TILDE: TILt-based Distributional Erasure for Concept Unlearning

Anonymous authors
Paper under double-blind review

Abstract

Concept unlearning in text-to-image diffusion and rectified-flow models is critical for safe and practical deployment: with rising privacy concerns, copyright disputes, trademark constraints, and safety regulations, deployed systems must reliably suppress unwanted concepts after training. Existing methods often remove the target concept effectively, but practical unlearning also requires an equally fundamental property: the unlearned model should retain quality, diversity, and semantic coverage on benign generation. The gold standard is a retain-only model trained from scratch without the unwanted data. However, common erasure objectives do not specify which post-unlearning distribution should approximate this reference, leaving retention as an implicit byproduct of heuristic updates. We propose TILDE, TILt-based Distributional Erasure, which formulates concept unlearning as a distributional alignment problem: the desired target is the minimum-deviation conditional distribution from the pretrained model under a forgetting constraint. This energy-tilted, anchor-free target downweights concept-expressing images while preserving benign relative mass for each prompt. We approximate this target with residual ∇-GFlowNet training for diffusion and rectified-flow generators, learning only a score correction relative to the pretrained sampler via a modular, differentiable concept energy. Across objects, artistic styles, characters and NSFW content, TILDE achieves strong forgetting while improving retention and distributional fidelity over prior baselines.

1 Introduction

Text-to-image generative models have become part of everyday creative and professional workflows, enabling high-fidelity image creation and editing across artistic styles, fictional characters, personal identities, products, and common scenes (Rombach et al., 2022; Esser et al., 2024). As these models become more widely used, the ability to edit their behavior after training is essential for legal compliance, privacy protection, and safety. Privacy regulations such as the GDPR recognize rights to erasure for personal data (European Parliament and Council of the European Union, 2016), while copyright, trademark, identity, and safety constraints can require suppressing specific concepts after training. Concept unlearning addresses this requirement by editing a pretrained text-to-image model so that it no longer generates images expressing an undesirable concept.

The central challenge in unlearning is that effective forgetting must not come at the cost of benign preservation: the edited model should maintain image quality and prompt alignment on benign prompts, without degrading related or general concepts. Yet, amid the push for higher benchmark scores, an urgent question arises: *are unlearned models truly forgetting, or merely satisfying superficial erasure proxies?* Standard evaluations often mistake degenerate failure modes such as mode collapse, visual distortion, or aggressive over-erasure of neighboring concepts for successful forgetting. The ideal reference is a retain-only model trained from scratch without the unwanted data, against which both forgetting and preservation can be assessed (Cho et al., 2025; Zhao et al., 2024). We study concept unlearning as an efficient approximation to this retain-only behavior.

Desiderata. This view yields three desiderata. **(D1) Effective forgetting:** the post-unlearning model should assign low probability to images expressing the target concept. **(D2) Local preservation:** benign prompts that are semantically close to the forgotten concept should remain close

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.

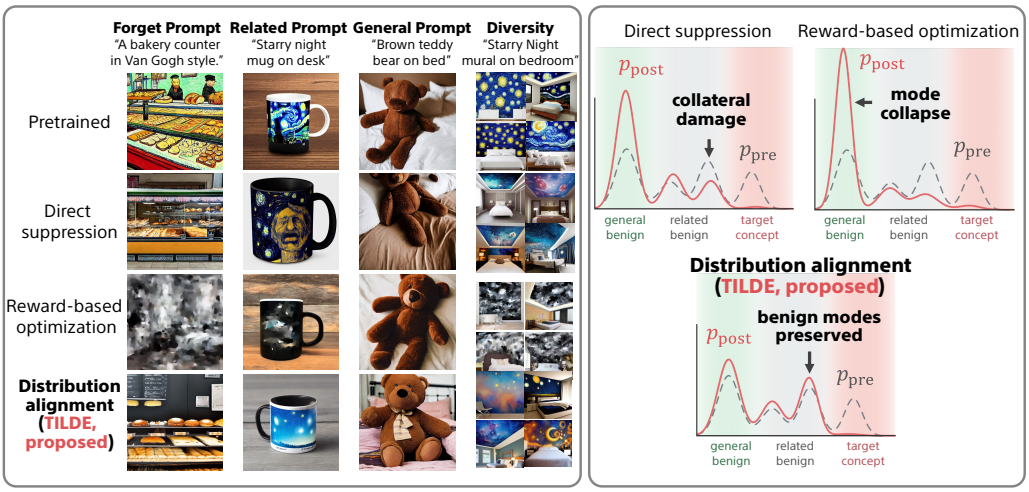

Figure 1: Concept unlearning failure modes and qualitative examples for Van Gogh style removal. Left: generations for forget, related, general, and diversity prompts under the pretrained model, direct suppression, reward-based optimization, and distribution alignment. Right: schematic post-unlearning behavior over output regions. Direct suppression can damage related benign modes, while reward-based optimization can collapse diversity onto a few safe modes. Distribution alignment suppresses the target concept while preserving benign modes.

to pretrained behavior, because this is where collateral damage is most likely. **(D3) Distributional fidelity:** outside the forget region, the post-unlearning distribution should stay close to the pretrained or retain-only reference at the per-prompt level, not just match coarse aggregate scores such as FID.

Existing methods rely on heuristic erasure proxies such as score suppression (Gandikota et al., 2023; Huang et al., 2024), anchor-based steering (Kumari et al., 2023; Gandikota et al., 2024; Bui et al., 2025), preference or reward-based optimization (Park et al., 2024), or trajectory-level objectives (Kusumba et al., 2025). While these objectives can suppress the target concept, they typically optimize ad-hoc losses without specifying what probability distribution the post-unlearning model ought to generate. Without specifying an explicit post-unlearning distribution, these ad-hoc objectives frequently sacrifice retention: strong erasure causes severe collateral damage to adjacent concepts (Amara et al., 2025), reward maximization collapses diversity onto safe modes (Uehara et al., 2024), and anchors induce prompt misalignment (George et al., 2025). Prior FADE evaluations show that methods scoring well on conventional metrics can drift *farther* from their retain-only reference in distribution than the original model (Cho et al., 2025). To uncover this, we take a broader view: evaluating retention across 60+ related and general concepts and measuring distributional fidelity via FADE. Figure 1 illustrates these common failure modes (collateral damage and mode collapse), contrasting them with our proposed minimum-deviation distributional target.

We argue that concept unlearning should be formulated by first defining the desired post-unlearning distribution, before choosing an optimization rule. Motivated by distributional evaluation (Cho et al., 2025), we formalize a minimal-deviation principle: among all conditional distributions that satisfy a forgetting constraint, select the one closest in KL divergence to the pretrained model. This casts unlearning as a constrained conditional distribution alignment problem, rather than heuristic score suppression, anchor replacement, or reward maximization. It also explains why retaining performance must be measured locally and globally: semantically nearby benign prompts are exactly where collateral damage is easiest to hide under coarse aggregate metrics.

We propose TILDE (TILt-based Distributional Erasure), which realizes this principle by tilting each prompt-conditioned pretrained distribution according to a forget energy. Under this target, concept-expressing outputs are down-weighted, while benign outputs retain their pretrained relative likelihoods up to normalization. Section 2 derives this closed-form, anchor-free Gibbs target. To train models toward this target distribution, we adopt GFlowNets (Bengio et al., 2023), which are designed to sample in proportion to unnormalized energy densities. Specifically, we use residual ∇-GFlowNet training (Liu et al., 2025): the pretrained sampler supplies the base distribution, and

the learner estimates only the score correction induced by the energy tilt. This formulation readily applies to both diffusion and rectified-flow architectures, and accommodates modular differentiable concept energies ranging from thresholded CLIP similarities to generative VLM token margins.

Contributions.

1. We formulate concept unlearning as constrained distributional alignment, making the post-unlearning distribution target explicit and deriving the corresponding anchor-free, prompt-conditioned minimal-deviation Gibbs form.
2. We introduce TILDE, which approximates the energy-tilted target via residual ∇-GFlowNet training, learning only a residual score correction relative to the frozen pretrained sampler.
3. We demonstrate that TILDE is modular across generative backbones (diffusion and rectified flow) and concept scorers (CLIP and VLMs), achieving effective erasure while significantly improving retention and distributional fidelity across objects, styles, and characters.

2 Target Distribution for Concept Unlearning

2.1 Prompt-Conditioned Target via Constrained KL Projection

We first make the desired post-unlearning conditional distribution explicit, instead of leaving it implicit in the choice of update rule. Let $\pi(y)$ denote the fixed prompt population, including prompts that may elicit the concept $\mathcal{C}$, related benign prompts, and general benign prompts. Let $E_{\mathcal{C}}(\mathbf{x}) \geq 0$ be a measurable forget energy that quantifies concept evidence: larger values indicate that image $\mathbf{x}$ more strongly expresses $\mathcal{C}$, whereas benign images, including semantically adjacent benign images, should have small or zero energy. We define the KL-projection target by

$$p^*_{\mathcal{C}} = \arg\min_{p} \mathbb{E}_{y\sim\pi}[\mathrm{KL}(p(\cdot \mid y) \,\|\, p_{\mathrm{pre}}(\cdot \mid y))] \quad \text{s.t.} \quad \mathbb{E}_{y\sim\pi,\, \mathbf{x}\sim p(\cdot|y)}[E_{\mathcal{C}}(\mathbf{x})] \leq \delta_{\mathcal{C}}, \tag{1}$$

Here, optimization is taken over conditional generators $p(\cdot \mid y)$ while the prompt distribution $\pi(y)$ is held fixed. The KL divergence penalizes unnecessary deviation from the pretrained base model, while the constraint bounds expected concept evidence to at most $\delta_{\mathcal{C}} > 0$. Under the mild regularity conditions of Proposition 1, this projection admits a unique closed-form solution:

$$p^*_{\mathcal{C}}(\mathbf{x} \mid y) = \frac{p_{\mathrm{pre}}(\mathbf{x} \mid y) \exp\big(-\beta E_{\mathcal{C}}(\mathbf{x})\big)}{Z_{\beta,\mathcal{C}}(y)}, \tag{2}$$

where $Z_{\beta,\mathcal{C}}(y)$ is the prompt-specific normalizing constant and $\beta \geq 0$ is the Lagrange multiplier associated with the forgetting constraint. The exponential factor downweights images with higher forget energy while preserving the pretrained relative probabilities of images assigned the same energy. In particular, for any benign region where $E_{\mathcal{C}}(\mathbf{x}) = 0$, the exponential reweighting factor is one. As a result, the target distribution leaves benign relative probabilities completely undistorted, preserving the pretrained model's internal diversity and mode structure.

The remaining design questions are how to instantiate $E_{\mathcal{C}}$ and how to approximate the resulting tilted distribution in a pretrained generator.

2.2 Differentiable Concept Energies and Surrogate Target

In practice, we require a differentiable energy function that quantifies concept evidence while providing informative gradients for generator training.

Measuring concept evidence via CLIP. In our primary experiments, we instantiate concept evidence using a frozen CLIP model. Let $\mathcal{Q}_{\mathcal{C}} = \{q_1, \ldots, q_M\}$ contain one or more short text descriptors of the target concept $\mathcal{C}$, such as "Pikachu" or "a photo of Pikachu". These descriptors identify the concept to suppress, whereas the generation prompt y specifies the image to generate. For an image $\mathbf{x}$, we define its CLIP concept score as the average cosine similarity to these descriptors:

$$s_{\mathcal{C}}(\mathbf{x}) = \frac{1}{|\mathcal{Q}_{\mathcal{C}}|} \sum_{q\in\mathcal{Q}_{\mathcal{C}}} \cos(\mathrm{CLIP}_{\mathrm{img}}(\mathbf{x}),\, \mathrm{CLIP}_{\mathrm{txt}}(q))\,.$$

Higher scores indicate stronger CLIP evidence of the target concept. The CLIP configuration and descriptor settings are given in Appendices E.1 and E.3.

From concept score to forget energy. We define the forget energy as

$$\widetilde{E}_{\mathcal{C}}(\mathbf{x}) = \begin{cases} 0, & \text{if } s_{\mathcal{C}}(\mathbf{x}) \le \tau, \\ \lambda_{\text{scale}} \left(\exp\big(\alpha(s_{\mathcal{C}}(\mathbf{x}) - \tau)\big) - 1 \right), & \text{if } s_{\mathcal{C}}(\mathbf{x}) > \tau. \end{cases} \tag{3}$$

Here τ is the CLIP-score threshold, $\alpha > 0$ controls the sharpness above the threshold, and $\lambda_{\text{scale}} > 0$ sets the overall penalty strength, including the Gibbs multiplier β.

The threshold is a critical design choice: without it, a continuous score penalty would exert non-zero gradients on benign generations. By contrast, Eq. (3) establishes a flat, zero-energy zone below τ while growing exponentially above it, ensuring continuity at the threshold ($s_{\mathcal{C}}(\mathbf{x}) = \tau \implies \widetilde{E}_{\mathcal{C}} = 0$) and sharp suppression for strong concept occurrences.

This yields the practical surrogate target

$$p_{\text{tar},\mathcal{C}}(\mathbf{x} \mid y) \propto p_{\text{pre}}(\mathbf{x} \mid y) \exp\big(-\widetilde{E}_{\mathcal{C}}(\mathbf{x})\big). \tag{4}$$

For outputs within the safe region ($s_{\mathcal{C}}(\mathbf{x}) \le \tau$), the reweighting factor is exactly one; above the threshold, it decreases as concept evidence increases. Thus the target preserves relative probabilities among images in the no-penalty region for each prompt while downweighting higher-scoring images. This target is anchor-free because it does not prescribe a replacement concept.

Modular extensions (VLMs and classifiers). The formulation is agnostic to the choice of scorer, provided it yields differentiable outputs with respect to image pixels. Beyond CLIP, we explore energies derived from frozen Vision-Language Models (e.g., next-token logit margins from Qwen2.5-VL) as well as task-specific classifiers. Appendix G details the VLM formulation and experiments, while Appendix F.1 describes classifier-based variants.

2.3 Implication for Optimization

The target to realize is the prompt-conditioned tilted density in Eq. (4), rather than to seek modes that maximize a scalar forgetting reward. Accordingly, the learning objective is to realize proportional sampling from this target, preserving benign-region coverage while down-weighting high-energy images. To this end, we adopt residual ∇-GFlowNet training (Liu et al., 2025). Because diffusion and rectified-flow models are parameterized via score functions (or velocity fields), the pretrained-times-energy structure of Eq. (4) allows the model to learn only the residual score correction induced by the energy tilt. The next section details this approximation.

3 Approximating the Target via Residual ∇-GFlowNet

We fix $\mathcal{C}$ throughout this section and use the practical thresholded energy $\widetilde{E}_{\mathcal{C}}$ from Section 2.2. We use residual ∇-GFlowNet training (Liu et al., 2025) to approximate the energy-tilted target in Eq. (4). Consider a latent diffusion model (Rombach et al., 2022). For a given prompt y, let $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ be the initial noise latent and let $\mathbf{z}_T \to \mathbf{z}_{T-1} \to \cdots \to \mathbf{z}_0$ denote the reverse denoising chain, with decoded image $\mathbf{x} = \mathcal{D}(\mathbf{z}_0)$. We denote by $P_F^{\theta}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)$ the adapted denoising transition (parameterized by LoRA weights θ) and by $P_F^{\text{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)$ the frozen pretrained transition. Following GFlowNet convention (Bengio et al., 2023), the subscript F denotes the forward generative (denoising) direction, distinct from the forward noising process in diffusion training (recalled in Appendix B.1).

