17.9 / 4.2
CA (Kumari et al., 2023)	60.82	92.70	96.01	79.69	46.67	81.97	90.11	67.08	**54.21**	734	10.1 / 4.2
SalUn (Fan et al., 2024)	86.26	95.08	90.39	90.43	86.91	**99.59**	96.35	**93.97**	61.05	667	30.8 / 4.0
EDiff (Wu et al., 2025)	92.42	98.93	73.91	87.06	86.67	48.48	94.03	70.09	81.42	1,567	27.8 / 4.0
SHS (Wu & Harandi, 2024)	95.84	43.27	80.42	65.25	80.73	67.99	81.15	76.11	119.34	1,223	31.2 / 4.0
Auxiliary SAE (Trained‡)											
SAeUron (Cywiński & Deja, 2025)	96.76	98.10	**98.70**	97.85‡	84.23	76.78	90.05	83.27	61.43	4,458‡	3.4 / 0.2
SAEmnesia (Cassano et al., 2026)	96.60	**99.30**	98.67	**98.18**‡	**94.65**	88.48	91.39	91.42‡	56.15	> 4,458‡	3.4 / 0.2
Inference-Time Steering											
SEOT (Li et al., 2024)	56.90	84.31	94.68	75.00	23.25	82.71	95.57	45.76	62.38	–	7.3 / 0.0
SAFREE (Yoon et al., 2025)	53.50	97.86	80.66	72.57	29.85	95.58	**98.82**	55.48	–	–	3.2 / 0.0
TraSCE (Jain et al., 2024)	72.35	96.84	76.06	80.42	49.85	95.59	97.87	73.65	–	–	3.2 / 0.0
EMIS (Ours, Decoupled)	92.70	95.90	85.20	**91.04**	**92.80**	93.60	95.40	**93.92**	75.07	–	6.0 / 0.0
EMIS (Ours, Oracle)	*100.00*	*98.10*	*96.70*	*98.25*	*98.90*	*96.20*	*97.10*	*97.39*	*73.63*	–	6.0 / 0.0

†Baseline results and metrics are compiled from the UNLEARNCANVAS benchmark (Zhang et al., 2024b) and related literature (Cywiński & Deja, 2025; Cassano et al., 2026; Lee et al., 2026). For EMIS, peak VRAM (6.0 GB) includes the ViT classifier during inference, with relative latency overhead scaling consistently with Table 1. ‡Amortized cost includes the one-time 86.7 h SAE dictionary pre-training phase on activations, requiring zero per-request optimization thereafter.

v1.4 across all 50 artistic styles and all 20 canonical object classes. Following the official protocol, we report Unlearning Accuracy (**UA** ↑), Cross-domain and In-domain Retain Accuracy (**CRA**, **IRA** ↑), their harmonic mean **H**(↑), and **Retain FID**(↓) on 5,000 reference images. To prevent evaluator leakage (where inference guidance exploits decision-boundary quirks of the test judge), our main EMIS configuration (**Decoupled**) steers sampling using a ViT classifier that followed the UNLEARNCANVAS training recipe for styles and objects, but initialized with a different random seed. We evaluate using the official benchmark ViT judges. To quantify how the operator performs when the unlearning specification aligns exactly with the target evaluator, we also report an **Oracle** configuration that queries the official evaluation ViT checkpoint during guidance. Section C.2 provides the full sensitivity sweep over the classifier logit threshold C and compares independent-seed vs evaluation-checkpoint guidance.

Quantitative results on UnlearnCanvas: As shown in Table 2, under independent-seed decoupling EMIS achieves $\mathbf{91.04}\%$ Style H, outperforming ESD (90.53%), SalUn (90.43%), and all training-free steering methods (SEOT 75%, SAFREE 72.57%, TraSCE 80.42%). On canonical objects, EMIS reaches $\mathbf{93.92}\%$ Object H, matching the strongest parameter fine-tuning baseline SalUn (93.97%) with zero weight updates and zero checkpoint storage. Style and object erasure examples are illustrated in Figure 3b, with extensive multi-seed grids in Section C.4. Crucially, while our method operates on the frozen base model, SalUn requires 667 s of per-concept optimization, 30.8 GB peak GPU memory, and a 4 GB checkpoint per concept (280 GB across all 70 tasks). Retain FID is preserved at 75.07 against the base model's 71.85. Furthermore, the **Oracle** row demonstrates what happens when the guidance predictor aligns exactly with the evaluation checkpoint: EMIS reaches *98.25%* Style H (100.0% UA, 98.1% CRA, 96.7% IRA) and *97.39%* Object H (98.90% UA) at *73.63* FID. This confirms that when given a precise specification of the target concept, posterior-odds steering executes the erasure almost losslessly, with the remaining gap between the Decoupled and Oracle rows reflecting disagreement between independently seeded classifiers across visually adjacent styles (e.g., *Expressionism* vs. *Fauvism*; see per-class breakdowns in Section C.3).

4.3 CROSS-TASK TRENDS: WEIGHT MODIFICATION VS. POSTERIOR-ODDS STEERING

Together, Tables 1 and 2 and Figure 4 highlight distinct trade-offs across unlearning families when moving between instances, artistic styles, and semantic objects. Plotting erasure against retention across target instances (Figure 4a), artistic styles (Figure 4b), and canonical objects (Figure 4c) reveals how each unlearning family occupies a distinct region of the trade-off plane, with EMIS pinned to the upper-right frontier across all three regimes. Inference-time prompt heuristics (SLD, Safe Denoiser, SPELL, SEOT, SAFREE, TraSCE; blue circles) cluster in the *upper-left*: lacking the retain-odds normalizer $(1 - \pi_t)^{-1}$ and probit dispersion τ_t, their repulsion either switches off

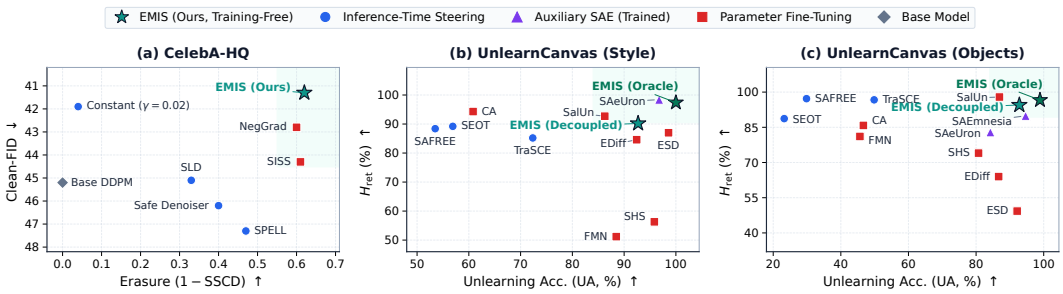

Figure 4: **Erasure–retention Pareto frontier (upper right is optimal)**. We plot **(a)** erasure ($1 - \text{SSCD} \uparrow$) vs. Clean-FID ($\downarrow$, inverted) on CelebA-HQ (M=50), and Unlearning Accuracy (UA $\uparrow$) vs. retain harmonic mean $H_{\text{ret}} \uparrow$ (CRA/IRA) across **(b)** 50 styles and **(c)** 20 objects on UnlearnCanvas. Heuristic steering (blue circles) clusters in the upper-left (high retention, incomplete erasure), while global weight edits (red squares: ESD, EDiff, FMN, SHS) drop into the lower-right for objects as retention degrades; EMIS (teal stars) remains in the shaded upper-right Pareto frontier across all three regimes without weight updates.

before trajectories escape target basins (0.53–0.67 SSCD on CelebA-HQ) or under-erases structured concepts (23.25–49.85% Object UA), preserving high retention (82.7–98.8%) only by leaving targets partially intact. Global weight-editing methods (ESD, EDiff, FMN, SHS; red squares) occupy the opposite extreme, drifting into the *lower-right* as task structure increases: while surface styles act largely as global texture filters that can be attenuated in cross-attention (90.53% and 87.06% Style H for ESD and EDiff), object categories share spatial-attention pathways across co-occurring classes, causing permanent edits to collapse retention on objects (CRA plunging from $93.96\% \rightarrow 44.23\%$ for ESD and $98.93\% \rightarrow 48.48\%$ for EDiff; FMN and SHS degrade Retain FID to 131.37 and 119.34).

EMIS (teal stars) avoids both failure modes, staying in the *upper-right Pareto region* across instances, styles, and objects. Unlike heuristic steering, the retain-odds denominator $(1 - \pi_t + \lambda)^{-1}$ amplifies repulsion whenever a trajectory enters a target basin ($\pi_t \rightarrow 1$), providing sufficient force to unlearn target data / objects. Unlike global weight editing, EMIS leaves generator parameters frozen and evaluates $\pi_t(x_t)$ per sample and timestep. For non-target trajectories ($\pi_t \rightarrow 0$), the guidance weight $w_\lambda(\pi_t)$ vanishes and generation proceeds unperturbed, preventing cross-class contamination on co-occurring objects. The only other methods remaining in the upper-right corner across both styles and objects require offline localization, either via gradient-saliency masks and a separate 4 GB checkpoint per concept (SalUn, 667 s/concept) or via 86.7 h of auxiliary SAE pretraining (SAeUron, SAEmnesia), whereas EMIS achieves this on the fly with zero training and storage cost.

