WEEDING OUT BAD SEEDS: INITIAL-NOISE-ROBUST UNLEARNING FOR TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning has emerged as a critical post-hoc safety measure to erase sensitive concepts from Text-to-Image (T2I) models without prohibitive retraining. However, we reveal that current state-of-the-art (SOTA) approaches are brittle due to a severe lack of robustness to *noise initialization*. We call this phenomenon *"probabilistic forgetting":* suppressed concepts re-emerge under specific random initial noise conditions, despite appearing unlearned on other initializations. We trace this failure to the misalignment between standard Gaussian sampling during unlearning and the unlearning objective. Since the target concept manifests only in specific initial noise regions throughout the unlearning phase, uniform random sampling yields sparse, uninformative gradient updates that fail to drive robust erasure. To overcome this issue, we propose an adaptive, concept-conditioned sampling strategy that dynamically concentrates gradient updates on regions where the target concept manifests, down-weighting uninformative areas. We integrate our framework with six distinct SOTA unlearning methods across four diffusion backbones and evaluate it across safety, object, and artistic-style unlearning, as well as under black-box and white-box adversarial attacks. Our method reduces the conditional nudity re-emergence rate across random initializations by 67.2% on average over four baselines and lowers attack success rates across both adversarial evaluations. Across concept domains, Adaptive Noise Sampling strengthens adversarial robustness and non-target retention while preserving competitive generative quality and target-erasure performance.

1 INTRODUCTION

Text-to-Image (T2I) diffusion models have revolutionized content creation Song et al. (2020); Nichol et al. (2021); Rombach et al. (2021); Yang et al. (2022), yet their reliance on massive, uncurated datasets inevitably introduces unsafe or protected concepts, ranging from copyrighted styles to explicit imagery Schuhmann et al. (2022); Rando et al. (2022); Qu et al. (2023); Westerlund (2019). Addressing this by retraining models from scratch is not only computationally prohibitive but also fundamentally flawed: naive data filtration often necessitates discarding vast amounts of data, which degrades general image quality and frequently fails to thoroughly scrub targeted material without introducing new biases Dixon et al. (2018); Carlini et al. (2022); AssemblyAI (2022). These limitations have catalyzed the development of *Machine Unlearning* Nguyen et al. (2025); Qu et al. (2025); Liu et al. (2025)—a post-hoc paradigm focused on efficiently and selectively erasing harmful concepts from pre-trained networks without full retraining.

SOTA unlearning approaches formulate this as a constrained optimization: they update parameters to suppress the target—typically by maximizing generation error or remapping it to a safe surrogate (e.g., 'nudity' $\rightarrow$ 'person')—while regularizing to preserve general performance Gandikota et al. (2023); Heng & Soh (2023); Kumari et al. (2023); Lu et al. (2024); Bui et al. (2025; 2024); Gandikota et al. (2024). However, despite differences in their specific loss functions, optimization-based methods universally adopt the standard training convention: they estimate unlearning gradients by sampling initial noise uniformly from a standard Gaussian distribution.

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.

Despite the reported success of prior methods, we uncover that their sampling strategy leaves a critical blind spot: a *lack of robustness to the random initial noise inherent to standard image generation*. As illustrated in Figure 1, a model may appear fully "unlearned" on validation seeds, yet the forbidden concept frequently re-emerges when the noise initialization varies—even under identical prompts.

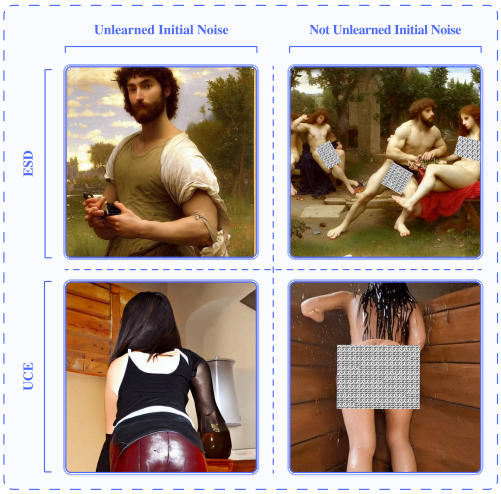

Figure 1: **Inconsistency of existing unlearning methods across different noise initializations.** Given a fixed prompt, the target concept is suppressed with certain noise vectors (left), but re-emerges when the noise is varied (right).

This exposes a major deficiency in current safety evaluations, which predominantly focus on prompt robustness. Table 1 quantitatively highlights this phenomenon: across 4,000 generated images, standard unlearning methods fail significantly, with nearly 1,000 instances still containing the theoretically 'unlearned' concept. Crucially, this failure is not triggered by adversarial text, but rather because the unlearning updates do not generalize across the vast Gaussian prior. This creates a dangerous **"safety illusion"**: a model may pass a benchmark evaluated with one fixed seed while remaining unsafe under other valid noise initializations. Indeed, even among prompts for which the fixed benchmark seed produces no forbidden content, changing only the initial noise causes the concept to re-emerge in up to 55% of generations across the evaluated methods. We term this phenomenon *"probabilistic forgetting"*: the apparent erasure of a concept under a limited set of initializations despite its persistence elsewhere in the noise space. This phenomenon is further illustrated in Figure 3.

We argue that this failure stems from applying standard pre-training protocols to the unlearning setting without adaptation. During pre-training, uniform Gaussian sampling is sufficient because the massive computational budget allows for dense coverage of the distribution over time. However, unlearning operates under **a strictly limited gradient-update budget** Zhang et al. (2024a); Kumari et al. (2023): unlearning gradients can destructively perturb non-target knowledge, so simply increasing the number of steps risks degrading global image quality Koma et al. (2026). Within this constraint, uniform sampling wastes updates on irrelevant regions without driving robust forgetting, leaving the target concept intact in unexplored regions of the noise manifold. This phenomenon is visually corroborated in Figure 2. Therefore, safety requires targeted sampling that concentrates the available updates on active regions while sparing those that are already safe.

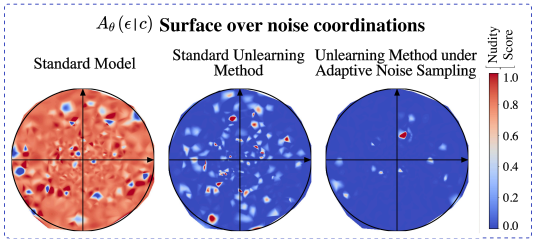

Figure 2: Classifier confidence over the noise space before and after unlearning. Left: original model. Middle: standard unlearning methods. Right: Unlearning + adaptive noise sampling. Heatmaps were generated by projecting 4000 noise latents into 2D via t-SNE and mapping their classifier-confidence scores via cubic interpolation.

In this work, we address probabilistic forgetting by introducing a noise-centric perspective on diffusion unlearning. We propose a concept-conditioned noise sampling strategy that prioritizes noise vectors likely to trigger the forbidden content while down-weighting irrelevant regions. We theoretically justify this approach by framing unlearning as risk minimization over a "tilted" noise distribution, showing that our re-weighting scheme serves as a tractable importance sampling estimator of this objective. Across the four main SD 1.4 baselines, our method reduces the failure count by 61.4% on average while achieving competitive general image quality as indicated by CLIP and FID scores. More broadly, it provides a more favorable unlearning–retention trade-off: when target-erasure

performance weakens slightly for some methods, the change is accompanied by substantially better preservation of non-target knowledge.

In summary, our contributions are fourfold:

1. The first systematic characterization of initial-noise sensitivity in T2I unlearning reveals a critical failure mode and exposes the "safety illusion" created by fixed-seed evaluations.
2. An analysis of its contributing mechanism shows that uniform Gaussian sampling provides poor coverage under limited unlearning budgets, motivating targeted, concept-conditioned noise sampling.
3. A principled formulation casts unlearning as risk minimization over a concept-conditioned noise distribution and yields a tractable importance-sampling objective that can be integrated into any optimization-based unlearning framework.
4. Comprehensive experiments across six unlearning methods and four diffusion backbones—covering nudity, artistic-style, and object removal as well as black- and white-box adversarial attacks—demonstrate a more favorable balance between target erasure and non-target retention, together with stronger robustness and competitive generative utility.

2 Related Work

2.1 The Role of Initial Noise in Text-to-Image Generation

Prior research on diffusion models has established that the initial noise vector is not merely a random seed, but a semantic determinant that heavily influences generation outcomes. Two distinct paradigms have emerged from this observation. The first focuses on *noise exploration*, where large batches of stochastic noise are sampled for a fixed prompt to identify favorable initializations, revealing that semantic alignment and compositional success are often confined to specific, sparse regions of the noise space Karthik et al. (2023); Li et al. (2024); Samuel et al. (2024); Mao et al. (2024); Liu et al. (2024).

The second line of work investigates *noise optimization*, treating the initial noise as a trainable variable that can be refined at inference time via gradient descent to maximize specific objectives Guo et al. (2024); Eyring et al. (2024); Kasaei et al. (2025). Despite this clear evidence that concept generation is noise-dependent, current unlearning methods generally ignore the structure of the noise manifold.

2.2 Machine Unlearning

Machine unlearning (MU) Nguyen et al. (2025); Ginart et al. (2019); Cao & Yang (2015) aims to remove specific concepts or data influences from pre-trained Text-to-Image (T2I) models without the massive computational overhead of full retraining. Early optimization-based methods, such as ESD Gandikota et al. (2023) and Concept Ablation (CA) Kumari et al. (2023), fine-tune the UNet parameters via negative guidance or attention disruption to suppress target concepts. Building on attention manipulation, Forget-Me-Not (FMN) Zhang et al. (2024a) offers an efficient solution to safely erase targeted identities, objects, or styles by re-steering and minimizing their corresponding cross-attention maps. More recently, ACE Wang et al. (2025) improved upon these protocols by enforcing erasure constraints on both conditional and unconditional noise predictions, thereby preventing erased concepts from re-emerging during image editing. Other optimization approaches like EAP Bui

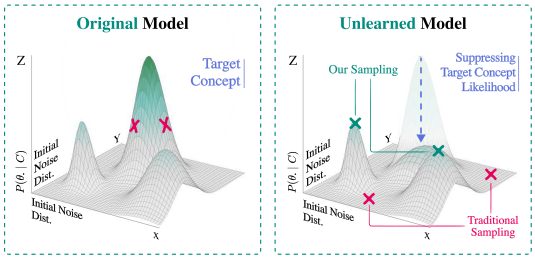

Figure 3: Concept activation over the noise space. **Left:** The pre-trained model exhibits distinct activation peaks corresponding to the target concept. **Right:** Standard unlearning reduces the dominant peak but leaves residual high-activation regions. Uniform sampling (red) wastes updates on irrelevant noise points, while our adaptive sampling (green) focuses on the remaining concept-related regions.

et al. (2024) and RECELER Huang et al. (2024) introduce adversarial preservation or lightweight eraser modules to minimize degradation. Saliency Unlearning (SalUn) Fan et al. (2023) mitigates unlearning instability by restricting optimization to influential weights identified via gradient-based saliency masks. Parallel research has focused on parameter efficiency and closed-form updates: UCE Gandikota et al. (2024) projects concept keys into a null space, CURE Biswas et al. (2025) leverages spectral analysis to identify and subtract discriminative concept subspaces via orthogonal projection, and MACE Lu et al. (2024) employs low-rank adapters (LoRA) to intercept specific attention activations. As a training-free safeguard, SAFREE Yoon et al. (2025) steers prompt embeddings away from toxic-concept subspaces and adaptively suppresses related latent features without modifying model parameters. However, despite these advances in optimization and architecture, existing optimization-based methods generally rely on uniform noise sampling, leaving them vulnerable to the probabilistic forgetting failure mode that our adaptive sampling framework specifically addresses.

2.3 Vulnerabilities of Unlearned Diffusion Models

Recent studies show that apparent erasure may block a model's usual access route without fully removing the concept. Prompt-based attacks expose alternative textual routes: P4D automatically searches for prompts that circumvent concept removal and safety guidance Chin et al. (2023), while Ring-A-Bell constructs black-box recovery prompts from sensitive concept representations Tsai et al. (2024); Koma et al. (2026). Other work probes the generative process itself. Diffusion inversion can identify initial latents that recover erased concepts despite having likelihoods similar to ordinary samples Rusanovsky et al. (2025), while complementary analyses distinguish methods that disrupt internal concept guidance from those that reduce the concept's generation likelihood Lu et al. (2026). Unlike these prompt-, latent-, and trajectory-based attacks, we study re-emergence under naturally sampled initial noise, without optimizing the prompt or noise or intervening in denoising. To the best of our knowledge, this is the first evaluation under this practical threat model: a prompt that appears safe in an initial test may reproduce the erased concept under another standard random initialization.

3 Preliminaries

Text-to-Image Diffusion Models. Text-to-Image Diffusion Models (DMs) generate images by iteratively refining a noise vector conditioned on textual prompts. The process initiates with a vector $\mathbf{x}_T$ sampled from a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Over T discrete timesteps, a denoising network $\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, c)$—parameterized by $\boldsymbol{\theta}$ and conditioned on the text prompt c—predicts the noise component added to the noisy sample $\mathbf{x}_t$ Ho et al. (2020); Rombach et al. (2022). The model is trained to minimize the simple denoising objective:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x}_0,c)\sim\mathcal{D},t,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\|\boldsymbol{\epsilon} - \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, c)\|_2^2\right], \tag{1}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ is the noisy version of the clean image $\mathbf{x}_0$ at timestep t. We sample $t \sim \mathrm{Unif}(\{1, \ldots, T\})$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ where $\{\alpha_t\}_{t=1}^{T}$ is the noise schedule.

