Locality-Aware Continual Unlearning for Diffusion Models

Naveen George[1,2†], Naoki Murata[2], Yuhta Takida[2], Konda Reddy Mopuri[1], and Yuki Mitsufuji[2,3]

[1] Indian Institute of Technology Hyderabad, India
[2] Sony AI, Japan
[3] Sony Group Corporation, Japan
ai23mtech12001@iith.ac.in,naoki.murata@sony.com

Abstract. Real-world deployment of text-to-image diffusion models requires continual concept removal as new privacy, copyright, or safety obligations arise over time. Existing unlearning methods, however, are designed for single-step deletion and collapse after only 3–5 sequential applications. We trace this instability to two compounding factors: *(i)* coarse mapping targets that cause degradation to accumulate unnecessarily across steps, and *(ii)* the absence of local protection for semantically neighboring concepts, whose shared internal representations make them the first to suffer collateral damage. Because this damage is strongest in the *local* semantic neighborhood of the forget concept, global replay alone cannot prevent it. Building on this analysis, we propose **Locality-Aware Continual Unlearning (LACU)**, a framework with two complementary mechanisms. **Locality-Aware Target Selection** chooses, for each forget prompt, the context-preserving mapping prompt that the diffusion model itself treats as most similar to the original prompt, measured by *score-prediction distance* (how differently the model denoises the same noisy image under two text conditions), ensuring each update is as small and targeted as possible. **Locality-Aware Replay** uses the same metric to identify the retain concepts closest to the forget concept in the model's own representation and replays them as a local functional regularizer, directly shielding the most vulnerable neighborhood. Combined with teacher-student distillation and lightweight ℓ_2 parameter regularization, LACU maintains stable unlearning over 10 sequential steps, preserving significantly higher related retention (RR_{acc}) and general retention (GR_{acc}) than recent baselines. The code is available at `https://github.com/SonyResearch/LACU`.

Keywords: Unlearning · Knowledge Distillation · Generative Replay

1 Introduction

Recent visual generative models such as Sora [6], Gemini [39], Imagen 3.0 [4], and DALL·E 3 [5] synthesize high-quality content at scale. Once deployed, these

† Work done during an internship at Sony AI.

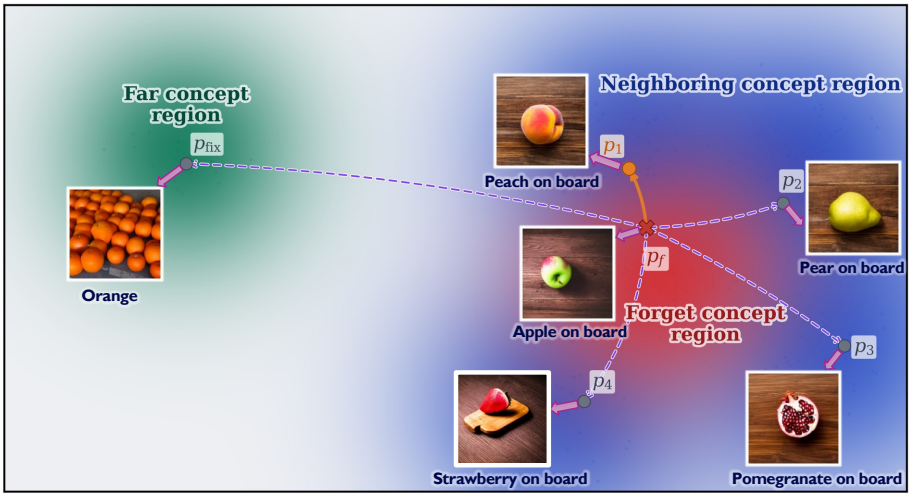

Fig. 1: Locality-Aware Target Selection in the model's score-prediction space. The landscape visualizes how the diffusion model itself perceives different prompts, where proximity is measured by *score-prediction distance* d_{score} (Eq. 2): prompts the model denoises similarly lie close together. **Red**: forget concept; **blue**: neighboring concepts; **teal**: far concept. Given a forget prompt p_f ("Apple on board" ✗), our method generates context-preserving candidates p_1–p_4 that replace only the forget concept while retaining the scene context, and selects p_1 ("Peach on board") as the mapping target because it has the smallest d_{score} to p_f. This proximity ensures only a small perturbation of the learned score field, minimizing collateral damage to neighbors. In contrast, traditional methods map all forget prompts to a fixed anchor p_{fix} ("Orange") that discards scene context, forcing a large displacement that disproportionately degrades neighboring concepts, compounding under sequential unlearning.

models need to accommodate deletion requests driven by privacy regulations [10], copyright claims, and safety policies. Because such requests arrive continuously rather than as a one-time batch, *continual unlearning* (CUL), i.e., sequentially removing concepts while preserving the remaining capabilities, is the operational baseline, not a special case.

Current unlearning methods target a single, one-shot deletion and degrade rapidly under repeated application: as we show empirically (Section 5), most baselines collapse after only 3–5 sequential steps. We hypothesize that the root cause is threefold. (1) Common mapping strategies redirect all forget prompts to the same fixed anchor (e.g., an empty string or a generic concept), which poorly approximates the ideal per-prompt mapping: different prompts containing the same forget concept can differ substantially in scene context, so a single surrogate forces unnecessarily *large score-prediction displacements* at each step. (2) These methods lack *local protection* for semantically neighboring concepts. Prior benchmarks [2, 45] show that the strongest collateral damage, called the *ripple effect*, falls on concepts that are semantically close to the forget concept because these

concepts share similar internal representations within the model. Each oversized update compounds this local distortion; after several steps, unprotected neighbors degrade into *retention collapse*. Global replay (i.e., replaying random general prompts without targeting the local neighborhood) or generic regularization cannot adequately defend these vulnerable neighbors: what is needed is a protection mechanism that focuses on the *local* semantic neighborhood of each forget concept. (3) Methods that do attempt more careful selection often rely on text-space similarity (e.g., CLIP embeddings), which does not necessarily reflect how the diffusion model itself processes different prompts; this mismatch further compounds errors under sequential application.

To address these issues, we propose **Locality-Aware Continual Unlearning (LACU)**, a framework whose two core components address locality through a single model-aware principle: measuring concept proximity relative to the forget target in the model's prediction space.

