TOWARDS BOUNDED CONTINUOUS CONCEPT ERASURE IN TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) diffusion models excel at generating high-fidelity images, yet they raise significant concerns regarding copyright infringement and unsafe content. Concept erasure mitigates these risks by removing specific semantics from a pretrained model. In real-world deployments, however, erasure requests often arrive sequentially rather than being known in advance. This **continuous concept erasure** setting requires a model to suppress each new target, retain all previous erasures, and preserve its generation ability on non-target concepts. Existing methods, primarily designed for isolated edits, struggle with the accumulated drift and historical erasure degradation caused by repeated updates. To address these challenges, we propose **L**yapunov-constrained **C**ontinuous **C**oncept **E**rasure (LCCE). LCCE formulates continuous erasure as an online control problem, using a dynamic virtual queue to accumulate preservation-budget violations and adaptively regulate subsequent edits. A historical retention loss further maintains the erased states established for previous targets. To provide the queue with a consistent preservation signal, LCCE identifies a semantic neighborhood for each target and incrementally consolidates these neighborhoods into a cumulative preservation bank. The resulting per-stage objective admits a closed-form solution, enabling efficient sequential editing without iterative fine-tuning. Comprehensive experiments demonstrate that LCCE outperforms state-of-the-art baselines. Even after sequentially erasing 100 artist concepts, LCCE reliably suppresses the target styles while maintaining high generative fidelity on non-target concepts. Our code will be released upon acceptance.

1 INTRODUCTION

In recent years, text-to-image (T2I) diffusion models, such as Stable Diffusion, have achieved remarkable progress in generating high-quality and diverse images, catalyzing rapid advancements in fields like creative design and virtual content generation (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Zhang et al., 2023; Ruiz et al., 2023; Nichol et al., 2021). However, these models are typically trained on massive, internet-derived datasets, inevitably learning and replicating copyrighted artistic styles, unsafe content, and biased concepts, thereby raising significant legal and ethical concerns (Carlini et al., 2023; Somepalli et al., 2023; Birhane et al., 2021; Cho et al., 2023; Luo et al., 2024; Li et al., 2025a; Chen et al., 2025b). In response, concept erasure has emerged, aiming to mitigate these risks by removing specific semantics while preserving the model's general generation capability. To efficiently erase concepts while retaining pre-trained knowledge, current mainstream methods typically adopt a joint optimization strategy, which employs an erasure loss to guide the model to maximally forget the target concept, alongside a preservation loss to constrain the representation deviation on non-target domains.

However, practical deployments increasingly demand continuous knowledge updates, as new artistic styles constantly emerge and legal definitions of harmful content shift over time (Zhang et al., 2024; Nguyen et al., 2025; Li et al., 2025b; Xie et al., 2025). Erasing all risky concepts in a single batch is highly impractical; instead, new erasure requests dynamically arise over time. Consequently, a practical system must support **continuous concept erasure** (Han et al., 2024; Lee et al., 2026), meaning the model must guarantee the persistent efficacy of each erasure request for the targeted concepts while strictly maintaining the generative fidelity of non-target domains. Existing methods, which primarily focus on the efficacy of single-step erasure, struggle severely in this continuous setting. Treating each sequential task as an isolated optimization neglects the cumulative distortion of the

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.

model's latent space; successive edits lead to an uncontrolled accumulation of errors, resulting in severe degradation of generative fidelity (Lee et al., 2026). Furthermore, choosing a new preservation set per request causes successive preservation losses to be measured over different semantic domains, making non-target drift hard to track and control consistently during editing. As a result, after multiple sequential erasures, models like ESD (Gandikota et al., 2023) and UCE (Gandikota et al., 2024) suffer from catastrophic forgetting, while MACE (Lu et al., 2024) inadvertently fails to reliably suppress historically erased targets due to a lack of long-term state locking. Recent continual unlearning methods mitigate these effects with stage-wise regularization or replay (Lee et al., 2026; George et al., 2026), yet each request still requires gradient-based optimization, and drift is controlled only within each stage rather than over the whole request stream.

To address these fundamental challenges, we propose **L**yapunov-constrained **C**ontinuous **C**oncept **E**rasure (LCCE), a framework designed to strictly govern the long-term generative integrity of diffusion models. We reformulate continuous concept erasure as an online control problem via Lyapunov optimization. By utilizing a dynamic virtual queue and a historical retention loss, LCCE adaptively balances ongoing erasure efficacy with historical state stability, thereby keeping the long-term average preservation loss within a prescribed budget. Furthermore, we introduce a cumulative preservation bank that incrementally merges target semantic neighborhoods and removes erased concepts, providing the queue with a consistent preservation signal without discarding non-target knowledge. Crucially, each stage admits a closed-form solution, so LCCE erases a concept in about 8 seconds, compared with 20 minutes for CUIG (Lee et al., 2026). We conduct extensive experiments on the widely-used Stable Diffusion v1.4 (Rombach et al., 2022) model across various continuous erasure tasks, including erasing artistic styles and object categories. Results demonstrate that LCCE achieves state-of-the-art performance. Even after continuously eradicating over 100 concepts, it robustly suppresses all targets while maintaining remarkably high generative fidelity for unrelated concepts, with minimal computational overhead. The main contributions of this paper are threefold:

- We formulate continuous concept erasure as a Lyapunov-constrained online control problem. A dynamic virtual queue records accumulated preservation violations and adaptively regulates non-target preservation across sequential editing stages.
- We introduce a cumulative preservation bank that consolidates the target-specific semantic neighborhoods observed throughout the request stream. Together with historical retention loss, this design preserves non-target knowledge while maintaining the erased states of previous targets.
- We conduct comprehensive experiments across multiple continuous concept erasure tasks. Results demonstrate that LCCE significantly outperforms state-of-the-art baselines, robustly suppressing all targets and preserving generative fidelity even after sequentially erasing over 100 concepts.

2 Related Work

Text-to-Image Diffusion Models. T2I diffusion models generate diverse and detailed images from text prompts (Saharia et al., 2022; Ramesh et al., 2022; Podell et al., 2024; Chen et al., 2023). Stable Diffusion (Rombach et al., 2022) denoises in the latent space of a pretrained autoencoder and serves as the standard open-source backbone for concept erasure.

Concept Erasure in T2I Diffusion Models. Despite the remarkable generative capabilities of T2I diffusion models, they inevitably remember and synthesize unauthorized or unsafe content, including copyrighted artwork and inappropriate images (Xie et al., 2025; Tsai et al., 2023). To mitigate these risks, concept erasure has emerged as a crucial mechanism. Existing approaches primarily fall into two paradigms. *Training-based methods* (Gandikota et al., 2023; Kumari et al., 2023; Fan et al., 2024; Shirkavand et al., 2025; Bui et al., 2025) continuously fine-tune model weights using specialized objective functions to eliminate target distributions. Conversely, *Editing-based methods* (Gandikota et al., 2024; Lu et al., 2024; Wang et al., 2025b; Gong et al., 2024; Lee et al., 2025) analytically update the key-value projection matrices within cross-attention layers, circumventing iterative training via closed-form solutions to significantly enhance efficiency. Other works use plug-in adapters (Huang et al., 2024) or text-encoder interventions (Chen et al., 2025a).