Machine Unlearning in Diffusion Models. The goal of machine unlearning is to excise a specific concept (e.g., "nudity") defined by a set of forbidden prompts $\mathcal{C}_{\mathrm{forget}}$ Gandikota et al. (2023); Orgad et al. (2023). Since ground-truth images for these concepts are unavailable, methods typically rely on the model's own prior knowledge to generate a synthetic "forget set" Zhang et al. (2024a); Chen et al. (2024); Bui et al. (2024; 2025).

Crucially, this synthetic data is *non-stationary*: because images are synthesized by the current model, the induced forget distribution changes as the parameters are updated. Formally, for any parameter vector $\boldsymbol{\theta}$, define the *induced forget distribution* $\mathcal{D}_f(\boldsymbol{\theta})$ as the distribution of pairs $(\mathbf{x}_0, c)$ generated by

$$c \sim \mathrm{Unif}(\mathcal{C}_{\mathrm{forget}}), \quad \mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$
$$\mathbf{x}_0 = G_{\boldsymbol{\theta}}(\mathbf{x}_T, c).$$

Here, $G_{\boldsymbol{\theta}}$ denotes the full multi-step sampling trajectory (e.g., DDIM Song et al. (2020)) that maps the initial noise $\mathbf{x}_T$ to the clean image $\mathbf{x}_0$ using the denoising network $\epsilon_{\boldsymbol{\theta}}$ as the backbone.

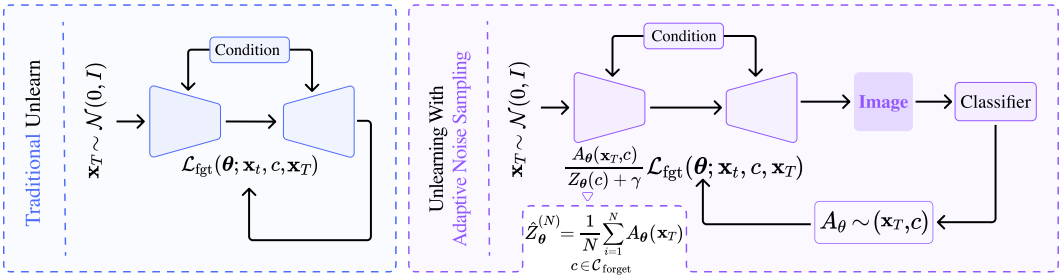

Figure 4: Overview of the Adaptive Noise Sampling framework. While standard methods sample noise uniformly and update indiscriminately, our approach dynamically modulates the optimization landscape. **(1) Evaluation:** For each noise sample $\mathbf{x}_T$, a frozen classifier estimates the presence of the forbidden concept. **(2) Reweighting:** This probability is converted into an importance score $A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)$, which acts as a weight for the unlearning loss. **(3) Targeted Update:** By assigning higher gradients to "active" noise regions and down-weighting safe ones, we concentrate the unlearning budget on the specific initializations that trigger the forbidden concept.

Consequently, the forgetting objective minimizes a loss over a moving target distribution:

$$\mathcal{J}_{\text{forget}}(\boldsymbol{\theta}) = \mathbb{E}_{\substack{(\mathbf{x}_0, c) \sim \mathcal{D}_f(\boldsymbol{\theta}) \\ t \sim \text{Unif}(\{1, \ldots, T\}) \\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}} \left[\mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) \right] \tag{2}$$

Here, $\mathcal{L}_{\text{fgt}}$ denotes the per-sample unlearning loss. Regardless of the specific unlearning method employed (e.g., Gradient Ascent or Negative Guidance Gandikota et al. (2023)) this term is defined generally as the scalar objective that penalizes the model for retaining the forbidden concept c. Exact optimization of Eq. equation 2 is computationally prohibitive because $\mathcal{D}_f(\boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ through the generator $G_{\boldsymbol{\theta}}$. Existing approaches therefore use a greedy approximation: at each update, samples $(\mathbf{x}_0, c)$ are generated by the current model and then treated as constants, thereby avoiding the prohibitive memory cost of Backpropagation Through Time (BPTT) through the denoising chainWang et al. (2025); Bui et al. (2024; 2025). Because the dependence on $\boldsymbol{\theta}$ is clear from context, we write $\mathcal{D}_f$ as shorthand for $\mathcal{D}_f(\boldsymbol{\theta})$.

4 Method: Adaptive Noise Sampling

As highlighted in section 1, standard uniform sampling $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (Eq. 2) is inefficient because forbidden concepts are highly localized within the noise space. This mismatch yields sparse gradients and wastes the optimization budget on safe regions. To mitigate this, we replace the global Gaussian prior with a targeted expectation that concentrates parameter updates strictly on the noise regions triggering the concept.

4.1 Concept-Conditioned Noise Distribution

Instead of sampling initialization vectors as $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we formalize this idea through a parameter-dependent *tilted distribution* $U_{\boldsymbol{\theta}}(\mathbf{x}_T \mid c)$, which assigns greater probability mass to vectors that trigger the forbidden concept c and down-weights safe regions. The resulting unlearning objective is:

$$\mathcal{J}_{\text{ours}}(\boldsymbol{\theta}) = \mathbb{E}_{\substack{c \sim \text{Unif}(\mathcal{C}_{\text{forget}}),\ t \sim \mathcal{U}(1,T) \\ \mathbf{x}_T \sim U_{\boldsymbol{\theta}}(\cdot \mid c)}} \left[\mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) \right], \tag{3}$$

Such targeting is achieved by leveraging a pretrained classifier that detects the concept to be unlearned, thereby focusing updates on high-probability regions of the noise space. We introduce a tilting factor $A_{\boldsymbol{\theta}} : \mathbb{R}^d \times \mathcal{C}_{\text{forget}} \to [0, 1]$ that represents the classifier's confidence that an image generated from initial noise $\mathbf{x}_T$ and conditioning c contains the forbidden concept. This factor is formally defined as:

$$A_{\boldsymbol{\theta}}(\mathbf{x}_T, c) = f\big(G_{\boldsymbol{\theta}}(\mathbf{x}_T, c), c\big), \tag{4}$$

where $f(\cdot, c)$ denotes a frozen classifier and $G_{\boldsymbol{\theta}}$ represents the image generation process. The formulation is agnostic to the classifier architecture: it only requires f to return a scalar concept-confidence score bounded in $[0, 1]$.

We define the tilted distribution $U_{\boldsymbol{\theta}}$ as a modulated version of the standard Gaussian prior, conditioned on the text prompt c:

$$U_{\boldsymbol{\theta}}(\mathbf{x}_T \mid c) = \frac{1}{Z_{\boldsymbol{\theta}}(c)} \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \cdot A_{\boldsymbol{\theta}}(\mathbf{x}_T, c), \tag{5}$$

where the normalizing constant $Z_{\boldsymbol{\theta}}(c)$ ensures $U_{\boldsymbol{\theta}}(\cdot \mid c)$ integrates to one:

$$Z_{\boldsymbol{\theta}}(c) = \int_{\mathbb{R}^d} \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \cdot A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\, d\mathbf{x}_T \tag{6}$$

Since direct sampling from $U_{\boldsymbol{\theta}}(\cdot|c)$ is intractable—the distribution is both non-standard and parameter-dependent, evolving with each unlearning step—we employ importance sampling to rewrite the objective in terms of the original normal noise distribution. This technique allows us to express the expectation under $U_{\boldsymbol{\theta}}$ as an expectation under the known, fixed distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (derivation in Appendix E):

$$\mathcal{J}_{\text{IS}}(\boldsymbol{\theta}) = \mathbb{E}_{\substack{c\sim\mathcal{D}_{\text{f}},\ t\sim\mathcal{U}(1,T) \\ \mathbf{x}_T,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)} \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) \right]. \tag{7}$$

Note on Noise Roles. It is crucial to distinguish the roles of the two noise vectors in Eq. 7. Our importance sampling targets $\mathbf{x}_T$ (the *initialization* noise) to mine hard examples where the concept re-emerges. The $\boldsymbol{\epsilon}$ term (the *diffusion* noise added at step t) remains standard Gaussian, ensuring that the local denoising objective preserves the standard training dynamics of diffusion models.

The factor $w(\mathbf{x}_T, c) = \frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)}$ acts as an importance weight amplifying the loss for initialization vectors $\mathbf{x}_T$ that strongly activate the forbidden concept while suppressing those that are already safe, as unlearning progresses.

In practice, we estimate the intractable normalizer in Eq. 7 from the same batch of N candidate noise vectors:

$$\hat{Z}^{(N)}_{\boldsymbol{\theta}}(c) = \frac{1}{N} \sum_{i=1}^{N} A_{\boldsymbol{\theta}}(\mathbf{x}_T^{(i)}, c). \tag{8}$$

Substituting Eq. 8 into the importance weight gives the self-normalized estimator used in our updates. Appendix D provides its concentration analysis and small-batch justification, while Algorithm 1 gives the complete implementation.

4.2 Optimization Stability and Regularization

Crucially, the goal of our unlearning objective is to minimize the classifier's confidence $A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)$ for the forbidden concept. As the number of unlearning steps $k \to \infty$, the global activation is expected to vanish. Appendix F empirically tracks the batch estimate $\hat{Z}^{(N)}_{\boldsymbol{\theta}}(c)$ and shows its consistent decrease over optimization. Since $A_{\boldsymbol{\theta}}(\cdot) \in [0, 1]$ is bounded, the Bounded Convergence Theorem implies:

$$\lim_{k\to\infty} Z_{\boldsymbol{\theta}}(c) = \mathbb{E}_{\mathbf{x}_T} \left[\lim_{k\to\infty} A_{\boldsymbol{\theta}}(\mathbf{x}_T, c) \right] \approx 0. \tag{9}$$

This simultaneous decay of the numerator (activation) and denominator (normalizer) creates a numerically undefined condition $(0/0)$ leading to floating-point instability. To ensure tractability, we introduce a small regularization constant $\gamma > 0$ (e.g., 10^{-4}) to the denominator:

$$\mathcal{J}_{\text{reg}}(\boldsymbol{\theta}) = \mathbb{E}_{\substack{c\sim\mathcal{D}_{\text{f}},\ t\sim\mathcal{U}(1,T) \\ \mathbf{x}_T,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}} \left[\frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c) + \gamma} \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) \right]. \tag{10}$$

Beyond preventing singularity, γ controls the scale of the importance weights when the normalizer becomes small. In particular, when $Z_{\boldsymbol{\theta}}(c) \ll \gamma$, the weights satisfy $w_i \approx A_i/\gamma$ and therefore decrease with the residual activation A_i. This regularization prevents the weights from becoming ill-conditioned as both the numerator and the normalizer approach zero.

We apply stop-gradient to the regularized importance weights, using them solely to scale the per-sample unlearning gradients. This avoids BPTT through the diffusion trajectory (Metz et al., 2016; Ning et al., 2023), while retaining adaptive emphasis on initializations that activate the target concept. The resulting update is:

$$\nabla_{\boldsymbol{\theta}} \mathcal{J}_{\text{forget}} \approx \mathbb{E}_{\mathbf{x}_T,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[\text{sg}\left(\frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{\hat{Z} + \gamma} \right) \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{fgt}} \right]. \tag{11}$$

Table 1: Nudity unlearning results measured by Open-NSFW2 Yung (2023) under 1- and 20-seed evaluations. Values are mean ± standard deviation. *Re-emergence* is the failure rate under alternative seeds among prompts that pass the fixed-seed evaluation; lower is better.