1. **Locality-Aware Target Selection.** For each forget prompt, we generate a set of context-preserving candidate replacements and select the one whose noise predictions are closest to those of the forget prompt on shared noisy states. By choosing the *nearest safe target*, each unlearning update becomes as small and precise as possible, reducing cumulative drift and limiting ripple effects on neighboring concepts.
2. **Locality-Aware Replay.** Using the same score-prediction-distance-based metric, we identify the retain concepts that are *closest to the forget concept in the model's own representation* and replay them as a local functional regularizer. This directly shields the most vulnerable neighborhood, precisely the concepts that benchmarks identify as first to break.
3. **Stable 10-step continual unlearning.** Combined with teacher-student distillation and lightweight ℓ_2 parameter regularization, LACU maintains stable performance across 10 sequential unlearning steps, preserving significantly higher related retention (RR_{acc}) and general retention (GR_{acc}) than recent baselines while effectively removing target concepts.

2 Related Work

One-Shot Unlearning in Diffusion Models. Machine unlearning has become a key post-hoc mechanism for aligning generative models with evolving safety, privacy, and copyright requirements without full retraining. Most existing approaches are single-step interventions that map forget concepts to fixed anchors. ESD [12] steers away from the target using classifier-free guidance; Concept Ablation [21] minimizes KL divergence toward a null anchor. More parameter-efficient variants target cross-attention layers [13, 44], use efficient concept erasure [16], train lightweight adapters [20, 27, 28], restrict updates via saliency [11], timestep selection [41], or reduce forget–retain gradient conflicts [32]. Complementary work analyzes gradient-ascent failures, whether erased concepts are truly removed, and how to make erasure more reliable or

adversarially robust [8, 26, 30, 38]. Despite their variety, these methods share a common weakness under repetition: *coarse replacement* strategies map diverse forget prompts to the same fixed anchor, forcing score-prediction displacements that accumulate and destabilize the model [15].

Continual Unlearning. Real deployments involve sequential removal requests, making CUL the practical setting. Recent CUL analyses identify knowledge erosion, forgetting reversal, and cumulative parameter drift as key sequential failure modes [1,17,23,31,40,43]. DUGE [40] and regularization-based methods [23] show that replay, distillation, and parameter regularization are useful backbones, but they mainly protect globally or in parameter space.

Target Selection and Mapping. The choice of *where to redirect* the forget concept substantially influences the magnitude of the resulting update. AGE [7] improves over fixed generic anchors by selecting an adaptive concept-level target from CLIP-filtered class labels through minimax optimization, supporting the view that forget concepts should be redirected to nearby non-synonym targets. However, its targets remain discrete labels chosen outside the current diffusion score field, and the update does not explicitly protect neighboring retain concepts where ripple effects concentrate. LACU addresses both issues by ranking prompt-level, context-preserving targets with score-prediction distance at each continual step and coupling the same locality measure with replay; AGE-style selection alone still yields high erasure but severe retention collapse in CUL (Section 5).

Replay and Distillation for Retention. Knowledge distillation from a frozen teacher to a student model is a proven retention mechanism in both continual learning [18, 29] and unlearning [9]. Existing replay strategies typically sample from a broad, global prompt pool, which distributes the regularization signal uniformly. However, prior benchmarks show that ripple effects concentrate in the *local neighborhood* of the forget concept [2, 45]; global replay therefore over-invests on already-safe regions and under-invests on the most fragile ones. A replay strategy that explicitly prioritizes the nearest neighbors of the forget concept can therefore provide stronger protection where it matters most.

3 Problem Formulation and Challenges

3.1 Continual Unlearning: Problem Setting

We consider continual unlearning for text-to-image diffusion models in the setting where *one forget concept is processed per step*. Let ϵ_{θ_0} be a pre-trained model and let $\{c_{\mathrm{f}}^{(1)}, c_{\mathrm{f}}^{(2)}, \ldots, c_{\mathrm{f}}^{(K)}\}$ be a sequence of forget concepts arriving over time. The model is updated sequentially as $\theta_i = \mathrm{Unlearn}(\theta_{i-1}, c_{\mathrm{f}}^{(i)})$ for $i \in \{1, \ldots, K\}$, while preserving a retain concept set C_{r} such that $c_{\mathrm{f}}^{(i)} \notin C_{\mathrm{r}}$ for all i. At each step, the update must simultaneously satisfy three requirements:

(i) *Forgetting*: prompts instantiating $c_{\mathrm{f}}^{(i)}$ should no longer produce recognizable target content.
(ii) *Retention*: non-target prompts remain behaviorally close to the previous model.

(iii) *Quality preservation*: overall image quality remains comparable to the original model.

These three axes define the design space for subsequent sections: each component of our method is motivated by which requirement it primarily serves.

3.2 Why Existing One-shot Methods Fail in CUL

We now discuss *why*, not merely how, existing unlearning methods, designed for single-step deletion, fail when applied sequentially. We identify three contributing factors, each of which motivates a specific component of our proposed framework.

Coarse Mapping Ignores Locality Most guidance-based unlearning methods (i.e., methods that steer the model's predictions away from the forget concept) employ a *prompt-to-phrase* mapping: all forget prompts are redirected to the same fixed anchor (e.g., an empty string or a generic concept) [11–13, 21, 42]. This is a poor substitute for per-prompt mapping because different prompts containing the same forget concept can differ substantially in scene context, attributes, and composition. Mapping them all to the same surrogate forces *large* score-prediction displacements, much larger than necessary if each prompt were redirected to a nearby, context-appropriate target. Under repetition, these oversized distortions accumulate into **cumulative degradation of the learned score field**, progressively eroding the model's generative fidelity [23, 40]. *Our Locality-Aware Target Selection (Section 4.4) addresses this by choosing the nearest safe mapping target for each forget prompt, minimizing the score-prediction displacement per step.*

Global Replay Cannot Protect Local Neighbors Even when methods include retention regularization (e.g., random replay or broad distillation), the regularization signal is spread uniformly across the concept space. However, the collateral damage from unlearning, the *ripple effect*, is not uniform: it falls disproportionately on concepts semantically close to the forget concept, because these concepts have similar internal representations and score predictions [2, 45]. Global replay over-invests on already-safe distant concepts and under-invests on the most vulnerable neighbors. Under repeated steps, the unprotected local neighborhood degrades fastest, causing **retention collapse**: related retention accuracy (RR_{acc}, the ability to generate concepts semantically close to the forgotten one) drops sharply even when general retention accuracy (GR_{acc}, generation quality on broadly unrelated concepts) appears relatively stable. *Our Locality-Aware Replay (Section 4.5) directly addresses this by identifying and reinforcing the concepts closest to the forget concept in the model's own representation.*

Text-Space Proximity $\neq$ Model-Space Proximity Text-space similarity (e.g., CLIP embeddings) doesn't reflect how similarly a diffusion model processes two prompts internally: prompts close in CLIP space may still induce very

different noise predictions. Moreover, each unlearning step alters the model's internal landscape, meaning proximity relationships valid before an update may no longer hold afterward. In CUL, where per-step precision is critical, such mismatches compound across steps. Our framework therefore uses score-prediction distance, computed directly in the model's noise-prediction space (Section 4.2), for all target and neighbor selection, and recomputes it at every continual step using the current frozen teacher to track the evolving model landscape.