Residual balance condition. Eq. (4) defines a clean-image tilt of the pretrained sampler. The remaining question is how this terminal reweighting should be propagated through the denoising step. We introduce a positive residual continuation weight $\widetilde{F}_t(\mathbf{z}_t, y)$. Intuitively, it is the benign mass still reachable from $\mathbf{z}_t$: large when continuations tend to end in low-energy images, and small when they tend to end in concept-expressing high-energy images. The adapted transition should therefore steer

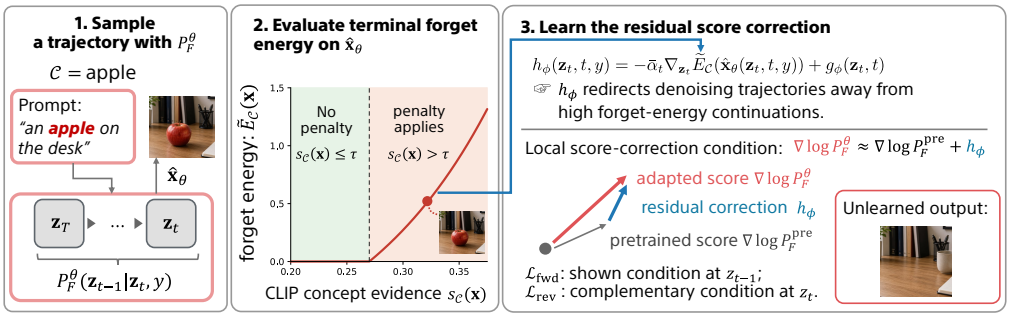

Figure 2: Overview of TILDE training in diffusion latent space. Given a target concept $\mathcal{C}$ and a sampled prompt y, the LoRA-adapted denoiser samples a trajectory $\mathbf{z}_T \to \cdots \to \mathbf{z}_0$, and evaluates the thresholded forget energy $\widetilde{E}_\mathcal{C}(\hat{\mathbf{x}})$ from CLIP concept evidence. Residual ∇-GFlowNet training penalizes the mismatch between the transition-score correction and the learned continuation-score estimate $h_\phi(\mathbf{z}_t, t, y)$, training the adapted sampler toward the energy-tilted target.

the pretrained transition toward states with higher continuation weight:

$$\frac{P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)}{P_F^{\text{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)} = \frac{\widetilde{F}_{t-1}(\mathbf{z}_{t-1}, y)}{\widetilde{F}_t(\mathbf{z}_t, y)}. \tag{5}$$

This condition is useful because it factorizes the target path-probability ratio locally. Along any trajectory $\mathbf{z}_T \to \cdots \to \mathbf{z}_0$, the intermediate $\widetilde{F}_t$ terms cancel, leaving only $\widetilde{F}_0(\mathbf{z}_0, y)/\widetilde{F}_T(\mathbf{z}_T, y)$. With $\widetilde{F}_0(\mathbf{z}_0, y) \propto \exp(-\widetilde{E}_\mathcal{C}(\mathcal{D}(\mathbf{z}_0)))$ and $\widetilde{F}_T(\mathbf{z}_T, y) = c_{\text{init}}(y)$ independent of $\mathbf{z}_T$, the terminal marginal is proportional to $p_{\text{pre}}(\mathbf{z}_0 \mid y) \exp(-\widetilde{E}_\mathcal{C}(\mathcal{D}(\mathbf{z}_0)))$, matching Eq. (4). Under these assumptions, the condition converts a terminal energy into proportional sampling: high-energy images lose mass, while images within the zero-energy region keep their pretrained relative proportions. Our fixed Gaussian prior, restricted transitions, and finite score-matching losses do not ensure these conditions; residual ∇-GFlowNet therefore approximates the target (Appendices C.2 and C.4).

Since $\widetilde{F}_t$ is unknown, we learn an approximation to its score

$$h_\phi(\mathbf{z}_t, t, y) \approx \nabla_{\mathbf{z}_t} \log \widetilde{F}_t(\mathbf{z}_t, y),$$

and use it to define the score-form training objective below. Training jointly optimizes the LoRA parameters and the auxiliary continuation-score parameters ϕ.

Score-form training objective. Directly enforcing Eq. (5) is inconvenient in diffusion models because it involves transition densities. Following residual ∇-GFlowNet (Liu et al., 2025), we use a score-matching surrogate based on local transition scores. Differentiating Eq. (5) with respect to $\mathbf{z}_{t-1}$ and approximating the unknown continuation score by h_ϕ gives the training condition:

$$\nabla_{\mathbf{z}_{t-1}} \log P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) - \nabla_{\mathbf{z}_{t-1}} \log P_F^{\text{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) \approx h_\phi(\mathbf{z}_{t-1}, t-1, y). \tag{6}$$

The left-hand side is the score correction introduced by finetuning, measuring the change in local probability flow relative to the pretrained denoiser. The right-hand side is the auxiliary estimate of the residual continuation score required by the same balance condition. We encourage this local consistency by penalizing the mismatch between the two vector fields, yielding the forward loss

$$\mathcal{L}_{\text{fwd}} = \mathbb{E}\Big[\big\|\nabla_{\mathbf{z}_{t-1}} \log P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) - \nabla_{\mathbf{z}_{t-1}} \log P_F^{\text{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) - h_\phi(\mathbf{z}_{t-1}, t-1, y)\big\|^2\Big]. \tag{7}$$

We similarly penalize deviations from the complementary score condition obtained by differentiating the residual balance relation with respect to $\mathbf{z}_t$, yielding the reverse loss $\mathcal{L}_{\text{rev}}$. Together, $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{rev}}$ encourage local consistency of the residual correction across denoising transitions. Unlike standard denoising-score training (Ho et al., 2020; Song et al., 2021), the target is not ground-truth noise or a data score, but the residual correction associated with the ideal energy-tilted sampler.

The residual field $h_\phi(\mathbf{z}_t, t, y)$ approximates the score of the continuation weight induced by the terminal forget energy. Since this score is not available in closed form at intermediate noisy states, we use the forward-looking residual parameterization

$$h_\phi(\mathbf{z}_t, t, y) = -\bar{\alpha}_t \nabla_{\mathbf{z}_t} \widetilde{E}_{\mathcal{C}}(\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, y)) + g_\phi(\mathbf{z}_t, t). \tag{8}$$

Here $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, y)$ is the one-step clean-image estimate predicted from the current noisy latent, and $\bar{\alpha}_t$ is the standard cumulative DDPM signal coefficient. The first term is the one-step pullback of the terminal-energy score, and g_ϕ parameterizes the remaining approximation residual of this continuation score.

Interpretation of h_ϕ. The residual field h_ϕ acts as an adaptive, state-dependent guide directing how the unlearned model should deviate from pretrained behavior. Instead of manually engineering opposing forget/retain prompts or prescribing surrogate anchors, TILDE lets the thresholded energy gradient dynamically guide this trajectory from the clean estimate $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, y)$. When the estimate exhibits high concept evidence, $\nabla_{\mathbf{z}_t} \widetilde{E}_{\mathcal{C}}(\hat{\mathbf{x}}_\theta)$ is active, producing an explicit score correction that steers the trajectory away from concept-expressing regions. Conversely, when the estimate falls below the threshold, the direct energy-pullback term is exactly zero, while the continuation term g_ϕ may remain active. Benign trajectories are therefore encouraged to remain close to pretrained behavior.

Overall loss and design rationale. We jointly optimize the LoRA update $\Delta\theta$ and the auxiliary continuation-score parameters ϕ by minimizing

$$\mathcal{L} = \mathcal{L}_{\text{fwd}} + \lambda_{\text{rev}} \mathcal{L}_{\text{rev}} + \lambda_{\text{term}} \mathcal{L}_{\text{term}}. \tag{9}$$

Here, λ_{rev} and λ_{term} are balancing weights, and $\mathcal{L}_{\text{term}}$ penalizes $\|g_\phi(\mathbf{z}_0, 0)\|_2^2$, encouraging the clean-end boundary condition to be determined solely by the forget energy. Closed-form Gaussian expressions for these transition scores are provided in Appendix C.

Our core contribution is formulating this unlearning-specific distributional target and approximating it in generative latent space, using residual ∇-GFlowNet as the optimization backbone. Each design choice in our pipeline systematically serves this target: **(1) Residual formulation:** because Eq. (4) is a multiplicative tilt over the base model, the learner estimates only the score correction relative to the pretrained sampler. **(2) Score-form objective:** matching vector fields directly aligns with the score and velocity parameterizations of diffusion and flow models. **(3) Forward-looking parameterization:** pulling the terminal concept signal back to intermediate noisy latents bridges the temporal credit-assignment gap across long reverse trajectories. **(4) Thresholded energy:** vanishing direct energy gradients in low-evidence regions reduce pressure to alter benign outputs. **(5) LoRA adapters:** restricting fine-tuning to low-rank adapters (Hu et al., 2022) imposes an orthogonal parameter-space minimal-update bias.

Extension to rectified-flow models. Rectified-flow models predict a velocity (Esser et al., 2024); the same residual formulation applies to generators like FLUX and SD3.5. Using the noise coordinate σ and interpolation convention $\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$, the one-step clean-latent estimate becomes

$$\widehat{\mathbf{z}}_0 = \mathbf{z}_\sigma - \sigma v_\theta(\mathbf{z}_\sigma, \sigma, y). \tag{10}$$

We decode this estimate to evaluate the same forget energy and replace $\bar{\alpha}_t$ in Eq. (8) with the heuristic signal weight $a(\sigma) = (1-\sigma)^2$. The residual balance relation and overall loss retain their form, using adapted and pretrained stochastic transitions with the same noise schedule and variance. These stochastic transitions supply the log-density scores required during training, while inference remains deterministic. The stochastic-chain guarantee does not directly establish exact sampling under deterministic inference. Full transition and score derivations appear in Appendix D.3.

4 Related Work

Current concept-unlearning methods. Existing approaches include direct suppression and selective finetuning (e.g., ESD, SalUn, and Selective Amnesia) (Gandikota et al., 2023; Fan et al., 2024; Heng & Soh, 2023), target-space editing frameworks such as UCE and MACE (Gandikota et al., 2024; Lu et al., 2024), and preference/reward-based optimization such as DUO, DDPO, and

DRaFT (Park et al., 2024; Black et al., 2024; Clark et al., 2024). These methods can remove target concepts effectively, but they often leave the final post-unlearning distribution implicit. While concept unlearning in rectified-flow architectures (such as FLUX) remains underexplored, recent efforts have explored flow-transformer erasure (EraseAnything; Gao et al., 2025a), trajectory alignment (EraseFlow; Kusumba et al., 2025), and sparse neuron masking (MCE; Zhang et al., 2025). We show that TILDE's distributional tilt natively extends to rectified-flow velocity fields.

Anchors, mapping, and dynamic targets. Many concept-erasure methods can be interpreted as choosing a destination representation and mapping the forget concept toward that target. Concept Ablation uses explicit anchor replacement (Kumari et al., 2023), while other methods use broader target mappings at the embedding or attention level (Gandikota et al., 2024; Lu et al., 2024). ACE argues that fixed generic anchors are often suboptimal and proposes adaptive/dynamic target selection based on concept context (Bui et al., 2025; George et al., 2026).

GFlowNet-based and distributional methods. Our work builds on ∇-GFlowNet (Liu et al., 2025), originally developed for reward-aligned sampling in diffusion models. EraseFlow (Kusumba et al., 2025) is the closest unlearning-specific GFlowNet baseline. It finetunes a pretrained diffusion model with LoRA using anchor trajectories and a trajectory-balance (TB) objective. Our approach uses residual ∇-DB to learn the score correction for an anchor-free energy tilt of the pretrained distribution.

Failures, robustness, and side effects. Recent evidence shows that apparent forgetting is often fragile: erased concepts can be revived or bypassed under adversarial prompting and model probing (Tsai et al., 2024; Zhang et al., 2024c), and relearning/revival remains a practical threat (Gao et al., 2025b; George et al., 2025). This has motivated robust unlearning methods that incorporate adversarial training and robustness-aware objectives (Zhang et al., 2024b; Kim et al., 2024; Srivatsan et al., 2025). Another persistent failure mode is the ripple effect on nearby concepts, where semantically adjacent benign concepts are unintentionally degraded after unlearning (Amara et al., 2025).

5 Experiments

5.1 Experimental Setup and Evaluation Metrics

Setup. Our primary experiments unlearn concepts from Stable Diffusion v1.5 (Rombach et al., 2022) and FLUX.1-dev by optimizing LoRA adapters (rank $r = 8$) via AdamW. The main benchmark comprises **12 unlearning targets**: five objects, four characters (celebrities and fictional entities), and three artistic styles (with nudity unlearning evaluated separately in the appendix). To detect localized collateral damage that narrow evaluations easily overlook, we evaluate retention across an expansive set of **60+ concepts** in total (spanning unlearning targets, semantically adjacent, and unrelated concepts) GPT-4o-mini generates 30 diverse training prompts $\mathcal{Y}_C$ per unlearning target and 20 held-out prompts per evaluation concept. We generate 8 images per evaluation prompt ($\sim$10,000 images per model checkpoint). More details in Appendices E.3.

Metrics. We use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) for zero-shot binary classification, querying generated images with prompts such as "Does this image contain $\langle$concept$\rangle$?" We report: (1) **Unlearning Accuracy** ($U_{\text{acc}} \uparrow$), measuring target concept suppression on forget prompts; (2) **Related Retention** ($RR_{\text{acc}} \uparrow$), quantifying collateral damage on semantically adjacent concepts; and (3) **General Retention** ($GR_{\text{acc}} \uparrow$), assessing preservation on unrelated prompts. To assess distributional fidelity beyond classifier-based scores, we report **FID** (Heusel et al., 2017) on retain prompts and **FADE** (Cho et al., 2025), the symmetric KL divergence to a retain-only finetuned reference: $\text{FADE}(y) := D_{\text{KL}}(p_r^y \,\|\, p_u^y) + D_{\text{KL}}(p_u^y \,\|\, p_r^y)$, estimated via weighted denoising-loss differences (Appendix E.7).

5.2 Experimental Results and Analysis

Q1: How well do existing methods retain overall model performance during unlearning? Table 1 illustrates how conventional accuracy metrics can mask severe distributional side-effects. Aggressive erasure methods like MACE (Lu et al., 2024) achieve near-complete forgetting but severely degrade related and general retention. Anchor-based approaches (CA, UCE) preserve

Table 1: Category averages. SD v1.5 base: accuracy 89%, CLIP score 32.6. **Objects**: apple, banana, golf ball, cat, dog; **characters**: Brad Pitt, Lionel Messi, Mickey Mouse, Pikachu; **styles**: cartoon, Van Gogh, Monet. FID averages FID of all 12 concept unlearnings. FADE averages Van Gigh, Picasso and Monet styles. Per-concept results: Appendix I.