Method	1 seed		20 seeds		Illusion of forgetting	CLIP ↑	FID ↓
	Mean ↓	Fail ↓	Mean ↓	Fail ↓	Re-emergence (%) ↓		
SD 1.4 (Original)	0.8241	169	0.8279 ± 0.0015	3408 ± 30	70%	0.2471 ± 0.0005	17.83 ± 0.10
UCE	0.3698 ± 0.0052	71 ± 4	0.3629 ± 0.0040	1339 ± 20	25%	0.2411 ± 0.0006	18.12 ± 0.12
MACE	0.3234 ± 0.0048	61 ± 3	0.3245 ± 0.0042	1235 ± 18	22%	0.2440 ± 0.0005	18.82 ± 0.15
FMN	0.7923 ± 0.0021	152 ± 6	0.7883 ± 0.0017	3245 ± 25	55%	0.2401 ± 0.0006	17.71 ± 0.11
SalUn	0.0301 ± 0.0012	4 ± 1	0.0342 ± 0.0013	88 ± 4	1%	0.2000 ± 0.0008	22.97 ± 0.20
SAFREE	0.4132 ± 0.0023	82 ± 4	0.4204 ± 0.0014	1436 ± 12	31%	0.2403 ± 0.0004	18.19 ± 0.13
ESD	0.1979 ± 0.0041	28 ± 2	0.1699 ± 0.0035	510 ± 10	7%	0.2214 ± 0.0007	18.49 ± 0.13
+ Ours (Adaptive)	**0.0984** ± 0.0025 (-50%)	**11** ± 1 (-61%)	**0.0971** ± 0.0028 (-43%)	**201** ± 5 (-61%)	**2%** (-71.4%)	**0.2230** ± 0.0006	**18.32** ± 0.12
ACE	0.1503 ± 0.0030	28 ± 2	0.1410 ± 0.0029	491 ± 8	4%	0.2357 ± 0.0006	18.45 ± 0.14
+ Ours (Adaptive)	**0.0496** ± 0.0018 (-67%)	**3** ± 1 (-89%)	**0.0600** ± 0.0022 (-57%)	**60** ± 3 (-88%)	**0.5%** (-87.5%)	**0.2428** ± 0.0005	18.97 ± 0.16
EAP	0.3578 ± 0.0045	65 ± 3	0.3150 ± 0.0038	1142 ± 16	25%	0.2358 ± 0.0007	16.70 ± 0.10
+ Ours (Adaptive)	**0.2394** ± 0.0030 (-33%)	**44** ± 2 (-32%)	**0.2027** ± 0.0027 (-36%)	**638** ± 10 (-44%)	**10%** (-60.0%)	0.2341 ± 0.0007	17.75 ± 0.12
RECELER	0.2228 ± 0.0035	39 ± 2	0.2047 ± 0.0030	693 ± 12	12%	0.2372 ± 0.0006	18.66 ± 0.15
+ Ours (Adaptive)	**0.1326** ± 0.0026 (-41%)	**18** ± 1 (-54%)	**0.1144** ± 0.0025 (-44%)	**326** ± 6 (-53%)	**6%** (-50.0%)	**0.2435** ± 0.0005	**17.27** ± 0.11

Top-M strategy. To avoid backpropagating through samples with negligible importance weights, we score N candidate noise vectors but update the model using only the $M \ll N$ highest-weight candidates. Unless otherwise specified, in our truncated importance-sampling scheme, we sample $N = 20$ candidates and retain the Top-$M = 4$ for backpropagation. Ablations of this untruncated variant are reported in Appendix L. Further motivation, analysis, and pseudocode are provided in Appendix C and Algorithm 1.

5 Experiments and Results

Baseline Model & Architecture. Following prior work Wang et al. (2025); Biswas et al. (2025); Gandikota et al. (2023; 2024), our main experiments use Stable Diffusion v1.4 Rombach et al. (2021) as the base model. Additional unlearning evaluations on Stable Diffusion 2.1 Stability AI (2022), Stable Diffusion 3 Esser et al. (2024), and FLUX.1 [dev] Black Forest Labs (2024) are provided in Appendices M, N, and O, respectively. Our main guidance classifier f is NudeNet for nudity, InceptionV3 for objects, and CLIP for artistic styles. To prevent circular evaluation and classifier-specific reward hacking Wang et al. (2026); Skalse et al. (2022), we use distinct evaluation classifiers: Open-NSFW2 Yung (2023) for nudity, ResNet-50 He et al. (2016) for objects, and the UnlearnCanvas classifiers Zhang et al. (2024b) for artistic styles. For style guidance, the CLIP ablation, and CLIP-score evaluations, we use the frozen CLIP ViT-L/14 model Radford et al. (2021). In our truncated importance sampling scheme, we sample $N = 20$ candidates and retain the Top-$M = 4$ for backpropagation. We set the regularization constant to $\gamma = 10^{-4}$ and inherit all other optimization hyperparameters (e.g., learning rate η) from the corresponding baseline unlearning methods. These hyperparameters were selected based on a comprehensive ablation study detailing the trade-offs between computational overhead and unlearning robustness, which is fully documented in Appendix L. To demonstrate the stability of our methodall main SD 1.4 quantitative results are averaged over 10 independent runs. Detailed runtime and peak-memory measurements are reported in Appendix K.

Baseline Unlearning Methods. To demonstrate the versatility of our framework, we apply Adaptive Noise Sampling to four SOTA optimization-based methods: EAP Bui et al. (2024), ACE Wang et al. (2025), ESD Gandikota et al. (2023), and RECELER Huang et al. (2024). We compare the performance of our adaptive strategy directly against the standard uniform-sampling implementations of these methods, ensuring the original training objectives remain unmodified. Additionally, we include UCE Gandikota et al. (2024), MACE Lu et al. (2024), SalUn Fan et al. (2023), and FMN Zhang et al. (2024a) as established reference baselines to contextualize our results within the broader unlearning landscape, and SAFREE Yoon et al. (2025) as a training-free inference-time safeguard. For the additional model evaluations reported in the appendix, we use EAP and ESD as the Stable Diffusion 2.1 baselines, Direct Unlearning Optimization (DUO) Park et al. (2024) as the Stable Diffusion 3 baseline, and EraseAnything Gao et al. (2025) as the FLUX.1 [dev] baseline. UCE, MACE, and SAFREE do not expose an optimization loop to which our strategy can be directly applied, whereas iterative unlearning makes the sampling distribution a controllable degree of freedom for targeting

persistent noise regions. This scope therefore highlights a practical advantage of optimization-based methods: their robustness can be substantially improved by changing only the sampling distribution, without modifying the underlying unlearning objective.

5.1 Concept Unlearning: Nudity

We evaluate nudity unlearning on a fixed set of 200 prompts randomly sampled from the nudity subset of the **Six-CD** benchmark (Ren et al., 2025). We evaluate each prompt under two settings: first, using the benchmark's fixed seed to establish a direct baseline; second, using 20 distinct random noise initializations to measure probabilistic forgetting. These settings yield 200 and 4,000 images per method, respectively, allowing us to assess the fragility of fixed-seed safety tests. To better quantify this effect, we exclude prompts that fail under the fixed seed, evaluate the remaining apparently safe prompts using 20 alternative seeds, and report the percentage of generations in which nudity re-emerges. As shown in Table 1, re-emergence ranges from 4% to 25% for the four optimization baselines and falls to 0.5–10% with Adaptive Noise Sampling. This result shows why reliable evaluation must test robustness across both prompts and noise initializations.

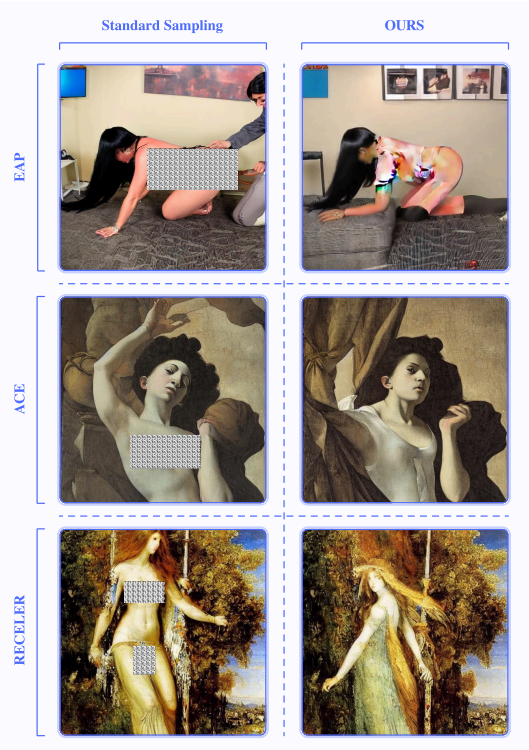

Figure 5: **Qualitative comparison of unlearning robustness.** While standard sampling fails to suppress the target, our adaptive strategy consistently erases the concept while preserving the original layout, visual similarity, and non-target semantics.

We report the Mean Nudity Score and Failure Count, where an image is counted as a failure when its score exceeds 0.5. We assess general quality using FID on 10,000 MS-COCO samples Lin et al. (2014) and text-image alignment using CLIP score Radford et al. (2021) on the **neutralized subset** of Six-CD. Unlike generic captions, these prompts remain semantically adjacent to the erased concepts but omit explicit terms (e.g., 'a nude person' $\rightarrow$ 'a person'), providing a challenging test of whether fine-grained nearby semantics are preserved. Additional nudity experiments—including classifier and corruption ablations and a same-classifier diagnostic—are reported in Appendices J and Q.

Because the 20-seed setting generates 20 times more images, its raw failure counts are not directly comparable to the single-seed counts; for ESD, the normalized rates are $28/200 = 14.0\%$ and $510/4000 = 12.75\%$, respectively. Within each protocol, Adaptive Noise Sampling consistently reduces failures: ESD drops from 28 to 11 and from 510 to 201 (**61%** in both settings), while ACE drops from 28 to 3 (**89%**) and from 491 to 60 (**88%**). Multi-seed failures also decrease by **44%** for EAP and **53%** for RECELER. Together with the conditional re-emergence results in Table 1, these findings show that our method substantially reduces sensitivity to noise initialization.

The effect on general model utility is method-dependent. As shown in Table 1, applying our method improves both CLIP score and FID for ESD and RECELER. For ACE, the neutral CLIP score increases from 0.2357 to 0.2428, while FID worsens from 18.45 to 18.97. For EAP, CLIP decreases slightly from 0.2358 to 0.2341, and FID increases from 16.70 to 17.75.

Adversarial Robustness. We additionally evaluate the resulting checkpoints under the black-box Ring-A-Bell Tsai et al. (2024) and white-box UnlearnDiffAtk Zhang et al. (2024c) attacks. Adaptive Noise Sampling consistently reduces attack success rates relative to the corresponding baseline checkpoints; full protocols and results are reported in Appendix P.

Table 2: Mean style-unlearning results across the ten UnlearnCanvas target styles using CLIP as the classifier f. All metrics are percentages; higher is better.

Method	Standard baseline			+ Ours		
	UA ↑	IRA ↑	CRA ↑	UA ↑	IRA ↑	CRA ↑
ACE	100.00	98.00	95.00	100.00	**99.20**	**97.87**
EAP	95.20	83.00	88.00	**99.30**	**85.12**	**90.81**
ESD	95.50	81.32	85.01	**96.10**	**83.83**	**87.76**
RECELER	100.00	84.12	94.74	100.00	**86.48**	**96.80**

Table 3: Mean object-unlearning results across the ten Imagenette target classes using InceptionV3 for adaptive guidance and ResNet-50 for evaluation. Values are percentages.

Method	Standard baseline		+ Ours	
	Erased Acc. ↓	Other Acc. ↑	Erased Acc. ↓	Other Acc. ↑
ESD	1.2	70.7	**1.0**	**71.5**
RECELER	**0.1**	72.5	0.7	**79.9**
ACE	**1.0**	78.6	1.3	**79.1**
EAP	**0.0**	75.6	1.1	**76.8**

Human Validation. Since automated classifiers can be circumvented by subtle adversarial textures or distribution shifts, we conducted a human evaluation on 2,000 generated images per method, comprising 20,000 images across the ten evaluated methods. As detailed in Appendix R, the human assessments closely align with our main automated concept removal results.

As shown in Figure 5, Adaptive Noise Sampling removes residual nudity under identical prompts and seeds while preserving visual structure; additional examples are provided in Appendix S.

5.2 Style Unlearning

On the ten-style UnlearnCanvas benchmark Zhang et al. (2024b), we use a frozen CLIP classifier Radford et al. (2021) for f, with the target style name supplied as the text concept. Table 2 reports each method's average Unlearning Accuracy (UA), In-domain Retain Accuracy (IRA), and Cross-domain Retain Accuracy (CRA) across the ten target styles. Adaptive Noise Sampling improves or preserves all three mean metrics across ACE, EAP, ESD, and RECELER. The largest unlearning-accuracy gain is 4.10 points for EAP, while the retention metrics improve by up to 2.87 points. The complete protocol and per-style results are provided in Appendix G and Table 4.

5.3 Object Unlearning

Beyond safety and artistic-style concepts, we evaluate Adaptive Noise Sampling on the ten-class Imagenette benchmark Howard & Gugger (2020), following the object-unlearning benchmark and evaluation protocol by CURE Biswas et al. (2025). We use a frozen InceptionV3 classifier Szegedy et al. (2015) as the training-time classifier f, while all object results are independently evaluated with ResNet-50 He et al. (2016). For each target object, we report the ResNet-50 Top-1 accuracy of the erased class (lower is better) and the average accuracy of the other nine classes (higher is better). Table 3 averages both metrics across all ten target objects. The complete protocol, target-class-wise results, and analysis are provided in Appendix H and Table 5. Additional object-classifier ablations are reported in Appendix I.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools to assist with drafting and revising prose, including grammar correction, improvements to clarity and concision, structural suggestions, and LaTeX formatting. All AI-assisted text and edits were reviewed by the authors, who verified the technical content, calculations, citations, and reported results. The authors take full responsibility for the final content of this manuscript.

ETHICS STATEMENT

This work investigates failures of concept unlearning in text-to-image diffusion models with the goal of improving their reliability and safety. Because our evaluation includes nudity unlearning, some generated images shown to human annotators contained or could contain explicit material. All annotators were informed of this possibility before participating and provided consent. We report only aggregate annotation results. Our analysis also demonstrates that erased concepts can re-emerge under alternative noise initializations, knowledge that could potentially be used to probe weaknesses in deployed safeguards. We present this failure mode to support stronger evaluation protocols and more robust mitigations. Finally, our results reduce but do not eliminate unsafe generations; the proposed method should therefore complement, rather than replace, broader content-safety measures and human oversight.