Together, these three factors (excessive updates from coarse mapping, unprotected local neighborhoods, and proxy-based selection) explain why existing methods collapse after 3–5 sequential steps (see Table 1). Our framework addresses each factor with a corresponding component.

3.3 Denoising Trajectory as an Interpretive Lens

To build intuition for why locality-aware design helps, we briefly appeal to the denoising-trajectory viewpoint. A text-to-image diffusion model can be viewed as a score-field estimator: its noise predictions $\epsilon_\theta(z_t, t, p)$ define a vector field that guides noisy latents z_t toward clean images. Different text prompts induce different trajectories through this field. When two prompts produce *similar* noise predictions on shared noisy states, their trajectories are close, and redirecting one toward the other requires only a small perturbation of the score field. Conversely, redirecting to a *distant* target requires a large field distortion that can inadvertently deflect neighboring trajectories.

This perspective explains the benefit of both our components: (i) choosing a mapping target with minimal score-prediction distance keeps the field perturbation small, limiting drift; (ii) replaying the concepts whose trajectories are nearest to the forget trajectory directly stabilizes the region of the score field most affected by the unlearning update.

4 Method: Locality-Aware Continual Unlearning

4.1 Overview

We frame each continual unlearning step as a multi-objective teacher-student process. The frozen previous-step model $\epsilon_{\hat{\theta}_{i-1}}$ serves as the teacher; the student ϵ_{θ_i} is initialized from it and updated to satisfy three complementary objectives:

- **Unlearn** ($\mathcal{L}_{\text{unlearn}}$): redirect the model's response to forget prompts toward a safe, nearby mapping target, editing *locally and minimally* (addresses coarse-mapping failure, Section 3.2).
- **Local replay** ($\mathcal{L}_{\text{retain}}$): preserve the denoising behavior of the concepts nearest to the forget concept, which are most vulnerable to the ripple effect (addresses neighborhood failure).
- **Parameter regularization** ($\mathcal{L}_{\text{reg}}$): a lightweight ℓ_2 penalty that limits cumulative drift over many steps (addresses parameter drift across sequential steps): $\mathcal{L}_{\text{reg}} = \|\theta_i - \hat{\theta}_{i-1}\|_2^2$.

The final objective combines all three:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{unlearn}}\mathcal{L}_{\text{unlearn}} + \lambda_{\text{retain}}\mathcal{L}_{\text{retain}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{1}$$

A key design principle unifies the first two components: both use *score-prediction distance* (defined next) to make locality-aware decisions in the model's own representation space. This coupling is the main design difference from using the ingredients independently. Target selection alone, including AGE-style adaptive replacement, can make the forget update smaller but does not preserve nearby non-target trajectories during training. Conversely, global replay or parameter regularization, as used in continual-unlearning backbones, constrains drift but can over-regularize the update while still under-protecting the most affected neighbors. LACU ties the two decisions together: the forget target and the retain neighborhood are selected by the same frozen teacher and the same score-prediction-distance criterion at each continual step.

4.2 A Model-Aware Locality Metric: Score-Prediction Distance

Central to our framework is a single metric that defines "close" in the *diffusion model's own terms*. Given two text prompts p_a and p_b, a frozen model $\epsilon_{\hat{\theta}}$, and a set of K shared random noisy-latent/timestep pairs $\{(z_{t_k}^{(k)}, t_k)\}_{k=1}^{K}$, the **score-prediction distance**[4] is:

$$d_{\text{score}}(p_a, p_b; \hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}\left\|\epsilon_{\hat{\theta}}(z_{t_k}^{(k)}, t_k, p_a) - \epsilon_{\hat{\theta}}(z_{t_k}^{(k)}, t_k, p_b)\right\|_2^2. \tag{2}$$

Here K denotes the number of sampled noisy states; in practice we use K_f samples for target selection (Section 4.4) and K_r samples for replay ranking (Section 4.5). Each noisy latent $z_{t_k}^{(k)}$ is constructed by first sampling a clean latent $z_0^{(k)}$ and then applying the forward diffusion process [19] at a randomly sampled timestep t_k: $z_{t_k}^{(k)} = \sqrt{\bar{\alpha}_{t_k}}\, z_0^{(k)} + \sqrt{1-\bar{\alpha}_{t_k}}\, \epsilon^{(k)}$, where $\epsilon^{(k)} \sim \mathcal{N}(0, \mathbf{I})$.

Why random latents suffice. We sample $z_0^{(k)} \sim \mathcal{N}(0, \mathbf{I})$ instead of generating DDIM-inverted latents [37], which would require a costly multi-step inference pass per prompt. This is valid because score-prediction distance is used for *relative ranking*: all candidates share the same (z_{t_k}, t_k) pairs, so content-independent bias cancels out. We also verify empirically that the selections agree with DDIM-latent ranking in the majority of cases (see Supplementary).

This metric measures how differently the model *denoises* the same noisy input under two conditioning signals. A small distance means the model treats the two prompts similarly; redirecting from p_a to p_b therefore requires only a small perturbation of the learned score field.

[4] Strictly speaking, this is the mean squared ℓ_2 distance; we use "distance" for brevity since only relative ranking matters in our application.