Method	Objects			Characters			Style			Avg. FADE ↓	Avg. FID ↓
	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑		
ESD-u	0.74	0.58	0.79	0.97	0.36	0.77	0.55	0.81	0.84	254	10.0
ESD-x	0.77	0.58	0.78	0.98	0.23	0.55	0.70	0.71	0.82	200	9.5
UCE	0.62	0.87	0.83	0.92	0.57	0.67	0.29	0.93	0.87	138	14.8
CA	0.61	0.91	0.90	0.97	0.59	0.87	0.86	0.72	0.87	152	7.3
MACE	0.44	0.79	0.83	0.99	0.17	0.22	1.00	0.03	0.01	67370	181.1
MCE	0.42	0.81	0.83	0.71	0.66	0.83	0.59	0.72	0.83	1598	15.0
DUO	0.91	0.76	0.88	0.97	0.46	0.86	0.56	0.49	0.87	358	9.0
EraseFlow	0.82	0.53	0.72	0.99	0.16	0.57	0.86	0.52	0.72	400	17.6
Meta	0.95	0.36	0.72	0.99	0.21	0.71	0.69	0.27	0.56	40435	21.9
SHS	0.84	0.30	0.35	0.85	0.34	0.32	0.80	0.17	0.30	150	171.3
EDiff	0.39	0.83	0.88	0.52	0.78	0.88	0.32	0.85	0.88	287	6.8
SPEED	0.41	0.89	0.89	0.80	0.80	0.89	0.25	0.93	0.88	135	3.2
ScaPre	0.92	0.60	0.82	0.98	0.44	0.80	0.52	0.78	0.84	149	11.2
SMem	0.51	0.85	0.88	0.90	0.62	0.87	0.23	0.86	0.89	126	3.5
TILDE	0.97	0.72	0.86	0.99	0.64	0.87	0.88	0.76	0.84	130	9.6

Table 2: Category averages for concept unlearning on **FLUX.1-dev**. More details reported in Table 21.

Method	Objects			Characters			Style		
	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑	U_{Acc} ↑	RR_{Acc} ↑	GR_{Acc} ↑
EraseAnything	0.10	0.87	0.93	0.62	0.54	0.94	0.59	0.44	0.93
EraseFlow	0.99	0.41	0.90	0.97	0.27	0.84	0.61	0.41	0.93
MCE	0.21	0.81	0.81	0.51	0.62	0.82	0.62	0.55	0.82
TILDE	0.97	0.85	0.93	0.98	0.68	0.94	0.84	0.78	0.93

benign quality through explicit retain supervision, but leave substantial concept leakage. Meanwhile, recent methods like SPEED, ScaPre, and MCE heavily constrain parameter drift to preserve benign generation. However, this parameter restriction severely compromises unlearning particularly for styles and objects, showing that high retention can simply be an artifact of incomplete erasure.

Q2: How well does TILDE perform? TILDE operates on the Pareto frontier: it achieves near-complete forgetting across all categories while maintaining strong related and general retention, without requiring explicit retain sets or surrogate anchors. Its low FADE score (130) indicates closer agreement with the retain-only finetuned reference than most baselines. With retain prompts optionally added, retention improves further(Q5).

Q3: Why does the GFlowNet formulation help? We ablate each ingredient of the residual ∇-GFlowNet objective on the Pikachu concept (Table 3); recall that (Liu et al., 2025) replaces intractable absolute log-flow targets with a gradient-form (score-matching) loss well-conditioned in latent space, and that the residual g_ϕ restricts learning to the score *correction* relative to the pre-trained model.

Table 3: Pikachu objective ablation with shared thresholded energy and LoRA parameterization.

Method	U_{acc} ↑	RR_{acc} ↑	GR_{acc} ↑
TILDE (resi ∇-GFlowNet)	0.99	0.76	0.88
w/o residual	0.96	0.79	0.87
w/o ∇ form	0.40	0.81	0.87
w/o g_ϕ	0.97	0.72	0.83
DDPO (reward max.)	0.98	0.61	0.33

The ∇ form is the primary driver of forgetting: dropping to 0th-order DB collapses unlearning to 0.40, since absolute-flow targets are too poorly conditioned in latent space to suppress the concept. The residual g_ϕ is the primary driver of retention: removing it forces the full score field to absorb the forget update, dropping general retention to 0.83. DDPO-style reward maximization keeps forgetting strong

but reduces general retention to 0.33, supporting the retention benefit of our residual score-matching objective over reward maximization.

Q4: How does the energy threshold τ affect unlearning? Figure 3 traces the accuracy and retention trade-off across threshold values τ (Eq. (3)) on the Pikachu concept. Removing the threshold entirely (not shown) causes catastrophic retention collapse: U_{acc}=1.00 but RR_{acc}=0.08, GR_{acc}=0.49, since every sample receives an energy penalty regardless of concept relevance. As τ rises from a low value, forgetting stays strong while retention steadily improves, reflecting the growing no-tilt region that shields benign samples from direct terminal-energy gradients. Our chosen τ=0.22 strikes the point where forgetting remains near-complete and retention is high; pushing τ higher causes forgetting to degrade sharply while retention gains diminish. The reason: below τ, $\widetilde{E}_{\mathcal{C}}(\mathbf{x}) = 0$ exactly, so the explicit terminal-energy force vanishes for those clean predictions, leaving only indirect effects through residual-flow propagation and shared LoRA parameters.

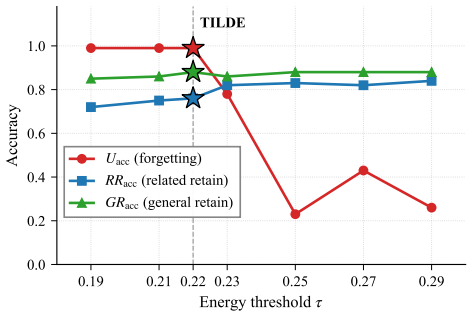

Figure 3: Pikachu threshold ablation; stars mark τ=0.22. Without thresholding (off-axis), $(U, RR, GR) = (1.00, 0.08, 0.49)$.

Q5: Does including retain prompts during unlearning help further? Since benign images below τ receive no direct terminal-energy penalty, general or related-retain prompts can be naturally mixed into the training batch as a regularizer toward pretrained benign behavior (Table 4). Both variants improve related-retain accuracy without measurably affecting unlearning, consistent with the thresholded-energy prediction: below-threshold samples receive no direct terminal-energy gradient. Related-retain prompts give a slightly larger gain than general ones, since they probe the exact neighborhood where collateral damage leaks in through shared LoRA parameters. This benefit is a byproduct of thresholding: unlike CA and UCE, our pipeline absorbs retain prompts as ordinary training samples with no extra loss term, so the Pareto improvement of Q2 extends rather than trades off when retain supervision is available.

Table 4: Effect of retain prompts on Pikachu unlearning and retention.

Setting	$U_{\text{acc}} \uparrow$	$RR_{\text{acc}} \uparrow$	$GR_{\text{acc}} \uparrow$
TILDE	0.99	0.76	0.88
TILDE + General retain	0.98	0.82	0.88
TILDE + Related retain	0.99	0.84	0.88

Q6: Does TILDE support different generators and concept scorers? Yes. Experiments confirm that TILDE is modular across both generative backbones and concept scorers: **Generators:** Beyond latent diffusion (SD v1.4, v1.5, v2), our residual formulation extends natively to rectified-flow transformers, including SD3 and FLUX.1-dev (Table 8). **Scorers:** A frozen Qwen VLM supplies a differentiable alternative to CLIP; compatible classifiers fit the same interface (Appendices G and F.1). At near-complete Pikachu erasure, the CLIP configuration retains more benign content, consistent with its flat zero-energy region.

6 Discussion

This work casts concept unlearning as distributional alignment: TILDE derives a minimum-deviation, energy-tilted target under a forgetting constraint and approximates it through residual ∇-GFlowNet training. The diffusion and rectified-flow realizations support replaceable differentiable concept energies. Experiments show strong forgetting with competitive retention.

Limitations and future work. The method requires a differentiable concept scorer, and its performance depends on how well that scorer distinguishes the target from related benign content. The energy threshold and scale would require calibration for new concepts or base models. Future work will explore broader concept coverage, robustness and continual concept unlearning while preserving generation quality.

AI USE STATEMENT

We used AI tools to assist with writing and polishing the manuscript and debugging research code. As part of our experimental methodology, we queried GPT-4o-mini through API calls to generate synthetic training and evaluation prompts and identify related concepts, as described in Section 5.1 and Appendices E.3 and E.7. The prompt-generation system instructions are provided in Appendix E.3. The authors take responsibility for the final manuscript, code, and experimental results, including material developed with AI assistance.

ETHICS STATEMENT

This work aims to improve the safety of text-to-image models by suppressing unwanted concepts while preserving benign generation. So our evaluation includes nudity unlearning and therefore involves NSFW prompts and generated images. This sensitive content is used to evaluate concept suppression and the behavior of the models under safety-relevant conditions. All experiments use publicly available pretrained models. These models can nevertheless produce harmful content, and the reported unlearning results do not constitute a guarantee of safe generation under all prompts.

REPRODUCIBILITY STATEMENT

To support reproduction, Appendix E documents the experimental setup, including the system instructions for generating concept-specific training prompts (Appendix E.3), optimization settings (Appendix E.4), and unlearning thresholds and energy hyperparameters for the primary configuration (Appendix E.6). Algorithm 1, together with these implementation details, specifies the training procedure for the Stable Diffusion v1.5 results in the main paper. Our evaluation follows a setup similar to LACU (George et al., 2026); metric definitions and evaluation protocols are provided in Appendix E.7. We use publicly released baseline implementations, train the baselines with the same unlearning prompts used for our method, and follow their documented procedures as closely as possible. FID is computed using the UnlearnCanvas evaluation code (Zhang et al., 2024a), while FADE uses the authors' released code (Cho et al., 2025) together with the UnlearnCanvas-based reference construction described in Appendix E.7. All base models are publicly available. Experiments with alternative concept scorers and generative backbones are reported in Appendix G and Appendix H, respectively. Appendices C and D provide the theoretical assumptions, proofs, derivations, and references to the underlying results.

REFERENCES

Ibtihel Amara, Ahmed Imtiaz Humayun, Ivana Kajic, Zarana Parekh, Natalie Harris, Sarah Young, Chirag Nagpal, Najoung Kim, Junfeng He, Cristina Nader Vasconcelos, Deepak Ramachandran, Golnoosh Farnadi, Katherine Heller, Mohammad Havaei, and Negar Rostamzadeh. Erasing more than intended? how concept erasure degrades the generation of non-target concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16420–16430, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Anh Tuan Bui, Thuy-Trang Vu, Long Tung Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tZdqL5FH7w.

Sungjun Cho, Dasol Hwang, Frederic Sala, Sangheum Hwang, Kyunghyun Cho, and Sungmin Cha. Reference-specific unlearning metrics can hide the truth: A reality check. *arXiv preprint arXiv:2510.12981*, 2025.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016. Official Journal of the European Union, 2016. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679. General Data Protection Regulation, Article 17: Right to erasure.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Rohit Gandikota, Joanna Materzyńska, 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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18470–18494. PMLR, 2025a.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2131–2141, 2025b.

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

Naveen George, Naoki Murata, Yuhta Takida, Konda Reddy Mopuri, and Yuki Mitsufuji. Locality-aware continual unlearning for diffusion models. In *European Conference on Computer Vision*, pp. 389–407. Springer, 2026.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Changhoon Kim, Kyle Min, and Yezhou Yang. R.A.C.E.: Robust adversarial concept erasure for secure text-to-image diffusion model. In *Computer Vision – ECCV 2024*, volume 15141 of *Lecture Notes in Computer Science*, pp. 461–478. Springer Nature Switzerland, 2024.

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

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. EraseFlow: Learning concept erasure policies via GFlowNet-driven alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70876–70911. Curran Associates, Inc., 2025. URL `https://openreview.net/forum?id=igB289kbej`.

Zhen Liu, Tim Z. Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed GFlowNets. In *International Conference on Learning Representations (ICLR)*, 2025.

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.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 80244–80267. Curran Associates, Inc., 2024.

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

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.

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

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M. Patel, and Karthik Nandakumar. STEREO: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23765–23774, 2025.

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 *12th International Conference on Learning Representations*, 2024.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M. Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

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.

Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75192–75215. PMLR, 2025.

Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Kompella, Xiaoming Liu, and Sijia Liu. UnlearnCanvas: A stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *Advances in Neural Information Processing Systems*, 2024a.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36748–36776. Curran Associates, Inc., 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.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12293–12333, 2024. URL `https://openreview.net/forum?id=QAbhLBF72K`.

Appendix

Table of Contents

A. Algorithm Pseudocode ... **15**

B. Notation and Background ... **16**

B.1 Diffusion Models ... 16

B.2 GFlowNets and Detailed Balance ... 16

C. Theoretical Analysis and Proofs ... **16**

C.1 From Detailed Balance to Residual ∇-DB Losses ... 16

C.2 Distributional Correctness ... 18

C.3 Adapting ∇-DB for Forget Energies ... 19

C.4 Optimization Caveat ... 19

D. Extended Derivations ... **19**

D.1 Closed-Form Conditional Gibbs Projection ... 19

D.2 Continuation Approximation and Endpoint Conventions ... 20

D.3 Rectified-Flow Transitions and Scores ... 21

E. Additional Experimental Details ... **23**

E.1 CLIP Model ... 23

E.2 Base Diffusion Model and Adapters ... 23

E.3 Prompt Sets and Generation Strategy ... 23

E.4 Optimization and Schedule ... 24

E.5 Compute ... 24

E.6 Hyperparameter Sensitivity ... 24

E.7 Evaluation Protocol ... 25

F. Nudity Unlearning with a Classifier-Based Concept Energy ... **27**

F.1 Classifier-Based Concept Energies ... 27

F.2 Nudity Unlearning Setup ... 27

G. Unlearning with a Vision–Language Model as the Concept-Energy Scorer ... **29**

G.1 From a Yes/No Question to a Forgetting Energy ... 29

G.2 Why We Use EOT-Softplus ... 29

G.3 Calibration and Differentiable Implementation ... 30

G.4 Pikachu Results and Comparison with CLIP ... 30

G.5 Reporting Across Concepts and Categories ... 31

H. Performance Across Diffusion and Rectified-Flow Architectures ... **33**

I. Per-Concept Quantitative Results ... **34**

A Algorithm Pseudocode

Algorithm 1 uses Gaussian transitions $P_F^\theta = \mathcal{N}(\mu_\theta, s_t^2\mathbf{I})$ and $P_F^{\mathrm{pre}} = \mathcal{N}(\mu^{\mathrm{pre}}, s_t^2\mathbf{I})$, with shared positive variance independent of both latent states. Diffusion uses $a_t = \bar{\alpha}_t$; SD3 uses Appendix D.3. The operator sg detaches gradient targets, and partial transition derivatives hold the other endpoint fixed. The calibrated energy is Eq. (3) or Eq. (47), with the threshold subgradient convention in Appendix C.3. Zero-variance transitions require a separate endpoint rule (Appendix D.2).

Algorithm 1 Forget-Energy-Tilted ∇-GFlowNet Training

Require: Frozen generator θ_0; transition P_F and clean predictor $\hat{\mathbf{x}}_\theta$; calibrated energy $\widetilde{E}_{\mathcal{C}}$; forget prompts $\mathcal{Y}_{\mathcal{C}}$; weights $a_t, \lambda_{\mathrm{rev}}, \lambda_{\mathrm{term}}$

Setup. Initialize rank-8 LoRA $\Delta\theta$ on attention projections (U-Net (W_Q, W_K, W_V, W_O) for diffusion).
Initialize g_ϕ with the architecture-specific configuration (diffusion channels $\{64, 128, 256, 256\}$).