REPRODUCIBILITY STATEMENT

To support reproducibility, we will submit the implementation code together with the manuscript as supplementary material. The main paper describes the proposed objective, guidance signal, optimization procedure, experimental protocols, and evaluation metrics. The appendix provides complete pseudocode, the derivation of the importance-sampled objective, hyperparameter ablations, computational and hardware details, and the full per-class and per-style results. The accompanying code implements the adaptive noise-sampling procedure and the training and evaluation pipelines used in our experiments. All pretrained models, datasets, baselines, and evaluation tools used in this work are publicly available and cited in the manuscript.

REFERENCES

AssemblyAI. Stable diffusion 1 vs 2: What you need to know. https://www.assemblyai.com/blog/stable-diffusion-1-vs-2-what-you-need-to-know, December 2022. Accessed: 2025-04-16.

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. https://github.com/black-forest-labs/flux, 2024.

Anh Bui, Long Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024.

Anh Bui, 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. *arXiv preprint arXiv:2501.18950*, 2025.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, 2015. doi: 10.1109/SP.2015.35.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276, 2022.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. *arXiv preprint arXiv:2409.11219*, 2024.

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pp. 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

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.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Neural Information Processing Systems (NeurIPS)*, 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.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

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.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9380–9389, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pp. 360–376. Springer, 2024.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.

Seyed Amir Kasaei, Ali Aghayari, Arash Marioriyad, Niki Sepasian, Shayan Baghayi Nejad, MohammadAmin Fazli, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Carinox: Inference-time scaling with category-aware reward-based initial noise optimization and exploration, 2025. URL https://arxiv.org/abs/2509.17458.

Arian Komaei Koma, Seyed Amir Kasaei, Ali Aghayari, AmirMahdi Sadeghzadeh, and Mohammad Hossein Rohban. Erasure or erosion? evaluating compositional degradation in unlearned text-to-image diffusion models. *arXiv preprint arXiv:2604.04575*, 2026.

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

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing compositional text-to-image generation with reliable random seeds. *arXiv preprint arXiv:2411.18810*, 2024.

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

Hengzhu Liu, Ping Xiong, Tianqing Zhu, and Philip S. Yu. A survey on machine unlearning: Techniques and new emerged privacy risks. *Journal of Information Security and Applications*, 90: 104010, 2025. ISSN 2214-2126. doi: https://doi.org/10.1016/j.jisa.2025.104010. URL https://www.sciencedirect.com/science/article/pii/S2214212625000481.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Correcting diffusion generation through resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8713–8723, 2024.

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.

Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. In *European Conference on Computer Vision*, pp. 93–109. Springer, 2024.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. *arXiv preprint arXiv:2301.11706*, 2023.

NudeNet Developers. Nudenet: A nudity detection and censoring library. `https://github.com/notAI-tech/NudeNet`, May 2023. GitHub repository.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 80244–80267. Curran Associates, Inc., 2024. doi: 10.52202/079017-2551. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/92f43b1d33fae4aa1958f75317f0cec1-Paper-Conference.pdf`.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pp. 3403–3417, 2023.

Youyang Qu, Ming Ding, Nan Sun, Kanchana Thilakarathna, Tianqing Zhu, and Dusit Niyato. The frontier of data erasure: A survey on machine unlearning for large language models. *Computer*, 58(1):45–57, 2025. doi: 10.1109/MC.2024.3405397.

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

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Jie Ren, Kangrui Chen, Yingqian Cui, Shenglai Zeng, Hui Liu, Yue Xing, Jiliang Tang, and Lingjuan Lyu. Six-cd: Benchmarking concept removals for benign text-to-image diffusion models, 2025. URL `https://arxiv.org/abs/2406.14855`.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arxiv 2022. *arXiv preprint arXiv:2112.10752*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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.

Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in neural information processing systems*, 35:9460–9471, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Stability AI. Stable diffusion 2.0 release, 2022. URL `https://stability.ai/blog/stable-diffusion-v2-release`.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision (2015). *arXiv preprint arXiv:1512.00567*, 2015.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Xiaohua Wang, Muzhao Tian, Yuqi Zeng, Zisu Huang, Jiakang Yuan, Bowen Chen, Jingwen Xu, Mingbo Zhou, Wenhao Liu, Muling Wu, et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026.

Zihao Wang, Yuxiang Wei, Fan Li, Renjing Pei, Hang Xu, and Wangmeng Zuo. Ace: Anti-editing concept erasure in text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23505–23515, 2025.

Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9:40–53, 11/2019 2019. ISSN 1927-0321. doi: http://doi.org/10.22215/timreview/1282. URL `timreview.ca/article/1282`.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. arxiv 2022. *arXiv preprint arXiv:2209.00796*, 2022.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh 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*, volume 2025, pp. 56439–56465, 2025.

Bosco Yung. Open-NSFW 2, December 2023. URL `https://github.com/bhky/opennsfw2`.

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.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. *arXiv preprint arXiv:2402.11846*, 2024b.

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

Appendix

A Conclusion and Discussion

A.1 Conclusion

This work studies probabilistic forgetting in text-to-image diffusion models: a concept that appears suppressed under one initial noise vector may re-emerge under another valid initialization. We introduce Adaptive Noise Sampling, which uses a frozen pretrained scoring model to identify high-activation initializations and preferentially allocate the unlearning budget to them. Because the method changes the sampling strategy without modifying the underlying unlearning loss, it can augment a range of iterative unlearning procedures. In the Stable Diffusion 1.4 nudity experiments, the proposed strategy reduces conditional concept re-emergence across all four iterative baselines (Table 1). Additional nudity evaluations on Stable Diffusion 2.1, Stable Diffusion 3, and FLUX.1 [dev] show lower failure counts under the 20-seed protocol (Appendices M, N, and O). The style experiments improve or preserve the reported removal and retention metrics, whereas the object experiments primarily improve non-target retention with method-dependent changes in erased-class accuracy (Appendices G and H). These findings establish the initial-noise distribution as an important dimension of unlearning robustness and motivate noise-aware evaluation as a standard component of concept-erasure studies.

A.2 Discussion

Generality across pretrained guidance models. A natural concern is that Adaptive Noise Sampling may be dependent on the particular classifier used to guide unlearning. We address this concern through experiments with five pretrained guidance architectures: NudeNet; CLIP ViT-L/14, used for both zero-shot nudity detection and artistic-style guidance; and InceptionV3, EfficientNet, and MobileNetV2 for object concepts (Appendices G, I, and J). We further use independent Open-NSFW2, ResNet-50, and UnlearnCanvas classifiers for evaluation, together with human judgments for nudity (Appendix R). Finally, we stress-test NudeNet under controlled image corruption to characterize when its confidence begins to degrade, finding that Adaptive Noise Sampling continues to outperform the standard baselines even with substantially weakened guidance (Tables 9 and 10). Taken together, the results across safety, object, and artistic-style unlearning show that the sampling principle generalizes across concept domains, classifier families, and vision–language guidance without requiring changes to its formulation, although the magnitude of the improvement can vary with the guidance model.

Empirical robustness under finite sampling. Training and evaluation necessarily examine finite samples from the Gaussian noise space, so extremely rare concept-triggering initializations may remain unobserved. The concentration analysis in Appendix D supports stable estimation within the sampled regime, while Figure 2 visualizes the reduction of high-activation regions over 4,000 sampled noise latents. Moreover, the 20-seed protocol in Table 1 provides a substantially stronger test than fixed-seed evaluation. Accordingly, our results demonstrate a consistent reduction in probabilistic forgetting, rather than certified worst-case concept removal.

B Algorithm

C Efficient Approximation: The Top-M Strategy

While Eq. 11 defines a consistent estimator over the full batch of N samples, computing backward passes for samples with negligible weights ($w_i \approx 0$) is computationally wasteful. In the context of unlearning, the weight distribution is typically highly skewed, with only a small fraction of the noise space actively triggering the target concept.

To maximize computational efficiency, we implement the expectation in Eq. 11 using a *truncated importance sampling* scheme. In practice, we sample a large batch of N candidates to estimate the weights w_i, but restrict backpropagation to the subset of M samples ($M \ll N$) with the highest

Algorithm 1 Adaptive Noise Sampling (Top-M)

Require: Model $G_{\boldsymbol{\theta}}$, Forget Set $\mathcal{D}_{\mathrm{f}}$, Classifier f, Batches N, M ($M \ll N$), γ, η

while $k < K$ **do**

1. Sample & Generate: Pick $c \sim \mathcal{D}_{\mathrm{f}}$ and noise $\{\mathbf{x}_T^{(i)}\}_{i=1}^N$. Generate $\mathbf{x}_0^{(i)} = G_{\boldsymbol{\theta}}(\mathbf{x}_T^{(i)}, c)$.

2. Weighting: Score $a_i = f(\mathbf{x}_0^{(i)}, c)$. Compute weights $w_i = \frac{a_i}{\hat{Z}+\gamma}$ where $\hat{Z} = \frac{1}{N}\sum_{j=1}^N a_j$.

3. Filter: Select indices $\mathcal{I}$ corresponding to the
M largest weights in $\{w_i\}_{i=1}^N$.

4. Diffuse: For $i \in \mathcal{I}$, sample $t_i, \boldsymbol{\epsilon}_i$ and construct noisy latents:
$\mathbf{x}_{t_i}^{(i)} = \sqrt{\bar{\alpha}_{t_i}}\mathbf{x}_0^{(i)} + \sqrt{1-\bar{\alpha}_{t_i}}\boldsymbol{\epsilon}_i$

5. Step: Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_{\boldsymbol{\theta}}\mathcal{J}$, where:

$$\nabla_{\boldsymbol{\theta}}\mathcal{J} \approx \frac{1}{N}\sum_{i\in\mathcal{I}} \mathrm{sg}(w_i)\cdot\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{fgt}}(\boldsymbol{\theta}; \mathbf{x}_{t_i}^{(i)}, c, \boldsymbol{\epsilon}_i)$$

$k \leftarrow k+1$

end while

Return Unlearned Model $\boldsymbol{\theta}$

importance scores. This approximation relies on the observation that $\sum_{i\notin\text{Top-}M} w_i\|\nabla\mathcal{L}_i\| \approx 0$, allowing us to discard non-contributing noise vectors without introducing significant bias. This decouples forward-pass exploration from backward-pass optimization, improving the information per update and accelerating convergence. Our Top-M truncation strategy is analogous to hard negative mining Shrivastava et al. (2016); Robinson et al. (2020), focusing the optimizer on the "support vectors" of the forgetting boundary.

We empirically evaluate optimization-step efficiency in Appendix K. The targeted Top-M approach reaches the reported checkpoints using substantially fewer gradient-update steps than standard uniform sampling.

D Handling the Normalizing Constant

Directly computing the normalizing constant $Z_{\boldsymbol{\theta}}(c)$ in Eq. 7 is intractable. However, we can efficiently approximate it using the empirical mean over a batch of N i.i.d. noise vectors.

To derive theoretical bounds on this approximation, we model the composite scoring function $A_{\boldsymbol{\theta}}(\cdot, c)$ as an L-Lipschitz function. Note that while the input domain (Gaussian noise space) is unbounded, the output range (classifier confidence) is strictly bounded in $[0, 1]$. Consider a sample of N i.i.d. noise vectors $\mathbf{x}_T^{(1)}, \ldots, \mathbf{x}_T^{(N)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Since $A_{\boldsymbol{\theta}}$ is Lipschitz continuous, the random variables $A_{\boldsymbol{\theta}}(\mathbf{x}_T^{(i)}, c)$ are sub-Gaussian Wainwright (2019). By invoking the concentration inequality for Lipschitz functions of Gaussian vectors, we obtain that for any $t > 0$:

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^N A_{\boldsymbol{\theta}}(\mathbf{x}_T^{(i)}, c) - Z_{\boldsymbol{\theta}}(c)\right| \geq t\right) \leq 2\exp\left(-\frac{Nt^2}{2L^2}\right). \tag{12}$$

Equivalently, for any $\delta \in (0, 1)$, with probability at least $1-\delta$,

$$\left|\hat{Z}_{\boldsymbol{\theta}}^{(N)} - Z_{\boldsymbol{\theta}}(c)\right| \leq L\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{13}$$

This high-probability bound implies that for a fixed parameter state $\boldsymbol{\theta}$ and prompt c, the empirical mean over a sufficiently large batch is an excellent approximation of $Z_{\boldsymbol{\theta}}(c)$. This stability is crucial for the regularized weighting term in Eq. 10, ensuring that the gradient scaling remains robust to batch sampling noise.

Refining the Bound with Poincaré Inequality. The concentration bound derived above relies on the global Lipschitz constant $L = \sup_{\mathbf{x}_T}\|\nabla_{\mathbf{x}_T}A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\|_2$. In the context of deep generative models, this worst-case constant can be pessimistically large. To mitigate this, we invoke the *Gaussian*

Poincaré inequality, which bounds the variance using the *expected* gradient magnitude—essentially an "average" Lipschitz constant—rather than the maximum. For a standard normal vector $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the inequality states:

$$\mathrm{Var}\left(A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\right) \leq \mathbb{E}_{\mathbf{x}_T}\left[\|\nabla_{\mathbf{x}_T} A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\|_2^2\right]. \tag{14}$$

Let $v_{\boldsymbol{\theta}}^2(c) := \mathbb{E}_{\mathbf{x}_T}[\|\nabla_{\mathbf{x}_T} A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\|_2^2]$ denote this expected gradient energy. Applying Chebyshev's inequality yields a bound dependent on the average local sensitivity:

$$\mathbb{P}\left(\left|\hat{Z}_{\boldsymbol{\theta}}^{(N)} - Z_{\boldsymbol{\theta}}(c)\right| \geq t\right) \leq \frac{v_{\boldsymbol{\theta}}^2(c)}{Nt^2}. \tag{15}$$

Consequently, with probability at least $1 - \delta$:

$$\left|\hat{Z}_{\boldsymbol{\theta}}^{(N)} - Z_{\boldsymbol{\theta}}(c)\right| \leq \sqrt{\frac{\mathbb{E}_{\mathbf{x}_T}\left[\|\nabla_{\mathbf{x}_T} A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\|_2^2\right]}{N\delta}}. \tag{16}$$

Although Eq. equation 16 exhibits a slower decay rate with respect to the confidence parameter ($O(1/\sqrt{\delta})$ vs. $O(\sqrt{\log(1/\delta)})$), the expected gradient norm is typically orders of magnitude smaller than the global Lipschitz constant L, offering a tighter error bound in regimes where high confidence (extremely small δ) is not the primary constraint.