Continuous Concept Erasure. In practical deployments, erasure requests arrive over time, so the same model must be edited repeatedly (Han et al., 2024; Lee et al., 2026). Recent studies have attempted to tackle this challenge. CCRT (Han et al., 2024) anchors the semantic space via LLM-guided prompt calibration, CUIG (Lee et al., 2026) constrains parameter drift via add-on regular-

ization and preserves semantically close concepts through gradient projection, and LACU (George et al., 2026) protects semantic neighbors through locality-aware target selection and replay. These methods regulate drift with a penalty defined at each stage and run gradient-based optimization for every request. LCCE instead enforces a preservation budget over the entire request stream through Lyapunov optimization, following bounded sequential editing in large language models (Wang et al., 2025a), and realizes each stage with a closed-form update.

3 PRELIMINARY

3.1 DIFFUSION MODELS AND CONDITIONAL INTERFACE

Diffusion models generate images by reversing a noise-adding process with a denoising network $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$, where $\mathbf{c}$ is the text condition. In a cross-attention layer, the text condition is projected into keys $\mathbf{k} = W_K\mathbf{c}$ and values $\mathbf{v} = W_V\mathbf{c}$, where W_K and W_V are the key and value projection matrices. LCCE sequentially updates these projections, denoted generically by W.

3.2 CONTINUOUS CONCEPT ERASURE

We consider a sequence of erasure requests indexed by $t \in \{1, \dots, T\}$. At stage t, the current request is represented by $\mathcal{T}_t = \{C_t, V_t^*\}$, where $C_t \in \mathbb{R}^{d\times n_t}$ contains the target text representations and $V_t^* \in \mathbb{R}^{d'\times n_t}$ denotes their desired safe responses, obtained by projecting a neutral safe target with the current weights $W(t-1)$ (*picture* for artists, *person* for celebrities, and an empty prompt for objects; Appendix B.3). Let $W(0)$ be the original projection and $W(t)$ the projection after the t-th edit.

Continuous concept erasure must satisfy two coupled requirements. First, the current target should be mapped to its safe response while previously erased concepts remain suppressed. Second, the edit should preserve the representations of relevant non-target concepts. At stage t, we therefore maintain an active anchor set $\mathcal{A}_t$, assembled from the local neighborhoods observed up to stage t, with representations $C_{\mathcal{A}_t} \in \mathbb{R}^{d\times m_t}$, and define its reference output using the original projection as $V_{\mathcal{A}_t} = W(0)C_{\mathcal{A}_t}$. The corresponding preservation loss, measured against $W(0)$, is

$$PL(t) = \frac{1}{m_t} \left\| W(t)C_{\mathcal{A}_t} - V_{\mathcal{A}_t} \right\|_F^2 . \tag{1}$$

The local anchor construction and the long-term control of this loss are described in Section 4.

3.3 LIMITATIONS OF MAINSTREAM EDITING METHODS

Mainstream methods such as UCE solve each edit as a single-step dual-objective problem:

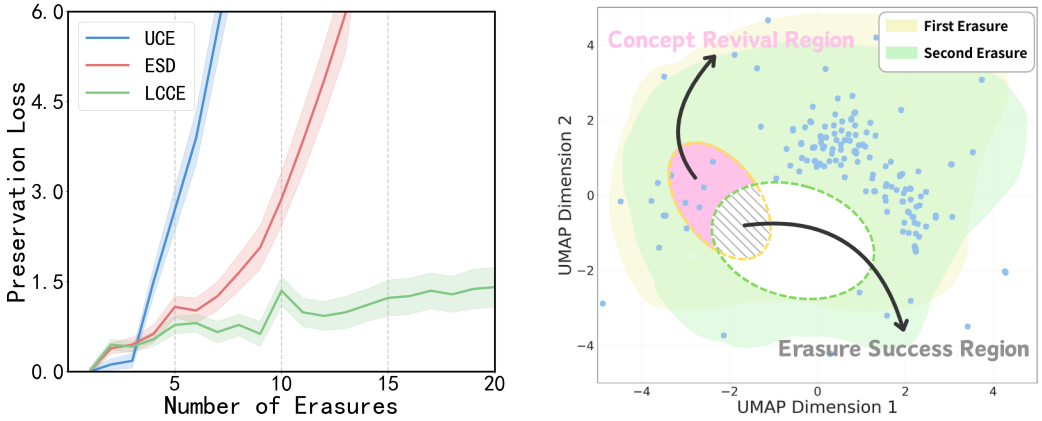
$$\min_{W(t)} \underbrace{\|W(t)C_t - V_t^*\|_F^2}_{EL(t)} + \lambda \underbrace{\|W(t)C_P - W(t-1)C_P\|_F^2}_{PL_{\text{prev}}(t)} . \tag{2}$$

While effective for a single edit, this scheme faces two challenges in continuous erasure:

- **Cumulative Parameter Drift:** Because $PL_{\text{prev}}(t)$ is only a soft constraint relative to the preceding model, the preservation reference itself moves with every edit, and residual errors accumulate across stages, progressively degrading non-target representations and generation quality (Fig. 1a), consistent with the drift observed in continual unlearning (Lee et al., 2026).
- **Weakening of Historical Erasures:** Each update focuses on the current target without explicitly preserving previous erasure responses. Consequently, later edits may weaken or reverse earlier erasures, causing removed concepts to re-emerge (Fig. 1b).

4 METHOD

Building on the continuous-erasure setting introduced above, LCCE controls preservation across the entire request sequence rather than treating it as an independent penalty at each stage. A virtual queue determines how strongly the next edit should protect non-target knowledge, while a cumulative preservation bank supplies the queue with a semantically continuous preservation signal. A historical retention loss prevents old concepts from re-emerging. All three components are combined in a closed-form update of the cross-attention key and value projections.



(a) (b)

Figure 1: **(a) Visualization of Preservation Loss.** Baselines (UCE and ESD) show a sharp increase in preservation loss as erasures grow, indicating that cumulative drift progressively degrades model stability. **(b) Visualization of Representation Distribution.** A second erasure can inadvertently shift previously removed concepts back into the active generative region, demonstrating that the effects of earlier tasks are unstable and can be undone by subsequent erasures.

4.1 Lyapunov-Constrained Sequential Editing

The failure modes in Fig. 1 arise because a fixed single-stage penalty does not account for preservation errors left by earlier edits. We instead bound the average preservation loss over the sequence:

$$\begin{aligned} \min_{\{W(t)\}} \ & \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\big(EL(t)+\beta BL(t)\big), \\ \text{s.t.} \ & \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} PL(t) \le D, \end{aligned} \tag{3}$$

where $BL(t)$ retains previous erasure responses and D is the long-term preservation budget. The layer-wise budget is calibrated from the preservation loss observed after the first edit.

Inspired by the Lyapunov controller of LyapLock (Wang et al., 2025a), we introduce a virtual queue

$$Z(t+1) = \max\big\{Z(t) + a\big(PL(t) - D\big), Z_{\min}\big\}. \tag{4}$$

The queue records preservation debt. When $PL(t) > D$, the excess increases $Z(t+1)$ and gives preservation a larger role in the next update. When the loss remains below the budget, the pressure is released toward the positive floor $Z_{\min}$. LCCE therefore adapts to the accumulated effect of previous edits instead of applying the same preservation penalty throughout the sequence.

At stage t, the queue turns the long-term problem into the following Lyapunov-constrained action:

$$\begin{aligned} \min_{W} \quad & \lambda\|W - W(t-1)\|_F^2 + \eta EL(t) + \beta BL(t) + a\rho Z(t) PL(t), \\ \text{s.t.} \quad & WS = W(0)S. \end{aligned} \tag{5}$$

Here, only the queue state varies across stages; the other coefficients stay fixed.