5 Conclusion and Limitations

In this work, we showed that the search for training-free unlearning methods has been a search for an empirical approximation to a quantity with an exact closed form: the continuous-time posterior odds. By recognizing that the safe retain distribution is already encoded by the frozen base model, EMIS unifies instance data unlearning and concept erasure under a single operator. Clipping the odds guarantees that reverse trajectories sample from an explicit normalized density, turning unlearning into an a-priori safety specification with predictable leakage guarantees and state-of-the-art performance across both concept and data unlearning tasks.

Limitations: Our guarantee constrains the output distribution that a serving pipeline emits, not the physical weights of the model. This is an appropriate guarantee for a service provider controlling inference, but insufficient wherever parameter-level deletion is strictly required. For that setting, regressing directly onto the mixture score identity yields an exact fine-tuning objective without synthetic retain data, which we leave to future evaluation. Furthermore, because the prior probability of generating forbidden content is unknown on complex, high-dimensional benchmarks, our closed-form leakage reduction acts as an a-priori design specification rather than an empirically measured statistical distance. Finally, under adversarial prompt search, the primary vulnerability is steering trajectories into regions where the membership gate is poorly estimated rather than where guidance is weak, making robust gate calibration a promising avenue for future work.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the drafting and typesetting of proofs, and implementing experimental methods. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing research methodology, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results; furthermore, generating synthetic datasets and assisting with translation are not applicable to this work. Additionally, we used generative AI tools for drafting parts of the manuscript, editing the text to improve readability and conciseness, formatting references, brainstorming, and sourcing/searching for literature. We have reviewed all AI-assisted work: all mathematical proofs and underlying derivations were thoroughly vetted, independently checked, and confirmed for correctness by the authors; all experimental scripts were tested; and all drafted text and citations were manually verified. We take full responsibility for the final content of this work, including all text, claims, derivations, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support the verification and reproducibility of our results, theoretical claims, algorithms, and experimental configurations are detailed throughout the paper and appendix. The complete mathematical derivations for the Bayesian score decomposition Equation (1) and the full proofs of Lemma 3.1 and Theorem 3.2 (properties of bounded force, exact score, VP-SDE consistency, and closed-form leakage bounds) are provided in Appendix D. The step-by-step EMIS inference procedure is formally specified in Algorithm 1 (Appendix A), with explicit closed-form implementations of the directional posterior variance schedules and multi-target repulsion scores detailed in Sections A.1 and A.2, alongside the probit integral marginalization in Appendix B. For experimental reproducibility, Section 4, Appendix C.1, and Appendix C.2 describe the benchmark datasets, base model checkpoints (DDPM on CelebA-HQ 256 and Stable Diffusion v1.4 on UNLEARNCANVAS), evaluation metrics (SSCD, Clean-FID, UA, CRA, IRA, Harmonic Mean H), and comprehensive hyperparameter sensitivity sweeps across bandwidths and decision margins.

REFERENCES

Silas Alberti, Kenan Hasanaliyev, Manav Shah, and Stefano Ermon. Data unlearning in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Brian D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pp. 5253–5270, 2023.

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. SAEmnesia: Erasing concepts in diffusion models with supervised sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2026.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Bartosz Cywiński and Kamil Deja. SAeUron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *International Conference on Machine Learning (ICML)*, 2025.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

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. In *International Conference on Learning Representations (ICLR)*, 2024.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. TraSCE: Trajectory steering for concept erasure. *arXiv preprint arXiv:2412.07658*, 2024.

Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Mingyu Kim, Young-Heon Kim, and Mijung Park. Safety-guided flow (SGF): A unified framework for negative guidance in safe generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Michael Kirchhof, James Thornton, Pierre Ablin, Louis Béthune, Eugène Ndiaye, and Marco Cuturi. Sparse repellency for shielded generation in text-to-image diffusion models. In *International Conference on Machine Learning (ICML)*, 2025.

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

Uichan Lee, Jeonghyeon Kim, and Sangheum Hwang. Localized concept erasure in text-to-image diffusion models via high-level representation misdirection. In *International Conference on Learning Representations (ICLR)*, 2026.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

David J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38(4):181–188, 1961.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, 2022.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14532–14542, 2022.

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

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Yuhao Sun, Lingyun Yu, Hao-Xiang Xu, Fengyuan Miao, Zhuoer Xu, and Hongtao Xie. Orthogonal concept erasure for diffusion models. In *International Conference on Machine Learning (ICML)*, 2026.

Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision (ECCV)*, pp. 367–384, 2024.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. EraseDiff: Erasing undesirable influence in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28263–28273, 2025.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1755–1764, 2024a.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, and Sijia Liu. UnlearnCanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024b.

A THE EMIS GUIDANCE PROCEDURE AND CLOSED-FORM IMPLEMENTATION

Algorithm 1 gives the full Exact Mixture Score (EMIS) sampling procedure. Differences from unguided reverse diffusion are confined to three lines: evaluating the calibrated membership gate $\pi_t(x_t)$ at the posterior-variance temperature τ_t (Lines 4–6), computing the clipped posterior-odds weight $w_\lambda(\pi_t)$ (Line 7), and shifting the noise prediction along the repulsion direction $\epsilon_\theta(x_t, t) - \epsilon_q(x_t, t)$ (Line 9). Notably, instead of tuning heuristic timestep windows or unnormalized drift scales, every hyperparameter in Algorithm 1 defines either the clean-space unlearning specification (σ_0, classifier margin C, and mixture regularization λ) or the zero-overhead Gaussian posterior approximation around $\hat{x}_0$ (κ). Below we detail how the two task-dependent primitives—the directional posterior variance $v_{0|t}$ and the repulsion target $\epsilon_q = -\sigma_t s_q$—are evaluated in closed form.

Algorithm 1 Exact Mixture Score (EMIS).

Require: Base model ϵ_θ, membership $m(x_0)$, clip λ, temperature (σ_0, κ), schedule $\{\alpha_t, \sigma_t\}$

1: $x_T \sim \mathcal{N}(0, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\hat{x}_0 \leftarrow (x_t - \sigma_t \epsilon_\theta(x_t, t))/\alpha_t$ ▷ Tweedie posterior mean, already computed
4: $v_{0|t} \leftarrow$ posterior dispersion from schedule $\{\alpha_t, \sigma_t\}$ ▷ Section A.1, closed form
5: $\tau_t \leftarrow \sqrt{\sigma_0^2 + \kappa^2 v_{0|t}}$ ▷ Probit temperature (Equation (9))
6: $\pi_t \leftarrow$ gate from $m(\hat{x}_0)$ at temperature τ_t ▷ Equation (8) and Section B.2
7: $w_\lambda(\pi_t) \leftarrow \pi_t/(1 - \pi_t + \lambda)$ ▷ Clipped posterior odds, bounded by $\mathcal{O}_{\max} = 1/\lambda$
8: $\epsilon_q \leftarrow$ repulsion target ▷ $-\sigma_t s_q$ (Equation (A7)), or $\epsilon_\theta(x_t, t, c^*)$
9: $\tilde{\epsilon} \leftarrow \epsilon_\theta(x_t, t) + w_\lambda(\pi_t)\,[\epsilon_\theta(x_t, t) - \epsilon_q]$ ▷ Equivalent to $\tilde{s}_\lambda = s + w_\lambda(\pi_t)[s - s_q]$
10: $x_{t-1} \leftarrow \text{ReverseStep}(x_t, \tilde{\epsilon}, t)$
11: **end for**
12: **return** x_0

A.1 CLOSED-FORM POSTERIOR DISPERSION FROM THE NOISE SCHEDULE

Evaluating the full network Jacobian $\mathbf{J}_\theta(x_t, t)$ in Miyasawa's posterior covariance identity (Equation (A8)) is unnecessary in practice: replacing $\Sigma_{0|t}(x_t)$ by its zero-cost isotropic schedule approximation eliminates any extra network passes during sampling. Dropping the curvature correction $\mathbf{J}_\theta$ in Equation (A8) yields the inverse-SNR dispersion

$$v_{0|t} = \frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t} = \frac{\sigma_t^2}{\alpha_t^2}, \qquad \tau_t = \sqrt{\sigma_0^2 + \kappa^2 \frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}}, \tag{A1}$$

which we deploy in our latent-space concept unlearning experiments and Figure 2 (where at $t = 880$, $v_{0|880} = 62.3$ and $\tau_{880} = \sqrt{1.0^2 + 3.8^2 \times 62.3} = 30.0$). Alternatively, substituting a standard normal prior score $\nabla_{x_t} s(x_t, t) = -\mathbf{I}$ into Equation (A8) cancels the $\bar{\alpha}_t^{-1}$ prefactor to give the bounded schedule

$$v_{0|t} = 1 - \bar{\alpha}_t = \sigma_t^2, \qquad \tau_t = \sqrt{\sigma_0^2 + \kappa^2(1 - \bar{\alpha}_t)}, \tag{A2}$$

which we use in our pixel-space DDPM experiments. Both schedules depend solely on the forward coefficients $\bar{\alpha}_t$ and add zero computational overhead to the reverse step.