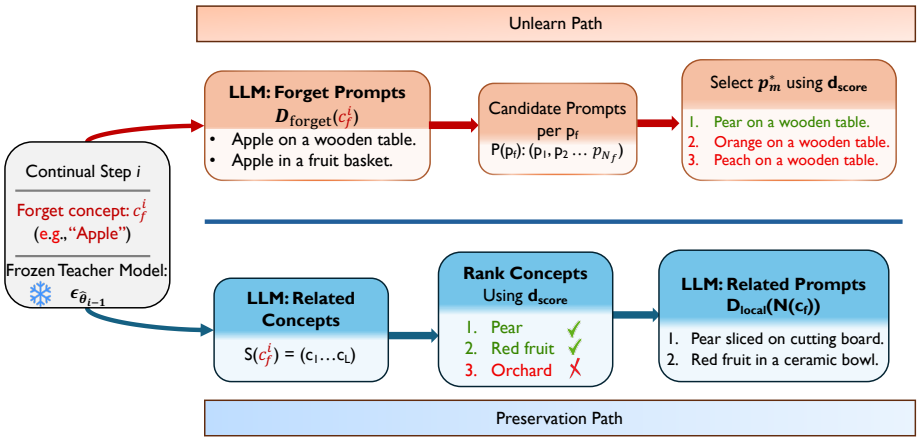

Fig. 2: Prompt generation and selection pipeline in LACU. At each continual step, the forget concept $c_{\text{f}}^{(i)}$ and the frozen teacher $\epsilon_{\hat{\theta}_{i-1}}$ drive two parallel paths. **Unlearn Path (top):** An LLM generates forget prompts and context-preserving mapping candidates per prompt; the optimal target p_{m}^* is selected via the smallest d_{score} (Eq. 3). **Preservation Path (bottom):** LLM-proposed related concepts are ranked by d_{score}; the top-N_r nearest form the local neighborhood $\mathcal{N}(c_{\text{f}})$ (✓), distant ones are discarded (✗), and replay prompts are generated for the selected neighbors.

Why not CLIP similarity? CLIP compares prompts in a text-embedding space that is disjoint from the model's conditioning and denoising pathways. Two prompts can be close in CLIP space yet induce very different noise predictions, and vice versa. Because unlearning directly modifies the model's parameters, the metric guiding target and neighbor selection should reflect the model's own noise predictions. Score-prediction distance satisfies this requirement: it operates on the same noisy latent states and model weights that the unlearning update will change, making it a *model-aware* measure of locality. We validate this empirically in Section 5, where score-prediction-distance-based selection outperforms CLIP-based ranking for both target selection and replay neighbor selection.

4.3 Prompt Design and Data Flow

At continual step i, $c_{\text{f}}^{(i)}$ denotes the **forget concept**. A **forget prompt** p_{f} instantiates it in a specific scene context; a **mapping target** p_{m}^* is the safe replacement selected for p_{f}; and a **replay prompt** p_{r} belongs to a concept in the local neighborhood of $c_{\text{f}}^{(i)}$. The complete data flow is illustrated in Fig. 2: for each step we generate forget prompts, select per-prompt mapping targets via d_{score} (Eq. 3), identify the nearest retain concepts as replay neighbors via Eq. 4, and alternate between unlearning and replay distillation with ℓ_2 regularization. LLM prompt templates and selection variants are detailed in Section 5.3 and Supplementary Section 8.

4.4 Component A: Locality-Aware Target Selection

This component addresses the coarse-mapping failure identified in Section 3.2: instead of redirecting all forget prompts to a single fixed anchor, we select a separate, context-preserving **mapping target** for each forget prompt via *prompt-to-prompt* mapping.

Example. When forgetting Apple, *"an Apple logo engraved on a silver laptop lid on a desk"* is best mapped to *"a tech company logo engraved on a silver laptop lid on a desk,"* whereas *"a sliced apple on a wooden cutting board in kitchen light"* is best mapped to *"a sliced red fruit on a wooden cutting board in kitchen light."* Although both contain the same concept name, their surrounding context differs, so prompt-level selection is more faithful than a single concept-level surrogate.

Two-Stage Procedure Selection is recomputed at every continual step using the frozen teacher $\epsilon_{\hat{\theta}_{i-1}}$, because the model's denoising landscape changes after each update; targets chosen from an earlier model can become misaligned with the current score field.

Stage 1: Candidate generation. An LLM proposes $|\mathcal{P}(p_\mathrm{f})|{=}M$ candidate replacement prompts that preserve the non-target scene content of p_f while removing the forget concept.

Stage 2: Model-aware selection. We select the candidate with the minimum score-prediction distance to the forget prompt (Eq. 2):

$$p_\mathrm{m}^*(p_\mathrm{f}) = \arg\min_{p\in\mathcal{P}(p_\mathrm{f})} d_\mathrm{score}(p_\mathrm{f},\, p;\, \hat{\theta}_{i-1}). \tag{3}$$

This yields the mapping target whose score predictions are closest to the forget prompt's, so the required score-prediction displacement is minimized.

Algorithm 1 $\mathcal{L}_\mathrm{unlearn}$: Locality-Aware Target Selection

Require: Student ϵ_{θ_i}, Teacher $\epsilon_{\hat{\theta}_{i-1}}$, forget prompts $\mathcal{D}_\mathrm{forget}$, candidate prompt generator $\mathcal{P}(\cdot)$, number of candidates M, sample count K_f

*// **Stage 1: Offline candidate selection (one-time, before training)***

for each $p_\mathrm{f} \in \mathcal{D}_\mathrm{forget}$ **do**
  Generate candidates $\mathcal{P}(p_\mathrm{f}) = \{p_1, \ldots, p_M\}$
  $p_\mathrm{m}^*(p_\mathrm{f}) \leftarrow \arg\min_{p\in\mathcal{P}(p_\mathrm{f})} d_\mathrm{score}(p_\mathrm{f},\, p;\, \hat{\theta}_{i-1})$ *// Eq. 3*

*// **Stage 2: Online training (per-iteration optimization)***

for each training iteration **do**
  Sample $p_\mathrm{f} \sim \mathcal{D}_\mathrm{forget}$; look up $p_\mathrm{m}^*(p_\mathrm{f})$
  $z_0^\mathrm{u} \sim \mathrm{DDIM}(\epsilon_{\hat{\theta}_{i-1}}, p_\mathrm{f})$
  Sample $t \sim \mathcal{U}(1,T)$, $\epsilon^\mathrm{u} \sim \mathcal{N}(0,\mathbf{I})$
  $z_t^\mathrm{u} \leftarrow \sqrt{\bar{\alpha}_t} z_0^\mathrm{u} + \sqrt{1-\bar{\alpha}_t}\epsilon^\mathrm{u}$
  $\epsilon_\mathrm{student}^\mathrm{u} \leftarrow \epsilon_{\theta_i}(z_t^\mathrm{u}, t, p_\mathrm{f})$, $\epsilon_\mathrm{teacher}^\mathrm{u} \leftarrow \epsilon_{\hat{\theta}_{i-1}}(z_t^\mathrm{u}, t, p_\mathrm{m}^*(p_\mathrm{f}))$
  $\mathcal{L}_\mathrm{unlearn} \leftarrow \|\epsilon_\mathrm{teacher}^\mathrm{u} - \epsilon_\mathrm{student}^\mathrm{u}\|_2^2$

4.5 Component B: Locality-Aware Replay

This component addresses the unprotected-neighborhood failure in Section 3.2: unlearning perturbs denoising behavior of nearby concepts most strongly, so we focus replay on the *local* neighborhood rather than distributing it globally.