Set adapted parameters $\theta \leftarrow \theta_0 + \Delta\theta$; freeze the decoder and concept scorer.

for epoch $= 1, \ldots, N_{\text{epochs}}$ **do**
  for batch $= 1, \ldots, N_{\text{batches}}$ **do**
    Sample forget prompts $\{y_i\}$ from $\mathcal{Y}_{\mathcal{C}}$

    Phase 1: sample prompt-conditioned trajectories.
    Sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    for $t = T, T-1, \ldots, 1$ **do**
      Compute the adapted noise prediction (diffusion) or velocity prediction (rectified flow).
      $\mathbf{z}_{t-1} \sim P_F^\theta(\cdot \mid \mathbf{z}_t, y_i)$ via the chosen stochastic transition
    end for
    Decode terminal latent: $\hat{\mathbf{x}} \leftarrow \mathcal{D}(\mathbf{z}_0)$

    Phase 2: evaluate the terminal forget energy.
    Evaluate $\widetilde{E}_{\mathcal{C}}(\hat{\mathbf{x}})$ using the chosen scorer and calibrated energy map.

    Phase 3: match residual ∇-DB scores.
    Sample a timestep subset $\mathcal{T}$ with $s_t^2 > 0$ using the configured selection rule.
    for $t \in \mathcal{T}$ **do**
      $\mathbf{s}_{\mathrm{fwd}} \leftarrow \big(\mu_\theta(\mathbf{z}_t, t, y_i) - \mu^{\mathrm{pre}}(\mathbf{z}_t, t, y_i)\big)/s_t^2$
      $\mathbf{s}_{\mathrm{rev}} \leftarrow \nabla_{\mathbf{z}_t} \log P_F^{\mathrm{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y_i) - \nabla_{\mathbf{z}_t} \log P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y_i)$
      $\mathbf{h}_{t-1} \leftarrow -a_{t-1}\,\mathrm{sg}[\nabla_{\mathbf{z}_{t-1}} \widetilde{E}_{\mathcal{C}}(\hat{\mathbf{x}}_\theta(\mathbf{z}_{t-1}, t-1, y_i))] + g_\phi(\mathbf{z}_{t-1}, t-1)$
      $\mathbf{h}_t \leftarrow -a_t\,\mathrm{sg}[\nabla_{\mathbf{z}_t} \widetilde{E}_{\mathcal{C}}(\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, y_i))] + g_\phi(\mathbf{z}_t, t)$
      Accumulate $\mathcal{L}_{\mathrm{fwd}}$ from $\|\mathbf{s}_{\mathrm{fwd}} - \mathbf{h}_{t-1}\|^2$ and $\mathcal{L}_{\mathrm{rev}}$ from $\|\mathbf{s}_{\mathrm{rev}} - \mathbf{h}_t\|^2$
    end for

    Phase 4: update parameters.
    $\mathcal{L}_{\mathrm{term}} \leftarrow \|g_\phi(\mathbf{z}_0, 0)\|^2$ {data-end boundary}
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{fwd}} + \lambda_{\mathrm{rev}}\mathcal{L}_{\mathrm{rev}} + \lambda_{\mathrm{term}}\mathcal{L}_{\mathrm{term}}$
    Update $\Delta\theta$ and ϕ with AdamW

  end for
end for
return LoRA-adapted generator $\theta_0 + \Delta\theta$; discard g_ϕ and the scorer.

B Notation and Background

We collect the diffusion notation and detailed-balance identity used below; Appendix C.1 gives the full residual ∇-GFlowNet derivation.

B.1 Diffusion Models

For a clean latent $\mathbf{z}_0$, define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ under a DDPM variance schedule $0 < \beta_t < 1$ (Ho et al., 2020; Song et al., 2021). The noising marginal and prompt-conditioned denoising transition are

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}\big)\,, \tag{11}$$

$$P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_t, t, y), \sigma_t^2\mathbf{I}\big)\,, \tag{12}$$

$$\mu_\theta(\mathbf{z}_t, t, y) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, y)\right). \tag{13}$$

The subscript F denotes the GFlowNet sampling direction, which is denoising. Equation (13) uses adjacent DDPM steps; other samplers require their actual transition means and variances. The residual losses use exact *conditional transition* scores with shared positive, state-independent variance, rather than assuming the learned noise predictor is the exact score of a sampler marginal.

In latent diffusion (Rombach et al., 2022), a frozen decoder $\mathcal{D}$ maps the terminal latent to an image. For $\bar{\alpha}_t > 0$, the intermediate clean-image estimate is

$$\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t, y) = \mathcal{D}\left(\frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, y)}{\sqrt{\bar{\alpha}_t}}\right). \tag{14}$$

This estimate supplies the image at which the concept energy is evaluated and differentiated in Eq. (8). The terminal Gaussian prior and clean-end conventions are distinguished from exact boundary identities in Appendices C.2 and D.2.

B.2 GFlowNets and Detailed Balance

GFlowNets learn a sampling policy whose terminal distribution is proportional to a nonnegative reward R, with finite nonzero normalizer $Z = \int R(\mathbf{x})\, d\mathbf{x}$ (Bengio et al., 2023). For a trajectory $s_0 \to \cdots \to s_T$, forward and backward policies P_F, P_B and a positive state flow F satisfy detailed balance when

$$P_F(s_{t+1} \mid s_t)F(s_t) = P_B(s_t \mid s_{t+1})F(s_{t+1}), \qquad F(s_T) = R(s_T). \tag{15}$$

With normalized backward policies and an initial distribution matching its normalized flow, this relation telescopes to the desired terminal distribution. In diffusion, a continuum of initial Gaussian noise states makes this compatibility condition substantive; the learned sampler's backward kernel also need not equal the DDPM noising kernel. Appendix C.1 derives the residual score equations of Liu et al. (2025) in our denoising notation, and Appendix C.2 states the resulting conditional sampling guarantee.

C Theoretical Analysis and Proofs

C.1 From Detailed Balance to the Trainable Residual ∇-DB Losses

We follow the residual ∇-DB derivation of Liu et al. (2025, Section 3.2, Eqs. (11)–(19)), with their scaled log-reward replaced by $-\widetilde{E}_{\mathcal{C}}$. Their sampling clock runs from x_0 to x_T; ours runs from $\mathbf{z}_T$ to $\mathbf{z}_0$. We give the derivation explicitly in our notation. We assume the positive transition densities and state flows are continuously differentiable on the domain where the score identities are used. Each transition-score derivative below holds the other endpoint fixed; at a threshold kink, the specified training subgradient does not assert classical differentiability.

For a fixed prompt y, let $P_B(\mathbf{z}_t|\mathbf{z}_{t-1}, y)$ denote a common backward kernel, while P_F denotes the denoising transition in the GFlowNet sampling direction. We suppress the prompt argument of P_B below. Assume, for this derivation, that the policies and positive state flows satisfy

$$P_F(\mathbf{z}_{t-1}|\mathbf{z}_t, y)\, F(\mathbf{z}_t, y) = P_B(\mathbf{z}_t|\mathbf{z}_{t-1})\, F(\mathbf{z}_{t-1}, y). \tag{16}$$

Exact time reversal of a DDPM corruption process satisfies this identity with its corruption marginals. A learned Gaussian denoiser generally does not satisfy it with the fixed DDPM noising kernel. The actual pretrained Markov chain instead has its own Bayes reverse kernel, which gives exact DB with its actual marginals. The terminal path tilt described in Appendix C.2 shares that reverse kernel. Taking logarithms gives

$$\log P_F(\mathbf{z}_{t-1}|\mathbf{z}_t, y) + \log F(\mathbf{z}_t, y) = \log P_B(\mathbf{z}_t|\mathbf{z}_{t-1}) + \log F(\mathbf{z}_{t-1}, y). \tag{17}$$

Differentiating with respect to the next denoised state $\mathbf{z}_{t-1}$ yields the forward ∇-DB equation

$$\nabla_{\mathbf{z}_{t-1}} \log P_F(\mathbf{z}_{t-1}|\mathbf{z}_t, y) = \nabla_{\mathbf{z}_{t-1}} \log P_B(\mathbf{z}_t|\mathbf{z}_{t-1}) + \nabla_{\mathbf{z}_{t-1}} \log F(\mathbf{z}_{t-1}, y), \tag{18}$$

because $F(\mathbf{z}_t, y)$ does not depend on $\mathbf{z}_{t-1}$. Differentiating instead with respect to the conditioning latent $\mathbf{z}_t$ yields the reverse equation

$$\nabla_{\mathbf{z}_t} \log P_B(\mathbf{z}_t|\mathbf{z}_{t-1}) - \nabla_{\mathbf{z}_t} \log P_F(\mathbf{z}_{t-1}|\mathbf{z}_t, y) = \nabla_{\mathbf{z}_t} \log F(\mathbf{z}_t, y), \tag{19}$$

because $F(\mathbf{z}_{t-1}, y)$ does not depend on $\mathbf{z}_t$.

Now let P_F^{pre} be the pretrained denoising policy and P_F^θ the finetuned one, both conditioned on the same prompt y. Subtracting the pretrained and finetuned forward equations cancels the shared backward policy P_B and gives

$$\nabla_{\mathbf{z}_{t-1}} \log P_F^\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - \nabla_{\mathbf{z}_{t-1}} \log P_F^{\mathrm{pre}}(\mathbf{z}_{t-1}|\mathbf{z}_t, y) = \nabla_{\mathbf{z}_{t-1}} \log \tilde{F}(\mathbf{z}_{t-1}, y), \tag{20}$$

with $\tilde{F} = F/F^{\mathrm{pre}}$. Subtracting the reverse equations gives

$$\nabla_{\mathbf{z}_t} \log P_F^{\mathrm{pre}}(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - \nabla_{\mathbf{z}_t} \log P_F^\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, y) = \nabla_{\mathbf{z}_t} \log \tilde{F}(\mathbf{z}_t, y). \tag{21}$$

These two equations define the ideal targets approximated by $\mathcal{L}_{\text{fwd}}$ and $\mathcal{L}_{\text{rev}}$; the same equations follow by differentiating the residual ratio directly, without the DDPM noising assumption.

With the forward-looking parameterization from Eq. (8), the same residual flow network is evaluated at the state appearing on the right-hand side:

- forward direction: evaluate $\nabla \log \tilde{F}$ at $(\mathbf{z}_{t-1}, y)$;
- reverse direction: evaluate $\nabla \log \tilde{F}$ at $(\mathbf{z}_t, y)$.

Replacing the ideal continuation scores by h_ϕ, the practical losses are

$$\mathcal{L}_{\text{fwd}} = \mathbb{E}\Big[\big\|\nabla_{\mathbf{z}_{t-1}} \log P_F^\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - \nabla_{\mathbf{z}_{t-1}} \log P_F^{\mathrm{pre}}(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - h_\phi(\mathbf{z}_{t-1}, t-1, y)\big\|^2\Big], \tag{22}$$

$$\mathcal{L}_{\text{rev}} = \mathbb{E}\Big[\big\|\nabla_{\mathbf{z}_t} \log P_F^{\mathrm{pre}}(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - \nabla_{\mathbf{z}_t} \log P_F^\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, y) - h_\phi(\mathbf{z}_t, t, y)\big\|^2\Big]. \tag{23}$$

The expectations use sampled prompts, transitions, and timesteps. The terminal penalty $\mathcal{L}_{\text{term}} = \mathbb{E}_{\mathbf{z}_0, y}[\|g_\phi(\mathbf{z}_0, 0)\|^2]$ encourages the data-end boundary condition under the exact clean-end convention stated in the main text.

Gaussian closed forms and JVP computation. For $P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) = \mathcal{N}(\mu_\theta(\mathbf{z}_t, t, y), \sigma_t^2 \mathbf{I})$ and its pretrained counterpart, assume differentiable means and a shared variance $\sigma_t^2 > 0$ independent of both latent states. With $\tilde{P}_F^\theta = P_F^\theta / P_F^{\mathrm{pre}}$, the forward residual and reverse conditioning scores are

$$\nabla_{\mathbf{z}_{t-1}} \log \tilde{P}_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) = \frac{\mu_\theta(\mathbf{z}_t, t, y) - \mu^{\mathrm{pre}}(\mathbf{z}_t, t, y)}{\sigma_t^2}, \tag{24}$$

$$\nabla_{\mathbf{z}_t} \log P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y) = \frac{1}{\sigma_t^2} J_{\mu_\theta}(\mathbf{z}_t, t, y)^\top \big(\mathbf{z}_{t-1} - \mu_\theta(\mathbf{z}_t, t, y)\big). \tag{25}$$

Replacing μ_θ by μ^{pre} gives the pretrained reverse score. Autodiff evaluates these Jacobian products without forming dense Jacobians, holding $\mathbf{z}_{t-1}$ fixed; differentiating through its reparameterized sampling path computes a different quantity. State-dependent covariance requires additional covariance and log-determinant derivatives, and zero variance makes the density score undefined.

C.2 Distributional Correctness

The following is the integrated residual-DB sampling implication underlying Liu et al. (2025, Proposition 4 and Appendices B.3–B.4), specialized to our terminal energy. We make the initial-noise compatibility condition explicit for the samplers considered here.

Theorem 1 (Terminal tilt under compatible boundaries)**.** *Fix a prompt y. Let the pretrained and adapted samplers have normalized Markov kernels with positive densities on common transition supports, the same initial prior $q_T = \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the same measurable deterministic decoder $\mathcal{D}$. Let $E \geq 0$ be measurable and finite almost surely under the pretrained terminal law, so $0 < Z(y) = \mathbb{E}_{\mathrm{pre}}[e^{-E(\mathcal{D}(\mathbf{z}_0))} \mid y] \leq 1$. Suppose positive continuation weights satisfy, almost everywhere on the pretrained path support,*

$$\frac{P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)}{P_F^{\mathrm{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)} = \frac{\tilde{F}_{t-1}(\mathbf{z}_{t-1}, y)}{\tilde{F}_t(\mathbf{z}_t, y)}, \qquad t = 1, \ldots, T, \tag{26}$$

with $\tilde{F}_0(\mathbf{z}_0, y) = e^{-E(\mathcal{D}(\mathbf{z}_0))}$ and $\tilde{F}_T(\mathbf{z}_T, y) = c(y) > 0$ independent of $\mathbf{z}_T$. Then

$$p_\theta(\mathbf{x} \mid y) = \frac{p_{\mathrm{pre}}(\mathbf{x} \mid y) e^{-E(\mathbf{x})}}{Z(y)}. \tag{27}$$

Proof. For the denoising trajectory $\mathbf{z}_T \to \mathbf{z}_{T-1} \to \cdots \to \mathbf{z}_0$, the Markov factorization gives

$$p_\theta(\mathbf{z}_{0:T} \mid y) = q_T(\mathbf{z}_T) \prod_{t=1}^{T} P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y), \tag{28}$$

$$p_{\mathrm{pre}}(\mathbf{z}_{0:T} \mid y) = q_T(\mathbf{z}_T) \prod_{t=1}^{T} P_F^{\mathrm{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y). \tag{29}$$

The common initial prior therefore cancels in the ratio. Substituting Eq. (26) and canceling successive continuation weights yields

$$\begin{aligned} \frac{p_\theta(\mathbf{z}_{0:T} \mid y)}{p_{\mathrm{pre}}(\mathbf{z}_{0:T} \mid y)} &= \prod_{t=1}^{T} \frac{P_F^\theta(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)}{P_F^{\mathrm{pre}}(\mathbf{z}_{t-1} \mid \mathbf{z}_t, y)} \\ &= \prod_{t=1}^{T} \frac{\tilde{F}_{t-1}(\mathbf{z}_{t-1}, y)}{\tilde{F}_t(\mathbf{z}_t, y)} = \frac{\tilde{F}_0(\mathbf{z}_0, y)}{\tilde{F}_T(\mathbf{z}_T, y)} = \frac{e^{-E(\mathcal{D}(\mathbf{z}_0))}}{c(y)}. \end{aligned} \tag{30}$$