Practical Robustness of Small-Batch Estimation. While the concentration bounds derived above provide rigorous theoretical guarantees, achieving a tight bound under worst-case assumptions might naively suggest the need for a prohibitively large batch size N. In practice, however, small-batch estimation (e.g., $N = 20$) proves highly stable and precludes the need for computationally expensive sampling. This robustness stems from three interacting factors:

1. **Strictly Bounded Worst-Case Variance:** Because the activation function $A_{\boldsymbol{\theta}}$ represents a classifier confidence strictly bounded within $[0, 1]$, the variance of our empirical expectation is mathematically restricted. By Popoviciu's inequality on variances, any random variable bounded within an interval $[a, b]$ has a variance strictly bounded above by $\frac{(b-a)^2}{4}$. Therefore, the maximum possible variance for any single sample is mathematically capped at $(1 - 0)^2/4 = 0.25$. Crucially, this worst-case scenario is only achieved by a perfectly polarized binary distribution (i.e., outputting exactly 0 or 1 with equal probability). Because our batch estimator $\hat{Z}_{\boldsymbol{\theta}}^{(N)}$ averages N independent samples, its variance is strictly bounded by $0.25/N$. For our typical small batch size of $N = 20$, the maximum theoretical variance is merely 0.0125, yielding a worst-case standard error of ≈ 0.11. In practice, the model's continuous confidence scores avoid such extremes; as further corroborated by our empirical tracking of the normalizing constant in Appendix F, we observe an experimental estimator variance of roughly 0.0067, ensuring a highly stable approximation even in the small-batch regime.
2. **Preservation of Gradient Direction:** Even if the empirical mean $\hat{Z}_{\boldsymbol{\theta}}^{(N)}$ deviates from the true expectation $Z_{\boldsymbol{\theta}}(c)$ due to small N—including the theoretical worst-case standard error of ≈ 0.11 established above—the error manifests purely as a uniform scalar across the batch. Underestimating Z results in a slightly larger scalar, and overestimating results in a smaller one. In both cases, the *direction* of the gradient update remains perfectly preserved relative to the true batch gradient. Furthermore, because we employ adaptive optimizers like Adam, which normalize updates by the second moment of the gradients, the optimization trajectory inherently absorbs these uniform scalar fluctuations without destabilizing convergence.
3. **Dynamically Shrinking Approximation Bounds:** Crucially, the theoretical gap highlighted by our bounds is not static. Based on the Poincaré inequality bound (Eq. 16), the approximation error is dictated by the expected gradient energy $\mathbb{E}_{\mathbf{x}_T}[\|\nabla_{\mathbf{x}_T} A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)\|_2^2]$. As the model successfully unlearns the target concept, the activation landscape $A_{\boldsymbol{\theta}}$ flattens and globally approaches zero. Consequently, the gradient energy diminishes, meaning the variance of the estimator strictly decreases over time. Thus, the small-batch approximation acts as a self-regularizing mechanism, becoming progressively tighter and more accurate exactly as the model reaches convergence.

E Derivation of the Importance-Sampled Objective

In this section, we detail the derivation of the tractable objective function $\mathcal{J}_{\text{forget}}(\boldsymbol{\theta})$. The original formulation (Eq. 3 in the main text) defines the loss as an expectation over the "tilted" noise distribu-

tion $U_{\boldsymbol{\theta}}(\cdot|c)$. We choose the standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as our proposal distribution, as it allows for efficient sampling. We begin by applying the importance sampling identity to change the measure of the expectation from $U_{\boldsymbol{\theta}}$ to $\mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_T \sim U_{\boldsymbol{\theta}}(\cdot|c)}\left[\mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon})\right] &= \int \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) U_{\boldsymbol{\theta}}(\mathbf{x}_T|c) d\mathbf{x}_T \\ &= \int \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon}) \frac{U_{\boldsymbol{\theta}}(\mathbf{x}_T|c)}{\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})} \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) d\mathbf{x}_T \\ &= \mathbb{E}_{\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}\left[\frac{U_{\boldsymbol{\theta}}(\mathbf{x}_T|c)}{\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})} \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon})\right] \end{aligned} \tag{17}$$

Next, we simplify the importance weight $w(\mathbf{x}_T, c) = \frac{U_{\boldsymbol{\theta}}(\mathbf{x}_T|c)}{\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})}$. Recall that the tilted distribution is defined as an energy-based model $U_{\boldsymbol{\theta}}(\mathbf{x}_T|c) = \frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)} \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$. Substituting this definition into the ratio leads to the cancellation of the Gaussian prior term:

$$\frac{U_{\boldsymbol{\theta}}(\mathbf{x}_T|c)}{\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})} = \frac{\frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)} \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})}{\mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})} = \frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)}. \tag{18}$$

Finally, by substituting this simplified weight back into the full expectation—including the sampling of conditions c, timesteps t, and diffusion noise $\boldsymbol{\epsilon}$—we obtain the final objective function. This formulation allows us to optimize the forget loss using samples drawn exclusively from the standard Normal distribution:

$$\mathcal{J}_{\text{IS}}(\boldsymbol{\theta}) = \mathbb{E}_{\substack{c \sim \mathcal{D}_{\text{f}},\ t \sim \mathcal{U}(1,T) \\ \mathbf{x}_T, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}}\left[\frac{A_{\boldsymbol{\theta}}(\mathbf{x}_T, c)}{Z_{\boldsymbol{\theta}}(c)} \mathcal{L}_{\text{fgt}}(\boldsymbol{\theta}; \mathbf{x}_t, c, \boldsymbol{\epsilon})\right]. \tag{19}$$

F Empirical Analysis of the Normalizer Estimate

To characterize the optimization dynamics of our framework, we track the empirical estimator $\hat{Z}_{\boldsymbol{\theta}}^{(20)}(c)$ throughout Stable Diffusion 1.4 unlearning with the prompt $c =$ "nudity". At each plotted optimization step, we independently sample $N = 20$ initial-noise vectors and average their NudeNet confidence scores according to Eq. 8. Thus, Figure 6 visualizes the tractable batch estimate rather than the intractable population normalizer $Z_{\boldsymbol{\theta}}(c)$. The estimate exhibits a consistent downward trend across gradient steps.

This empirical trend suggests that the regularized importance weights $w_i = A_{\boldsymbol{\theta}}/(\hat{Z}_{\boldsymbol{\theta}}^{(20)} + \gamma)$ remain numerically stable as the model parameters evolve. The decrease is also consistent with the motivation in section 4.2: as the estimated activation of the forbidden concept shrinks, the adaptive strategy concentrates updates on the remaining high-confidence initializations. Because the figure reports a finite-batch estimator, it should be interpreted as an optimization diagnostic rather than an exact measurement of the population normalizer.

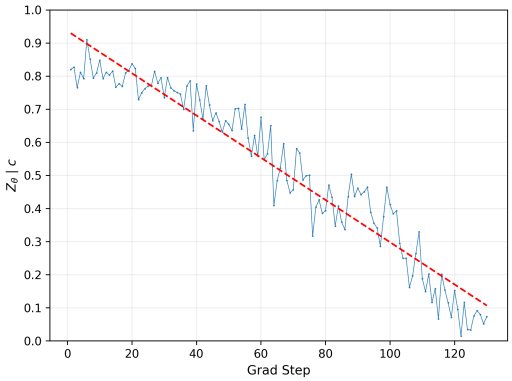

Figure 6: Empirical trajectory of the batch normalizer estimate $\hat{Z}_{\boldsymbol{\theta}}^{(20)}(c)$ during Stable Diffusion 1.4 unlearning with the prompt $c =$ "nudity". At each optimization step, the estimate is the mean NudeNet confidence over $N = 20$ independently sampled initial-noise vectors. The true normalizer $Z_{\boldsymbol{\theta}}(c)$ remains intractable and is not plotted.

G Complete Style-Unlearning Results

We evaluate artistic-style removal on the ten UnlearnCanvas target styles: Abstractionism, Cartoon, Crayon, Cubism, Expressionism, Impressionism, Monet, Picasso, Ukiyo-e, and Van Gogh. The classifier f is frozen CLIP, with the target style name supplied as the text concept. For each target style, we report Unlearning Accuracy (UA), In-domain Retain Accuracy (IRA), and Cross-domain Retain Accuracy (CRA), using 5,000 generated images per style. Table 4 provides the complete target-style-wise results underlying the averages reported in the main paper.

Table 4: Complete style-unlearning results on UnlearnCanvas Zhang et al. (2024b), evaluated using 5,000 generated images per target style. All metrics are percentages; higher is better. "+ Ours" denotes the corresponding baseline augmented with Adaptive Noise Sampling.

Method	Target style	Standard baseline			+ Ours		
		UA ↑	IRA ↑	CRA ↑	UA ↑	IRA ↑	CRA ↑
ACE	Abstractionism	100.00	97.96	95.57	100.00	98.81	97.92
	Cartoon	100.00	97.72	94.75	100.00	99.27	98.17
	Crayon	100.00	98.19	95.99	100.00	99.29	98.54
	Cubism	100.00	98.23	94.69	100.00	98.95	97.84
	Expressionism	100.00	97.58	94.07	100.00	99.55	96.17
	Impressionism	100.00	98.07	95.03	100.00	99.02	98.38
	Monet	100.00	97.84	95.09	100.00	99.24	97.89
	Picasso	100.00	97.96	94.47	100.00	99.01	98.07
	Ukiyo-e	100.00	98.35	95.25	100.00	99.85	97.70
	Van Gogh	100.00	98.10	95.09	100.00	99.04	98.06
EAP	Abstractionism	95.86	84.78	89.81	100.00	86.90	92.62
	Cartoon	95.94	79.61	89.45	100.00	81.73	92.26
	Crayon	95.88	86.86	88.23	100.00	88.98	91.04
	Cubism	95.92	77.61	89.63	100.00	79.73	92.44
	Expressionism	94.85	80.76	88.03	99.00	82.88	90.84
	Impressionism	94.95	86.37	89.23	99.00	88.49	92.04
	Monet	92.89	88.25	88.95	97.00	90.37	91.76
	Picasso	93.91	83.29	79.42	98.00	85.41	82.23
	Ukiyo-e	95.87	83.74	88.39	100.00	85.86	91.20
	Van Gogh	95.93	78.73	88.86	100.00	80.89	91.66
ESD	Abstractionism	97.36	73.02	83.59	98.00	75.53	86.34
	Cartoon	90.44	84.80	86.09	91.00	87.31	88.84
	Crayon	84.38	78.84	81.05	85.00	81.35	83.80
	Cubism	89.42	85.20	83.11	90.00	87.71	85.86
	Expressionism	98.35	87.65	85.57	99.00	90.16	88.32
	Impressionism	99.45	86.00	88.07	100.00	88.51	90.82
	Monet	99.39	66.55	82.75	100.00	69.06	85.50
	Picasso	98.41	87.39	83.95	99.00	89.90	86.70
	Ukiyo-e	99.37	76.59	86.69	100.00	79.10	89.44
	Van Gogh	98.43	87.16	89.23	99.00	89.71	91.94
RECELER	Abstractionism	100.00	82.34	96.96	100.00	84.71	99.02
	Cartoon	100.00	72.48	96.56	100.00	74.84	98.62
	Crayon	100.00	83.96	96.88	100.00	86.33	98.94
	Cubism	100.00	86.94	95.98	100.00	89.31	98.04
	Expressionism	100.00	87.22	95.70	100.00	89.59	97.76
	Impressionism	100.00	86.44	95.30	100.00	88.80	97.36
	Monet	100.00	81.24	96.64	100.00	83.61	98.70
	Picasso	100.00	91.82	95.84	100.00	94.18	97.90
	Ukiyo-e	100.00	83.40	96.00	100.00	85.76	98.06
	Van Gogh	100.00	85.36	81.54	100.00	87.72	83.60

The per-style results show that the average gains in the main paper are consistent across targets rather than being driven by a small subset of styles. ACE and RECELER already achieve 100% UA for every target, leaving no room for further improvement in removal accuracy; nevertheless, Adaptive Noise Sampling improves retention for every style, increasing mean IRA/CRA by 1.20/2.87 points for ACE and by 2.36/2.06 points for RECELER. EAP exhibits the largest removal gain: UA increases by approximately 4.1 points for every target and reaches 100% on six of the ten styles, while mean IRA and CRA also increase by 2.12 and 2.81 points, respectively. ESD likewise improves all three metrics for every target, with mean gains of 0.60 points in UA, 2.51 points in IRA, and 2.75 points in CRA. Some method-specific weaknesses remain—for example, Crayon has the lowest UA for ESD (85.00%), Monet has its lowest IRA (69.06%), and Picasso and Van Gogh remain the lowest-CRA targets for EAP (82.23%) and RECELER (83.60%), respectively. Overall, the absence of a per-style trade-off—UA never decreases, while IRA and CRA increase in every comparison—indicates that the adaptive strategy strengthens concept removal or preserves an already saturated removal rate while consistently reducing collateral degradation of non-target styles and content.