A.2 CLOSED-FORM REPULSION SCORE AND MULTI-TARGET GATE

Under the protected-region definition $q_0(x_0) = m(x_0)p_0(x_0)/\pi$ of Equation (7), the diffused forbidden density at horizon t is

$$q_t(x_t) = \frac{1}{\pi} \int p_{0t}(x_t \mid x_0)\, m(x_0)\, p_0(x_0)\, \mathrm{d}x_0 = \frac{p_t(x_t)}{\pi}\, \mathbb{E}_{p(x_0|x_t)}[m(x_0)], \tag{A3}$$

where the second equality uses Bayes' rule $p(x_0 \mid x_t) = p_{0t}(x_t \mid x_0)p_0(x_0)/p_t(x_t)$. Recognizing the expectation as $\pi_t(x_t)$ recovers $\pi q_t(x_t) = \pi_t(x_t)p_t(x_t)$ (Equation (3)). Differentiating with respect to x_t and applying Tweedie's formula gives the forbidden score as a membership-reweighted posterior mean,

$$s_q(x_t, t) = \frac{\alpha_t\, \hat{x}_0^{(q)}(x_t, t) - x_t}{\sigma_t^2}, \qquad \hat{x}_0^{(q)}(x_t, t) = \frac{\mathbb{E}_{p(x_0|x_t)}\big[m(x_0)\, x_0\big]}{\mathbb{E}_{p(x_0|x_t)}\big[m(x_0)\big]}. \tag{A4}$$

Under the Gaussian posterior approximation $p(x_0 \mid x_t) \approx \mathcal{N}(\hat{x}_0, \Sigma_{0|t})$ and the Gaussian instance membership function of Equation (8) with target x^* and bandwidth $h^2 \triangleq D\sigma_0^2$, both expectations are Gaussian integrals and the reweighted mean is a precision-weighted shrinkage blend,

$$\hat{x}_0^{(q)} = \big(\Sigma_{0|t}^{-1} + h^{-2}\mathbf{I}\big)^{-1}\big(\Sigma_{0|t}^{-1}\hat{x}_0 + h^{-2}x^*\big), \tag{A5}$$

shrinking Tweedie's estimate toward x^* according to the ratio of posterior dispersion to protected-region bandwidth.

Deployed multi-target instance form and concept form. In the broad-posterior regime $v_{0|t} \gg \sigma_0^2$, which holds over the active repulsion window, Equation (A5) reduces to $\hat{x}_0^{(q)} \to x^*$. For a multi-target cohort $\{x_0^{(m)}\}_{i=1}^M$ with clean-space log-kernels $\ell_{t,m} \triangleq -\|\hat{x}_0(x_t, t) - x_0^{(m)}\|_2^2/(2D\tau_t^2)$ and softmax responsibilities $\tilde{\omega}_m(x_t, t) \triangleq \mathrm{softmax}_m(\ell_{t,\cdot})$, the aggregate gate and responsibility-weighted target are

$$\pi_t(x_t) = \sum_{i=1}^{M} \tilde{\omega}_i(x_t, t)\exp(\ell_{t,i}), \qquad \hat{x}_0^{(q)}(x_t, t) = \sum_{i=1}^{M} \tilde{\omega}_i(x_t, t)\, x_0^{(i)}. \tag{A6}$$

The corresponding diffused Gaussian mixture score is closed form,

$$s_q(x_t, t) = \sum_{i=1}^{M} \omega_i(x_t, t)\left(\frac{\alpha_t x_0^{(i)} - x_t}{\sigma_t^2}\right), \qquad \omega_i(x_t, t) \triangleq \mathrm{softmax}_i\left(-\frac{\|x_t - \alpha_t x_0^{(i)}\|_2^2}{2\sigma_t^2}\right), \tag{A7}$$

with equivalent noise prediction $\epsilon_q(x_t, t) = -\sigma_t s_q(x_t, t)$. Equation (A5) and Equation (A7) agree to first order in $\sigma_0^2/v_{0|t}$ and differ only in the final steps of sampling where the gate has already closed. Conversely, for conditional concept erasure of target c^*, $q_0(x_0) = p_0(x_0 \mid c^*)$, so $\epsilon_q(x_t, t) = \epsilon_\theta(x_t, t, c^*)$ is the standard conditional denoiser output and $\pi_t(x_t) = \sigma(\ell(\hat{x}_0)/\tau_t - C)$ evaluates the classifier's one-vs-rest log-odds $\ell(\hat{x}_0)$ against decision margin C (Section B.2).

B Derivation of Uncertainty-Aware Temperature Scaling

B.1 Posterior Covariance and 1D Directional Projection

Under the forward transition kernel $p_{0t}(x_t \mid x_0) = \mathcal{N}(x_t;\, \alpha_t x_0,\, \sigma_t^2\mathbf{I})$ (with $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t^2 = 1 - \bar{\alpha}_t$), Tweedie's first-order formula yields the posterior mean $\hat{x}_0(x_t, t) \triangleq \mathbb{E}[x_0 \mid x_t] = (x_t + \sigma_t^2 s(x_t, t))/\alpha_t$, while Miyasawa's second-order identity (Miyasawa, 1961; Efron, 2011) yields the exact posterior covariance $\Sigma_{0|t}(x_t) \triangleq \mathrm{Cov}(x_0 \mid x_t)$:

$$\Sigma_{0|t}(x_t) = \frac{\sigma_t^2}{\alpha_t}\nabla_{x_t}\hat{x}_0(x_t, t) = \frac{\sigma_t^2}{\alpha_t^2}\left(\mathbf{I} + \sigma_t^2\nabla_{x_t}s(x_t, t)\right) = \frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}\left(\mathbf{I} - \sqrt{1 - \bar{\alpha}_t}\,\mathbf{J}_\theta(x_t, t)\right), \tag{A8}$$

where $\mathbf{J}_\theta(x_t, t) \triangleq \nabla_{x_t}\epsilon_\theta(x_t, t)$. Expanding the clean membership function $m(x_0)$ to second order around $\hat{x}_0$ and taking the expectation under $p(x_0 \mid x_t)$ annihilates the linear term ($\mathbb{E}[x_0 - \hat{x}_0 \mid x_t] = 0$), isolating the leading-order posterior-variance discrepancy between the exact gate $\pi_t(x_t) = \mathbb{E}[m(x_0) \mid x_t]$ and the zero-variance plug-in estimate $m(\hat{x}_0)$ (Chung et al., 2023):

$$\pi_t(x_t) - m(\hat{x}_0) = \frac{1}{2}\,\mathrm{Tr}\left(\nabla_{x_0}^2 m(\hat{x}_0)\,\Sigma_{0|t}(x_t)\right) + \mathcal{O}(\sigma_t^4). \tag{A9}$$

Early in reverse diffusion ($\sigma_t^2 \to 1$), $\Sigma_{0|t}$ is large and this second-order curvature term dominates, explaining why deterministic plug-in approximations suffer from severe estimation bias and premature guidance deflections.

For both membership specifications in Equation (8), local variation around $\hat{x}_0$ is governed by a single unit direction $u(x_t)$—the radial displacement $u = (\hat{x}_0 - x^*)/\|\hat{x}_0 - x^*\|_2$ for instance unlearning, and the classifier logit gradient $u = \nabla_{x_0}\ell(\hat{x}_0)/\|\nabla_{x_0}\ell(\hat{x}_0)\|_2$ for concept erasure. Projecting the Gaussian posterior $p(x_0 \mid x_t) \approx \mathcal{N}(x_0; \hat{x}_0, \Sigma_{0|t})$ onto u reduces the D-dimensional expectation $\mathbb{E}[m(x_0) \mid x_t]$ to a 1D integral governed by the scalar directional posterior variance:

$$v_{0|t}(x_t) \triangleq u^\top \Sigma_{0|t}(x_t)\, u. \tag{A10}$$

B.2 Closed-Form Probit Marginalization for Instances and Concepts

Instance unlearning via Gaussian convolution. Modeling clean membership around target x^* as the Gaussian RBF potential $m(x_0) = \exp(-\|x_0 - x^*\|_2^2/(2D\sigma_0^2))$ and projecting the posterior dispersion along $u = (\hat{x}_0 - x^*)/\|\hat{x}_0 - x^*\|_2$ reduces $\pi_t(x_t) = \int m(x_0)\,\mathcal{N}(x_0; \hat{x}_0, \Sigma_{0|t})\,\mathrm{d}x_0$ to the convolution of a clean target ball of variance $\sigma^2_{\text{target}} = D\sigma_0^2$ with a directional posterior variance $\sigma^2_{\text{post}}(t) = D\kappa^2 v_{0|t}$. Because the convolution of two Gaussians sums their variances ($\sigma^2_{\text{total}}(t) = D(\sigma_0^2 + \kappa^2 v_{0|t}) = D\tau_t^2$), normalizing the convolved kernel to unit peak at $\hat{x}_0 = x^*$ yields:

$$\pi_t(x_t) = \exp\left(-\frac{\|\hat{x}_0(x_t, t) - x^*\|_2^2}{2D\tau_t^2}\right), \quad \text{where} \quad \tau_t = \sqrt{\sigma_0^2 + \kappa^2 v_{0|t}}. \tag{A11}$$

Concept and style classifiers via MacKay's probit integral. Consider erasing a style or object concept $c \in \{1, \ldots, K\}$ scored by a $(K+1)$-class clean-space classifier $f : \mathbb{R}^D \to \mathbb{R}^{K+1}$. Under a multiclass softmax, the target probability reduces to a logistic sigmoid $\sigma(z) = (1 + e^{-z})^{-1}$ on the one-vs-rest log-odds. Introducing a decision margin threshold C (where $C = \log K$ corresponds to the uniform prior log-odds $-\log K$ over K retain classes, and calibrating C adjusts the decision boundary margin analogously to the regularization margin in SVMs) defines the clean-space log-likelihood ratio $\ell(x_0)$ and membership $m(x_0)$ at base temperature σ_0:

$$\ell(x_0) \triangleq f_c(x_0) - \log \sum_{j \neq c} \exp f_j(x_0), \qquad m(x_0) = \sigma\left(\frac{\ell(x_0)}{\sigma_0} - C\right). \tag{A12}$$