Here the last equality uses both the terminal-energy boundary and the initial continuation weight being independent of $\mathbf{z}_T$. Integrating the path density fixes its normalization:

$$1 = \int p_{\mathrm{pre}}(\mathbf{z}_{0:T} \mid y) \frac{e^{-E(\mathcal{D}(\mathbf{z}_0))}}{c(y)} \, d\mathbf{z}_{0:T} = \frac{Z(y)}{c(y)}. \tag{31}$$

Thus $c(y) = Z(y)$. Since the remaining weight depends only on $\mathbf{z}_0$, integrating out $\mathbf{z}_1, \dots, \mathbf{z}_T$ gives

$$p_\theta(\mathbf{z}_0 \mid y) = \frac{p_{\mathrm{pre}}(\mathbf{z}_0 \mid y) e^{-E(\mathcal{D}(\mathbf{z}_0))}}{Z(y)}. \tag{32}$$

Finally, for every measurable image set A, the shared decoder gives

$$\begin{aligned} \Pr_\theta(\mathcal{D}(\mathbf{z}_0) \in A \mid y) &= \frac{1}{Z(y)} \int_{\mathcal{D}^{-1}(A)} e^{-E(\mathcal{D}(\mathbf{z}_0))} p_{\mathrm{pre}}(\mathbf{z}_0 \mid y)\, d\mathbf{z}_0 \\ &= \frac{1}{Z(y)} \int_A e^{-E(\mathbf{x})} p_{\mathrm{pre}}(d\mathbf{x} \mid y). \end{aligned} \tag{33}$$

This is the claimed image-distribution tilt. The pushforward argument covers non-injective decoders and image measures without pixel-space Lebesgue densities. □

Initial-state compatibility. Writing $w(\mathbf{z}_0) = e^{-E(\mathcal{D}(\mathbf{z}_0))}$, define

$$H_t(\mathbf{z}_t, y) = \mathbb{E}_{\mathrm{pre}}[w(\mathbf{z}_0) \mid \mathbf{z}_t, y], \qquad H_0 = w. \tag{34}$$

Normalizing Eq. (26) forces $\tilde{F}_t = \mathbb{E}_{P_F^{\mathrm{pre}}}[\tilde{F}_{t-1} \mid \mathbf{z}_t, y] = H_t$, so the theorem's boundaries with the unchanged Gaussian prior require $H_T(\mathbf{z}_T, y) = Z(y)$ almost surely. Finite-step diffusion and flow samplers do not ensure this compatibility, and our implementation retains the Gaussian prior. Recovering the integrated ratio from the score identities also requires an integrable continuation field and both identities to hold across a connected product domain; finite sampled losses do not ensure these properties.

C.3 Adapting ∇-DB for Forget Energies

For the thresholded energy in Eq. (3), with $\lambda_{\text{scale}}, \alpha > 0$, differentiation away from $s_{\mathcal{C}} = \tau$ gives

$$\frac{\partial \widetilde{E}_{\mathcal{C}}}{\partial s_{\mathcal{C}}} = \begin{cases} \lambda_{\text{scale}} \alpha \exp\big(\alpha(s_{\mathcal{C}} - \tau)\big), & s_{\mathcal{C}} > \tau, \\ 0, & s_{\mathcal{C}} < \tau. \end{cases} \tag{35}$$

At the threshold the scalar subdifferential is $[0, \lambda_{\text{scale}}\alpha]$; training uses the zero subgradient. The image gradient follows by the chain rule wherever $s_{\mathcal{C}}$ is differentiable. Finite energy gives a positive terminal weight $e^{-\widetilde{E}_{\mathcal{C}}}$, with normalized ratio $e^{-\widetilde{E}_{\mathcal{C}}}/Z(y)$. Scores below τ receive zero direct energy gradient, but identifying them as benign depends on detector accuracy, and their continuation scores and intermediate transitions need not remain unchanged (Section 3).

C.4 Optimization Caveat

The nonconvex LoRA and auxiliary-field objective provides empirical score-matching surrogates, with no general equivalence between zero loss and distributional correctness or global gradient-descent convergence guarantee. Theorem 1 requires the integrated ratio relation, the terminal-energy boundary, and continuation weights constant over the shared initial prior; finite sampled losses do not establish these conditions or field integrability. Moreover, the Gaussian forward residual in Eq. (24) is independent of $\mathbf{z}_{t-1}$, so exact matching across full Gaussian support requires $\log H_{t-1}$ to be affine: fixed-covariance transitions generally approximate a nonlinear terminal tilt even with unlimited mean-network capacity. Initial-prior incompatibility and training on only a subset of prompts further limit the guarantee for both diffusion and stochastic flow realizations.

D Extended Derivations

D.1 Closed-Form Conditional Gibbs Projection

The main text uses a prompt-conditioned target, with fixed prompt population $\pi(y)$. The optimization is therefore over conditional distributions $p(\cdot \mid y)$, while $\pi(y)$ is held fixed.

Proposition 1 (Conditional Gibbs projection for concept unlearning)**.** *Consider Eq.* (1) *over conditional distributions $p(\cdot \mid y)$ that are absolutely continuous with respect to $p_{\mathrm{pre}}(\cdot \mid y)$ for π-almost every y and have finite prompt-averaged energy. Assume $E_{\mathcal{C}}$ is nonnegative, measurable, and finite almost surely under the pretrained conditionals, with finite pretrained prompt-averaged energy $\bar{e}_0 = \mathbb{E}_{\pi,p_{\mathrm{pre}}}[E_{\mathcal{C}}] < \infty$. Assume the feasible set contains a strictly feasible conditional distribution. Then the minimizer is unique up to π-null prompt sets and has the form*

$$p^*_{\mathcal{C}}(\mathbf{x} \mid y) = \frac{p_{\mathrm{pre}}(\mathbf{x} \mid y)\exp(-\beta E_{\mathcal{C}}(\mathbf{x}))}{Z_{\beta,\mathcal{C}}(y)}, \qquad Z_{\beta,\mathcal{C}}(y) \coloneqq \int p_{\mathrm{pre}}(\mathbf{x} \mid y)\exp(-\beta E_{\mathcal{C}}(\mathbf{x}))\, d\mathbf{x}, \tag{36}$$

for some finite shared $\beta \geq 0$ with $0 < Z_{\beta,\mathcal{C}}(y) \leq 1$ for π-almost every y. Moreover, β satisfies the KKT conditions

$$\beta \geq 0, \qquad \mathbb{E}_{y\sim\pi,\,\mathbf{x}\sim p^*_{\mathcal{C}}(\cdot|y)}[E_{\mathcal{C}}(\mathbf{x})] \leq \delta_{\mathcal{C}}, \qquad \beta\Big(\mathbb{E}_{y\sim\pi,\,\mathbf{x}\sim p^*_{\mathcal{C}}(\cdot|y)}[E_{\mathcal{C}}(\mathbf{x})] - \delta_{\mathcal{C}}\Big) = 0. \tag{37}$$

Proof. Write p_β for this normalized tilt, $e_y(\beta) = \mathbb{E}_{p_\beta(\cdot|y)}[E_{\mathcal{C}}]$, and $\bar{e}(\beta) = \mathbb{E}_\pi[e_y(\beta)]$. For π-almost every prompt, the pretrained mean $e_y(0)$ is finite. Dominated convergence gives continuity at $\beta = 0$. For $\beta > 0$, exponential damping makes all tilted energy moments finite, so differentiation of the normalized mean gives

$$\frac{de_y(\beta)}{d\beta} = -\operatorname{Var}_{p_\beta(\cdot|y)}(E_{\mathcal{C}}) \leq 0. \tag{38}$$

Thus $0 \leq e_y(\beta) \leq e_y(0)$. To establish its limit, let $m(y) = \operatorname{ess\,inf}_{p_{\mathrm{pre}}(\cdot|y)} E_{\mathcal{C}}$. For any $\varepsilon > 0$, the mass $A_y(\varepsilon) = \Pr_{\mathrm{pre}}(E_{\mathcal{C}} \leq m(y) + \varepsilon/2 \mid y)$ is positive. Splitting the tilted mean at $m(y) + \varepsilon$ and bounding its normalizer below by $A_y(\varepsilon)e^{-\beta(m(y)+\varepsilon/2)}$ yields

$$0 \leq e_y(\beta) - m(y) \leq \varepsilon + \frac{e_y(0) - m(y)}{A_y(\varepsilon)}e^{-\beta\varepsilon/2}. \tag{39}$$

Taking $\beta \to \infty$ and then $\varepsilon \downarrow 0$ proves $e_y(\beta) \to m(y)$. Since $e_y(0)$ is integrable over prompts, dominated convergence gives continuity of $\bar{e}$ on $[0,\infty)$ and $\lim_{\beta\to\infty}\bar{e}(\beta) = \mathbb{E}_\pi[m(y)]$. Any admissible conditional law has mean energy at least $m(y)$, so strict feasibility implies $\mathbb{E}_\pi[m(y)] < \delta_{\mathcal{C}}$. If $\bar{e}_0 \leq \delta_{\mathcal{C}}$, choose $\beta = 0$; otherwise continuity supplies a finite $\beta > 0$ with $\bar{e}(\beta) = \delta_{\mathcal{C}}$.

For completeness, let $K(p) = \mathbb{E}_\pi[\mathrm{KL}(p(\cdot \mid y)\|p_{\mathrm{pre}}(\cdot \mid y))]$. Jensen's inequality gives $-\mathbb{E}_\pi \log Z_{\beta,\mathcal{C}} \leq \beta\bar{e}_0 < \infty$, so $K(p_\beta)$ is finite. For any feasible p with finite KL, the exact identity

$$K(p) - K(p_\beta) = \mathbb{E}_\pi[\mathrm{KL}(p(\cdot \mid y)\|p_\beta(\cdot \mid y))] + \beta\big(\bar{e}(\beta) - \mathbb{E}_{\pi,p}[E_{\mathcal{C}}]\big) \geq 0 \tag{40}$$

certifies optimality: either $\beta = 0$ or $\bar{e}(\beta) = \delta_{\mathcal{C}}$. Infinite-KL candidates cannot improve it. Equality requires $p = p_\beta$ almost surely, proving uniqueness and the stated complementary-slackness conditions. □

Thus β is tied to the active constraint level in the stated constrained problem. When β is instead chosen directly as a hyperparameter, one is solving the relaxed penalized objective rather than fixing δ in advance. Strict feasibility excludes the limiting minimum-energy constraint, which can require $\beta \to \infty$ and is not covered by the finite-multiplier claim. In the thresholded practical surrogate used in the main method, this scalar strength is absorbed into $\widetilde{E}_{\mathcal{C}}$ through λ_{scale}, while the target remains prompt-conditioned through $p_{\mathrm{pre}}(\mathbf{x} \mid y)$ and $Z_{\beta,\mathcal{C}}(y)$.

D.2 Continuation Approximation and Endpoint Conventions

The exact residual continuation score is $\nabla \log H_t$, where H_t is the conditional expectation in Eq. (34). The auxiliary term g_ϕ is the forward-looking continuation residual of Eq. (8). It differs from the additional transition-dependent correction considered in Appendix C of Liu et al. (2025), and is not a proved correction for violation of DDPM detailed balance. The auxiliary network is conditioned on prompt embeddings in the concept-unlearning trainer; this argument is suppressed in $g_\phi(\mathbf{z}_t, t)$.

Even with a differentiable scorer, a freely parameterized vector network need not be the gradient of a scalar potential. Detaching the energy gradient avoids second derivatives of the scorer, but exact parameter differentiation of the conditioning-state transition score would still involve mixed derivatives of the denoiser.

The clean-end identity requires both $a_0 = 1$ and $\hat{\mathbf{x}}(\mathbf{z}_0, 0, y) = \mathcal{D}(\mathbf{z}_0)$. A scheduler index named zero can still have nonzero noise; evaluating another denoiser prediction at that index is not this exact boundary. Likewise, a zero-variance final transition has no ordinary Gaussian log-density score. Flooring its variance defines a different stochastic kernel, whereas omitting its score does not establish the full-chain ratio condition in Theorem 1. Finite terminal penalties encourage, rather than ensure, the boundary condition.

D.3 Rectified-Flow Transitions and Scores

For SD3, the adapted network predicts a velocity v_θ and receives sequence and pooled text embeddings. We use Eq. (10) to reconstruct a clean latent, divide by the VAE scaling factor, add its shift factor, and decode before evaluating the energy. LoRA updates the image-attention and joint-attention projections; the transformer backbone, VAE, and available text encoders remain frozen.

Interior drift and sign convention. For $0 < \sigma < 1$, let $q_\sigma(\mathbf{z} \mid y)$ be the marginal of $\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$, with independent $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and let $u_\sigma(\mathbf{z}, y) = \mathbb{E}[\boldsymbol{\epsilon} - \mathbf{z}_0 \mid \mathbf{z}_\sigma = \mathbf{z}, y]$ be its exact marginal velocity. Differentiating the Gaussian corruption density gives

$$\ell_\sigma(\mathbf{z}, y) \coloneqq \nabla_{\mathbf{z}} \log q_\sigma(\mathbf{z} \mid y) = -\frac{\mathbb{E}[\boldsymbol{\epsilon} \mid \mathbf{z}, y]}{\sigma} = -\frac{\mathbf{z} + (1-\sigma)u_\sigma(\mathbf{z}, y)}{\sigma}. \tag{41}$$

The last equality follows by taking conditional expectations in the interpolation identity. Its marginals obey the continuity equation $\partial_\sigma q_\sigma = -\nabla \cdot (u_\sigma q_\sigma)$. For a stochastic process run in *decreasing* σ, let b_σ be the drift and let the Brownian increment have covariance $-d\sigma\,\mathbf{I}$. Equivalently, in the increasing clock $r = 1 - \sigma$, the drift is $-b_{1-r}$ and the diffusion coefficient is $\gamma(1-r)$. Converting its Fokker–Planck equation back to σ gives

$$\begin{aligned} \partial_\sigma q_\sigma &= -\nabla \cdot (b_\sigma q_\sigma) - \frac{\gamma(\sigma)^2}{2}\Delta_{\mathbf{z}} q_\sigma, \\ b_\sigma &= u_\sigma - \frac{\gamma(\sigma)^2}{2}\ell_\sigma = \frac{\gamma(\sigma)^2}{2\sigma}\mathbf{z} + \left(1 + \frac{\gamma(\sigma)^2(1-\sigma)}{2\sigma}\right) u_\sigma. \end{aligned} \tag{42}$$

Substitution cancels the diffusion term because $\ell_\sigma q_\sigma = \nabla q_\sigma$. Thus the continuous process shares the interpolation marginals on interior intervals, provided it starts from the matching marginal and the coefficients and density ensure a well-posed SDE and a unique Fokker–Planck evolution (Song et al., 2021). Both correction coefficients have positive signs in b_σ; the sampling step multiplies them by $\Delta\sigma < 0$.