Closed-form editing may also perturb conditioning directions that are shared across many prompts. Following SPEED (Li et al., 2025c), we preserve a compact invariant subspace $S = [s_{\text{SOT}}, \mu_1, \mu_2, \mu_3]$, where s_{SOT} is the start-of-text representation and $\{\mu_i\}_{i=1}^{3}$ are cluster centers of the null-prompt hidden states. We require $W(t)S = W(0)S$ at every stage, preventing sequential edits from altering these shared conditioning responses. This constraint is incorporated into the closed-form solution, with its derivation given in Appendix A.6.

Under the standard boundedness and feasibility assumptions of Lyapunov optimization, mean-rate stability of $Z(t)$ ensures that the long-term preservation budget is satisfied (Appendix A).

4.2 Cumulative Preservation Bank

Concepts in the semantic neighborhood of an erasure target are particularly vulnerable to collateral changes during model editing (Lee et al., 2026; George et al., 2026). For each target e_t, we use an LLM to obtain related concepts and rank them by their cosine similarity to e_t in the frozen text-encoder space. The top-K_0 concepts form the stage-local neighborhood $\mathcal{N}_t$. Since LCCE edits the key and value projections that act directly on these representations, proximity in this space identifies the inputs most affected by an edit without any diffusion forward pass.

Using $\mathcal{N}_t$ independently at each stage would change the semantic domain observed by the virtual queue, making its accumulated state incomparable across requests. We therefore let the bank inherit all previous anchors, add the current neighborhood, and remove erased concepts:

$$\mathcal{A}_t = \big(\mathcal{A}_{t-1} \cup \mathcal{N}_t\big) \setminus \mathcal{E}_{1:t}, \qquad \mathcal{A}_0 = \varnothing, \tag{6}$$

where $\mathcal{E}_{1:t}$ collects the erased concepts and their aliases, and duplicate anchors are merged.

The preservation loss in Eq. 1 is evaluated on $\mathcal{A}_t$ using the original projection $W(0)$ as the fixed reference. Its normalization by $m_t = |\mathcal{A}_t|$ ensures that the queue consistently receives the average non-target drift over the semantic domain observed so far. This incremental construction therefore gives the queue a comparable preservation signal across stages while allowing the active domain to evolve. The corresponding theoretical result is provided in Appendix A.

4.3 Closed-Form Update with Historical Retention Loss

The virtual queue regulates non-target drift, whereas the historical retention loss addresses the second failure mode in Fig. 1. After stage i is completed, LCCE records the realized erased response $\bar{V}_i = W(i)C_i$. Later edits preserve this response instead of recomputing it from a modified model. The resulting historical retention loss $BL(t)$ prevents later requests from undoing earlier erasures.

The required history can be summarized by two fixed-size sufficient statistics:

$$H_{VC}^{t-1} = \sum_{i<t} \frac{\bar{V}_i C_i^\top}{n_i}, \qquad H_{CC}^{t-1} = \sum_{i<t} \frac{C_i C_i^\top}{n_i}. \tag{7}$$

Thus, the number of stored matrices does not grow with the request sequence.

Eq. 5 then yields the unconstrained solution $\widetilde{W}(t) = A(t)B(t)^{-1}$, where

$$\begin{aligned} A(t) &= \lambda W(t-1) + \frac{\eta}{n_t} V_t^* C_t^\top + \beta H_{VC}^{t-1} + \frac{a\rho Z(t)}{m_t} V_{\mathcal{A}_t} C_{\mathcal{A}_t}^\top, \\ B(t) &= \lambda I + \frac{\eta}{n_t} C_t C_t^\top + \beta H_{CC}^{t-1} + \frac{a\rho Z(t)}{m_t} C_{\mathcal{A}_t} C_{\mathcal{A}_t}^\top. \end{aligned} \tag{8}$$

The four terms respectively keep the solution near the preceding model, erase the current target, replay historical erasure responses, and preserve the active non-target domain. In particular, $Z(t)$ connects the closed-form editor to the long-term controller: accumulated preservation debt directly changes the action selected at the next stage.

Finally, we impose the invariant condition in Eq. 5 by analytically correcting $\widetilde{W}(t)$ under the metric induced by $B(t)$. The explicit correction is given in Appendix A.6. We solve the resulting linear systems directly and apply the update to the key and value projections in each cross-attention layer.

5 Experiments

We describe the setup (Section 5.1), evaluate LCCE on continuous artist-style erasure (Section 5.2) and object erasure (Section 5.3), and analyze runtime and components (Section 5.4). The appendix further reports celebrity erasure, sensitivity analysis, visualizations, and concept revival.

5.1 Experimental Setup

Base Model and Generation Settings. All experiments are conducted on the widely adopted Stable Diffusion v1.4 (SD-v1.4) (Rombach et al., 2022). Unless otherwise specified, images are generated

Table 1: **Artist-style erasure results across continuous erasure stages.** Each row is a method. Columns are grouped by the number of erased artists (4/8/16). RR-CLS, RR-LLM, CLIP, and VQA are reported; higher is better. Best in **bold**, second best underlined.

Method	4 Artists				8 Artists				16 Artists			
	RR-CLS	RR-LLM	CLIP	VQA	RR-CLS	RR-LLM	CLIP	VQA	RR-CLS	RR-LLM	CLIP	VQA
UCE	0.609	0.760	28.69	**50.76**	**0.675**	0.910	17.50	15.50	0.685	0.965	12.20	5.40
ESD	0.594	0.880	27.14	49.49	0.608	0.954	24.05	45.10	0.601	0.960	21.15	31.20
CCRT	**0.629**	**0.990**	24.87	41.48	0.612	0.960	22.40	38.50	0.625	0.920	19.50	29.10
MACE	0.495	0.670	28.49	42.31	0.465	0.665	27.85	40.20	0.450	0.690	25.90	37.10
SPEED	0.503	0.640	28.65	35.40	0.500	0.650	28.05	39.50	0.515	0.695	25.80	40.20
CUIG GradProj	0.535	0.840	28.62	46.17	0.613	0.832	26.84	42.35	0.692	0.932	24.89	39.16
LCCE (Ours)	0.610	0.950	**28.75**	47.12	0.638	**0.989**	**28.42**	**46.21**	**0.704**	**0.982**	**27.43**	**46.05**

using the DDIM sampler with 50 denoising steps. For a fair comparison, all methods use the same concept order, evaluation prompts, and random seeds within each experiment.

Baselines. We compare LCCE against six representative concept erasure methods: (1) the fine-tuning-based approach **ESD** (Gandikota et al., 2023) and **MACE** (Lu et al., 2024), (2) closed-form optimization methods including **UCE** (Gandikota et al., 2024), and **SPEED** (Li et al., 2025c), (3) **CCRT** (Han et al., 2024) and **CUIG** (Lee et al., 2026), designed for continuous erasure.

Evaluation Tasks and Prompts. We consider two principal continuous-erasure settings. The artist-style task sequentially removes a predefined sequence of artists and evaluates both the remaining target-style signals and the semantic quality of the generated images. The object-category task sequentially erases object classes while monitoring both erased and retained categories. The task-specific concept sets and prompt protocols are described in their corresponding subsections.

Evaluation Metrics. We evaluate three complementary aspects:

(1) **Erasure Efficacy** measures whether the requested concepts remain suppressed after sequential editing. For artist styles, we report RR-CLS and RR-LLM following CCRT. For objects, we report the accuracy of a pretrained ResNet on the erased classes (lower is better).

(2) **Generative Fidelity** measures whether the edited model retains its general image-generation capability. We use the CLIP score to evaluate text–image consistency and the VQA score to measure the preservation of visual content and fine-grained attributes relative to the original SD-v1.4 model.