Linearizing $\ell(x_0) \approx \ell(\hat{x}_0) + g_0^\top(x_0 - \hat{x}_0)$ with $g_0 \triangleq \nabla_{x_0}\ell(\hat{x}_0)$ under $x_0 \sim \mathcal{N}(\hat{x}_0, \Sigma_{0|t})$ implies that $Z = \ell(x_0)/\sigma_0$ conditioned on x_t is univariate normal, $Z \mid x_t \sim \mathcal{N}(\ell(\hat{x}_0)/\sigma_0,\ \|g_0\|_2^2 v_{0|t}/\sigma_0^2)$. Approximating the logistic sigmoid on the data evidence Z by the cumulative normal distribution $\sigma(z) \approx \Phi(\lambda_{\mathrm{M}} z)$ (MacKay, 1992) turns the evidence expectation into an exact Gaussian convolution that attenuates the noisy log-odds $\ell(\hat{x}_0)$ by τ_t while keeping the prior margin C unattenuated:

$$\pi_t(x_t) \approx \sigma\left(\sigma^{-1}\left[\int_{-\infty}^{\infty} \sigma(z)\,\mathcal{N}\left(z;\ \frac{\ell(\hat{x}_0)}{\sigma_0},\ \frac{\|g_0\|_2^2 v_{0|t}}{\sigma_0^2}\right)\mathrm{d}z\right] - C\right) \tag{A13}$$

$$\approx \sigma\left(\frac{\ell(\hat{x}_0)}{\sqrt{\sigma_0^2 + \lambda_{\mathrm{M}}^2\|g_0\|_2^2 v_{0|t}}} - C\right). \tag{A14}$$

Absorbing the local Lipschitz factor $\lambda_{\mathrm{M}}^2\|g_0\|_2^2$ into the scalar hyperparameter κ^2 yields $\pi_t(x_t) = \sigma(\ell(\hat{x}_0)/\tau_t - C)$ with $\tau_t = \sqrt{\sigma_0^2 + \kappa^2 v_{0|t}}$, recovering the exact same temperature schedule as instance unlearning (Equation (A11)) without backpropagating through the classifier or VAE decoder. Crucially, when $v_{0|t}$ is large early in sampling ($\tau_t \gg \sigma_0$), $\ell(\hat{x}_0)/\tau_t \to 0$ and $\pi_t(x_t) \to \sigma(-C)$ (equal to the uniform base rate $\frac{1}{K+1} \approx 0.02$ for $K = 50$ styles when $C = \log K \approx 3.91$; Figure 2), preventing premature false-positive deflections on ambiguous high-noise states.

C Experimental Protocols, Parameter Sweeps, and Per-Class Breakdowns

C.1 CelebA-HQ 256: Parameter Sweeps and Bandwidth Progression

We evaluate unconditional training data unlearning on `google/ddpm-celebahq-256` (50-step stochastic DDPM sampler) across the cohort of $M = 6$ target celebrity identities (noise seeds

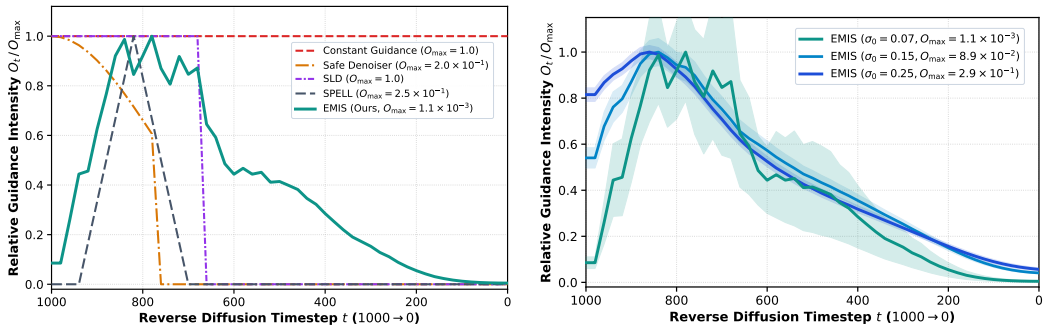

Figure A1: **Empirical temporal guidance profile across reverse timesteps ($M = 50$, 50-step DDPM, $t \in [1000, 0]$). (a)** Normalized trajectory guidance intensity $\mathcal{O}_t(x_t)/\max_t \mathcal{O}_t$ for EMIS ($\sigma_0 = 0.07$), contrasted against Constant Guidance, Safe Denoiser ($t_{\min} = 780$ cutoff), SLD, and SPELL (legend reports peak trajectory odds across the cohort). EMIS automatically concentrates repulsion across the semantic transition ($t \in [600, 950]$) and quenches naturally during texture refinement ($t < 400$). **(b)** Bandwidth progression across $\sigma_0 \in \{0.07, 0.15, 0.25\}$ (mean ± 1 SEM over 50 seeds), demonstrating how tightening σ_0 sharpens the automatic guidance envelope without manual timestep cutoffs.

$\{852, 3279, 3479, 3906, 4166, 11396\}$) and the expanded cohort of $M = 50$ simultaneous identities (target IDs 10000–10049). Figure A1 gives the empirical temporal guidance profile across reverse timesteps ($t = 1000 \to 0$) and probit bandwidths $\sigma_0 \in \{0.07, 0.15, 0.25\}$, alongside sensitivity sweeps over temporal truncation windows $[t_{\min}, t_{\max}]$ and kernel scale η for Safe Denoiser (Kim et al., 2025) (Table A1a) and barrier radii r and guidance weight γ for SPELL (Kirchhof et al., 2025) (Table A1b).

EMIS probit bandwidth sensitivity ($\sigma_0 \in \{0.07, 0.15, 0.25\}$). As shown in Figure A1b, the clean-space radius σ_0 in $\tau_t = \sqrt{\sigma_0^2 + \kappa^2 v_{0|t}}$ controls how rapidly the gate $\pi_t(x_t) = \exp(-\|\hat{x}_0(x_t, t) - x^*\|_2^2/(2D\tau_t^2))$ closes once a trajectory is deflected away from a protected target x^*:

- $\sigma_0 = 0.07$ **(Default Configuration):** During the earliest high-noise steps ($t \in [750, 950]$), posterior uncertainty dominates ($v_{0|t} \gg \sigma_0^2$), so $\tau_t \approx \kappa\sqrt{v_{0|t}}$ activates repulsion across the initial semantic bifurcation window and deflects $\hat{x}_0(x_t, t)$ out of the target basin. As $v_{0|t}$ contracts below $t \approx 650$, $\tau_t \to \sigma_0 = 0.07$: once the per-dimension displacement $d = \|\hat{x}_0 - x^*\|_2/\sqrt{D}$ exceeds ≈ 0.12, the gate $\pi_t(x_t)$ and clipped weight $w_\lambda(\pi_t)$ vanish, leaving the frozen base network unperturbed to synthesize high-frequency facial textures (Target SSCD $= 0.40 \pm 0.10$, Clean-FID $= 45.0$).
- $\sigma_0 = 0.15$ **and** $\sigma_0 = 0.25$**:** Larger clean-space radii keep τ_t wider during intermediate steps, sustaining higher cohort-averaged trajectory odds (peaks 8.9×10^{-2} and 2.9×10^{-1}, respectively) over a broader spatial neighborhood (Target SSCD $= 0.48$ at $\sigma_0 = 0.15$) while still quenching automatically before terminal pixel synthesis.

C.2 UnlearnCanvas: Classifier Decoupling, Margin Calibration, and Retain FID

Independent-seed classifier decoupling vs. evaluation-checkpoint Oracle. Throughout, $H = 3/(\text{UA}^{-1}+\text{CRA}^{-1}+\text{IRA}^{-1})$ denotes the official three-way harmonic mean, and our decoupled guidance classifier uses a logit margin threshold (Equation (A13)) of $C = 2.5$ on styles and $C = 2.0$ on objects. To prevent evaluator leakage, all main EMIS trajectories (**Decoupled**) on UnlearnCanvas are steered by an independently seeded ViT classifier trained using the official UnlearnCanvas training recipe (achieving 99.95% validation accuracy on the 50 styles and 99.41% on the 20 objects) and evaluated against the untouched official benchmark judges (`style50.pth` and `style50_cls.pth`). In high-dimensional classification spaces ($K = 51$ classes), subtle stylistic overlaps between visually related styles (e.g., *Expressionism* vs. *Fauvism*, or *Monet* vs. *Mosaic*) can

Table A1: **Hyperparameter sensitivity sweeps on CelebA-HQ 256 ($M = 6$, 50-step DDPM). Left (Table A1a):** Safe Denoiser (Kim et al., 2025) sweep over guidance window $[t_{\min}, t_{\max}]$ and kernel scale η. Extending guidance into $t < 780$ causes Clean-FID to surge ($82.7 \rightarrow 215.3$), whereas truncating at $t_{\min} = 780$ stalls Target SSCD at 0.58. **Right (Table A1b):** SPELL (Kirchhof et al., 2025) sweep over barrier radius r and guidance weight γ. Selected baseline configurations reported in Table 1 are highlighted in gray.