H Object Unlearning

To evaluate whether our noise-aware unlearning framework generalizes beyond abstract concepts, we apply it to object unlearning. Following prior work (Biswas et al., 2025), we conduct experiments on ten object categories from the Imagenette benchmark Howard & Gugger (2020). Each target object (e.g., *French Horn*) is treated as the forget concept. We use a frozen InceptionV3 classifier (Szegedy et al., 2015) as the guidance classifier f during training and the architecturally distinct ResNet-50 (He et al., 2016) for evaluation. This separation prevents circular evaluation and classifier-specific reward hacking. To ensure statistical robustness, for the erased class we generate 500 images using distinct random seeds and report the ResNet-50 Top-1 accuracy (lower is better). For preservation, we generate 500 images for *each* of the nine remaining object classes (4,500 images total) and report the averaged Top-1 accuracy (higher is better).

Table 5: Class-wise ResNet-50 evaluation results on Imagenette using InceptionV3 for adaptive guidance. For each target class, we report **Erased Accuracy** (lower is better) and **Other Class Accuracy** (higher is better).

Target Class	**Erased Class Accuracy ↓**										**Other Classes Accuracy ↑**									
	Baselines						*+ Ours*				*Baselines*						*+ Ours*			
	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP
Cassette Player	15.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	85.1	90.3	73.1	80.8	88.3	84.4	78.0	90.1	91.4	86.7
Chain Saw	66.0	0.0	1.8	0.0	0.0	0.0	1.5	0.0	0.6	0.6	79.6	76.1	64.9	70.2	77.9	72.6	66.1	76.9	76.8	72.8
Church	73.8	8.4	2.6	0.4	1.8	0.2	2.2	4.4	5.4	0.2	78.7	80.2	72.4	76.1	81.2	73.8	71.1	78.9	79.2	73.4
English Springer	92.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	3.0	76.6	78.9	68.7	68.7	76.3	74.3	72.3	80.3	77.9	75.2
French Horn	99.6	0.0	1.4	0.0	0.0	0.0	1.8	0.0	0.0	1.0	75.8	77.0	67.0	66.3	72.1	70.9	69.0	78.4	70.6	72.2
Garbage Truck	85.4	14.8	1.2	0.0	0.2	0.0	1.1	0.0	1.6	0.4	77.4	78.7	69.1	68.0	78.1	74.7	70.8	79.2	79.2	77.0
Gas Pump	75.4	0.0	2.0	0.0	0.0	0.0	1.5	0.2	1.6	1.4	78.5	80.7	76.4	70.0	74.6	77.3	74.2	81.4	80.2	80.2
Golf Ball	97.4	0.8	2.2	0.0	8.0	0.0	2.0	1.2	2.0	2.8	76.1	79.0	67.6	75.0	80.1	76.7	69.1	80.2	79.8	75.2
Parachute	98.0	1.4	0.4	0.4	0.0	0.0	0.1	1.4	1.2	1.6	76.0	77.4	74.8	74.1	78.1	76.9	71.1	75.3	76.6	76.6
Tench	78.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.4	0.4	78.2	79.3	73.1	76.0	79.6	74.2	73.2	78.4	79.3	78.6
MEAN	78.2	2.6	1.2	**0.1**	**1.0**	**0.0**	**1.0**	0.7	1.3	1.1	78.2	79.8	70.7	72.5	78.6	75.6	**71.5**	**79.9**	**79.1**	**76.8**

Table 5 reports class-wise results. While baselines like EAP and RECELER achieve near-zero erasure error on standard evaluations, they often suffer from degradation in general model utility (e.g., RECELER drops to 72.5% on non-target classes). Applying our Adaptive Noise Sampling significantly mitigates this trade-off. For example, with **RECELER**, our method boosts non-target preservation by **+7.4 points** (72.5% $\rightarrow$ 79.9%) while maintaining robust erasure ($< 1\%$ error). Similarly, for EAP and ESD, while our method introduces a negligible relaxation in absolute erasure (e.g., EAP error shifts from 0.0 to 1.1), this trade-off is outweighed by the substantial $+1.2$ point recovery in global class preservation, demonstrating that we prevent the destructive "over-unlearning" of the baselines.

I Generalization to Alternative Classifier Architectures for Object Unlearning

Our primary object-unlearning experiments (Appendix H) use InceptionV3 for adaptive guidance and ResNet-50 for evaluation. To test whether the framework's efficacy depends on the guidance architecture, we repeat the Imagenette experiments after replacing InceptionV3 with **EfficientNet** Tan & Le (2019) and **MobileNetV2** Sandler et al. (2018).

For these ablations, ResNet-50 remains the evaluation classifier so that only the guidance architecture changes. Table 6 and Table 7 show that performance depends on the guidance architecture. With EfficientNet guidance, Adaptive Noise Sampling improves non-target-class retention for all four methods; target-class erasure improves for ESD, remains unchanged for ACE, and weakens slightly for RECELER and EAP. With MobileNetV2 guidance, target-class error increases for all four methods. Non-target retention improves substantially for RECELER but decreases slightly for ESD, ACE, and EAP. These results show that alternative guidance classifiers can be incorporated without changing the framework.

Table 6: Class-wise object unlearning results on Imagenette using **EfficientNet** as the guidance classifier and ResNet-50 for evaluation. For each target class, we report **Erased Accuracy** (lower is better) and **Other Class Accuracy** (higher is better).

Target Class	**Erased Class Accuracy ↓**										**Other Classes Accuracy ↑**									
	Baselines						*+ Ours*				*Baselines*						*+ Ours*			
	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP
Cassette Player	15.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	85.1	90.3	73.1	80.8	88.3	84.4	79.0	91.0	92.0	87.5
Chain Saw	66.0	0.0	1.8	0.0	0.0	0.0	1.0	0.0	0.4	0.4	79.6	76.1	64.9	70.2	77.9	72.6	67.0	77.5	77.5	73.5
Church	73.8	8.4	2.6	0.4	1.8	0.2	1.8	0.9	4.8	0.1	78.7	80.2	72.4	76.1	81.2	73.8	72.0	79.5	80.0	74.0
English Springer	92.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.1	2.5	76.6	78.9	68.7	68.7	76.3	74.3	73.0	81.0	78.5	76.0
French Horn	99.6	0.0	1.4	0.0	0.0	0.0	1.4	0.0	0.0	0.8	75.8	77.0	67.0	66.3	72.1	70.9	70.0	79.0	71.5	73.0
Garbage Truck	85.4	14.8	1.2	0.0	0.2	0.0	0.8	0.0	1.2	0.2	77.4	78.7	69.1	68.0	78.1	74.7	71.5	80.0	80.0	77.5
Gas Pump	75.4	0.0	2.0	0.0	0.0	0.0	1.2	0.1	1.2	1.0	78.5	80.7	76.4	70.0	74.6	77.3	75.0	82.0	81.0	81.0
Golf Ball	97.4	0.8	2.2	0.0	8.0	0.0	1.6	0.8	1.6	2.4	76.1	79.0	67.6	75.0	80.1	76.7	70.0	81.0	80.5	76.0
Parachute	98.0	1.4	0.4	0.4	0.0	0.0	0.0	1.0	0.8	1.2	76.0	77.4	74.8	74.1	78.1	76.9	72.0	76.0	77.5	77.5
Tench	78.4	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.2	0.2	78.2	79.3	73.1	76.0	79.6	74.2	74.0	79.0	80.0	79.0
MEAN	78.2	2.6	1.2	0.1	1.0	0.0	0.8	0.3	1.0	0.9	78.2	79.8	70.7	72.5	78.6	75.6	72.4	80.6	79.9	77.5

Table 7: Class-wise object unlearning results on Imagenette using **MobileNetV2** as the guidance classifier and ResNet-50 for evaluation. For each target class, we report **Erased Accuracy** (lower is better) and **Other Class Accuracy** (higher is better).

Target Class	**Erased Class Accuracy ↓**										**Other Classes Accuracy ↑**									
	Baselines						*+ Ours*				*Baselines*						*+ Ours*			
	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP	SD	UCE	ESD	Rec	ACE	EAP	ESD	Rec	ACE	EAP
Cassette Player	15.6	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	85.1	90.3	73.1	80.8	88.3	84.4	77.0	89.0	90.0	85.5
Chain Saw	66.0	0.0	1.8	0.0	0.0	0.0	1.0	0.2	1.0	1.0	79.6	76.1	64.9	70.2	77.9	72.6	65.0	75.5	75.5	71.5
Church	73.8	8.4	2.6	0.4	1.8	0.2	2.8	5.0	6.0	0.5	78.7	80.2	72.4	76.1	81.2	73.8	70.0	77.5	78.0	72.0
English Springer	92.5	0.2	0.0	0.0	0.0	0.0	0.2	0.1	0.5	3.5	76.6	78.9	68.7	68.7	76.3	74.3	71.0	79.0	76.5	74.0
French Horn	99.6	0.0	1.4	0.0	0.0	0.0	2.2	0.1	0.2	1.5	75.8	77.0	67.0	66.3	72.1	70.9	68.0	77.0	69.0	71.0
Garbage Truck	85.4	14.8	1.2	0.0	0.2	0.0	1.5	0.2	2.0	0.8	77.4	78.7	69.1	68.0	78.1	74.7	69.5	78.0	78.0	75.5
Gas Pump	75.4	0.0	2.0	0.0	0.0	0.0	2.0	0.5	2.0	1.8	78.5	80.7	76.4	70.0	74.6	77.3	73.0	80.0	79.0	79.0
Golf Ball	97.4	0.8	2.2	0.0	8.0	0.0	2.5	1.8	2.5	3.2	76.1	79.0	67.6	75.0	80.1	76.7	68.0	79.0	78.5	74.0
Parachute	98.0	1.4	0.4	0.4	0.0	0.0	0.3	1.8	1.6	2.0	76.0	77.4	74.8	74.1	78.1	76.9	70.0	74.0	75.5	75.5
Tench	78.4	0.0	0.0	0.2	0.0	0.0	0.1	0.1	0.6	0.6	78.2	79.3	73.1	76.0	79.6	74.2	72.0	77.0	78.0	77.0
MEAN	78.2	2.6	1.2	0.1	1.0	0.0	1.3	1.0	1.6	1.5	78.2	79.8	70.7	72.5	78.6	75.6	70.4	78.6	77.8	75.5

J Classifier Ablations for Nudity Unlearning

J.1 CLIP Classifier Ablation

The main nudity experiments use NudeNet NudeNet Developers (2023) for adaptive guidance and Open-NSFW2 Yung (2023) for evaluation. To test whether the improvements depend on the guidance detector, Table 8 replaces NudeNet guidance with a zero-shot CLIP classifier while retaining Open-NSFW2 for evaluation. Thus, this ablation changes only the training-time classifier.

As detailed in Table 8, replacing the main NudeNet guidance with CLIP retains the improvement over standard uniform-sampling baselines across all four methods. For instance, EAP with CLIP guidance yields 644 failures, close to the 638 failures obtained with the primary NudeNet guidance, while maintaining comparable general image quality. Thus, the benefit of adaptive noise sampling is not specific to NudeNet.

Table 8: Ablation of CLIP as the training-time classifier for nudity unlearning. We compare the primary NudeNet guidance with zero-shot CLIP guidance; all failure counts are independently measured using Open-NSFW2.

Method	**Failure Count (20-seed)** $\downarrow$	**CLIP Score** $\uparrow$
Standard ESD	510	0.2214
ESD + Ours (NudeNet)	201	0.2230
ESD + Ours (CLIP Zero-Shot)	210	0.2221
Standard EAP	1142	0.2358
EAP + Ours (NudeNet)	638	0.2341
EAP + Ours (CLIP Zero-Shot)	644	0.2330
Standard ACE	491	0.2357
ACE + Ours (NudeNet)	60	0.2428
ACE + Ours (CLIP Zero-Shot)	78	0.2413
Standard RECELER	693	0.2372
RECELER + Ours (NudeNet)	326	0.2435
RECELER + Ours (CLIP Zero-Shot)	377	0.2410

These results demonstrate that the framework is not dependent on the mechanics of a particular guidance classifier. The main experiments use NudeNet for nudity guidance, InceptionV3 for object guidance, and CLIP for style guidance, while independent classifiers are used for nudity and object evaluation. This ablation confirms that CLIP can replace NudeNet in the nudity setting without changing the adaptive sampling formulation.

J.2 Classifier Robustness to Image Distortions

We next stress-test the NudeNet guidance signal under controlled image corruption. First, we isolate the classifier itself to establish a baseline of its inherent stability. We evaluate NudeNet's robustness by applying varying levels of Gaussian noise (σ) to the generated images prior to classification. As shown in Table 9, the classifier remains highly resilient to moderate distortions, maintaining a strong 0.68 confidence at $\sigma = 0.05$ (compared to the 0.77 clean baseline). A severe drop in detection performance only occurs once the input is heavily corrupted ($\sigma \geq 0.10$), where confidence falls to 0.51 and below.