(3) **Concept Locality** measures collateral damage to non-target concepts. For the object task, it is quantified by the ResNet classification accuracy on retained classes, where higher accuracy indicates better locality. For artists, we apply the fidelity metrics to fixed non-target prompts.

Detailed metric definitions and calculation procedures are provided in Appendix B.4.

5.2 Continuous Artist Style Erasure

We sequentially erase 16 artists with distinctive styles selected from the WikiArt dataset. For each artist, we use 50 text prompts and generate one image per prompt after completing the corresponding editing stage. Following CCRT, we evaluate erasure efficacy using RR-CLS and RR-LLM. RR-CLS measures the degree to which the target style has been removed, using artist-specific binary classifiers, while RR-LLM uses GPT-4o-mini to judge whether the generated images are free of the erased style. We additionally report CLIP-Score and VQA-Score to evaluate text–image alignment and visual fidelity. Table 1 reports the quantitative results, and Figs. 3 and 2 show qualitative comparisons.

Sustained Erasure Efficacy. As the editing sequence progresses, methods that prioritize preservation, such as MACE and SPEED, retain substantial traces of the target styles, as reflected by their inferior RR-CLS and RR-LLM scores. CUIG GradProj provides a stronger sequential baseline by reducing interference between consecutive edits and better maintaining previously established erasures. Nevertheless, LCCE achieves the strongest overall erasure performance across the editing stages. Its historical retention loss prevents later updates from weakening earlier erasures, enabling the requested artist styles to remain suppressed throughout the sequence. Qualitatively, LCCE removes distinctive cues such as Van Gogh's brushwork and Vermeer's lighting.

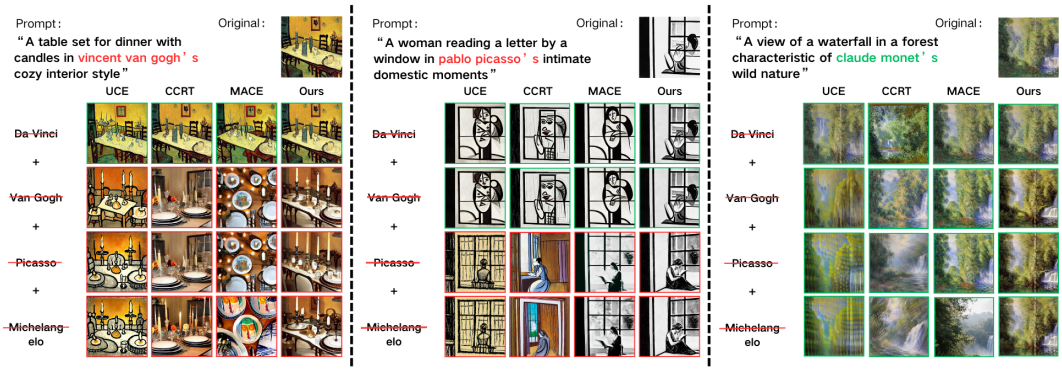


Figure 2: **Qualitative comparison during sequential artist-style erasure.** Each row adds one erased artist, and each panel shows one prompt with its original SD v1.4 generation. Red boxes mark generations whose target style should be removed, and green boxes mark generations that should be preserved. LCCE suppresses a target style once it is erased and preserves both the prompt content and styles that have not yet been targeted; competing methods either retain visible target-style cues or introduce increasing content and quality degradation.

Preservation of Generative Fidelity. Methods that pursue aggressive target removal, including UCE, ESD, and CCRT, suffer substantial degradation in text–image alignment and visual quality as sequential edits accumulate. Although MACE and SPEED preserve image quality more effectively, this advantage is accompanied by weaker target suppression. CUIG GradProj achieves a better balance than these conventional baselines, but still exhibits greater fidelity degradation than LCCE. In contrast, LCCE maintains consistently strong CLIP and VQA scores while retaining coherent semantic content and visual structure. These results indicate that the virtual queue, supported by the cumulative preservation bank, effectively regulates accumulated non-target drift without sacrificing the persistence of previous erasures. Stage-wise trajectories (Appendix C.2) further show that artists not yet targeted keep their styles until their own erasure stage.

Sustained Robustness under Extreme Stress. To further evaluate scalability and long-term stability, we conducted a stress test by extending continuous concept erasure to 100 distinct artists using UCE, SPEED, MACE, and LCCE. Other fine-tuning-based methods are excluded from this stress test because their per-concept optimization cost accumulates to several days or weeks over 100 sequential requests. As illustrated in the right panel of Figure 3, the performance of baseline methods deteriorates significantly as the number of erased concepts increases. Specifically, the generative capability of UCE and MACE drops sharply until it completely collapses. While SPEED exhibits better resilience than UCE, its performance remains consistently inferior to our method. In contrast, LCCE maintains remarkable consistency throughout the entire 100-step process, preserving high generative fidelity and text-image alignment with minimal degradation, thereby mitigating the catastrophic forgetting observed in existing frameworks.

5.3 Continuous Object Erasure

Following the evaluation protocol of UCE, we sequentially erase ten object classes from Imagenette and use ten retained ImageNet classes to evaluate non-target preservation. After all ten erasure stages, we generate 50 images for each erased and retained class and classify them using a pretrained ResNet-50. We report the mean accuracy on erased (Acc_e) and retained (Acc_p) classes.

To summarize the trade-off between erasure and preservation, we report H_0, the harmonic mean of the erasure success rate $1 - \mathrm{Acc}_e$ and preservation accuracy Acc_p. A higher H_0 therefore indicates a better balance between removing target classes and retaining non-target knowledge.

Comparison with Existing Methods. Table 2 shows that existing methods fail in different directions. UCE, ESD, CCRT, and MACE remove the erased classes but severely damage the retained ones, whereas SPEED preserves the retained classes yet leaves much of the target information.

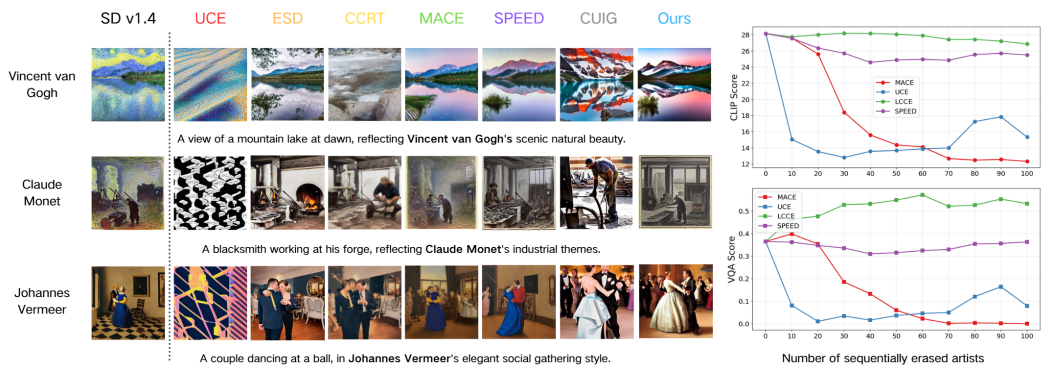


Figure 3: **Evaluation of continuous artist-style erasure.** (Left) Qualitative comparison after sequentially erasing ten artists (Leonardo da Vinci, Vincent van Gogh, etc.). The first column shows the unedited SD v1.4 as a style reference; the remaining columns show each method after all ten erasures, using the prompt below each row. (Right) CLIP and VQA scores over 100 erasures.

Table 2: **Performance comparison on continuous object erasure.** Acc_e denotes the accuracy of erased classes (lower is better), while Acc_p denotes the accuracy of unrelated preserved classes (higher is better). H_0 represents the harmonic mean, reflecting the overall balance. The results of the remaining seven categories are provided in Appendix C.1.