Implemented Gaussian step. Let $\Delta\sigma = \sigma_{\text{next}} - \sigma < 0$. Replacing u_σ by the learned velocity and taking an Euler–Maruyama step gives the implemented interior update

$$\begin{aligned} \gamma(\sigma) &= \eta\sqrt{\frac{\sigma}{1-\sigma}}, \qquad s_\sigma = \gamma(\sigma)\sqrt{-\Delta\sigma}, \\ \mu_\theta(\mathbf{z}_\sigma, \sigma, y) &= \left(1 + \frac{\gamma(\sigma)^2}{2\sigma}\Delta\sigma\right)\mathbf{z}_\sigma + \left(1 + \frac{\gamma(\sigma)^2(1-\sigma)}{2\sigma}\right) v_\theta(\mathbf{z}_\sigma, \sigma, y)\Delta\sigma, \\ \mathbf{z}_{\text{next}} &= \mu_\theta(\mathbf{z}_\sigma, \sigma, y) + s_\sigma\boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \end{aligned} \tag{43}$$

Here $\eta > 0$ controls the flow transition noise. The drift correction accompanies the injected noise; simply adding noise to a deterministic Euler step gives a different transition. The coefficient γ is singular at $\sigma = 1$, so Eq. (43) does not define an unguarded Euler step there. Near that endpoint, the implementation uses the next scheduler noise level to guard the noise coefficient, floors divisors, and floors s_σ at 10^{-5}. These guards define a modified discrete kernel, not an exact extension of the

continuous marginal argument. Setting $\eta = 0$ without a variance floor gives a deterministic kernel with no ordinary density score; retaining the floor still adds noise.

The adapted and frozen reference transitions use the same noise schedule and variance. Their next-state score difference is therefore

$$\nabla_{\mathbf{z}_{\text{next}}} \log P_F^{\theta} - \nabla_{\mathbf{z}_{\text{next}}} \log P_F^{\text{pre}} = \frac{\mu_\theta - \mu_{\text{pre}}}{s_\sigma^2}, \tag{44}$$

while the current-state derivatives hold $\mathbf{z}_{\text{next}}$ fixed and differentiate each mean. The pretrained velocity is evaluated with LoRA disabled. These are exact scores of the defined Gaussian kernels, even when the learned velocity is approximate. In the forward-looking parameterization, the DDPM signal coefficient $\bar{\alpha}_t$ is replaced by $a(\sigma) = (1-\sigma)^2$, with numerical clipping of $1 - \sigma$. The terminal penalty encourages the auxiliary residual to vanish at the clean endpoint.

These transitions define the stochastic reference sampler for residual training. A learned or classifier-free-guided velocity need not satisfy Eq. (41) for its actual sampler marginals. This approximation, finite discretization, endpoint guards, approximate continuation scores, initial-prior compatibility, and the use of a deterministic exported sampler prevent a guarantee that inference samples the same tilted distribution. In particular, p_{pre} in the ratio theorem is the terminal law of the specified stochastic reference, which need not equal the deterministic pretrained SD3 law. Theorem 1 remains conditional on its stated assumptions.

E Additional Experimental Details

E.1 CLIP Model

The primary configuration uses `openai/clip-vit-large-patch14` (ViT-L/14, 224×224 input, output dimension 768) for both the forget energy and all CLIP evaluation scores. These scores share an encoder with the training signal and are not independent validation. Substituting another differentiable scorer requires recalibrating its energy boundary and scale; Appendix G describes Qwen.

Descriptors may average short, scene-independent paraphrases (e.g., "Pikachu" and "a photo of Pikachu"). Unless stated otherwise, experiments use the canonical descriptor (e.g., "Van Gogh style"), as specified in Section 5.1.

E.2 Base Diffusion Model and Adapters

The primary diffusion experiments adapt Stable Diffusion v1.5 (`runwayml/stable-diffusion-v1-5`). LoRA adapters are inserted into the U-Net cross- and self-attention projections (W_Q, W_K, W_V, W_O) at rank $r = 8$, with the standard $\alpha = r$ scaling. Only the LoRA weights $\Delta\theta$ and the auxiliary residual parameterization g_ϕ (channels $\{64, 128, 256, 256\}$, time-embedded) receive gradient updates; the U-Net backbone, latent autoencoder, text encoder, and CLIP model are all frozen. We use 50-step DDPM sampling at training and inference time for this configuration.

E.3 Prompt Sets and Generation Strategy

Prompt generation strategy. We use GPT-4o-mini to generate varied, visually grounded scene prompts under the following system instructions.

Forget Prompt Generation – System Prompt

```
You are a dataset prompt writer for diffusion models (e.g., Stable
Diffusion). Your job is to produce short, simple, natural-language
prompts that depict a clearly visible instance of the specified
concept in-frame.
GENERAL STYLE: 4-10 words per prompt; one simple main clause; present
tense preferred; everyday vocabulary only; 0-1 adjectives; no
tags/parameters (no colons, aspect ratios, seeds, CFG, hashtags);
minimal punctuation.
VISUAL GROUNDING (MANDATORY): The concept MUST be a visible, concrete
subject in the image. If the concept is polysemous, use the physical
object sense. The scene must make the concept clearly depictable and
in-frame. Match number and determiners consistently
(a/an/one/two/many).
DIVERSITY REQUIREMENTS: Across prompts, balance indoor vs. outdoor,
day vs. night, weather conditions, everyday activities, solitary vs.
small-group scenes, and near vs. far framing. Avoid near-duplicates
and repeated sentence structure.
```

Retain prompt generation. For the optional retain-prompt augmentation in Section 5.2 (Q5), the LLM proposes $\sim$50 related concepts per target: siblings, hypernyms, co-occurring concepts, and cross-domain analogues, excluding target synonyms, subtypes, and parts. CLIP text–text similarity (or denoiser-based score-prediction distance) narrows this pool; the LLM then generates 10 scene prompts per selected neighbor:

Retain Prompt Generation – System Prompt

```
You are a prompt generator for Stable Diffusion models.
INPUT: UNLEARN_CONCEPT (the concept being unlearned; must NOT appear
in any prompt) and RELATED_CONCEPT (the target subject for retention).
CONSTRAINTS: 5-10 words per prompt; present tense; simple, everyday
vocabulary; ≤ 1 comma and ≤ 1 adjective per prompt; no special tags or
parameters; do not include UNLEARN_CONCEPT or any obvious
synonym/nickname.
CONTENT: Describe simple, plausible scenes involving RELATED_CONCEPT.
Ensure variety:  mix indoor/outdoor settings, day/night, weather,
activities, solitary vs. group, near vs. far framing.  Do not repeat
sentence structure; avoid near-duplicates.  Keep all prompts safe and
neutral in tone.
OUTPUT: Exactly 10 prompts, one per line.  No numbering, bullet
points, section titles, or concept labels.
```

Prompt sets used in experiments. The four sets cover training forget prompts, held-out forget prompts, nearby benign concepts, and unrelated benign content. Representative counts are:

- **Van Gogh forget:** 30 training prompts (e.g., "a painting in the style of Van Gogh", "Van Gogh starry night", "Van Gogh self-portrait") and 20 disjoint held-out evaluation prompts.
- **Van Gogh benign:** Prompts on various subjects covering landscapes, portraits, animals, and objects in non-Van Gogh styles(Total 60+ concept and for each concept 20 prompts each).
- **Pikachu forget (training/held-out):** 30 + 20 prompts (e.g., "pikachu", "a picture of pikachu", "pikachu in a forest").
- **Pikachu benign:** Various prompts on non-Pikachu characters and general objects (Total 60+ concept and for each concept 20 prompts each).

Retention prompts are shared across methods.

E.4 Optimization and Schedule

For the primary diffusion configuration, LoRA parameters and ϕ are jointly optimized with AdamW at learning rate 1×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 10^{-2}, and gradient clipping at norm 1.0. We use a batch size of 4 forget prompts, with one denoising trajectory per prompt and $|\mathcal{T}| = 0.1T = 5$ score-matched timesteps per trajectory. Training runs for 1,000 batches per concept. The loss weights in Eq. (9) are $\lambda_{\text{rev}} = 1.0$ and $\lambda_{\text{term}} = 0.1$.

E.5 Compute

Primary diffusion training on SDv1.5 takes 2–5 hours per concept on 2 NVIDIA H100's (80 GB), dominated by trajectory rollout and the reverse-score Jacobian-vector product (Phases 1 and 3 of Algorithm 1). Inference uses standard 50-step DDPM sampling with the LoRA correction folded into the U-Net forward pass, adding no cost beyond the base model.

E.6 Hyperparameter Sensitivity

SD v1.5 threshold selection. CLIP similarity can remain nonzero even when the target concept is absent. To characterize this residual similarity, we adapt Six-CD's dual-version idea by removing the target expression from each prompt while preserving its context (Ren et al., 2025, Section 4, Table 3). We generate 16 images per benign prompt with frozen SD v1.5 and average their CLIP ViT-L/14 similarities to the paired *full original prompts* to obtain reference scores.

Using these reference scores as a guide, together with empirical hyperparameter experiments, we select shared thresholds of $\tau = 0.21$ for Objects, 0.22 for Characters, and 0.20 for Styles. During unlearning, the energy uses *concept descriptors* rather than full prompts, so these thresholds are

empirical settings informed by the reference measurements. The threshold marks the onset of the energy penalty; scores at or below it receive zero direct penalty. Applying the method to new concepts or base models requires threshold calibration. Training duration also depends on concept specificity: individual characters tend to unlearn faster, whereas broader categories such as Cat and Dog may require more epochs for strong forgetting.

- λ_{scale} **(tilt magnitude).** The most sensitive parameter: small values leave residual concept content; large values destabilize training. We sweep $\lambda_{\text{scale}} \in \{0.5, 1, 2, 5\}$ and select the smallest value yielding a stable forget rate.
- α **(slope).** Controls energy growth above τ and is less sensitive than the other two parameters; we use $\alpha = 10$ throughout.

E.7 Evaluation Protocol

VLM-based automated evaluation. We use **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025) as a zero-shot binary classifier without labeled examples, asking, e.g., *"Does this image contain ⟨concept⟩?"* Unlearning accuracy aggregates No answers; related and general retention aggregate Yes answers for the requested concept. In the Qwen-energy experiment, the evaluator shares its model with the training scorer (Appendix G).

Why not classifiers. Fixed-vocabulary classifiers restrict concept coverage (e.g., ImageNet's 1000 classes); UnlearnCanvas heads (Zhang et al., 2024a) are also tied to their dataset's object and style distributions. On original Stable Diffusion outputs and out-of-distribution images, we observed inconsistent classifier labels, including poor generalization to canonical artistic styles, while Qwen judgments aligned better with manual inspection.

Evaluation datasets and scale. Each checkpoint uses three disjoint, held-out sets: forget $\mathcal{D}_{\text{eval_forget}}$, related-retain $\mathcal{D}_{\text{eval_related}}$, and general-retain $\mathcal{D}_{\text{eval_general}}$. The benchmark contains **12 unlearning targets** (five objects, four characters, three styles) and **60+ distinct concepts** including retention controls. With **20 unique evaluation prompts per concept** and **8 images per prompt**, evaluation produces approximately **10,000 images per model checkpoint**. The general-retain set is fixed and unrelated to all unlearning targets.

Related concept identification. Following EraseBench (Amara et al., 2025), GPT-4o-mini identifies related concepts with the following system-prompt excerpt:

Related Concept Identification – System Prompt (Excerpt)

```
Your main task is to help identify concepts for evaluating
text-to-image models.  The key idea is to identify 3-4 concepts that
are semantically entangled with the Given Concept...
```

Examples include *Pikachu* → *Squirtle, Charmander* and *Van Gogh* → *Impressionism*, used to measure the ripple effect (Amara et al., 2025; Zhang et al., 2024a).

Metric definitions. We report target absence U_{acc} and requested-concept recognition on related and general prompts, RR_{acc} and GR_{acc}. Complementary CLIP scores measure graded text–image alignment: U_{clip} evaluates the forget prompt's benign context as a proxy for "in-prompt retainability" (Ren et al., 2025), while RR_{clip} and GR_{clip} evaluate the corresponding retention prompts. High unlearning accuracy and context alignment indicate erasure with context preservation according to these metrics; the shared CLIP encoder limitation is described in Appendix E.1. The saved VLM-energy results instead use full-prompt CLIP alignment, as specified in Appendix G.5.

Generative quality. We compute **Fréchet Inception Distance (FID)** (Heusel et al., 2017) with the UnlearnCanvas evaluation script and methodology (Zhang et al., 2024a), comparing unlearned and original Stable Diffusion images on the same evaluation prompts. The pretrained-model distribution serves as the benign-generation reference.

FADE computation and retain-only reference. **FADE** (Cho et al., 2025) is the symmetric KL divergence in Section 5.1: non-negative and zero exactly when the conditional distributions agree almost everywhere. We approximate it with weighted denoising-MSE differences between unlearned and retain-only models, assuming comparable variational-bound gaps. Sampling noise and likelihood approximation can make raw directional estimates negative.

Following UnlearnCanvas (Zhang et al., 2024a), the reference construction has two stages:

1. **Unlearning starting point.** Finetune Stable Diffusion v1.5 on the *full* UnlearnCanvas dataset (all styles), then apply concept unlearning to that checkpoint.
2. **Retain-only reference.** Separately finetune a fresh pretrained Stable Diffusion v1.5 copy on UnlearnCanvas$\setminus\mathcal{C}_{\mathrm{style}}$ for each target style. Downstream data excludes that style, but the shared SD v1.5 initialization may already encode it; the reference is not trained from scratch without concept exposure.

We construct references for **Van Gogh style**, **Picasso style**, and **Monet style**, which UnlearnCanvas supports. **Avg. FADE** in Table 1 averages these three per-style comparisons, subject to the shared-pretraining limitation. FADE therefore measures agreement with this downstream retain-only reference; it does not certify equivalence to a model never exposed to the target concept.

F NUDITY UNLEARNING WITH A CLASSIFIER-BASED CONCEPT ENERGY

Nudity is a composite visual concept involving different exposed body parts and visual contexts. Our initial approach uses the binary NudeNet classifier to measure image-level unsafe content. It performs well on curated held-out prompts under the binary unsafe metric while preserving benign prompt alignment, but leaves substantial residual nudity on I2P. We therefore introduce a more comprehensive concept energy that combines the image-level score with explicit evidence of exposed body parts.

F.1 CLASSIFIER-BASED CONCEPT ENERGIES

A frozen differentiable classifier whose vocabulary covers $\mathcal{C}$ can supply a score over mutually exclusive target labels $\mathcal{K}_{\mathcal{C}}$:

$$s_{\mathcal{C}}^{\mathrm{cls}}(\mathbf{x}) = \sum_{k \in \mathcal{K}_{\mathcal{C}}} p_{\mathrm{cls}}(k \mid \mathbf{x}).$$

For a binary detector, this is the positive-class probability. Substitution for CLIP in Eq. (3) preserves the nonnegative thresholded energy and target form in Eq. (4), but requires classifier-specific threshold and scale calibration.

Gradients pass through class probabilities and preprocessing to the image; argmax or detached labels remove this signal. Preservation requires distinguishing the target from related benign content, and the flat region preserves relative probabilities only among zero-energy images.

F.2 NUDITY UNLEARNING SETUP

Combined concept energy. We combine the binary classifier with a frozen NudeNet body-part detector:

$$\widetilde{E}_{\mathrm{nude}}(\mathbf{x}) = E_{\mathrm{binary}}(\mathbf{x}) + \sum_{k \in \mathcal{K}} E_k(\mathbf{x}). \tag{45}$$

Here $\mathcal{K}$ is the set of target body-part categories, E_{binary} penalizes image-level unsafe evidence, and E_k penalizes exposure of category k. Each penalty is zero below its threshold and grows with stronger evidence. The body-part terms can therefore penalize exposure even when the binary classifier predicts safe. We use the negative combined energy as the log reward, retaining the existing residual ∇-GFlowNet objective.

Training setup. For comparison with EraseFlow (Kusumba et al., 2025), we adapt Stable Diffusion v1.4 using this combined energy. We train U-Net LoRA adapters and the auxiliary flow model while keeping the pretrained model weights and concept scorers frozen.