(a) Safe Denoiser window & scale sweep.

Window $[t_{\min}, t_{\max}]$	**Scale** η	**Target SSCD (↓)**	**Clean-FID (↓)**
[780, 1000] (Default)	0.05	0.81 ± 0.04	45.2
[780, 1000]	0.10	0.69 ± 0.07	45.3
[780, 1000]	0.20	$\mathbf{0.58 \pm 0.09}$	**45.8**
[780, 1000]	0.50	0.54 ± 0.10	48.6
[780, 1000]	1.00	0.52 ± 0.11	54.2
[500, 1000]	0.10	0.52 ± 0.09	58.4
[500, 1000]	0.20	0.44 ± 0.11	82.7
[0, 1000] (Full)	0.10	0.39 ± 0.11	142.6
[0, 1000]	0.20	0.29 ± 0.13	215.3

(b) SPELL barrier radius & scale sweep.

Radius r	**Scale** γ	**Target SSCD (↓)**	**Clean-FID (↓)**
$r = 0.05$	0.5	0.78 ± 0.05	45.2
$r = 0.05$	1.0	0.71 ± 0.06	45.4
$r = 0.05$	2.0	0.67 ± 0.08	45.9
$r = 0.10$	0.5	0.65 ± 0.07	45.6
$r = 0.10$	1.0	0.59 ± 0.08	45.8
$r = 0.10$	2.0	0.55 ± 0.09	46.8
$r = 0.20$ (Default)	0.5	0.59 ± 0.08	45.9
$r = 0.20$	1.0	$\mathbf{0.55 \pm 0.09}$	**46.2**
$r = 0.20$	2.0	0.51 ± 0.10	48.5
$r = 0.30$	0.5	0.55 ± 0.09	46.9
$r = 0.30$	1.0	0.51 ± 0.11	49.8
$r = 0.30$	2.0	0.47 ± 0.12	56.4

Table A2: **Probit variance-scaling sensitivity ($\kappa \in \{0.0, 0.25, 0.5, 1.0\}$ on the 5-style validation cohort at uncalibrated prior offset $C = \log K = 3.91$) and SVM-style logit margin sweep ($C \in \{3.5, 3.0, 2.5\}$ across all 50 Styles & 20 Objects) on UNLEARNCANVAS.** Setting $\kappa > 0$ (Equation (9)) scales the effective temperature $\tau_t = \sqrt{\sigma_0^2 + \kappa^2 v_{0|t}}$ with Tweedie posterior uncertainty, improving both unlearning accuracy (+5.0 pp UA) and in-domain retention (+1.0 pp IRA, +5.4 pp H) over the uncalibrated plug-in score ($\kappa = 0.0$). Official Retain FID is defined across the full 50-style benchmark (1,000 retain images vs. 5,000 reference images; '–' denotes sub-cohorts or non-style splits).

Domain / Cohort	**Configuration (κ or Margin C)**	**UA (↑)**	**CRA (↑)**	**IRA (↑)**	**Harmonic Mean H (↑)**	**Retain FID (↓)**
5-Style Validation	Uncalibrated Plug-in ($\kappa = 0.0$, $C = \log K$)	31.00%	97.00%	98.00%	56.85%	–
5-Style Validation	Probit Variance Scaling ($\kappa = 0.25$, $C = \log K$)	29.00%	97.00%	99.00%	54.65%	–
5-Style Validation	Probit Variance Scaling ($\kappa = 0.50$, $C = \log K$)	33.00%	97.00%	99.00%	59.16%	–
5-Style Validation	**Probit Variance Scaling ($\kappa = 1.00$, $C = \log K$)**	**36.00%**	**97.00%**	**99.00%**	**62.26%**	–
50 Styles (Full)	Decoupled Guidance ($C = 3.5$)	95.70%	95.00%	78.50%	88.98%	76.01
50 Styles (Full)	Decoupled Guidance ($C = 3.0$)	94.40%	95.50%	83.20%	90.68%	75.87
50 Styles (Full)	**Decoupled Guidance ($C = 2.5$, Main)**	92.70%	**95.90%**	**85.20%**	**91.04%**	75.07
50 Styles (Full)	*Oracle Guidance (Official Eval Judge)*	*100.00%*	*98.10%*	*96.70%*	*98.25%*	*73.63*
20 Objects (Full)	**Decoupled Guidance ($C = 2.0$, Main)**	**92.80%**	**93.60%**	**95.40%**	**93.92%**	–
20 Objects (Full)	*Oracle Guidance (Official Eval Judge)*	*98.90%*	*96.20%*	*97.10%*	*97.39%*	–

produce small non-zero tail probabilities ($p \approx 0.005$) on retain prompts. As shown in Table A2, whereas fixing the prior offset at the uncalibrated uniform value $C = \log K \approx 3.91$ on the 5-style validation cohort over-suppresses the gate (yielding $31.00\% \rightarrow 36.00\%$ UA as κ increases from $0.0 \rightarrow 1.0$), calibrating the decision margin $C \in \{3.5, 3.0, 2.5\}$ across the full 50-style benchmark balances erasure and retention, raising In-Domain Retain Accuracy (IRA) from $78.50\% \rightarrow 85.20\%$ and achieving a Harmonic Mean of $H = 91.04\%$ at $C = 2.5$. When EMIS instead queries the official evaluation ViT checkpoint directly during guidance (**Oracle**), performance reaches 100.00% Style UA, 98.10% CRA, and 96.70% IRA ($H = 98.25\%$), and 98.90% Object UA, 96.20% CRA, and 97.10% IRA ($H = 97.39\%$), confirming that when the target specification matches the evaluator, posterior-odds steering executes concept erasure almost losslessly.

Retain FID dynamics and comparison with weight-tuning baselines. All Retain FID values in Table 2 are computed strictly under the official UNLEARNCANVAS evaluation pipeline (Inception-v3 `Mixed_7c` activations with TTUR covariance matching on 5,000 reference images). While certain fine-tuning baselines report numerically lower FID than the unmodified base SD v1.4 generator (71.85), such as CA (54.21) and SEOT (62.38), this reflects two well-documented artifacts: (i) *under-unlearning* (CA and SEOT achieve only 56.9%–60.8% Style UA and 23.3%–46.7% Object UA, leaving the generator largely unperturbed), and (ii) *retain-prompt mode shrinkage*, where hundreds of gradient descent steps on the fixed retain prompt templates reduce generative variance across seeds. Conversely, when weight-editing methods push for thorough erasure (UA > 88%), cross-attention

interference causes severe distributional drift (EDiff 81.42, SHS 119.34, FMN 131.37). Operating at inference time with frozen weights, EMIS achieves $> 92.7\%$ UA across both styles and objects while keeping Retain FID at 75.07 (Decoupled) and 73.63 (Oracle)—within $\Delta \approx +1.8$ to $+3.2$ of the unmodified base generator (71.85).

C.3 Complete 50-Style and 20-Object Per-Class Breakdowns

Table A3 and Table A4 report the complete per-class evaluation of EMIS across all 50 artistic styles ($C = 2.5$) and all 20 canonical object categories ($C = 2.0$) under independent-seed classifier decoupling.

Table A3: **Complete Per-Style Breakdown of EMIS (Decoupled, $C = 2.5$) Across All 50 Artistic Styles in UNLEARNCANVAS** (20 forget samples + 20 retain samples per style).