Procedure For each forget concept $c_{\mathrm{f}}^{(i)}$, we first obtain a candidate pool $\mathcal{S}(c_{\mathrm{f}}^{(i)}) = \{c_1, \ldots, c_L\}$ via an LLM that enumerates semantically related concepts (the LLM provides breadth; model-aware ranking provides precision). We then rank each candidate concept c_l $(l = 1, \ldots, L)$ by its score-prediction distance to the forget concept, using the concept names directly as conditioning prompts:

$$d_l = d_{\mathrm{score}}\big(c_{\mathrm{f}},\, c_l;\, \hat{\theta}_{i-1}\big), \tag{4}$$

where d_{score} is computed with K_r sampled noisy states (Eq. 2, using K_r in place of K). We keep the N_r nearest concepts:

$$\mathcal{N}(c_{\mathrm{f}}) = \underset{c_l \in \mathcal{S}(c_{\mathrm{f}})}{\text{Top-N}_{\mathrm{r}}}\big(-d_l\big), \tag{5}$$

and generate replay prompts from $\mathcal{N}(c_{\mathrm{f}})$ for local retention distillation.

Algorithm 2 $\mathcal{L}_{\mathrm{retain}}$: Locality-Aware Replay

Require: Student ϵ_{θ_i}, Teacher $\epsilon_{\hat{\theta}_{i-1}}$, forget concept c_{f}, candidate pool $\mathcal{S}(c_{\mathrm{f}}) = \{c_1, \ldots, c_L\}$, sample count K_r, neighbor count N_r

*// **Stage 1: Offline neighbor selection (one-time, before training)***

for each $c_l \in \mathcal{S}(c_{\mathrm{f}})$ **do**
  $d_l \leftarrow d_{\mathrm{score}}\big(c_{\mathrm{f}},\, c_l;\, \hat{\theta}_{i-1}\big)$ *// Eq. 2*
$\mathcal{N}(c_{\mathrm{f}}) \leftarrow \underset{c_l \in \mathcal{S}(c_{\mathrm{f}})}{\text{Top-N}_{\mathrm{r}}}(-d_l)$
Construct local replay prompt set $\mathcal{D}_{\mathrm{local}}(\mathcal{N}(c_{\mathrm{f}}))$

*// **Stage 2: Online training (per-iteration optimization)***

for each training iteration **do**
  Sample $p_{\mathrm{r}} \sim \mathcal{D}_{\mathrm{local}}(\mathcal{N}(c_{\mathrm{f}}))$
  $z_0^{\mathrm{r}} \sim \mathrm{DDIM}(\epsilon_{\hat{\theta}_{i-1}}, p_{\mathrm{r}})$
  Sample $s \sim \mathcal{U}(1, T)$, $\epsilon^{\mathrm{r}} \sim \mathcal{N}(0, \mathbf{I})$
  $z_s^{\mathrm{r}} \leftarrow \sqrt{\bar{\alpha}_s} z_0^{\mathrm{r}} + \sqrt{1 - \bar{\alpha}_s}\epsilon^{\mathrm{r}}$
  $\epsilon_{\mathrm{student}}^{\mathrm{r}} \leftarrow \epsilon_{\theta_i}(z_s^{\mathrm{r}}, s, p_{\mathrm{r}})$, $\epsilon_{\mathrm{teacher}}^{\mathrm{r}} \leftarrow \epsilon_{\hat{\theta}_{i-1}}(z_s^{\mathrm{r}}, s, p_{\mathrm{r}})$
  $\mathcal{L}_{\mathrm{retain}} \leftarrow \|\epsilon_{\mathrm{teacher}}^{\mathrm{r}} - \epsilon_{\mathrm{student}}^{\mathrm{r}}\|_2^2$

5 Experiments

We evaluate whether locality-aware design, using score-prediction distance for both target selection and replay, improves sequential unlearning stability. Specifically, we test two hypotheses: (i) locality-aware target selection reduces cumulative drift, and (ii) locality-aware replay improves neighborhood retention. This section details the setup and metrics used to test both.

Table 1: Quantitative comparison of LACU against existing unlearning methods under continual unlearning. All methods start from the same Stable Diffusion v1.5 checkpoint (**SD v1.5 base: Accuracy 89%, CLIP score 32.6**) and unlearn the same 10 concepts; results are averaged over 5 random orderings to eliminate order dependence. Baselines that achieve high U_{acc} (e.g., EraseFlow, ANT, UCE, AGE) suffer severe retention collapse ($RR_{\text{acc}}/GR_{\text{acc}} \to 0$) over sequential steps, while methods that preserve retention (e.g., DUGE, Sculpt. Mem) sacrifice unlearning accuracy. **LACU balances both retention and unlearning.**

Method	After 1 Unlearning				After 5 Unlearning				After 10 Unlearning			
	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$
ESD-u [12]	0.97	29.1	0.60	0.75	0.83	24.7	0.30	0.35	0.93	21.3	0.13	0.15
ESD-x [12]	0.99	27.3	0.55	0.77	0.93	23.0	0.29	0.28	0.96	19.0	0.09	0.08
UCE [13]	0.99	21.2	0.65	0.39	0.86	21.5	0.30	0.26	0.96	19.5	0.08	0.05
MACE [27]	0.99	28.2	0.69	0.83	0.94	18.5	0.06	0.03	0.97	19.4	0.03	0.02
Meta [14]	0.97	27.76	0.45	0.78	0.95	20.1	0.10	0.09	1.00	18.2	0.02	0.01
EraseFlow [22]	0.98	23.9	0.27	0.53	0.92	18.0	0.07	0.02	0.97	19.4	0.03	0.02
ANT [25]	0.98	27.2	0.49	0.75	1.00	18.1	0.02	0.01	1.00	20.5	0.00	0.00
CA [21]	0.95	28.1	0.73	0.84	0.89	27.5	0.58	0.64	0.90	22.8	0.34	0.34
AGE [7]	0.99	25.2	0.49	0.82	0.99	21.4	0.36	0.55	0.99	19.1	0.15	0.23
DUGE [40]	0.96	29.7	0.70	0.85	0.69	32.6	0.73	0.80	0.74	30.7	0.62	0.73
SMem [24]	0.96	31.4	0.70	0.84	0.64	31.5	0.72	0.78	0.73	29.3	0.58	0.67
Grad Proj [23]	0.99	26.3	0.68	0.85	0.86	28.7	0.69	0.78	0.83	28.1	0.61	0.71
LACU (Ours)	0.99	30.0	0.84	0.89	0.89	29.7	0.80	0.87	0.90	29.1	0.72	0.87

5.1 Experimental Settings

Base Model and Training. All experiments are performed on the Stable Diffusion v1.5 model [35], which is the commonly used base model for experimentation (see Supplementary for other SD models and full training details). Each unlearning step is trained for 250–350 optimizer steps using AdamW on 3 NVIDIA H100 GPUs across the 10-concept sequence; each unlearning update takes approximately 60 minutes. Timesteps t and s are sampled uniformly from 0 to 600.