Method	$\text{Acc}_e \downarrow$				$\text{Acc}_p \uparrow$				$H_0 \uparrow$
	golf ball	english springer	garbage truck	**Avg. 10**	light-house	pizza	suit	**Avg. 10**	**Avg. 10**
SD v1.4 (Base)	0.990	0.930	0.780	0.759	0.980	0.990	0.870	0.924	0.382
UCE	0.040	**0.000**	**0.000**	0.004	0.070	0.000	0.250	0.127	0.225
ESD	**0.000**	**0.000**	**0.000**	0.001	0.100	0.000	0.040	0.018	0.035
CCRT	**0.000**	**0.000**	0.010	0.001	0.010	0.040	0.010	0.008	0.016
MACE	**0.000**	**0.000**	**0.000**	**0.000**	0.000	0.000	0.000	0.000	0.000
SPEED	0.760	0.190	0.410	0.334	**0.990**	0.990	0.880	**0.925**	0.775
CUIG GradProj	**0.000**	**0.000**	**0.000**	0.005	0.950	**1.000**	**0.910**	0.802	0.888
LCCE (Ours)	**0.000**	**0.000**	**0.000**	0.002	0.910	0.990	0.900	0.860	**0.924**

CUIG GradProj is the strongest sequential baseline, combining effective erasure with considerably better preservation.

Erasure–Preservation Balance. LCCE matches the near-complete erasure of the aggressive editors while keeping most of the base model's accuracy on retained classes, and it therefore attains the highest H_0, avoiding the collateral damage of aggressive editing. The qualitative trajectories in Fig. 4 further show that each target object remains generable before its designated erasure stage and is effectively removed once the corresponding edit is applied.

5.4 More Analysis

Time Cost. Table 3 reports the per-concept runtime of the three continuous erasure methods on the same hardware, excluding image generation and evaluation. LCCE is two to three orders of magnitude faster than both baselines. CCRT spends most of its time on a genetic search that repeatedly queries the diffusion model and an LLM, and CUIG GradProj still runs gradient-based optimization for every request. LCCE only precomputes local anchors and applies a closed-form update, taking about 8 seconds per concept. This gap compounds over long erasure sequences, where per-concept costs accumulate.

Ablation Study. We ablate the three principal components of LCCE on the sequential erasure of 16 artist styles, reporting CLIP-Score for generation fidelity and RR-LLM for target-style erasure.

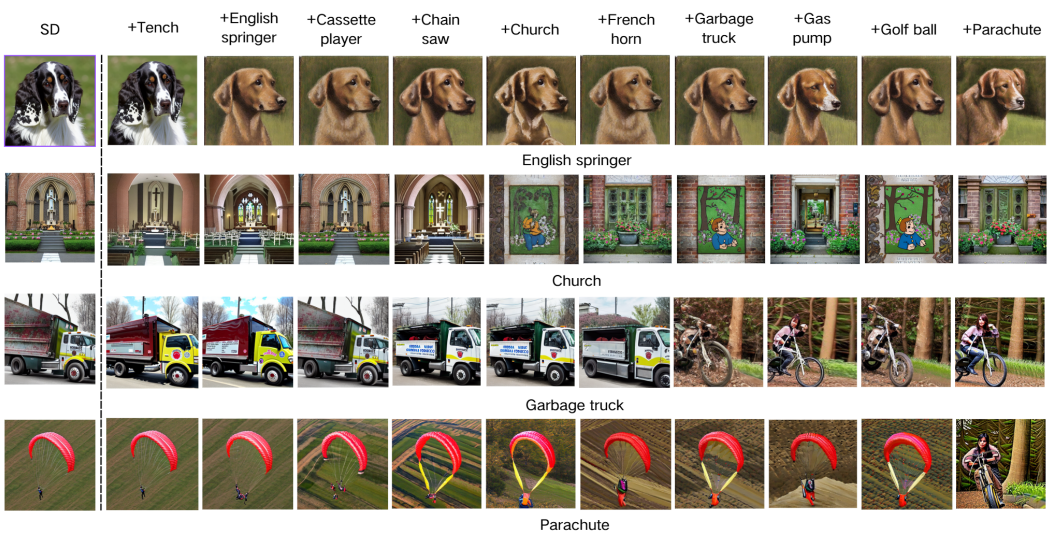


Figure 4: **Qualitative results of continuous object erasure.** Each target object remains generable before its designated erasure stage, while the corresponding concept is no longer reflected in the generated images after erasure.

Table 3: **Average runtime per erased concept.** Preprocessing covers concept-specific preparation before editing; erasure is the time to update the model parameters.

Method	Preprocessing operation	Preprocess	Erasure	Total
CCRT	Calibration-entity search	200 min	60 min	260 min
CUIG GradProj	Auxiliary-concept construction	5 s	20 min	20 min 5 s
LCCE (Ours)	Local-anchor precomputation	**2 s**	**6 s**	**8 s**

Variant	CLIP ↑	RR-LLM ↑
w/o Cumulative Preservation Bank	26.52	0.956
w/o Dynamic Queue	25.10	**1.000**
w/o Historical Retention Loss	**28.74**	0.926
LCCE	27.43	0.982

Table 4: **Ablation study on 16 sequential artist-style erasures.** Each variant removes one component of LCCE.

Table 4 shows that each component contributes a distinct form of stability. Removing the cumulative preservation bank lowers both fidelity and erasure, because changing the local anchors at each stage discards previously observed semantic neighborhoods. Fixing the preservation pressure in place of the dynamic queue yields the strongest erasure but the largest fidelity loss, so aggressive erasure alone increases non-target degradation. Removing the historical retention loss has the opposite effect: fidelity improves, but erasure weakens because later updates are no longer constrained to retain earlier erased responses. Only the complete LCCE keeps both fidelity and erasure high.

6 Conclusion

We introduce Lyapunov-constrained Continuous Concept Erasure (LCCE), a closed-form framework for sequential concept erasure in text-to-image diffusion models. By formulating continuous editing as an online control problem, LCCE uses a dynamic virtual queue to accumulate preservation violations and adaptively regulate non-target drift across editing stages. A cumulative preservation bank provides a consistent semantic domain for this long-term control, while the historical retention loss helps maintain the states established by previous erasures. LCCE maintains high generative fidelity and robustly suppresses historical targets, even after sequentially erasing over 100 concepts. We discuss the scope of LCCE and directions beyond it in Appendix H.

AI USE STATEMENT

Required disclosure. We did not use generative AI tools for any task that requires disclosure. No AI tool generated data, formulated or proved the theoretical results, proposed hypotheses, designed the method or the experiments, implemented the method, or interpreted the results. **Recommended disclosure.** We used LLM-based assistants for (i) writing, including editing for readability and grammar, suggesting structure, and drafting passages from author-provided content, all of which the authors revised; (ii) literature search, to locate and summarize related work, with every cited paper checked by the authors; (iii) LaTeX formatting; and (iv) experiment operations, where coding agents launched and monitored GPU jobs and collected their outputs into tables under author-specified configurations. The authors reviewed all AI-assisted content and take full responsibility for this paper.

ETHICS STATEMENT

This work aims to make text-to-image models safer and more respectful of creators' rights by removing copyrighted styles, identities, and other undesired concepts as removal requests arrive. Because erasure changes what a model can generate, it should be applied under transparent governance, since the same mechanism could be used to suppress legitimate content. Our experiments use publicly available models and concept lists: artist names from WikiArt, object classes from Imagenette and ImageNet, and celebrity names from the benchmark used by MACE. Images of artist styles and real individuals were generated solely for evaluating erasure and are not released. The human study (Appendix F) asked participants only to rank generated images, and no personal information about participants was collected.