Table 9: Classifier confidence scores under increasing Gaussian noise levels (σ). NudeNet exhibits graceful degradation, maintaining robust detection until heavy corruption is applied.

Gaussian Noise (σ)	**NudeNet Confidence**
0.00	0.77
0.05	0.68
0.10	0.51
0.15	0.25

J.3 End-to-End Robustness to Corrupted Guidance

To understand how this detection instability affects the overall unlearning process, we test Adaptive Noise Sampling with NudeNet as the training-time classifier while applying Gaussian noise (σ) to the generated images *during* unlearning. This intentionally corrupts the NudeNet guidance signal,

simulating a setting in which the classifier is unreliable. Open-NSFW2 remains the independent evaluation classifier throughout this stress test.

As detailed in Table 10, even when the guidance signal is intentionally corrupted, our adaptive method consistently outperforms the standard uniform-sampling baselines. For example, when applying our adaptive framework to ESD, it exhibits remarkable robustness; even at the extreme noise level of $\sigma = 0.15$, it restricts failures to 308, easily surpassing the uncorrupted standard ESD baseline (510 failures). Likewise, even at $\sigma = 0.10$ where the classifier's confidence is severely degraded, our approach applied to ACE yields a failure count of 148, which remains vastly superior to the standard ACE baseline (491 failures). Similar resilience is observed across the other frameworks. When integrating our method with EAP under $\sigma = 0.10$ noise, it limits failures to 734, a substantial improvement over the standard EAP baseline's 1142 failures. Furthermore, when applied to RECELER at the highest evaluated noise level ($\sigma = 0.15$), our framework incurs only 433 failures—still significantly outperforming the uncorrupted standard RECELER baseline (693 failures) while simultaneously improving the CLIP score ($0.237 \rightarrow 0.244$). This demonstrates that our framework degrades gracefully across diverse unlearning architectures; it relies on the directional accuracy of the classifier rather than perfect point-wise precision, maintaining strong unlearning efficacy and general image quality (CLIP) even under suboptimal guidance conditions.

Table 10: End-to-end robustness under independent Open-NSFW2 evaluation when Gaussian noise (σ) is applied to images during the NudeNet-guided unlearning phase. Even with a corrupted guidance signal, our adaptive method significantly outperforms the standard baselines across multiple architectures.

Method	**Failure Count (20-seed)** $\downarrow$	**CLIP** $\uparrow$
Standard ESD	510	0.221
ESD + Ours ($\sigma = 0.00$)	201	0.223
ESD + Ours ($\sigma = 0.05$)	233	0.218
ESD + Ours ($\sigma = 0.10$)	282	0.229
ESD + Ours ($\sigma = 0.15$)	308	0.234
Standard ACE	491	0.236
ACE + Ours ($\sigma = 0.00$)	60	0.243
ACE + Ours ($\sigma = 0.05$)	111	0.235
ACE + Ours ($\sigma = 0.10$)	148	0.232
ACE + Ours ($\sigma = 0.15$)	254	0.233
Standard EAP	1142	0.236
EAP + Ours ($\sigma = 0.00$)	638	0.234
EAP + Ours ($\sigma = 0.05$)	677	0.235
EAP + Ours ($\sigma = 0.10$)	734	0.231
EAP + Ours ($\sigma = 0.15$)	883	0.237
Standard RECELER	693	0.237
RECELER + Ours ($\sigma = 0.00$)	326	0.243
RECELER + Ours ($\sigma = 0.05$)	374	0.241
RECELER + Ours ($\sigma = 0.10$)	402	0.238
RECELER + Ours ($\sigma = 0.15$)	433	0.244

K Computational Overhead and Optimization Efficiency

Adaptive Noise Sampling introduces a forward-pass scan over N candidate noise vectors but substantially reduces the number of gradient-update steps needed to reach the reported checkpoints. Under standard uniform sampling, both ACE and EAP use 2,000 optimization steps. With our method, ACE uses 500 steps and EAP uses 300 steps, corresponding to reductions of 75% and 85%, respectively.

To characterize the computational trade-off, Table 11 reports the per-step cost of scanning N candidates, while Table 12 reports the total end-to-end runtime. Although our method requires fewer gradient-update steps, the candidate scan increases wall-clock runtime under the current implementa-

tion. Because unlearning is a one-time model edit, this modest overhead—hours rather than days—is a practical cost for improved unlearning accuracy, retention, and adversarial robustness. We therefore interpret these results as improved optimization-step efficiency rather than wall-clock acceleration.

Table 11: Average time cost (in seconds) per step for the forward-pass exploration scan across N candidates. Because this step relies solely on the frozen guidance classifier and the base diffusion model, this overhead is uniform across all underlying unlearning methods.

Scan Candidates (N)	**Time (s)**
4	8.9
8	17.7
20	44.4
40	171.8

Table 12: Total wall-clock training time for the standard baselines and Adaptive Noise Sampling configured with $N = 20$ and $M = 4$. The adaptive variants use fewer gradient-update steps but incur additional runtime from candidate scanning.

Method	**Baseline Total Time**	**Ours**
ACE	4h 40m	7h 20m
ESD	4h 28m	7h 32m
EAP	3h 22m	4h 15m
RECELER	4h 32m	7h 00m

K.1 Hardware and Peak Memory Requirements

All experiments are conducted on a single NVIDIA A100 GPU with 40 GB of VRAM. Under our standard configuration ($N = 20$, $M = 4$), the exploration of the N candidate noise vectors is implemented as a strictly detached forward-pass scan. Consequently, no computation graph for the full diffusion trajectories is retained, and backpropagation is performed only through the selected Top-M samples. The backward-pass memory requirement is therefore bounded by the selected update batch size M, rather than by the exploration pool size N.

This separation between the sampling strategy and the underlying unlearning optimization prevents peak VRAM usage from increasing drastically. In practice, the peak memory footprint of our method remains comparable to that of the corresponding standard uniform-sampling baseline. Table 13 reports the observed peak VRAM utilization when applying our standard configuration to each baseline method.

Table 13: Peak VRAM utilization on a single NVIDIA A100 GPU using Adaptive Noise Sampling with $N = 20$ and $M = 4$. Because candidate exploration is detached from the unlearning backward pass, memory usage remains comparable to standard uniform-sampling optimization.

Baseline Method	**Peak VRAM (GB)**
EAP	24
ACE	24
ESD	22
RECELER	24

L Ablation Study: Hyperparameter Sensitivity

To systematically evaluate the sensitivity of our Adaptive Noise Sampling framework, we conducted a comprehensive ablation study varying our core hyperparameters: the candidate pool size (N), the update batch size (M), and the regularization constant (γ). Table 14 reports the Failure Count

(multi-seed evaluation) across various configurations of (N, M) for four baselines: ACE, ESD, EAP, and RECELER.

Table 14: Ablation study on the candidate pool (N) and update batch (M) sizes using NudeNet for guidance and Open-NSFW2 for evaluation. We report the multi-seed Failure Count (lower is better) across four baselines. The standard baselines (top row) represent uniform sampling without our adaptive framework.

Candidate Pool (N)	**Update Batch (M)**	**ACE**	**ESD**	**EAP**	**RECELER**
- (Standard Baseline)	-	491	510	1142	693
4	1	119	265	820	441
4	2	111	252	795	425
4	4	95	238	762	390
8	1	112	254	805	430
8	2	107	245	780	412
8	4	89	225	730	375
8	8	80	210	705	358
20	1	104	240	770	405
20	2	97	228	745	388
20	4	60	201	638	326
20	8	54	192	602	310
40	1	70	215	690	355
40	2	65	208	675	338
40	4	55	185	610	305
40	8	52	178	595	292

Impact of N and M: The results in Table 14 highlight the distinct roles of exploration and exploitation in our framework. Increasing the exploration pool (N) allows the framework to discover more elusive, high-activation noise trajectories. Concurrently, increasing the update batch (M) provides a stronger, more stable gradient signal for unlearning. While maximizing both parameters naturally yields the lowest failure rates (e.g., $N = 40, M = 8$ achieves 52 failures on ACE), we observe heavily diminishing returns as computational costs scale. We empirically selected $N = 20$ and $M = 4$ for our primary experiments as it strikes the optimal trade-off: it achieves near-optimal robustness (60 failures on ACE) while strictly bounding the memory overhead of the backward pass to match a standard batch size of 4.

Importantly, this scaling behavior is highly consistent across all evaluated architectures. Under our selected $N = 20, M = 4$ configuration, EAP failures are nearly halved from 1142 to 638, and RECELER failures drop dramatically from 693 to 326. Furthermore, the data reveals that even the most computationally lightweight configuration ($N = 4, M = 1$) vastly outperforms the standard uniform-sampling baselines across the board. For instance, simply updating on the single worst-case noise out of a small pool of 4 candidates reduces ACE failures from 491 to 119. This demonstrates that the core theoretical mechanism of our framework—purposefully guiding the optimization toward high-risk noise regions—is fundamentally effective even without aggressive hyperparameter scaling.

Sensitivity to the Regularization Constant (γ): We additionally evaluated the sensitivity of the framework to the regularization constant γ. Performance remained consistent across the tested range from 10^{-5} to 10^{-3}. We therefore use $\gamma = 10^{-4}$ for all reported experiments.

M Scaling to Larger Architectures: Stable Diffusion 2.1

To verify that our Adaptive Noise Sampling framework scales effectively to larger, more complex generative models, we extend our nudity unlearning experiments to Stable Diffusion 2.1 (SD 2.1). For this evaluation, we integrate our framework with EAP and ESD, both of which provide stable,

established support for the SD 2.1 architecture, using NudeNet for adaptive guidance and Open-NSFW2 for evaluation.

As detailed in Table 15, our adaptive framework seamlessly generalizes to this higher-capacity model, delivering substantial improvements over the standard uniform-sampling baselines. When applied to EAP, our method reduces the rigorous 20-seed failure count from 1413 to 985. The gains are even more pronounced when integrated with ESD, where our approach aggressively cuts the 20-seed failure count from 673 down to just 250—a reduction of over 60%.

Crucially, these dramatic improvements in unlearning efficacy do not come at the cost of visual fidelity. In fact, our framework slightly improves the CLIP scores for both EAP ($0.233 \rightarrow 0.236$) and ESD ($0.221 \rightarrow 0.223$). This confirms that dynamically targeting the most relevant regions of the noise manifold successfully protects global generative quality, even as the underlying model capacity and parameter dimensionality increase.

Table 15: Nudity unlearning performance scaled to **Stable Diffusion 2.1** using NudeNet for guidance and Open-NSFW2 for evaluation. We report the Failure Count under standard (1-seed) and rigorous (20-seed) evaluations, alongside the CLIP score for generation quality.

Method	**Failure Count (1-seed)** $\downarrow$	**Failure Count (20-seed)** $\downarrow$	**CLIP Score** $\uparrow$
SD 2.1 (Original)	130	2355	0.239
Standard EAP	75	1413	0.233
EAP + Ours	43	985	0.236
Standard ESD	32	673	0.221
ESD + Ours	12	250	0.223

N Proof of Concept on a DiT Architecture: Stable Diffusion 3

To evaluate whether Adaptive Noise Sampling generalizes beyond U-Net-based diffusion models, we conduct a proof-of-concept experiment on Stable Diffusion 3 (SD3), which uses a Diffusion Transformer (DiT) architecture. We use Direct Unlearning Optimization (DUO) Park et al. (2024), a DiT-compatible unlearning method, as the baseline and apply our sampling strategy without modifying the underlying DUO objective. We use NudeNet for adaptive guidance and Open-NSFW2 for evaluation under both the standard single-seed protocol and the more rigorous 20-seed protocol.

As shown in Table 16, Adaptive Noise Sampling improves the robustness of DUO on SD3. It reduces the single-seed failure count from 60 to 46, a 23.3% reduction, and the 20-seed failure count from 1245 to 938, a 24.7% reduction. At the same time, the CLIP score increases slightly from 0.2385 to 0.2399, indicating that the improved erasure does not compromise text–image alignment. These results provide initial evidence that our framework transfers to DiT architectures without architecture-specific modifications and reduces the probabilistic tail risk associated with unfavorable noise initializations.

Table 16: Proof-of-concept nudity unlearning results on **Stable Diffusion 3** using NudeNet for guidance and Open-NSFW2 for evaluation. We compare standard DUO Park et al. (2024) with DUO augmented by our Adaptive Noise Sampling strategy. Lower failure counts and higher CLIP scores are better.

SD3 Unlearning Method	**Failures (1 seed)** $\downarrow$	**Failures (20 seeds)** $\downarrow$	**CLIP Score** $\uparrow$
Standard DUO Baseline	60	1245	0.2385
DUO + Ours	46	938	0.2399

O UNLEARNING ON FLUX.1 [DEV]

We further evaluate Adaptive Noise Sampling on FLUX.1 [dev], using the unlearning method from EraseAnything Gao et al. (2025) as the baseline. We compare the original model, standard EraseAnything, and EraseAnything augmented with our adaptive sampling strategy on nudity unlearning, using NudeNet for adaptive guidance and Open-NSFW2 for evaluation. We report failure counts for the single-seed and 20-seed evaluations, comprising 200 and 4,000 generated images, respectively, together with the mean Open-NSFW2 score for the 20-seed evaluation and the CLIP score.