Prompt Generation and Mapping. To construct the datasets for unlearning and preservation, we use an automated prompt-construction pipeline to generate diverse and semantically rich prompts.

- **Forget and Candidate Mapping Prompts:** For each concept targeted for unlearning, we generated N_f=100 unique forget prompts ($\mathcal{D}_{\text{forget}}$). For each forget prompt, we construct a candidate pool of M=10 context-preserving mapping prompts. The LLM prompt templates explicitly instruct exclusion of the forget concept while preserving context. We then select the final prompt-level mapping target using Eq. 3 with K_f scoring samples.
- **Retain Prompts:** For each forget concept, we obtain related candidate concepts via an LLM, then rank their proximity in noise-prediction space using Eq. 4 with K_r scoring samples. We keep the top-N_r=10 nearest concepts, generate replay prompts from them, and mix these with random global prompts for broader coverage. In ablations, we compare this score-prediction-distance-based ranking against CLIP-based neighbor ranking.

The LLM is used only to populate candidate pools; the LACU objective itself is agnostic to how these candidates are produced. In settings where LLM use is undesirable, the same score-prediction-distance selection can operate over templates, curated vocabularies, or human-provided candidate prompts and related concepts. In a no-LLM variant, we replace the candidate generator with template-based forget/mapping prompts and a same-category pool of related concepts. Its 10-step U_{acc}, RR_{acc}, and GR_{acc} are 0.85, 0.81, and 0.87, respectively, indicating that model-aware scoring and local replay drive the retention stability.

Unlearning Targets. We evaluated our method by sequentially unlearning a diverse set of 10 concepts, comprising objects, artistic styles, and famous individuals: *Pikachu, Brad Pitt, Golf Ball, Van Gogh Style, Apple, Spiderman, Lionel Messi, Cartoon Style, Banana* and *Mickey Mouse*. To eliminate dependence on a particular deletion order, we run each method over 5 different random permutations of these 10 concepts and report the average metrics.

5.2 Evaluation Metrics

We evaluate across three axes: unlearning efficacy, neighborhood retention, and global retention. Our evaluation extends UnlearnCanvas and EraseBench protocols [2, 45].

Unlearning and Retention Accuracy. To automate and standardize evaluation, we use the Qwen2.5-VL-7B-Instruct model [3] as a VLM judge, querying it with binary questions on generated images. The VLM is used only for evaluation; target and replay selection are decided by score-prediction distance in the diffusion model's prediction space. VLM-based evaluation has become standard practice in recent T2I unlearning studies, with SEE, EraseBench, CA, SA, and UnGuide using VLMs as classifiers or judges [2, 17, 21, 33, 36]. We found VLMs to be substantially more reliable than ImageNet-style classifiers or UnlearnCanvas heads, which struggle with diverse concepts (celebrities, artistic styles, etc.). A judge-sensitivity check with alternative VLMs changed the key rows by at most $\pm 2\%$ and preserved the relative ranking of LACU and the baselines. All evaluation prompts are unseen during training.

- **Unlearning Accuracy (U_{acc}) & CLIP Score (U_{clip}):** U_{acc} measures forgetting effectiveness. U_{clip} measures "in-prompt retainability" [34], ensuring benign parts of the prompt are still generated correctly. The ideal outcome is both high U_{acc} and high U_{clip}.
- **Related Retention Accuracy (RR_{acc}) & CLIP Score (RR_{clip}):** RR_{acc} measures the unintended impact on semantically adjacent concepts [2, 45] (e.g., testing "Impressionism" after unlearning "Van Gogh"). RR_{clip} ensures text-to-image alignment for these concepts is preserved.
- **General Retention Accuracy (GR_{acc}) & CLIP Score (GR_{clip}):** GR_{acc} assesses overall knowledge preservation on general unrelated prompts; GR_{clip} verifies text-to-image alignment remains intact.

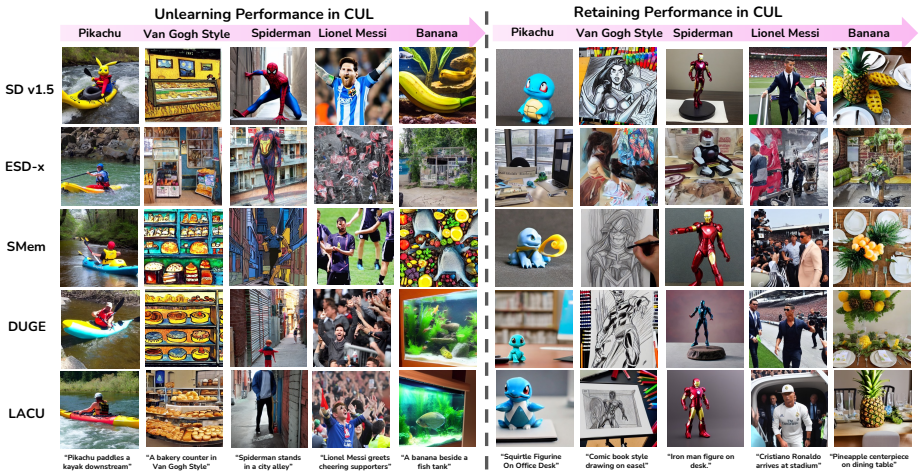

Fig. 3: Qualitative comparison after 10 sequential unlearning steps. Baselines like ESD-x [12] suffer catastrophic retention collapse on retained concepts (right), while DUGE [40] and SMem [24] avoid total collapse but still degrade. Our method (bottom row) effectively unlearns target concepts (left) while preserving generative quality on retained concepts (right).

The main tables report the compact accuracy view, while the Supplementary reports the full CLIP-score breakdown (U_{clip}, RR_{clip}, and GR_{clip}) for all steps and methods.

Generation Quality: We adopt Fréchet Inception Distance (FID) to evaluate image quality and diversity, following UnlearnCanvas benchmark methodology.

5.3 Experimental Results and Analysis

We have conducted extensive evaluations to assess the effectiveness of our locality-aware framework in the continual unlearning setup.