REPRODUCIBILITY STATEMENT

The continuous erasure setting and the LCCE objective are defined in Sections 3 and 4, and each stage is solved in closed form (Eq. 8), with the equality-constrained correction derived in Appendix A.6. The theoretical results in Appendix A state their assumptions and include complete proofs. Section 5.1 specifies the base model, sampler, and the concept order, prompts, and seeds shared by all methods. Appendix B lists the concept sets, including the full 100-artist sequence, the safe targets, and the definitions of all metrics, including the classifiers and the prompt used by the GPT-4o-mini judge. All hyperparameter values are listed in Appendix B.2 and their sensitivity is analyzed in Appendix G, and per-concept runtimes in Table 3. We will release our code upon acceptance.

REFERENCES

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Anh Bui, Trang Vu, Long Vuong, Trung Le, Paul Montague, Tamas Abraham, Junae Kim, and Dinh Phung. Fantastic targets for concept erasure in diffusion models and where to find them. *arXiv preprint arXiv:2501.18950*, 2025.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Hao Chen, Tianshu Zhu, Li Wang, Xin Yu, Long Gao, and Wei Zhou. Safe and reliable diffusion models via subspace projection. *arXiv preprint arXiv:2503.16835*, 2025a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Xin Chen, Gillian Dobbie, Xinyu Wang, Feng Liu, Di Wang, and Jingfeng Zhang. Robust learning of diffusion models with extremely noisy conditions. *arXiv preprint arXiv:2510.10149*, 2025b.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3043–3054, 2023.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5111–5120, 2024.

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, 2026.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 73–88. Springer, 2024.

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

Chung-Ping Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Reeler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 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.

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

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

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26256–26264, 2025a.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *arXiv preprint arXiv:2503.07392*, 2025c.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

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.

Zihao Luo, Xilie Xu, Feng Liu, Yun Sing Koh, Di Wang, and Jingfeng Zhang. Privacy-preserving low-rank adaptation against membership inference attacks for latent diffusion models. *arXiv preprint arXiv:2402.11989*, 2024.

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

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 36479–36494, 2022.

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

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023.

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? *arXiv preprint arXiv:2310.10012*, 2023.

Peng Wang, Biyu Zhou, Xuehai Tang, Jizhong Han, and Songlin Hu. LyapLock: Bounded knowledge preservation in sequential large language model editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6434–6459, 2025a.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768, 2025b. doi: 10.1109/CVPR52734.2025.02678.

Yiwei Xie, Ping Liu, and Zheng Zhang. Erasing concepts, steering generations: A comprehensive survey of concept suppression. *arXiv preprint arXiv:2505.19398*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Yi Zhang, Zhen Chen, Chih-Hong Cheng, Wenjie Ruan, Xiaowei Huang, Dezong Zhao, David Flynn, Siddartha Khastgir, and Xingyu Zhao. Trustworthy text-to-image diffusion models: A timely and focused survey. *arXiv preprint arXiv:2409.18214*, 2024.

A Theoretical Analysis

This appendix provides the theoretical justification for the main components of LCCE. We first establish the finite-horizon and long-term preservation bounds induced by the virtual queue. We then show that the same argument applies to the cumulative, time-varying preservation bank. Finally, we prove that the historical retention loss admits fixed-size sufficient statistics and that the resulting quadratic update has a unique equality-constrained closed-form solution.

A.1 Queue Stability and Long-Term Preservation

For a fixed editable projection, let

$$g(t) = PL(t) - D.$$

We assume that the preservation loss is finite and uniformly bounded:

$$0 \le PL(t) \le P_{\max}, \qquad D \le P_{\max}, \tag{9}$$

for some finite constant $P_{\max}$. The queue is updated by

$$Z(t+1) = \max\left\{Z(t) + a\big(PL(t) - D\big), Z_{\min}\right\}, \qquad a > 0, \tag{10}$$

with $Z(1) \ge Z_{\min} > 0$.

Lemma A.1 (One-step queue inequality)**.** *For every stage t,*

$$Z(t+1) \ge Z(t) + a\big(PL(t) - D\big). \tag{11}$$

Proof. The right-hand side of Eq. 10 is the maximum of $Z(t) + a(PL(t) - D)$ and $Z_{\min}$. It is therefore no smaller than the first argument of the maximum, which proves Eq. 11. □

Theorem A.2 (Finite-horizon preservation bound)**.** *For every horizon $T \ge 1$,*

$$\frac{1}{T}\sum_{t=1}^{T} PL(t) \le D + \frac{Z(T+1) - Z(1)}{aT}. \tag{12}$$

Proof. Summing Eq. 11 over $t = 1, \dots, T$ gives

$$\sum_{t=1}^{T} Z(t+1) \ge \sum_{t=1}^{T} Z(t) + a\sum_{t=1}^{T}\big(PL(t) - D\big).$$

The two queue sums telescope, yielding

$$Z(T+1) - Z(1) \ge a\sum_{t=1}^{T}\big(PL(t) - D\big).$$

Rearranging and dividing by aT gives Eq. 12. □

Corollary 1 (Long-term preservation)**.** If the queue is mean-rate stable,

$$\lim_{T\to\infty} \frac{Z(T)}{T} = 0,$$

then

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} PL(t) \le D. \tag{13}$$

Proof. By Theorem A.2,

$$\frac{1}{T}\sum_{t=1}^{T} PL(t) \le D + \frac{Z(T+1) - Z(1)}{aT}.$$

Since $Z(1)$ is finite and mean-rate stability implies $Z(T+1)/T \to 0$, the second term converges to zero. Taking the limsup proves Eq. 13. □

Remark

The virtual queue controls the long-term average preservation loss. It does not require $PL(t) \leq D$ at every stage and does not provide an individual guarantee for every anchor. Mean-rate stability of the queue is the condition that converts the finite-horizon bound into the long-term constraint.

A.2 Lyapunov Drift Bound

To relate the queue to the per-stage objective, define the shifted queue

$$Q(t) = Z(t) - Z_{\min} \geq 0.$$

Then Eq. 10 becomes

$$Q(t+1) = \max\left\{Q(t) + a\big(PL(t) - D\big), 0\right\}.$$

Let

$$L(Q(t)) = \frac{\rho}{2}Q(t)^2, \qquad \rho > 0.$$

Lemma A.3 (Lyapunov drift bound)**.** *Under Eq. 9,*

$$\begin{aligned} L(Q(t+1)) - L(Q(t)) \leq \rho a Q(t)\big(PL(t) - D\big) \\ + \frac{\rho a^2}{2} G_{\max}^2, \end{aligned} \tag{14}$$

where

$$G_{\max} = \max\{D, P_{\max} - D\}.$$

Proof. Since $\max\{x, 0\}^2 \leq x^2$ for every real x,

$$\begin{aligned} Q(t+1)^2 &\leq \big[Q(t) + a\big(PL(t) - D\big)\big]^2 \\ &= Q(t)^2 + 2aQ(t)\big(PL(t) - D\big) + a^2\big(PL(t) - D\big)^2. \end{aligned}$$

The boundedness assumption gives

$$\big(PL(t) - D\big)^2 \leq G_{\max}^2.$$

Multiplying by $\rho/2$ and subtracting $L(Q(t))$ proves Eq. 14. □

The drift term contains $a\rho Q(t)PL(t)$. The implementation uses $a\rho Z(t)PL(t)$ in Eq. 5. Since $Z(t) = Q(t) + Z_{\min}$,

$$a\rho Z(t)PL(t) = a\rho Q(t)PL(t) + a\rho Z_{\min}PL(t).$$

Thus, the implemented objective consists of the standard queue-dependent Lyapunov penalty together with a fixed positive preservation bias induced by the queue floor. The queue telescoping result is unaffected by this additional nonnegative term.