#	Style	UA	CRA	IRA	H	Retain Partner	#	Style	UA	CRA	IRA	H	Retain Partner
1	Abstractionism	85.0	90.0	100.0	91.25	Artist Sketch	26	Magic Cube	95.0	100.0	100.0	98.28	Meta Physics
2	Artist Sketch	95.0	100.0	100.0	98.28	Blossom Season	27	Meta Physics	100.0	90.0	100.0	96.43	Meteor Shower
3	Blossom Season	90.0	100.0	100.0	96.43	Bricks	28	Meteor Shower	100.0	90.0	100.0	96.43	Monet
4	Bricks	95.0	100.0	100.0	98.28	Byzantine	29	Monet	90.0	95.0	5.0	13.54	Mosaic
5	Byzantine	100.0	100.0	95.0	98.28	Cartoon	30	Mosaic	95.0	95.0	45.0	69.32	Neon Lines
6	Cartoon	100.0	95.0	25.0	49.57	Cold Warm	31	Neon Lines	100.0	95.0	100.0	98.28	On Fire
7	Cold Warm	100.0	95.0	95.0	96.61	Color Fantasy	32	On Fire	95.0	100.0	85.0	92.91	Pastel
8	Color Fantasy	100.0	100.0	85.0	94.44	Comic Etch	33	Pastel	100.0	100.0	90.0	96.43	Pencil Drawing
9	Comic Etch	100.0	95.0	100.0	98.28	Crayon	34	Pencil Drawing	95.0	95.0	100.0	96.61	Picasso
10	Crayon	100.0	100.0	100.0	100.00	Cubism	35	Picasso	100.0	90.0	100.0	96.43	Pop Art
11	Cubism	80.0	100.0	100.0	92.31	Dadaism	36	Pop Art	100.0	95.0	60.0	80.66	Red Blue Ink
12	Dadaism	95.0	95.0	100.0	96.61	Dapple	37	Red Blue Ink	100.0	95.0	90.0	94.82	Rust
13	Dapple	95.0	90.0	100.0	94.82	Defoliation	38	Rust	95.0	95.0	5.0	13.57	Sketch
14	Defoliation	65.0	100.0	100.0	84.78	Early Autumn	39	Sketch	95.0	90.0	100.0	94.82	Sponge Dabbed
15	Early Autumn	95.0	100.0	100.0	98.28	Expressionism	40	Sponge Dabbed	100.0	100.0	90.0	96.43	Structuralism
16	Expressionism	95.0	90.0	15.0	33.97	Fauvism	41	Structuralism	100.0	90.0	95.0	94.82	Superstring
17	Fauvism	95.0	100.0	100.0	98.28	French	42	Superstring	90.0	100.0	60.0	79.41	Surrealism
18	French	55.0	90.0	100.0	76.35	Glowing Sunset	43	Surrealism	100.0	95.0	40.0	65.90	Ukiyoe
19	Glowing Sunset	100.0	100.0	100.0	100.00	Gorgeous Love	44	Ukiyoe	100.0	90.0	90.0	93.10	Van Gogh
20	Gorgeous Love	90.0	95.0	85.0	89.81	Greenfield	45	Van Gogh	100.0	100.0	100.0	100.00	Vibrant Flow
21	Greenfield	100.0	90.0	100.0	96.43	Impressionism	46	Vibrant Flow	55.0	95.0	100.0	77.50	Warm Love
22	Impressionism	50.0	95.0	90.0	72.05	Ink Art	47	Warm Love	90.0	100.0	95.0	94.82	Warm Smear
23	Ink Art	95.0	100.0	100.0	98.28	Joy	48	Warm Smear	90.0	85.0	90.0	88.27	Watercolor
24	Joy	75.0	95.0	100.0	88.60	Liquid Dreams	49	Watercolor	100.0	100.0	65.0	84.78	Winter
25	Liquid Dreams	100.0	100.0	70.0	87.50	Magic Cube	50	Winter	100.0	100.0	95.0	98.28	Abstractionism
Macro Average Across All 50 Styles							**Overall**		**92.70**	**95.90**	**85.20**	**91.04**	–

Table A4: **Complete Per-Object Breakdown of EMIS (Decoupled $C = 2.0$ and Oracle) Across All 20 Canonical Objects in UNLEARNCANVAS** (50 forget samples + 50 retain samples per object across all 50 styles).

#	Object Class	Decoupled Evaluation ($C = 2.0$, Independent Seed)				Oracle Evaluation (Official Eval Judge)				Retain Partner
		UA (↑)	CRA (↑)	IRA (↑)	H (↑)	UA (↑)	CRA (↑)	IRA (↑)	H (↑)	
1	Architectures	90.0	96.0	98.0	94.54	100.0	98.0	100.0	99.32	Bears
2	Bears	96.0	96.0	80.0	90.00	100.0	98.0	86.0	94.24	Birds
3	Birds	94.0	100.0	98.0	97.27	100.0	100.0	100.0	100.00	Butterfly
4	Butterfly	78.0	96.0	100.0	90.26	100.0	94.0	96.0	96.60	Cats
5	Cats	100.0	98.0	100.0	99.32	100.0	100.0	100.0	100.00	Dogs
6	Dogs	88.0	84.0	96.0	89.06	88.0	88.0	98.0	91.10	Fishes
7	Fishes	94.0	94.0	96.0	94.66	100.0	100.0	96.0	98.63	Flame
8	Flame	90.0	90.0	88.0	89.32	100.0	92.0	92.0	94.52	Flowers
9	Flowers	98.0	92.0	100.0	96.54	100.0	98.0	100.0	99.32	Frogs
10	Frogs	100.0	88.0	100.0	95.65	100.0	100.0	98.0	99.32	Horses
11	Horses	88.0	98.0	96.0	93.80	100.0	100.0	96.0	98.63	Human
12	Human	90.0	88.0	100.0	92.38	94.0	90.0	100.0	94.49	Jellyfish
13	Jellyfish	90.0	96.0	96.0	93.91	100.0	100.0	100.0	100.00	Rabbits
14	Rabbits	98.0	100.0	98.0	98.66	100.0	100.0	98.0	99.32	Sandwiches
15	Sandwiches	86.0	98.0	100.0	94.24	100.0	90.0	100.0	96.43	Sea
16	Sea	92.0	84.0	80.0	85.05	100.0	90.0	84.0	90.87	Statues
17	Statues	98.0	92.0	96.0	95.27	100.0	100.0	98.0	99.32	Towers
18	Towers	96.0	96.0	98.0	96.66	100.0	96.0	100.0	98.63	Trees
19	Trees	90.0	94.0	100.0	94.49	96.0	90.0	100.0	95.15	Waterfalls
20	Waterfalls	100.0	92.0	88.0	93.07	100.0	100.0	100.0	100.00	Architectures
–	**Macro Average**	**92.80**	**93.60**	**95.40**	**93.92**	**98.90**	**96.20**	**97.10**	**97.39**	–

C.4 Qualitative Visual Comparison Grids

Figure A2, Figure A3, and Figure A4 provide expanded multi-seed and multi-prompt qualitative comparisons across unconditional instance unlearning and conditional concept erasure. On CelebA-HQ 256 (Figure A2), EMIS replaces the unlearning target identities while maintaining photorealistic facial structure and natural lighting. On UnlearnCanvas (Figures A3 and A4), EMIS removes target artistic brushwork or canonical object categories while preserving retained partner styles and objects.

D Mathematical Proofs

D.1 Derivation of the Bayesian Score Decomposition Identity

We provide the complete step-by-step derivation of the exact mixture score identity (Equation (1)). Let $p_0(x_0) = (1-\pi)r_0(x_0) + \pi q_0(x_0)$ denote the clean mixture distribution with retained density r_0, forbidden density q_0, and prior forbidden weight $\pi \in (0,1)$. By linearity of the Gaussian forward kernel $p_{0t}(x_t \mid x_0) = \mathcal{N}(x_t;\, \alpha_t x_0,\, \sigma_t^2 \mathbf{I})$, the diffused marginal at any $t \in [0,T]$ is

$$p_t(x_t) = \int p_{0t}(x_t \mid x_0)\, p_0(x_0)\, \mathrm{d}x_0 = (1-\pi)\, r_t(x_t) + \pi\, q_t(x_t). \tag{A15}$$

Taking ∇_{x_t} on both sides and substituting $\nabla_x p = p \nabla_x \log p$ with $s \triangleq \nabla_{x_t} \log p_t$, $s_r \triangleq \nabla_{x_t} \log r_t$, and $s_q \triangleq \nabla_{x_t} \log q_t$ gives

$$p_t(x_t)\, s(x_t,t) = (1-\pi)\, r_t(x_t)\, s_r(x_t,t) + \pi\, q_t(x_t)\, s_q(x_t,t). \tag{A16}$$

Isolating $(1-\pi) r_t s_r$ and dividing by $(1-\pi) r_t(x_t)$ yields

$$s_r(x_t,t) = \frac{p_t(x_t)}{(1-\pi) r_t(x_t)}\, s(x_t,t) - \frac{\pi q_t(x_t)}{(1-\pi) r_t(x_t)}\, s_q(x_t,t). \tag{A17}$$

Substituting $p_t(x_t) = (1-\pi) r_t(x_t) + \pi q_t(x_t)$ into the first ratio and defining the posterior odds $\mathcal{O}_t(x_t) \triangleq \frac{\pi q_t(x_t)}{(1-\pi) r_t(x_t)}$ (cf. Equation (1)), we obtain $\frac{p_t(x_t)}{(1-\pi) r_t(x_t)} = 1 + \mathcal{O}_t(x_t)$, so Equation (A17) becomes

$$s_r(x_t,t) = \big(1 + \mathcal{O}_t(x_t)\big)\, s(x_t,t) - \mathcal{O}_t(x_t)\, s_q(x_t,t) = s(x_t,t) + \mathcal{O}_t(x_t)\, [s(x_t,t) - s_q(x_t,t)]\,, \tag{A18}$$

completing the derivation of Equation (1). □

Remark (Duality with Classifier Guidance and Clean-Space Denoiser Displacement). Two immediate corollaries connect Equation (1) to prior formulations: (i) By Bayes' rule, the posterior retention probability is $p_t(y = \text{keep} \mid x_t) = (1-\pi) r_t(x_t) / p_t(x_t)$. Taking $\nabla_{x_t} \log$ on both sides yields $s_r(x_t,t) - s(x_t,t) = \nabla_{x_t} \log p_t(y = \text{keep} \mid x_t) = \mathcal{O}_t(x_t)[s(x_t,t) - s_q(x_t,t)]$, establishing that the mixture score identity is mathematically identical to continuous-time Bayesian classifier guidance (Dhariwal & Nichol, 2021; Song et al., 2021) without requiring spatial differentiation of a noisy classifier. (ii) Multiplying $s_r(x_t,t) - s(x_t,t) = \mathcal{O}_t(x_t)[s(x_t,t) - s_q(x_t,t)]$ by σ_t^2/α_t and applying Tweedie's formula $\hat{x}_0 = (x_t + \sigma_t^2 s)/\alpha_t$ directly recovers the clean-space denoiser displacement $\hat{x}_0^{(r)}(x_t,t) = \hat{x}_0(x_t,t) + \mathcal{O}_t(x_t)[\hat{x}_0(x_t,t) - \hat{x}_0^{(q)}(x_t,t)]$ of Kim et al. (2025) (Theorem 3.2).