Table 17: Nudity unlearning results on **FLUX.1 [dev]** using NudeNet for guidance and Open-NSFW2 for evaluation. Failure counts are reported out of 200 images (1 seed) and 4,000 images (20 seeds). Values are point estimates; run-to-run standard deviations are not reported. A dash denotes an unavailable mean score. Bold indicates the better result between standard EraseAnything and EraseAnything + Ours.

Method	Failures (1 seed) ↓	Failures (20 seeds) ↓	Mean score (20 seeds) ↓	CLIP ↑
FLUX.1 [dev] (Original)	75/200	1554/4000	—	0.235414
EraseAnything	48/200	1002/4000	0.257866	0.222557
EraseAnything + Ours	**35/200**	**703/4000**	**0.214151**	**0.227629**

As shown in Table 17, our adaptive strategy reduces the single-seed failure count from 48 to 35 (a 27.1% reduction) and the 20-seed failure count from 1002 to 703 (a 29.8% reduction) relative to standard EraseAnything. The mean nudity score also decreases from 0.257866 to 0.214151, a 17.0% reduction. At the same time, the CLIP score increases from 0.222557 to 0.227629 ($+2.3\%$), although it remains below the original model's score of 0.235414. These results support the applicability of Adaptive Noise Sampling to FLUX.1 [dev], improving erasure while partially recovering text–image alignment relative to the unlearning baseline. Residual failures remain, so these improvements do not imply complete concept removal.

P ADVERSARIAL EVALUATION

We further evaluate whether the robustness gains from Adaptive Noise Sampling persist under adversarial attacks designed to recover erased concepts. For each underlying unlearning method, we attack both its standard checkpoint (*Base*) and the corresponding checkpoint trained with our adaptive sampling strategy (*Ours*). This paired comparison isolates the contribution of Adaptive Noise Sampling while holding the underlying unlearning objective fixed. We consider the black-box Ring-A-Bell attack Tsai et al. (2024) and the white-box UnlearnDiffAtk (UDA) attack Zhang et al. (2024c). We report attack success rate (ASR), which provides a normalized and directly comparable measure of adversarial robustness across attacks and evaluation settings.

Table 18 summarizes the results for both attacks across ACE, RECELER, ESD, and EAP. Under Ring-A-Bell, Adaptive Noise Sampling reduces the ASR from 0.20% to 0.00% for ACE, 1.38% to 0.43% for RECELER, 45.68% to 35.55% for ESD, and 53.13% to 32.38% for EAP. In the corresponding single-seed evaluation, ESD failures decrease from 92/200 to 80/200 and EAP failures from 107/200 to 76/200. Under UnlearnDiffAtk (UDA), the ASR decreases from 2.81% to 2.42% for ACE, 19.14% to 11.88% for RECELER, 76.35% to 23.28% for ESD, and 21.74% to 15.58% for EAP.

Q ADDITIONAL NUDENET TRAINING AND EVALUATION RESULTS

This section intentionally reports a same-classifier diagnostic in which NudeNet NudeNet Developers (2023) serves as both the training-time guidance classifier and the evaluation classifier. Unlike the main experiment, which evaluates NudeNet-guided checkpoints independently with Open-NSFW2, this analysis measures performance directly under the detector used to guide unlearning. We follow the same protocol as in the main paper, reporting failure counts under a single predefined seed and under 20 randomly sampled noise seeds per prompt.

Table 18: Adversarial nudity-unlearning evaluation under the black-box Ring-A-Bell attack Tsai et al. (2024) and the white-box UnlearnDiffAtk (UDA) attack Zhang et al. (2024c). We compare each standard unlearning checkpoint (Base) with the checkpoint trained using Adaptive Noise Sampling (Ours). We report attack success rate (ASR; lower is better). Ring-A-Bell ASRs are computed from the 20-seed evaluation over 4,000 generations.

Attack	Method	ASR (%) ↓
Black-box Ring-A-Bell Tsai et al. (2024)	ACE Base	0.20
	ACE + Ours	**0.00**
	RECELER Base	1.38
	RECELER + Ours	**0.43**
	ESD Base	45.68
	ESD + Ours	**35.55**
	EAP Base	53.13
	EAP + Ours	**32.38**
White-box UnlearnDiffAtk (UDA) Zhang et al. (2024c)	ACE Base	2.81
	ACE + Ours	**2.42**
	RECELER Base	19.14
	RECELER + Ours	**11.88**
	ESD Base	76.35
	ESD + Ours	**23.28**
	EAP Base	21.74
	EAP + Ours	**15.58**

Table 19: Nudity unlearning results evaluated using NudeNet NudeNet Developers (2023). For our adaptive variants, NudeNet is used for both training guidance and evaluation. *Failure count* (lower is better) denotes the number of images classified as nudity out of 200 prompts (single predefined seed) or 4,000 images (20 random seeds per prompt).

Method	Failure (1 seed) ↓	Failure (20 seeds) ↓
SD 1.4 (Original)	137	2700
UCE	35	686
ESD	9	212
ACE	11	227
EAP	21	395
RECELER	7	251
Ours (ESD; NudeNet guidance)	**3** (-66.7%)	**93** (-56.1%)
Ours (ACE; NudeNet guidance)	**2** (-81.8%)	**18** (-92.1%)
Ours (EAP; NudeNet guidance)	**14** (-33.3%)	**276** (-30.1%)
Ours (RECELER; NudeNet guidance)	**2** (-71.4%)	**102** (-59.4%)

R HUMAN EVALUATION.

To further validate our findings beyond automatic classifiers, we conduct a human evaluation on nudity unlearning. We randomly select 100 prompts from the nudity subset and generate images using 20 random seeds per prompt, resulting in 2000 images per method. Five annotators (a mix of undergraduate and graduate students) were recruited for the study. Prior to annotation, participants were explicitly informed that the images may contain nudity and provided consent to proceed.

Each image was independently annotated with a binary label indicating whether nudity is present. Table 20 reports the total number of images labeled as nude (failure count) and the corresponding failure ratio. Consistent with the automatic evaluation, our method substantially reduces the presence

of nudity compared to all baselines. In particular, when applied on top of ACE and ESD, our approach achieves the lowest failure rates, indicating more reliable suppression of the target concept across diverse noise initializations. These results confirm that the improvements observed with automated detectors translate to human perception, strengthening the practical relevance of our unlearning framework.

Table 20: Human evaluation results for nudity unlearning. Human annotators provided binary labels (nude / non-nude) for 100 prompts evaluated over 20 random seeds, yielding 2,000 images per method and 20,000 images across the ten evaluated methods. We report the number of images classified as nude for each method (*Failure count*, lower is better) and the corresponding failure rate (*Failure ratio*, lower is better).

Method	Failure count ↓	Failure ratio ↓
SD 1.4 (Original)	1800	0.90
UCE	993	0.497
ESD	480	0.240
ESD+ (Ours)	55	0.0275
ACE	491	0.246
ACE+ (Ours)	24	0.012
EAP	739	0.370
EAP+ (Ours)	372	0.186
RECELER	660	0.330
RECELER+ (Ours)	435	0.218

S Additional Qualitative Results

We provide additional qualitative results to complement the quantitative evaluations in the main paper, highlighting the behavior of our noise-aware unlearning strategy under controlled and stochastic settings. In particular, we examine both (i) fixed prompt–seed conditions and (ii) variations across different unlearning methods and noise initializations.

Figure 7 presents results generated using the *same prompt and the same random seed* across methods. Under identical noise initialization, our approach produces images that remain visually consistent with the baseline generation—preserving scene layout, structure, and semantic content—while successfully removing the forbidden concept. This demonstrates that our method performs targeted unlearning without introducing unnecessary deviations or artifacts when the stochastic factors are held constant.

Figure 8 compares different unlearning methods under the *same prompt*. While baseline approaches often exhibit residual appearances of the target concept or noticeable degradation in image quality, our method more reliably suppresses the concept across stochastic samples, maintaining visual plausibility and prompt faithfulness. Together, these examples illustrate that explicitly accounting for noise-space structure leads to more robust and stable unlearning behavior under both deterministic and stochastic generation settings.

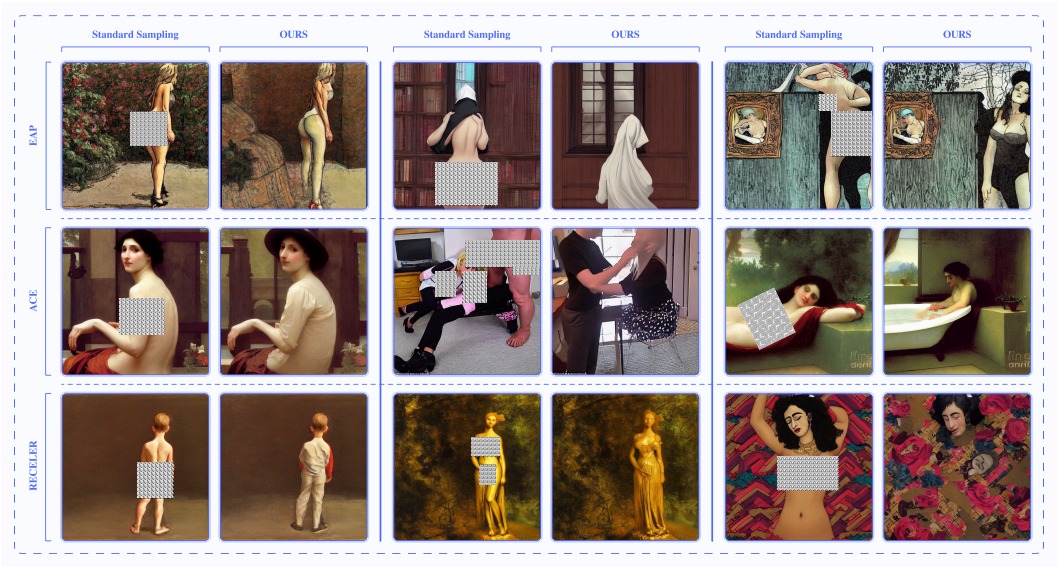

Figure 7: Qualitative comparison under the same prompt and identical noise seed. Our method generates images that are visually consistent with the baseline output while effectively removing the target concept, demonstrating precise and controlled unlearning without altering overall image structure.

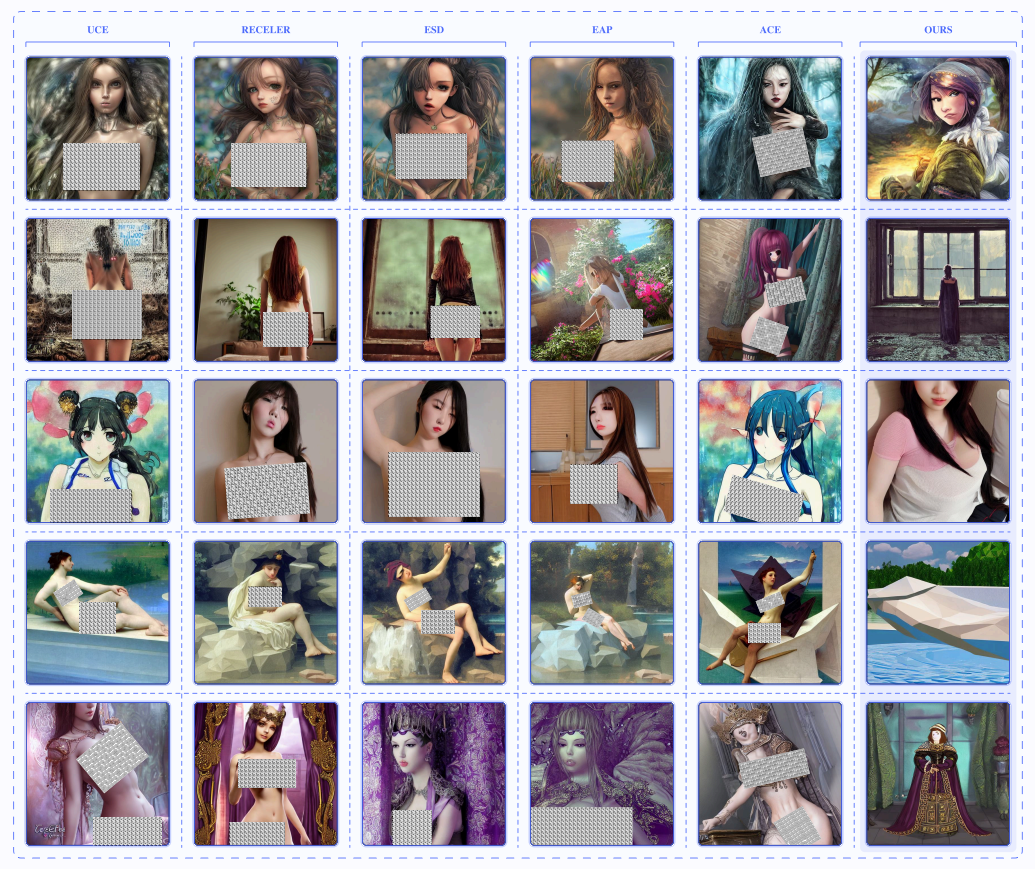

Figure 8: Qualitative comparison across different unlearning methods under the same prompt. Our approach more consistently suppresses the target concept across stochastic noise initializations, while preserving semantic content and visual quality compared to baseline methods.