A.3 Time-Varying Cumulative Preservation Bank

The cumulative preservation bank is updated as

$$\mathcal{A}_t = \big(\mathcal{A}_{t-1} \cup \mathcal{N}_t\big) \setminus \mathcal{E}_{1:t}.$$

Let

$$m_t = |\mathcal{A}_t|, \qquad C_{\mathcal{A}_t} = [c_{t,1}, \ldots, c_{t,m_t}],$$

and define the original-model reference

$$V_{\mathcal{A}_t} = W(0)C_{\mathcal{A}_t}.$$

The preservation loss at stage t is

$$PL(t) = \frac{1}{m_t}\left\|W(t)C_{\mathcal{A}_t} - V_{\mathcal{A}_t}\right\|_F^2. \tag{15}$$

Proposition 1 (Queue validity for a time-varying domain)**.** The finite-horizon and long-term preservation results remain valid when $\mathcal{A}_t$ changes across stages, provided that the scalar loss in Eq. 15 is finite and satisfies Eq. 9.

Proof. At each stage, Eq. 15 produces one scalar value $PL(t)$. The queue recursion in Eq. 10 depends only on this scalar and does not depend on the identities, cardinality, or nesting of the anchors used to compute it. Therefore, Lemma A.1 and its telescoping sum remain unchanged. The finite-horizon bound follows from Theorem A.2, and the long-term result follows from Corollary 1. □

Remark

The cumulative bank gives each queue update the same interpretation: average drift from the original model over the currently active non-target domain. The theory therefore applies to the evolving active domain, rather than claiming an individual guarantee for a fixed vocabulary or for every anchor.

A.4 Historical Retention Loss

After stage i, LCCE stores the realized response

$$\bar{V}_i = W(i)C_i.$$

The historical retention loss at stage t is

$$BL(t;W) = \sum_{i<t} \frac{1}{n_i} \left\| WC_i - \bar{V}_i \right\|_F^2 . \tag{16}$$

Define the accumulated statistics

$$H_{VC}^{t-1} = \sum_{i<t} \frac{\bar{V}_i C_i^\top}{n_i}, \qquad H_{CC}^{t-1} = \sum_{i<t} \frac{C_i C_i^\top}{n_i}.$$

Proposition 2 (Sufficient statistics for historical retention)**.** The loss in Eq. 16 can be written as

$$\begin{aligned} BL(t;W) = {} & \operatorname{tr}\left(W H_{CC}^{t-1} W^\top\right) - 2\operatorname{tr}\left(W (H_{VC}^{t-1})^\top\right) \\ & + \sum_{i<t} \frac{\|\bar{V}_i\|_F^2}{n_i}. \end{aligned}$$

Consequently, H_{VC}^{t-1} and H_{CC}^{t-1} preserve the same minimizer as the full historical loss.

Proof. For one historical request,

$$\begin{aligned} \frac{1}{n_i}\|WC_i - \bar{V}_i\|_F^2 &= \frac{1}{n_i} \operatorname{tr}\left[(WC_i - \bar{V}_i)(WC_i - \bar{V}_i)^\top\right] \\ &= \frac{1}{n_i} \operatorname{tr}(WC_i C_i^\top W^\top) - \frac{2}{n_i} \operatorname{tr}(WC_i \bar{V}_i^\top) + \frac{1}{n_i}\|\bar{V}_i\|_F^2. \end{aligned}$$

Summing over $i < t$ and using the cyclic property of the trace gives

$$\sum_{i<t} \frac{1}{n_i} \operatorname{tr}(WC_i C_i^\top W^\top) = \operatorname{tr}\left(W H_{CC}^{t-1} W^\top\right),$$

and

$$\sum_{i<t} \frac{1}{n_i} \operatorname{tr}(WC_i \bar{V}_i^\top) = \operatorname{tr}\left(W (H_{VC}^{t-1})^\top\right).$$

The final term is independent of W, which proves the result. □

A.5 Closed-Form Update

For a fixed stage t, define

$$\gamma_t = \frac{a\rho Z(t)}{m_t}.$$

Ignoring terms independent of W, the stage objective is

$$\begin{aligned} J_t(W) = \lambda\|W - W(t-1)\|_F^2 + \frac{\eta}{n_t}\|WC_t - V_t^*\|_F^2 \\ + \beta BL(t;W) + \gamma_t\|WC_{\mathcal{A}_t} - V_{\mathcal{A}_t}\|_F^2. \end{aligned} \tag{17}$$

Proposition 3 (Unique unconstrained closed-form update)**.** Assume $\lambda > 0$ and nonnegative objective weights. Then $J_t(W)$ is strictly convex and has the unique minimizer

$$\widetilde{W}(t) = A(t)B(t)^{-1},$$

where

$$\begin{aligned} A(t) &= \lambda W(t-1) + \frac{\eta}{n_t}V_t^*C_t^\top + \beta H_{VC}^{t-1} + \gamma_t V_{\mathcal{A}_t}C_{\mathcal{A}_t}^\top, \\ B(t) &= \lambda I + \frac{\eta}{n_t}C_tC_t^\top + \beta H_{CC}^{t-1} + \gamma_t C_{\mathcal{A}_t}C_{\mathcal{A}_t}^\top. \end{aligned}$$

Proof. Each term in Eq. 17 is a convex quadratic in W. The matrix λI is positive definite because $\lambda > 0$, while all remaining Gram-matrix contributions are positive semidefinite. Hence

$$B(t) \succ 0.$$

Expanding the objective and collecting terms that depend on W gives

$$J_t(W) = \operatorname{tr}\left(WB(t)W^\top\right) - 2\operatorname{tr}\left(A(t)W^\top\right) + \text{const}.$$

The matrix derivative is

$$\nabla_W J_t(W) = 2WB(t) - 2A(t).$$

The stationary condition is therefore

$$WB(t) = A(t).$$

Since $B(t)$ is positive definite and invertible,

$$W = A(t)B(t)^{-1}.$$

Strict convexity makes this stationary point the unique global minimizer. □

A.6 Equality-Constrained Closed-Form Correction

The invariant constraint is

$$WS = Y, \qquad Y = W(0)S.$$

The constrained problem is

$$\begin{aligned} \min_W \quad & \operatorname{tr}\left(WB(t)W^\top\right) - 2\operatorname{tr}\left(A(t)W^\top\right), \\ \text{s.t.} \quad & WS = Y. \end{aligned} \tag{18}$$

Proposition 4 (Equality-constrained solution)**.** Assume $B(t) \succ 0$ and that S has full column rank. Let

$$\widetilde{W}(t) = A(t)B(t)^{-1}.$$

Then the unique solution of Eq. 18 is

$$W^*(t) = \widetilde{W}(t) + \left(Y - \widetilde{W}(t)S\right)\left(S^\top B(t)^{-1}S\right)^{-1}S^\top B(t)^{-1}. \tag{19}$$

Proof. Since $\widetilde{W}(t)B(t) = A(t)$, completing the square gives

$$\begin{aligned}&\operatorname{tr}(WB(t)W^\top) - 2\operatorname{tr}(A(t)W^\top)\\ &= \operatorname{tr}\left[(W-\widetilde{W}(t))B(t)(W-\widetilde{W}(t))^\top\right] + \text{const.}\end{aligned}$$