How well do existing methods perform under sequential requests? Across 10 sequential steps, existing methods exhibit a clear stability–efficacy trade-off. Aggressive methods (ESD, ANT, EraseFlow) show near-perfect unlearning but destroy retention via compounding parameter drift. Conservative methods (DUGE, SMem) preserve retention but sacrifice unlearning efficacy. This confirms that one-shot objectives do not transfer to sequential operation. Most one-shot baselines use fixed anchors or mapping concepts, so repeated updates accumulate coarse redirections. AGE [7] improves target choice with an adaptive target, but Table 1 shows that target selection alone still performs poorly in CUL because nearby non-target score trajectories remain unprotected.

Does our framework improve continual unlearning? Yes. Our results show consistent gains across all core criteria. In $GR_{\text{acc}}/RR_{\text{acc}}$ trend curves over

Table 2: Comparison of mapping-target and replay-neighbor selection variants after 1, 5, and 10 unlearning steps. Naive/far mappings deteriorate rapidly, while score-prediction-distance-based selection (bottom row) provides the strongest long-term balance, notably the highest RR_{acc} and GR_{acc} at 10 steps.

Method	After 1 Unlearning Step				After 5 Unlearning Steps				After 10 Unlearning Steps			
	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$	$U_{\text{acc}}\uparrow$	$U_{\text{clip}}\uparrow$	$RR_{\text{acc}}\uparrow$	$GR_{\text{acc}}\uparrow$
LACU (naive)	0.89	24.2	0.88	0.88	0.84	22.6	0.57	0.57	0.86	22.9	0.41	0.55
LACU (LLM)	0.99	28.7	0.86	0.87	0.92	29.5	0.79	0.80	0.95	28.1	0.60	0.76
LACU (CLIP)	0.94	30.3	0.81	0.87	0.88	29.6	0.77	0.81	0.90	28.2	0.61	0.78
LACU (score)	1.00	30.0	0.84	0.89	0.90	29.8	0.81	0.87	0.92	29.2	0.74	0.87

10 steps, our method remains substantially flatter and higher than baselines, indicating stronger stability over sequential steps. These trends support the claim that locality-aware target selection and replay reduce cumulative distortion under repeated unlearning.

The dynamic nature of LACU also avoids a failure case of fixed-anchor methods. If a concept previously used as a mapping target later becomes a forget concept, a fixed mapping can degenerate into an ineffective self-mapping. LACU instead re-scores candidate trajectories with the current frozen teacher at each step and selects a new nearest safe target, so the model can still remove the newly requested concept; in our runs this produced no degradation in U_{acc}, with only slightly slower convergence.

How does the choice of mapping target and replay neighbors affect continual unlearning? Table 2 compares four selection strategies within the same LACU pipeline: naive/fixed anchors, direct LLM replacements, CLIP-ranked, and score-prediction-distance-ranked (ours). Naive mappings degrade fastest because a single generic anchor forces large score-prediction displacements that compound over steps. CLIP-ranked variants improve over naive and LLM but still drift as CLIP similarity is decoupled from the model's denoising dynamics. Score-prediction-distance selection achieves the strongest long-term balance because it directly measures how differently the model processes two prompts, ensuring each update is minimally disruptive. This confirms that *model-aware* selection is essential for stability under repeated unlearning.

Table 3: Ablation after 10 CUL steps. Replay and regularization together are needed to balance unlearning and retention.

Method	U_{acc}	U_{clip}	RR_{acc}	GR_{acc}
$\mathcal{L}_{\text{unl}}$ only	0.94	27.2	0.32	0.67
$\mathcal{L}_{\text{unl}}+\mathcal{L}_{\text{ret}}$	0.91	28.6	0.61	0.85
$\mathcal{L}_{\text{unl}}+\mathcal{L}_{\text{reg}}$	0.92	28.0	0.39	0.78
LACU	0.92	29.2	0.74	0.87

How does our method achieve this, and what is the contribution of each objective? Table 3 ablates each component after 10 CUL steps, with all variants using score-prediction-distance-based selection. $\mathcal{L}_{\text{unlearn}}$ **alone** forgets effectively but suffers retention collapse from compounding ripple effects. Adding replay ($\mathcal{L}_{\text{retain}}$) directly counters this by forcing the student to mimic the

teacher on nearby concepts, providing functional stability in the local neighborhood. Adding only regularization ($\mathcal{L}_{\text{reg}}$) prevents cumulative parameter drift but provides weaker retention recovery since it does not explicitly protect vulnerable neighbors. The **full LACU** combines both: replay shields the local neighborhood while regularization constrains long-term parameter drift, achieving the best overall balance across all metrics.

6 Conclusion

We presented Locality-Aware Continual Unlearning (LACU), a framework that stabilizes sequential concept removal in diffusion models by grounding both target selection and replay in a single model-aware principle: score-prediction distance. Locality-Aware Target Selection chooses the nearest safe mapping target for each forget prompt, keeping the score-prediction displacement as small as possible and limiting cumulative degradation of the learned score field. Locality-Aware Replay identifies and reinforces the retain concepts closest to the forget concept in the model's noise-prediction space, directly shielding the neighborhood where collateral damage concentrates. Together with teacher-student distillation and ℓ_2 regularization, these components yield stable 10-step continual unlearning with substantially higher related retention (RR_{acc}) and general retention (GR_{acc}) than existing baselines, demonstrating that locality-aware design is key to stability over sequential steps.

Limitations. Our implementation uses an LLM to generate candidate prompts and related concepts, although the core score-prediction-distance selection and training objectives do not depend on LLMs. The quality and coverage of the candidate pool, whether generated automatically or curated manually, bounds the effectiveness of downstream model-aware selection. Local replay also introduces an inherent precision trade-off when the forget concept and retained neighbors share heavily overlapping visual features. LACU mitigates this by regularizing only the nearest neighbor trajectories rather than the full retain distribution, which keeps object and identity unlearning strong, but style concepts remain harder because their features are more broadly shared. More broadly, current continual unlearning methods, including ours, have been evaluated on a limited number of sequential deletions, whereas real-world deployment may require hundreds or thousands of removals over a model's lifetime; developing methods that scale gracefully to such regimes while maintaining strong retention remains an important open challenge. Finally, the teacher-student distillation framework requires maintaining a frozen copy of the model alongside the trainable student, increasing memory and computation relative to single-model approaches.

Acknowledgements

We sincerely thank Satoshi Hayakawa for providing valuable feedback prior to submission.