D.2 Proof of Lemma 3.1

By the law of total expectation and the Markov property $y \perp\!\!\!\perp x_t \mid x_0$, $\pi_t(x_t) = \mathbb{E}_{x_0 \mid x_t}\big[\mathbb{E}[\mathbf{1}\{y = \text{forget}\} \mid x_0, x_t]\big] = \mathbb{E}_{x_0 \sim p(x_0 \mid x_t)}[m(x_0)]$. Bayes' rule then gives $(1-\pi) r_t(x_t) = p_t(y = \text{keep} \mid x_t) p_t(x_t) = \big(1 - \pi_t(x_t)\big) p_t(x_t)$, canceling both $p_t(x_t)$ and π in $\mathcal{O}_t = \pi q_t / [(1-\pi) r_t]$. □

D.3 Proof of Theorem 3.2

We restate the setting. Let $p_t(x_t) = (1-\pi)\, r_t(x_t) + \pi\, q_t(x_t)$ with $\pi \in (0,1)$ be the diffused marginal mixture at horizon $t \in [0,T]$, let $\pi_t(x_t) = \pi q_t(x_t)/p_t(x_t) \in [0,1]$ be the posterior forget

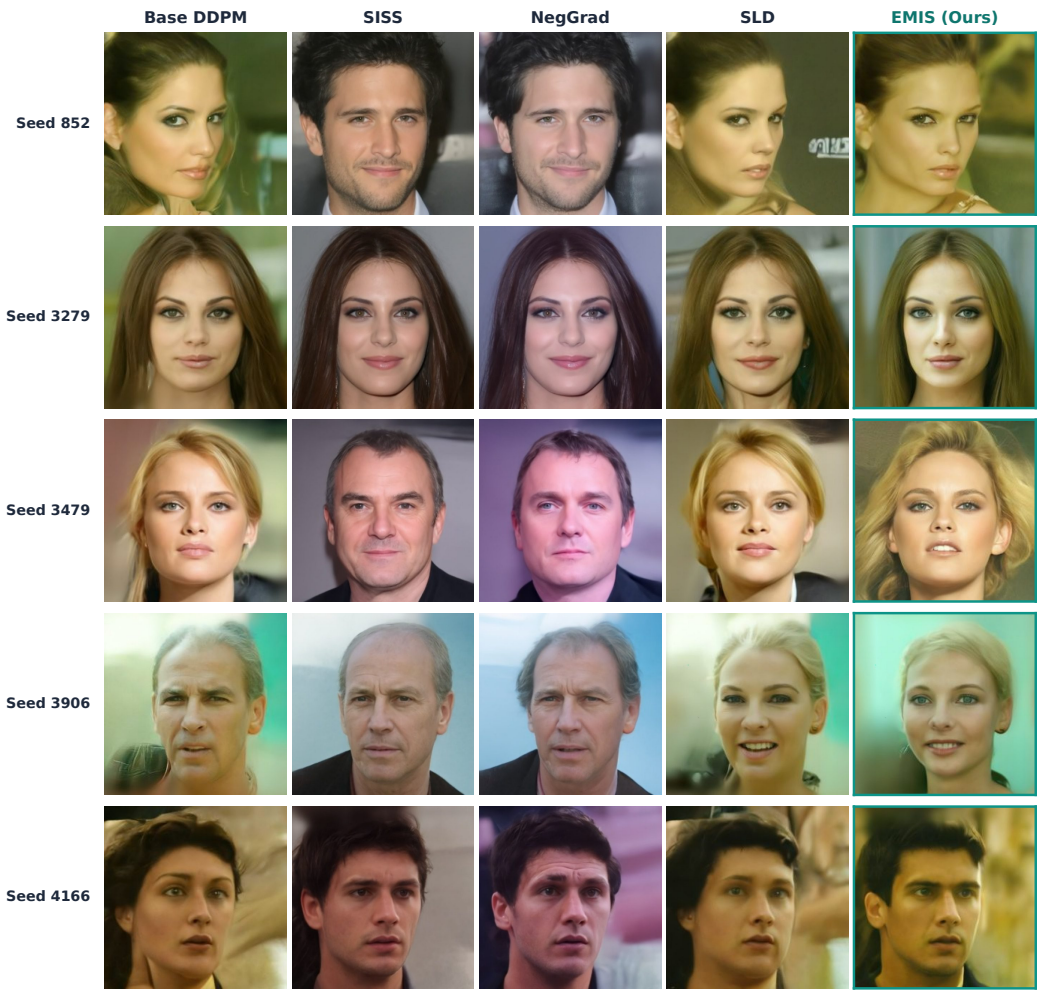

Figure A2: **Qualitative comparison of multi-target instance unlearning on CelebA-HQ 256 ($M = 50$, 50-step DDPM).** Visual generations for 5 seeds $(852, 3279, 3479, 3906, 4166)$ across the unmodified Base DDPM model, fine-tuning baselines (SISS, NegGrad), and inference-time steering methods (SLD and EMIS).

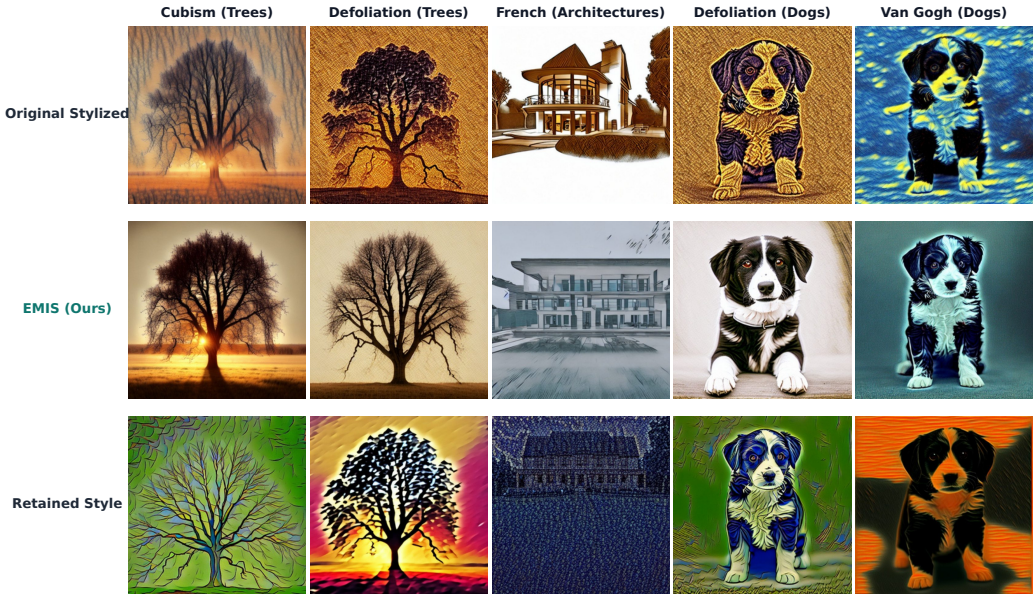

Figure A3: **Qualitative concept unlearning across artistic styles on UNLEARNCANVAS (Stable Diffusion v1.4).** Five representative benchmark styles evaluated across three rows: **Row 1 (Original Stylized):** Unmodified Base SD v1.4 generations; **Row 2 (EMIS):** Generations under active EMIS guidance, removing the target stylistic brushwork; **Row 3 (Retained Style):** Non-targeted partner styles generated under active guidance.

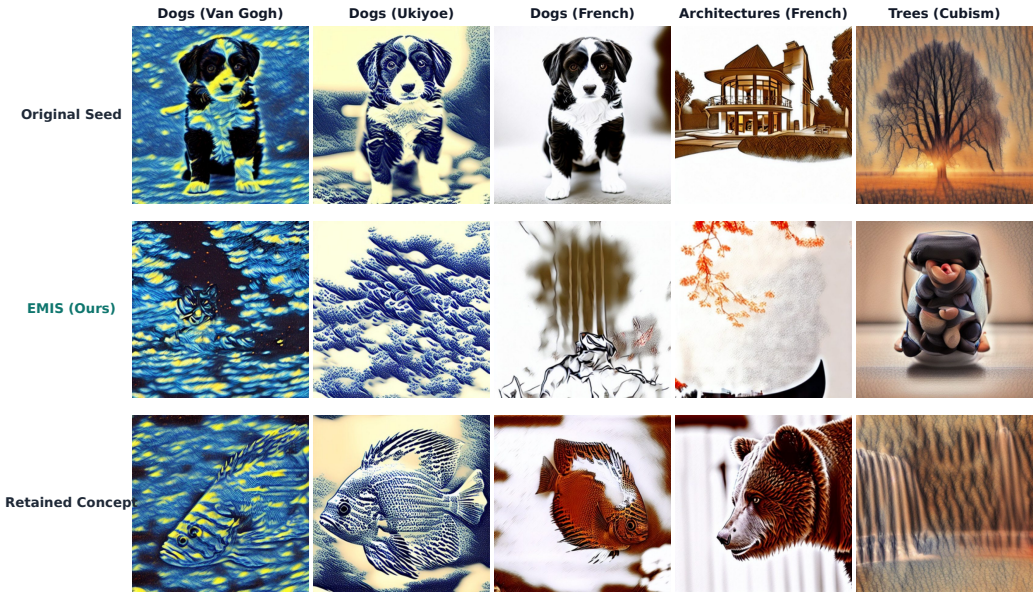

Figure A4: **Qualitative concept unlearning across canonical object categories on UNLEARNCANVAS (Stable Diffusion v1.4).** Five representative object unlearning tasks evaluated across three rows: **Row 1 (Original Seed):** Base SD v1.4 generations depicting targeted object classes; **Row 2 (EMIS):** Generations with EMIS actively suppressing the target object class; **Row 3 (Retained Concept):** Partner retain objects generated under active guidance.

probability, fix $\lambda > 0$ with $\mathcal{O}_{\max} \triangleq \lambda^{-1}$, and let

$$\tilde{s}_\lambda(x_t, t) \triangleq s(x_t, t) + \frac{\pi_t(x_t)}{1 - \pi_t(x_t) + \lambda}\big[s(x_t, t) - s_q(x_t, t)\big]. \tag{A19}$$

We prove each claim in turn.