Evaluation. We evaluate on the I2P subset released with EraseFlow and on held-out nudity and benign prompts. The main I2P metric is attack success rate (ASR): the percentage of generated images flagged by NudeNet for exposed body parts. We also measure the binary unsafe rate, use Qwen for an independent assessment, and evaluate benign prompt alignment and diversity with CLIP-based metrics.

Results. Table 5 summarizes the published I2P results for prior methods, as reported in EraseFlow (Kusumba et al., 2025). Our method achieves a strict nudity rate of 0.70% on the evaluated I2P subset, excluding detections of exposed feet, belly, and armpits.

Table 5: Nudity unlearning on I2P. Baseline results are taken from EraseFlow (Kusumba et al., 2025). Lower ASR is better

Method	I2P ASR (%) ↓
SD v1.4	93.66
ESD	13.30
UCE	19.71
MACE	6.30
DUO	16.90
EraseFlow	2.80
Adversarial methods	
R.A.C.E (Kim et al., 2024)	2.80
AdvUnlearn (Zhang et al., 2024b)	1.40
Inference time intervention	
SAFREE (Yoon et al., 2025)	21.83
TILDE (ours)	4.30

G UNLEARNING WITH A VISION–LANGUAGE MODEL AS THE CONCEPT-ENERGY SCORER

TILDE needs a differentiable measure of how strongly an image depicts the concept to forget. CLIP provides this measure in our main experiments, but the training objective also accepts other image-dependent energies. Here we train an SD v1.5 generator using a frozen Qwen2.5-VL-7B-Instruct model (Bai et al., 2025) in place of the CLIP training scorer. We first explain how a Yes/No question becomes a continuous forgetting signal, then motivate the EOT-softplus variant and compare its forgetting and retention behavior with CLIP.

G.1 FROM A YES/NO QUESTION TO A FORGETTING ENERGY

Let $\mathbf{x}$ be a generated image and $\mathcal{C}$ the target concept, such as Pikachu. We ask Qwen: *"Does this image contain or depict Pikachu? Answer Yes or No only."* The generation prompt is not included in this question, so the score is based on visual evidence rather than on the prompt naming the target. A decoded Yes/No answer is discrete; the token probabilities before decoding provide the continuous signal needed for training.

Write $q_{\mathrm{yes}}(\mathbf{x}, \mathcal{C})$ and $q_{\mathrm{no}}(\mathbf{x}, \mathcal{C})$ for the total next-token probabilities assigned to valid Yes and No spellings. For example, we group `Yes`, `yes`, and their single-token capitalization and leading-space variants, and similarly for No. This grouping reduces dependence on the spelling of an otherwise equivalent answer. Our concept-evidence score is

$$m_{\mathcal{C}}(\mathbf{x}) = \log \frac{q_{\mathrm{yes}}(\mathbf{x}, \mathcal{C})}{q_{\mathrm{no}}(\mathbf{x}, \mathcal{C})}. \tag{46}$$

A positive score favors Yes, a negative score favors No, and a larger score indicates stronger evidence of the target. These are probabilities of the next answer token, not probabilities of a complete generated response.

We turn this score into a nonnegative forgetting energy using softplus, the smooth function $\mathrm{softplus}(u) = \log(1 + e^u)$:

$$\widetilde{E}_{\mathcal{C}}^{\mathrm{VLM}}(\mathbf{x}) = \rho T_{\mathrm{VLM}}\, \mathrm{softplus}\left(\frac{m_{\mathcal{C}}(\mathbf{x}) - m_0}{T_{\mathrm{VLM}}} \right). \tag{47}$$

Here m_0 sets the boundary between weak and strong concept evidence, and $\rho > 0$ sets the overall penalty strength. The parameter $T_{\mathrm{VLM}} > 0$, called the *temperature*, controls the smoothness of the softplus penalty: a smaller value makes the transition around m_0 sharper, while a larger value makes it more gradual. Well above the boundary, the energy grows approximately linearly with the excess evidence. Far below the boundary, it approaches zero. As in Eq. (4), the pretrained image density is reweighted by $\exp(-\widetilde{E})$, so images with stronger target evidence are suppressed more strongly.

Softplus also makes the gradient more selective. The derivative of the energy with respect to the evidence score is

$$\frac{\partial \widetilde{E}_{\mathcal{C}}^{\mathrm{VLM}}}{\partial m_{\mathcal{C}}} = \rho\, \mathrm{sigmoid}\left(\frac{m_{\mathcal{C}} - m_0}{T_{\mathrm{VLM}}} \right). \tag{48}$$

Here $\mathrm{sigmoid}(u) = 1/(1 + e^{-u})$, so low-evidence images receive a much smaller direct forgetting signal. A raw-margin energy keeps pushing the score downward even after the model already favors No. Softplus reduces this pressure, although its derivative is not exactly zero at any finite margin.

G.2 WHY WE USE EOT-SOFTPLUS

A single image presentation can make the forgetting gradient sensitive to fine visual details. We therefore use a two-view form of *expectation over transformations* (EOT). Let $A(\mathbf{x})$ denote the image

after bilinear downsampling from 224×224 to 112×112 and upsampling back to 224×224, with antialiasing. The EOT energy averages the softplus energies of the original and transformed views:

$$\widetilde{E}_{\mathcal{C}}^{\mathrm{EOT}}(\mathbf{x}) = \tfrac{1}{2}\left[\widetilde{E}_{\mathcal{C}}^{\mathrm{VLM}}(\mathbf{x}) + \widetilde{E}_{\mathcal{C}}^{\mathrm{VLM}}(A(\mathbf{x}))\right]. \tag{49}$$

Both views remain differentiable. The optimizer therefore receives feedback about target evidence at two levels of image detail, while softplus attenuates the signal when that evidence is already weak. EOT is a deterministic average over these two views; it does not require sampling a new transformation at every step.

This design combines two useful properties: a nonnegative energy that weakens below the calibrated boundary, and a forgetting signal that considers more than one image presentation. The Pikachu results in Table 6 also support its use: EOT-softplus retains more related and general concepts than either Qwen alternative at both reported epochs. At epoch 20, this comes with less complete forgetting. We use EOT-softplus for the broader VLM study because retention is central to selective unlearning; the current evidence supports this empirical trade-off, without establishing robustness to transformations beyond the two used during training.

G.3 Calibration and Differentiable Implementation

Setting the energy parameters. Before unlearning, we calibrate the energy using target images and reference controls. We choose the boundary m_0 to distinguish these groups and set the smoothness T_{VLM} from the separation between their scores. We then adjust the strength ρ so that the initial forgetting gradients are comparable in magnitude to those from the CLIP energy. Thus, Qwen supplies the training score, while CLIP provides an offline reference for its strength. The controls include outputs of an earlier CLIP-unlearned model, treated as negative examples without individual verification.

Computing image gradients while keeping Qwen frozen. Images remain tensors throughout resizing and preprocessing. Qwen's parameters are frozen, but gradients pass through its computations to the image, the VAE decoder, and the predicted clean latent. The training log-potential is $\Phi = -\widetilde{E}$; its latent gradient enters the existing residual-score objective and is then detached, so optimization does not require second derivatives through Qwen. Runtime checks verify finite energies and latent gradients and confirm that Qwen receives no parameter gradients.

For implementation, if ℓ_k is the next-token logit for token k, and $\mathcal{V}_{\mathrm{yes}}$ and $\mathcal{V}_{\mathrm{no}}$ are the validated, deduplicated token sets described above, Eq. (46) can be evaluated as

$$m_{\mathcal{C}}(\mathbf{x}) = \log \sum_{k \in \mathcal{V}_{\mathrm{yes}}} e^{\ell_k(\mathbf{x})} - \log \sum_{k \in \mathcal{V}_{\mathrm{no}}} e^{\ell_k(\mathbf{x})}.$$

The common softmax normalizer cancels, so only the relevant language-model-head rows are needed. The raw-margin diagnostic instead uses the literal Yes/No logit difference; its signed energy is a diagnostic outside the nonnegative-energy assumptions of the main construction.

G.4 Pikachu Results and Comparison with CLIP

Each checkpoint uses 20 prompts per concept and eight images per prompt: 160 target images, 480 related images across Charmander, Squirtle, and Jigglypuff, and 3,200 general images across 20 concepts. We use a single unlearning metric, U_{acc}: the fraction of target-prompt images for which the direct Yes/No evaluator reports that the target is absent. Related and general retention, RR_{acc} and GR_{acc}, are the fractions for which the requested retain concept is present.

EOT-softplus improves related retention over single-view softplus by 13.75 percentage points at epoch 20 and 10.42 points at epoch 40. General retention also improves at both checkpoints. Continuing EOT training increases U_{acc} from 88.13% to 98.75%, while related retention falls from 80.83% to 56.88%. Thus, stronger target removal can still damage semantically related benign concepts; EOT reduces this problem relative to the other Qwen energies but does not eliminate it.

Table 6: **Qwen forgetting energies on Pikachu.** Grouping checkpoints by epoch exposes the forgetting–retention trade-off. Bold values are the largest within each column. Raw margin uses literal Yes/No tokens; the two softplus variants aggregate equivalent token forms.

Energy	Epoch 20			Epoch 40		
	$U_{\text{acc}} \uparrow$	$RR_{\text{acc}} \uparrow$	$GR_{\text{acc}} \uparrow$	$U_{\text{acc}} \uparrow$	$RR_{\text{acc}} \uparrow$	$GR_{\text{acc}} \uparrow$
Raw margin	0.96	0.50	0.80	**1.00**	0.19	0.73
Softplus	**0.99**	0.67	0.82	0.99	0.46	0.83
EOT-softplus	0.88	**0.81**	**0.84**	0.99	**0.57**	**0.83**

Why CLIP remains our primary energy. The CLIP-based Pikachu result in Table 18 reports approximately 100% unlearning, 76% related retention, and 88% general retention. At similarly near-complete forgetting, EOT-softplus at epoch 40 gives 98.75%, 56.88%, and 83.34%, respectively. CLIP also has higher related- and general-prompt alignment scores: $33.0/32.1$, compared with $31.93/31.10$ for that EOT checkpoint. These results favor the CLIP configuration for preserving benign concepts and prompt alignment while removing the target. EOT at epoch 20 has higher related retention than the CLIP result, but its target removal is substantially less complete.

The energy shape provides a reason for this retention advantage. The thresholded CLIP energy in Eq. (3) is exactly zero below its concept-evidence threshold. Images in this region receive no direct terminal-energy forgetting gradient, and their relative probabilities are unchanged by the ideal Gibbs reweighting. This reduces pressure to alter benign content that the detector already regards as target-free. Softplus only approaches zero, so weak residual pressure can remain on such images; EOT averages that pressure across views without creating an exactly flat region. Shared model parameters and approximate optimization can still change benign outputs under either energy. The comparison concerns the complete calibrated training configurations, rather than establishing that CLIP is intrinsically a better visual recognizer than Qwen.

G.5 Reporting Across Concepts and Categories

Table 7 reports EOT-softplus results for all 12 targets, grouped into objects, characters, and styles, together with category and overall averages. The six columns pair unlearning and retention accuracies with their corresponding CLIP scores. For these saved VLM experiments, CLIP scores use the *full evaluation prompt*, including the target name in U_{clip}. Consequently, U_{clip} is a text–image alignment diagnostic, not a separate measurement of the benign portion of a forget prompt: target presence can also affect it. We report this distinction explicitly when interpreting benign-content preservation.

The training scorer and semantic evaluator share the same Qwen snapshot, so these accuracies are not independent semantic validation. CLIP scores provide a complementary measurement; no repeated-training uncertainty estimates are available.

Table 7: **EOT-softplus results across concepts.** Category and concept wise evaluation numbers across all concepts.

Category	Concept	Forget		Related retain		General retain	
		$U_{\text{acc}} \uparrow$	$U_{\text{clip}} \uparrow$	$RR_{\text{acc}} \uparrow$	$RR_{\text{clip}} \uparrow$	$GR_{\text{acc}} \uparrow$	$GR_{\text{clip}} \uparrow$
Objects	Apple	0.99	28.0	0.44	28.3	0.85	30.9
	Banana	0.99	27.0	0.53	30.0	0.79	30.8
	Golf Ball	0.96	27.9	0.60	30.8	0.86	31.4
	Dog	1.00	24.9	0.48	28.5	0.79	30.7
	Cat	0.92	25.9	0.89	32.0	0.86	31.4
Characters	Mickey Mouse	0.91	31.7	0.50	32.1	0.81	30.9
	Pikachu	0.88	29.7	0.81	32.5	0.84	31.1
	Brad Pitt	1.00	25.1	0.36	30.1	0.87	31.4
	Lionel Messi	0.91	29.5	0.63	32.0	0.85	31.4
Styles	Van Gogh	0.96	30.5	0.79	31.6	0.83	30.9
	Monet	0.92	28.2	0.64	31.7	0.84	31.3
	Cartoon	0.99	26.8	0.66	28.6	0.76	30.9
Objects average		0.97	26.7	0.59	29.9	0.83	31.1
Characters average		0.93	29.0	0.58	31.7	0.84	31.2
Styles average		0.96	28.5	0.70	30.6	0.81	31.0
Overall average		**0.95**	**27.9**	**0.61**	**30.7**	**0.83**	**31.1**

H Performance Across Other Diffusion and Rectified-Flow Architectures

We evaluate concept unlearning across diffusion architectures and the SD3 and FLUX.1-dev rectified-flow transformers using the same evaluation protocol. The FLUX model is `black-forest-labs/FLUX.1-dev`. Table 8 reports category averages for objects, characters, and styles, with SD v1.5 as the reference model. Forgetting accuracy U_{acc} and all retention metrics are higher-is-better.

Table 8: **Category averages across base models.** Performance of TILDE on SD v1.4, SD v1.5, SD v2, SD3, and FLUX.1-dev, averaged within each concept category.

Base model	Category	Forget		Related retain		General retain	
		$U_{\mathrm{acc}} \uparrow$	$U_{\mathrm{clip}} \uparrow$	$RR_{\mathrm{acc}} \uparrow$	$RR_{\mathrm{clip}} \uparrow$	$GR_{\mathrm{acc}} \uparrow$	$GR_{\mathrm{clip}} \uparrow$
SD v1.4	Objects	0.98	0.24	0.69	0.29	0.84	0.32
	Characters	0.99	0.26	0.52	0.30	0.83	0.31
	Styles	0.84	0.25	0.65	0.29	0.82	0.31
SD v1.5	Objects	0.97	0.24	0.72	0.29	0.86	0.32
	Characters	0.99	0.28	0.64	0.32	0.87	0.32
	Styles	0.88	0.28	0.76	0.31	0.84	0.32
SD v2	Objects	0.99	0.27	0.70	0.30	0.85	0.32
	Characters	0.98	0.26	0.61	0.31	0.86	0.32
	Styles	0.87	0.25	0.59	0.30	0.85	0.32
SD3	Objects	0.96	0.25	0.82	0.32	0.93	0.33
	Characters	0.98	0.26	0.50	0.31	0.92	0.33
	Styles	0.87	0.25	0.68	0.31	0.92	0.33
FLUX.1-dev	Objects	0.97	0.29	0.85	0.31	0.93	0.32
	Characters	0.98	0.28	0.68	0.31	0.94	0.33
	Styles	0.84	0.30	0.78	0.31	0.93	0.33

These results show that TILDE transfers to other models and maintains strong general retention across objects, characters, and styles. The experiment uses preliminary SD3 hyperparameters that were not exhaustively tuned for this architecture. These settings may therefore not represent the best achievable operating point, and architecture-specific tuning may further improve the forgetting-retention trade-off.