References

1. Adhikari, S., Kumaravelu, V., Srijith, P.: An unlearning framework for continual learning. arXiv preprint arXiv:2509.17530 (2025)
2. Amara, I., Humayun, A.I., Kajic, I., Parekh, Z., Harris, N., Young, S., Nagpal, C., Kim, N., He, J., Vasconcelos, C.N., Ramachandran, D., Farnadi, G., Heller, K., Havaei, M., Rostamzadeh, N.: 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 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
5. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. OpenAI **2**(3), 8 (2023), `https://cdn.openai.com/papers/dall-e-3.pdf`, accessed 1 March 2026
6. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)
7. Bui, A.T., Vu, T.T., Vuong, L.T., Le, T., Montague, P., Abraham, T., Kim, J., Phung, D.: Fantastic targets for concept erasure in diffusion models and where to find them. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=tZdqL5FH7w`, accessed 1 March 2026
8. Chen, R., Guo, H., Wang, L., Zhang, C., Nie, W., Liu, A.A.: Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18927–18936 (2025)
9. Chen, T., Zhang, S., Zhou, M.: Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models. In: The Thirteenth International Conference on Learning Representations (2025), `https://openreview.net/forum?id=gjwhDHeAsz`, accessed 1 March 2026
10. European Parliament and Council of the European Union: Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (2016), `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, general Data Protection Regulation (GDPR), applicable from 25 May 2018; accessed 1 March 2026
11. Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In: The Twelfth International Conference on Learning Representations (2024), `https://openreview.net/forum?id=gn0mIhQGNM`, accessed 1 March 2026
12. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2426–2436 (2023)
13. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5111–5120 (2024)

14. Gao, H., Pang, T., Du, C., Hu, T., Deng, Z., Lin, M.: Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2131–2141 (2025)
15. George, N., Dasaraju, K.N., Chittepu, R.R., Mopuri, K.R.: 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)
16. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models. In: Computer Vision – ECCV 2024. pp. 73–88 (2024). https://doi.org/10.1007/978-3-031-73668-1_5, https://doi.org/10.1007/978-3-031-73668-1_5, accessed 1 March 2026
17. Heng, A., Soh, H.: Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems **36**, 17170–17194 (2023)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Huang, C.P., Chang, K.P., Tsai, C.T., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In: European Conference on Computer Vision. pp. 360–376. Springer (2024)
21. Kumari, N., Zhang, B., Wang, S.Y., Shechtman, E., Zhang, R., Zhu, J.Y.: Ablating concepts in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22691–22702 (2023)
22. Kusumba, N.S.A., Patel, M., Min, K., Kim, C., Baral, C., Yang, Y.: Eraseflow: Learning concept erasure policies via GFlownet-driven alignment. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=igB289kbej, accessed 1 March 2026
23. Lee, J., Mai, Z., Yoo, J., Fan, C., Zhang, C., Chao, W.L.: Continual unlearning for text-to-image diffusion models: A regularization perspective. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=BsY2Or9FOM, accessed 1 March 2026
24. Li, G., Xiao, Y., Ji, J., Deng, K., Hui, B., Guo, L., Ma, X.: Sculpting memory: Multi-concept forgetting in diffusion models via dynamic mask and concept-aware optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19659–19668 (2025)
25. Li, L., Lu, S., Ren, Y., Kong, A.W.K.: Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9257–9266 (2025)
26. Lu, K., Kriplani, N., Gandikota, R., Pham, M., Bau, D., Hegde, C., Cohen, N.: When are concepts erased from diffusion models? In: 39th Conference on Neural Information Processing Systems (NeurIPS) (2025)
27. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6430–6440 (2024)
28. Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., Ding, G.: One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7559–7568 (2024)

29. Masip, S., Rodriguez, P., Tuytelaars, T., Ven, G.M.v.d.: Continual learning of diffusion models with generative distillation. In: Lomonaco, V., Melacci, S., Tuytelaars, T., Chandar, S., Pascanu, R. (eds.) Proceedings of The 3rd Conference on Lifelong Learning Agents. Proceedings of Machine Learning Research, vol. 274, pp. 431–456. PMLR (29 Jul–01 Aug 2025), https://proceedings.mlr.press/v274/masip25a.html, accessed 1 March 2026
30. Mavrothalassitis, I., Puigdemont, P., Levi, N., Cevher, V.: Ascent fails to forget. Advances in Neural Information Processing Systems **38**, 45330–45365 (2025)
31. Park, E.J., Shin, Y., Woo, S.S.: Robust continual unlearning against knowledge erosion and forgetting reversal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7978–7987 (2026)
32. Patel, G., Qiu, Q.: Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4211–4221 (2025)
33. Polowczyk, A., Polowczyk, A., Malarz, D., Kasymov, A., Mazur, M., Tabor, J., Spurek, P.: Unguide: Learning to forget with lora-guided diffusion models. arXiv preprint arXiv:2508.05755 (2025)
34. Ren, J., Chen, K., Cui, Y., Zeng, S., Liu, H., Xing, Y., Tang, J., Lyu, L.: Six-cd: Benchmarking concept removals for text-to-image diffusion models. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28769–28778 (2025)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
36. Saha, S., Saha, S., Gaur, M., Gokhale, T.: Side effects of erasing concepts from diffusion models. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 14991–15007. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.findings-emnlp.810, https://aclanthology.org/2025.findings-emnlp.810/, accessed 1 March 2026
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP, accessed 1 March 2026
38. Srivatsan, K., Shamshad, F., Naseer, M., Patel, V.M., Nandakumar, K.: Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23765–23774 (2025)
39. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
40. Thakral, K., Glaser, T., Hassner, T., Vatsa, M., Singh, R.: Continual unlearning for foundational text-to-image models without generalization erosion. arXiv preprint arXiv:2503.13769 (2025)
41. Wu, J., Harandi, M.: Scissorhands: Scrub data influence via connection sensitivity in networks. In: European Conference on Computer Vision. pp. 367–384. Springer (2024)
42. Wu, J., Le, T., Hayat, M., Harandi, M.: Erasing undesirable influence in diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28263–28273 (2025)

43. Wuerkaixi, A., Wang, Q., Cui, S., Xu, W., Han, B., Niu, G., Sugiyama, M., Zhang, C.: Adaptive localization of knowledge negation for continual llm unlearning. In: Forty-second International Conference on Machine Learning (2025)
44. Zhang, G., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1755–1764 (2024)
45. Zhang, Y., Fan, C., Zhang, Y., Yao, Y., Jia, J., Liu, J., Zhang, G., Liu, G., Kompella, R.R., Liu, X., Liu, S.: Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In: Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)