Part (1): bounded force. Consider $w_\lambda : [0,1] \to \mathbb{R}$, $w_\lambda(u) = u/(1 - u + \lambda)$ for fixed $\lambda > 0$. Differentiating,

$$w'_\lambda(u) = \frac{(1 - u + \lambda) + u}{(1 - u + \lambda)^2} = \frac{1 + \lambda}{(1 - u + \lambda)^2} > 0 \qquad \forall u \in [0,1], \tag{A20}$$

so w_λ is strictly increasing. Its minimum is $w_\lambda(0) = 0$ and its maximum is

$$\sup_{u \in [0,1]} w_\lambda(u) = w_\lambda(1) = \frac{1}{\lambda} = \mathcal{O}_{\max}. \tag{A21}$$

Since $\pi_t(x_t) \in [0,1]$, we conclude $0 \le w_\lambda(\pi_t(x_t)) \le \mathcal{O}_{\max}$. The same computation gives the Lipschitz constant $\sup_{u \in [0,1]} |w'_\lambda(u)| = w'_\lambda(1) = (1 + \lambda)/\lambda^2$.

Part (2): exact score of an explicit density. We first verify that $\rho_t^{(\lambda)}$ of Equation (5) is a probability density. Substituting the mixture into the numerator,

$$\begin{aligned}(1 - \pi)\, r_t(x_t) + \lambda\, p_t(x_t) &= (1 - \pi)\, r_t(x_t) + \lambda\big[(1 - \pi)\, r_t(x_t) + \pi\, q_t(x_t)\big] \\ &= (1 - \pi)(1 + \lambda)\, r_t(x_t) + \lambda\pi\, q_t(x_t),\end{aligned} \tag{A22}$$

which is nonnegative. Integrating over $\mathbb{R}^D$ with $\int r_t = \int q_t = 1$,

$$\int_{\mathbb{R}^D} \big[(1 - \pi)\, r_t + \lambda\, p_t\big]\, \mathrm{d}x_t = (1 - \pi)(1 + \lambda) + \lambda\pi = (1 - \pi) + \lambda, \tag{A23}$$

so $\rho_t^{(\lambda)}$ integrates to one.

For the score, the denominator $(1 - \pi) + \lambda$ is a positive constant in x_t, so

$$\nabla_{x_t} \log \rho_t^{(\lambda)}(x_t) = \frac{(1 - \pi)\, \nabla_{x_t} r_t(x_t) + \lambda\, \nabla_{x_t} p_t(x_t)}{(1 - \pi)\, r_t(x_t) + \lambda\, p_t(x_t)}. \tag{A24}$$

Using $\nabla_{x_t} r_t = r_t\, s_r$ and $\nabla_{x_t} p_t = p_t\, s$,

$$\nabla_{x_t} \log \rho_t^{(\lambda)}(x_t) = \frac{(1 - \pi)\, r_t\, s_r + \lambda\, p_t\, s}{(1 - \pi)\, r_t + \lambda\, p_t}. \tag{A25}$$

From the mixture, $(1 - \pi) r_t = p_t - \pi q_t$, so dividing by $p_t > 0$,

$$\frac{(1 - \pi)\, r_t(x_t)}{p_t(x_t)} = 1 - \frac{\pi\, q_t(x_t)}{p_t(x_t)} = 1 - \pi_t(x_t). \tag{A26}$$

Dividing numerator and denominator of Equation (A25) by p_t and substituting Equation (A26),

$$\nabla_{x_t} \log \rho_t^{(\lambda)}(x_t) = \frac{\big(1 - \pi_t(x_t)\big)\, s_r(x_t, t) + \lambda\, s(x_t, t)}{1 - \pi_t(x_t) + \lambda}. \tag{A27}$$

Multiplying the exact identity Equation (1) by $(1 - \pi_t(x_t))$ clears its denominator,

$$\big(1 - \pi_t\big) s_r = \big(1 - \pi_t\big) s + \pi_t\big[s - s_q\big], \tag{A28}$$

and substituting into the numerator of Equation (A27),

$$\begin{aligned}\nabla_{x_t} \log \rho_t^{(\lambda)} &= \frac{\big(1 - \pi_t\big) s + \pi_t\big[s - s_q\big] + \lambda\, s}{1 - \pi_t + \lambda} \\ &= \frac{\big(1 - \pi_t + \lambda\big) s + \pi_t\big[s - s_q\big]}{1 - \pi_t + \lambda} = s + \frac{\pi_t}{1 - \pi_t + \lambda}\big[s - s_q\big] \equiv \tilde{s}_\lambda,\end{aligned} \tag{A29}$$

which proves part (2).

Part (3): forward and reverse consistency. Let $\mathcal{T}_t$ denote the forward Gaussian convolution operator, $(\mathcal{T}_t\mu_0)(x_t) \triangleq \int p_{0t}(x_t \mid x_0)\,\mu_0(x_0)\,\mathrm{d}x_0$. Because $\mathcal{T}_t$ is linear and $r_t = \mathcal{T}_t[r_0]$, $p_t = \mathcal{T}_t[p_0]$, applying it to $\rho_0^{(\lambda)} = \frac{(1-\pi)r_0+\lambda p_0}{(1-\pi)+\lambda}$ gives

$$\mathcal{T}_t\left[\rho_0^{(\lambda)}\right] = \frac{(1-\pi)\,\mathcal{T}_t[r_0] + \lambda\,\mathcal{T}_t[p_0]}{(1-\pi)+\lambda} = \frac{(1-\pi)\,r_t + \lambda\,p_t}{(1-\pi)+\lambda} = \rho_t^{(\lambda)}. \tag{A30}$$

Hence $\{\rho_t^{(\lambda)}\}$ is the exact marginal evolution of the forward VP-SDE initialized at $\rho_0^{(\lambda)}$. By Anderson's reverse-time theorem (Anderson, 1982; Song et al., 2021), the reverse-time VP-SDE driven by $\nabla_{x_t}\log\rho_t^{(\lambda)} = \tilde{s}_\lambda$ reverses this evolution and yields $x_0 \sim \rho_0^{(\lambda)}$.

Part (4): leakage in closed form and total variation convergence. Evaluating Equation (5) at $t = 0$ and using Equation (A22) expresses the terminal law as a convex combination of the clean retain and forbidden components,

$$\rho_0^{(\lambda)}(x_0) = \underbrace{\left(\frac{(1-\pi)(1+\lambda)}{(1-\pi)+\lambda}\right)}_{1-\pi^{\mathrm{out}}(\lambda)} r_0(x_0) \;+\; \underbrace{\left(\frac{\lambda\pi}{(1-\pi)+\lambda}\right)}_{\pi^{\mathrm{out}}(\lambda)} q_0(x_0), \tag{A31}$$

so a draw from $\rho_0^{(\lambda)}$ is generated from the forbidden component ($y = \mathrm{forget}$) with probability $\pi^{\mathrm{out}}(\lambda) = \lambda\pi/((1-\pi)+\lambda)$. Taking the ratio of the two component weights gives the output odds

$$\mathcal{O}^{\mathrm{out}}(\lambda) = \frac{\pi^{\mathrm{out}}(\lambda)}{1-\pi^{\mathrm{out}}(\lambda)} = \frac{\lambda\pi}{(1-\pi)(1+\lambda)} = \left(\frac{\lambda}{1+\lambda}\right)\mathcal{O}^{\mathrm{prior}} = \left(\frac{1}{1+\mathcal{O}_{\max}}\right)\mathcal{O}^{\mathrm{prior}}, \tag{A32}$$

where $\mathcal{O}^{\mathrm{prior}} = \pi/(1-\pi)$ and $\mathcal{O}_{\max} = \lambda^{-1}$. Finally, subtracting $r_0(x_0)$ from Equation (A31) gives $\rho_0^{(\lambda)}(x_0) - r_0(x_0) = \pi^{\mathrm{out}}(\lambda)\big(q_0(x_0) - r_0(x_0)\big)$, so the total variation distance to r_0 is

$$\left\|\rho_0^{(\lambda)} - r_0\right\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathbb{R}^D}\left|\rho_0^{(\lambda)}(x_0) - r_0(x_0)\right|\mathrm{d}x_0 = \pi^{\mathrm{out}}(\lambda)\,\|q_0 - r_0\|_{\mathrm{TV}} \le \frac{\lambda\pi}{(1-\pi)+\lambda} \xrightarrow{\lambda\to 0^+} 0, \tag{A33}$$

completing the proof of Theorem 3.2. □