I PER-CONCEPT QUANTITATIVE RESULTS

This section provides the full per-concept breakdown of the SD v1.5 quantitative results summarised in Table 1. FLUX.1-dev results are reported as category averages in Table 21. Each SD v1.5 table reports results for a single unlearning target across all methods, including unlearning accuracy ($U_{\text{Acc}} \uparrow$), related-retain accuracy ($RR_{\text{Acc}} \uparrow$), general-retain accuracy ($GR_{\text{Acc}} \uparrow$), and the corresponding CLIP scores ($\uparrow$).

Table 9: Concept unlearning results for **Apple**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.88	28.7	0.66	29.4	0.78	31.7
ESD-x	0.82	29.4	0.69	29.5	0.80	31.9
UCE	0.75	29.4	0.91	29.9	0.81	31.3
CA	0.69	29.6	0.97	30.7	0.90	32.8
MACE	0.31	31.4	0.84	29.6	0.80	31.9
MCE	0.44	31.8	0.83	30.8	0.83	32.5
DUO	0.98	29.4	0.85	30.7	0.89	32.7
EraseFlow	0.76	30.2	0.72	29.1	0.82	32.3
Meta	0.98	25.0	0.42	26.6	0.72	30.6
SHS	0.76	25.0	0.56	26.1	0.53	27.8
EDiff	0.51	29.9	0.82	30.9	0.88	32.5
SPEED	0.54	31.7	0.88	31.2	0.89	32.8
ScaPre	0.95	29.2	0.67	30.3	0.83	32.0
SMem	0.56	32.2	0.88	31.4	0.89	32.9
TILDE	0.97	26.8	0.75	28.4	0.85	31.9

Table 10: Concept unlearning results for **Banana**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.99	27.3	0.41	29.5	0.76	31.5
ESD-x	0.97	27.4	0.55	30.6	0.80	31.9
UCE	0.93	26.7	0.88	31.0	0.85	31.8
CA	0.91	27.3	0.94	30.9	0.90	32.6
MACE	0.69	26.8	0.72	29.7	0.81	31.8
MCE	0.79	28.9	0.79	30.7	0.81	32.2
DUO	1.00	29.4	0.85	32.2	0.88	32.7
EraseFlow	0.85	29.1	0.72	30.3	0.83	32.2
Meta	0.92	26.1	0.33	27.8	0.73	30.6
SHS	0.82	22.9	0.48	25.9	0.49	27.4
EDiff	0.66	29.3	0.89	32.1	0.88	32.5
SPEED	0.88	29.4	0.93	31.9	0.89	32.7
ScaPre	0.99	26.8	0.52	29.9	0.82	31.7
SMem	0.76	30.6	0.87	32.1	0.88	32.7
TILDE	0.98	22.6	0.78	29.4	0.86	31.9

Table 11: Concept unlearning results for **Brad Pitt**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.96	28.3	0.41	30.2	0.79	31.4
ESD-x	0.94	28.3	0.47	31.3	0.77	31.1
UCE	0.82	32.2	0.88	34.6	0.87	32.5
CA	0.99	26.9	0.70	33.7	0.88	32.4
MACE	1.00	17.2	0.00	16.6	0.02	19.5
MCE	0.94	29.5	0.42	31.5	0.83	32.5
DUO	1.00	26.2	0.28	29.2	0.86	32.4
EraseFlow	0.98	25.5	0.06	25.1	0.58	28.9
Meta	0.98	25.5	0.12	26.9	0.75	31.0
SHS	0.73	26.1	0.50	27.6	0.43	26.3
EDiff	0.39	32.4	0.88	34.0	0.89	32.4
SPEED	0.98	30.0	0.88	35.0	0.89	32.5
ScaPre	0.94	28.7	0.67	32.6	0.86	32.4
SMem	0.87	31.5	0.83	34.5	0.88	32.5
TILDE	0.98	27.2	0.55	30.3	0.88	32.2

Table 12: Concept unlearning results for **Cartoon Style**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.70	28.5	0.80	30.1	0.83	32.1
ESD-x	0.89	28.4	0.80	30.2	0.84	32.1
UCE	0.24	32.1	0.94	31.8	0.89	32.8
CA	0.97	28.4	0.79	30.5	0.87	32.5
MACE	1.00	21.4	0.00	20.9	0.02	19.3
MCE	0.79	29.3	0.73	31.1	0.84	32.5
DUO	0.23	29.4	0.78	29.8	0.88	32.5
EraseFlow	0.89	26.1	0.35	25.8	0.71	30.1
Meta	0.93	25.6	0.13	22.8	0.48	28.2
SHS	0.97	23.9	0.26	25.8	0.48	27.3
EDiff	0.49	29.6	0.88	31.2	0.87	32.5
SPEED	0.31	31.8	0.84	32.0	0.89	32.8
ScaPre	0.76	29.5	0.78	31.2	0.86	32.4
SMem	0.16	32.3	0.84	32.0	0.89	32.7
TILDE	0.82	28.9	0.80	30.4	0.82	31.8

Table 13: Concept unlearning results for **Cat**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.49	28.7	0.77	31.0	0.80	31.8
ESD-x	0.61	27.8	0.77	30.6	0.77	31.3
UCE	0.27	31.2	0.99	32.6	0.88	32.6
CA	0.33	31.3	0.99	32.6	0.89	32.9
MACE	0.12	31.2	0.95	32.2	0.85	32.4
MCE	0.12	31.0	0.90	32.4	0.83	32.5
DUO	0.96	27.9	0.85	32.1	0.87	32.6
EraseFlow	0.94	25.4	0.27	27.0	0.53	28.7
Meta	0.95	24.4	0.46	27.5	0.72	30.3
SHS	0.88	18.8	0.11	21.1	0.16	22.8
EDiff	0.18	30.9	0.98	32.4	0.87	32.5
SPEED	0.39	31.7	0.97	32.5	0.88	32.7
ScaPre	0.91	25.6	0.92	31.6	0.83	31.6
SMem	0.24	31.7	0.95	32.9	0.88	32.7
TILDE	0.96	26.3	0.82	30.9	0.87	32.1

Table 14: Concept unlearning results for **Dog**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.39	28.9	0.81	31.7	0.82	31.8
ESD-x	0.46	27.8	0.77	31.0	0.79	31.3
UCE	0.84	25.9	0.96	30.2	0.73	29.5
CA	0.23	30.8	1.00	33.2	0.91	32.8
MACE	0.21	30.0	0.94	32.8	0.86	32.4
MCE	0.07	30.8	0.95	33.3	0.85	32.7
DUO	0.81	27.5	0.76	31.6	0.87	32.5
EraseFlow	0.61	27.6	0.60	29.8	0.68	30.4
Meta	0.90	25.3	0.52	28.0	0.75	30.4
SHS	0.94	21.5	0.26	22.1	0.21	23.7
EDiff	0.12	29.6	0.95	32.6	0.88	32.4
SPEED	0.14	30.8	0.97	33.2	0.88	32.7
ScaPre	0.75	27.2	0.81	31.7	0.81	31.2
SMem	0.19	30.5	0.96	33.3	0.89	32.7
TILDE	0.94	21.8	0.59	27.4	0.80	31.2

Table 15: Concept unlearning results for **Golf Ball**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.97	27.5	0.26	28.3	0.79	31.5
ESD-x	1.00	24.4	0.13	26.5	0.75	30.9
UCE	0.30	32.9	0.65	30.5	0.88	32.6
CA	0.92	31.7	0.66	30.4	0.89	32.8
MACE	0.85	29.4	0.48	29.7	0.83	32.2
MCE	0.71	32.1	0.57	30.9	0.83	32.5
DUO	0.77	31.4	0.48	30.4	0.89	32.7
EraseFlow	0.96	29.1	0.33	28.9	0.75	31.1
Meta	1.00	24.3	0.09	25.4	0.70	30.2
SHS	0.83	23.5	0.09	23.3	0.35	25.6
EDiff	0.48	30.1	0.53	29.8	0.89	32.6
SPEED	0.09	33.2	0.71	30.5	0.90	32.8
ScaPre	0.99	27.0	0.10	27.7	0.82	32.0
SMem	0.77	31.9	0.61	30.6	0.89	32.8
TILDE	0.99	23.4	0.67	28.4	0.88	31.9

Table 16: Concept unlearning results for **Lionel Messi**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	1.00	25.9	0.31	30.4	0.77	31.7
ESD-x	0.99	24.9	0.26	28.7	0.70	30.6
UCE	1.00	22.5	0.25	26.0	0.49	26.7
CA	0.95	26.8	0.61	31.7	0.87	32.2
MACE	1.00	16.5	0.00	17.2	0.01	19.4
MCE	0.47	31.6	0.67	32.1	0.84	32.4
DUO	1.00	27.0	0.50	31.7	0.87	32.6
EraseFlow	0.99	21.3	0.08	24.3	0.45	26.9
Meta	1.00	23.9	0.19	28.3	0.74	31.1
SHS	0.94	19.4	0.12	21.5	0.26	23.5
EDiff	0.43	30.0	0.72	32.2	0.89	32.5
SPEED	0.98	29.2	0.69	32.8	0.89	32.7
ScaPre	0.99	26.2	0.44	30.1	0.75	30.3
SMem	0.83	29.0	0.54	32.0	0.85	32.6
TILDE	0.99	25.8	0.69	32.3	0.88	32.5

Table 17: Concept unlearning results for **Mickey Mouse**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.97	28.0	0.21	30.7	0.75	31.5
ESD-x	0.99	26.7	0.17	29.8	0.72	30.9
UCE	0.87	30.3	0.46	32.0	0.85	31.6
CA	0.96	29.6	0.32	32.4	0.87	32.6
MACE	1.00	20.8	0.00	17.0	0.02	19.5
MCE	0.64	33.3	0.74	34.3	0.82	32.4
DUO	0.93	29.8	0.43	32.4	0.85	32.4
EraseFlow	1.00	25.0	0.09	27.5	0.60	29.3
Meta	0.99	26.3	0.07	27.9	0.67	30.4
SHS	0.88	22.1	0.35	25.5	0.33	24.7
EDiff	0.53	30.7	0.75	34.1	0.87	32.3
SPEED	0.23	34.5	0.82	34.4	0.88	32.6
ScaPre	0.99	26.6	0.09	29.5	0.78	31.6
SMem	0.93	29.7	0.46	33.0	0.87	32.6
TILDE	0.97	28.7	0.56	31.8	0.85	31.9

Table 18: Concept unlearning results for **Pikachu**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.96	27.9	0.50	30.8	0.75	31.5
ESD-x	1.00	0.0	0.00	0.0	0.00	0.0
UCE	1.00	21.8	0.68	27.0	0.47	26.1
CA	0.97	27.9	0.74	33.8	0.87	32.4
MACE	0.96	27.9	0.66	33.6	0.85	32.3
MCE	0.78	32.6	0.80	34.6	0.83	32.4
DUO	0.96	30.9	0.63	33.2	0.85	32.3
EraseFlow	0.97	26.6	0.42	29.7	0.64	30.2
Meta	0.97	26.2	0.45	29.2	0.69	30.0
SHS	0.83	23.6	0.38	25.1	0.27	24.3
EDiff	0.72	30.8	0.78	34.2	0.87	32.2
SPEED	0.97	30.7	0.83	34.7	0.89	32.5
ScaPre	0.99	26.0	0.57	30.4	0.81	31.1
SMem	0.94	30.4	0.65	33.5	0.87	32.5
TILDE	1.00	29.9	0.76	33.0	0.88	32.1

Table 19: Concept unlearning results for **Van Gogh Style**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.76	30.1	0.78	32.9	0.84	32.1
ESD-x	0.81	29.4	0.75	32.2	0.81	31.7
UCE	0.49	32.5	0.91	33.6	0.89	32.6
CA	0.95	27.6	0.81	32.5	0.85	32.4
MACE	1.00	17.9	0.00	18.1	0.00	20.7
MCE	0.77	29.7	0.77	32.9	0.82	32.4
DUO	0.68	21.8	0.57	29.5	0.85	32.1
EraseFlow	0.84	29.2	0.61	30.9	0.71	30.4
Meta	0.43	20.2	0.43	24.2	0.49	25.8
SHS	0.68	20.9	0.00	21.4	0.04	19.7
EDiff	0.29	31.7	0.81	31.0	0.88	31.6
SPEED	0.25	33.3	0.98	33.8	0.88	32.6
ScaPre	0.55	31.5	0.84	33.1	0.86	32.5
SMem	0.35	33.9	0.91	33.5	0.88	32.6
TILDE	0.95	27.1	0.75	31.5	0.86	31.4

Table 20: Concept unlearning results for **Monet Style**.

Method	$U_{\text{Acc}} \uparrow$	$U_{\text{CLIP}} \uparrow$	$RR_{\text{Acc}} \uparrow$	$RR_{\text{CLIP}} \uparrow$	$GR_{\text{Acc}} \uparrow$	$GR_{\text{CLIP}} \uparrow$
ESD-u	0.17	30.3	0.87	31.1	0.84	31.1
ESD-x	0.41	29.5	0.57	31.0	0.81	31.6
UCE	0.12	32.0	0.93	33.5	0.82	32.0
CA	0.66	28.3	0.56	31.2	0.88	32.4
MACE	1.00	20.4	0.07	20.9	0.01	19.4
MCE	0.22	30.8	0.65	31.0	0.82	32.3
DUO	0.77	25.5	0.12	26.7	0.88	32.5
EraseFlow	0.84	28.3	0.60	29.4	0.73	30.2
Meta	0.71	24.5	0.26	26.0	0.69	29.9
SHS	0.76	22.9	0.26	24.8	0.36	25.6
EDiff	0.18	32.5	0.85	33.1	0.87	32.4
SPEED	0.19	32.5	0.97	34.0	0.88	32.6
ScaPre	0.26	32.5	0.72	32.8	0.81	32.3
SMem	0.18	32.2	0.82	32.9	0.89	32.6
TILDE	0.87	27.3	0.74	30.9	0.83	31.5

Table 21: **Category-level results on FLUX.1-dev.** Accuracy averages match Table 2. CLIP scores use the decimal presentation of Table 8; baseline CLIP scores are averaged equally across concepts and divided by 100.

Category	**Method**	**Forget**		**Related retain**		**General retain**	
		$U_{\text{acc}} \uparrow$	$U_{\text{clip}} \uparrow$	$RR_{\text{acc}} \uparrow$	$RR_{\text{clip}} \uparrow$	$GR_{\text{acc}} \uparrow$	$GR_{\text{clip}} \uparrow$
Objects	EraseAnything	0.10	0.33	0.87	0.31	0.93	0.31
	EraseFlow	0.99	0.25	0.41	0.27	0.90	0.30
	MCE	0.21	0.32	0.81	0.32	0.81	0.32
	TILDE	0.97	0.29	0.85	0.31	0.93	0.32
Characters	EraseAnything	0.62	0.31	0.54	0.31	0.94	0.31
	EraseFlow	0.97	0.23	0.27	0.26	0.84	0.30
	MCE	0.51	0.32	0.62	0.32	0.82	0.32
	TILDE	0.98	0.28	0.68	0.31	0.94	0.33
Styles	EraseAnything	0.59	0.29	0.44	0.29	0.93	0.31
	EraseFlow	0.61	0.28	0.41	0.29	0.93	0.31
	MCE	0.62	0.29	0.55	0.31	0.82	0.32
	TILDE	0.84	0.30	0.78	0.31	0.93	0.33