Let

$$\Delta W = W - \widetilde{W}(t), \qquad R = Y - \widetilde{W}(t)S.$$

The constrained problem becomes

$$\min_{\Delta W} \operatorname{tr}\left(\Delta W B(t) \Delta W^\top\right), \qquad \text{s.t.} \quad \Delta W S = R.$$

The Lagrangian is

$$\mathcal{L}(\Delta W, \Lambda) = \operatorname{tr}\left(\Delta W B(t) \Delta W^\top\right) + 2\operatorname{tr}\left[\Lambda^\top(\Delta W S - R)\right].$$

Differentiating with respect to ΔW gives

$$2\Delta W B(t) + 2\Lambda S^\top = 0,$$

so

$$\Delta W = -\Lambda S^\top B(t)^{-1}.$$

Substituting into $\Delta W S = R$ yields

$$-\Lambda S^\top B(t)^{-1} S = R.$$

Because $B(t) \succ 0$ and S has full column rank, $S^\top B(t)^{-1} S \succ 0$ and is invertible. Thus,

$$\Lambda = -R\left(S^\top B(t)^{-1} S\right)^{-1}.$$

Therefore,

$$\Delta W = R\left(S^\top B(t)^{-1} S\right)^{-1} S^\top B(t)^{-1},$$

which gives Eq. 19. The objective is strictly convex because $B(t) \succ 0$, so the feasible minimizer is unique. □

Corollary 2 (Exact invariant preservation)**.** The solution in Eq. 19 satisfies

$$W^*(t)S = W(0)S.$$

Proof. Multiplying Eq. 19 by S gives

$$\begin{aligned}W^*(t)S &= \widetilde{W}(t)S + \left(Y - \widetilde{W}(t)S\right)\left(S^\top B(t)^{-1} S\right)^{-1} S^\top B(t)^{-1} S\\ &= \widetilde{W}(t)S + Y - \widetilde{W}(t)S\\ &= Y = W(0)S.\end{aligned}$$

□

Remark

The historical retention statistics and the equality-constrained correction are exact closed-form constructions. The former preserves the optimizer of the full historical quadratic loss, while the latter preserves the selected conditioning responses without iterative fine-tuning.

A.7 Stability Condition

The queue recursion alone proves that mean-rate stability is sufficient for the long-term preservation constraint. Establishing mean-rate stability itself requires an additional feasibility condition: there must exist a policy that can keep the expected preservation loss below the budget with a nonzero margin. Under such a condition, standard Lyapunov drift-plus-penalty results apply when the per-stage action minimizes the corresponding drift bound exactly or up to a uniformly bounded error.

Accordingly, the theoretical guarantee of LCCE should be stated as the following implication:

$$\text{mean-rate stability of } Z(t) \quad \Longrightarrow \quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} PL(t) \le D.$$

The feasibility and bounded-approximation assumptions are not automatic for arbitrary request sequences; they are the standard conditions under which the queue is expected to be stable.

B EXPERIMENT DETAILS

B.1 DATASETS

Our experiments evaluate the performance of LCCE in three continual concept erasure tasks, using either specially constructed or curated datasets. For the Continuous Artist Style Erasure experiments, we curated 100 artists from the Wikiart dataset and performed continuous erasure on 16 artists and all 100 artists in a specific order as defined in Table 5. In the Continuous Object Erasure experiments, the erasure concepts are derived from the ten classes of the Imagenette dataset, while the retention concepts are sampled from the ImageNet dataset. Finally, the erasure concepts used in the Continuous Celebrity Erasure experiments are directly sourced from the MACE celebrity dataset.

Number	Target Concepts
100	*Leonardo da Vinci, Vincent van Gogh, Pablo Picasso, Michelangelo, Rembrandt, Claude Monet, Andy Warhol, Salvador Dali, Johannes Vermeer, Jackson Pollock, Nicholas Roerich, Pierre Auguste Renoir, Pyotr Konchalovsky, Camille Pissarro, Albrecht Durer, John Singer Sargent, Marc Chagall, Gustave Dore, Boris Kustodiev, Edgar Degas, Paul Cezanne, Ivan Aivazovsky, Martiros Saryan, Eugene Boudin, Childe Hassam, Ilya Repin, Ivan Shishkin, Raphael Kirchner, Henri Matisse, Camille Corot, Alfred Sisley, Odilon Redon, Isaac Levitan, James Tissot, Paul Gauguin, Maurice Prendergast, Peter Paul Rubens, Francisco Goya, William Merritt Chase, Ernst Ludwig Kirchner, Henri De Toulouse Lautrec, Zinaida Serebriakova, Joaquín Sorolla, Konstantin Makovsky, Amedeo Modigliani, Utagawa Kuniyoshi, Sam Francis, Konstantin Korovin, Thomas Eakins, David Burliuk, Mary Cassatt, Henri Fantin Latour, Egon Schiele, Vasily Surikov, Gustave Courbet, Lucian Freud, Fernand Leger, Gustave Loiseau, Konstantin Somov, Titian, Katsushika Hokusai, Ferdinand Hodler, Ivan Bilibin, Vasily Polenov, Berthe Morisot, Aubrey Beardsley, Giovanni Boldini, Georges Braque, Aleksey Savrasov, John Henry Twachtman, Kuzma Petrov Vodkin, Tintoretto, Orest Kiprensky, Edouard Manet, Jan Matejko, William Turner, Vasily Vereshchagin, Lucas Cranach The Elder, Valentin Serov, Joshua Reynolds, Edouard Cortes, Ilya Mashkov, Gustave Caillebotte, Nikolay Bogdanov Belsky, Sir Lawrence Alma Tadema, Juan Gris, Dante Gabriel Rossetti, Viktor Vasnetsov, Koloman Moser, Felix Vallotton, Raphael, Vasily Perov, Theo Van Rysselberghe, Paolo Veronese, Hieronymus Bosch, Edvard Munch, Arkhip Kuindzhi, Bartolome Esteban Murillo, Frans Hals, Pietro Perugino*

Table 5: Complete list of 100 artists used in Continuous Artist Style Erasure experiments.

B.2 IMPLEMENTATION DETAILS AND HYPERPARAMETER SETTINGS

We use $\lambda = 0.5$ and $K_0 = 50$ in all experiments. The remaining coefficients are chosen according to the concept type, as summarized in Table 6. For each edited layer ℓ, the preservation budget is initialized as $D^{(\ell)} = D^{(\ell)}_{\text{base}}$, where $D^{(\ell)}_{\text{base}}$ is calibrated from the preservation loss after the first edit. We then set

$$Z^{(\ell)}_{\min} = Z^{(\ell)}(1) = \sqrt{D^{(\ell)}_{\text{base}}}, \qquad a^{(\ell)} = \frac{1}{\sqrt{D^{(\ell)}_{\text{base}}}}.$$

These layer-wise values are computed independently for each edited projection.

For all baseline methods, we use their publicly released implementations and follow the preprocessing, optimization, and inference procedures described in the corresponding papers. Unless otherwise specified, we use the author-recommended or best-performing hyperparameter configurations

Table 6: **Hyperparameter settings used in LCCE.**

Setting	η	β	ρ
Artist erasure	1	1	0.003
Celebrity erasure	1	1	0.003
Object erasure	50	100	0.001
$\lambda = 0.5, \quad K_0 = 50$ for all settings			

Table 7: **Effect of the safe target on 16 sequential artist-style erasures.** Higher values indicate better performance for all metrics.

Safe target	CLIP ↑	VQA ↑	RR-CLS ↑	RR-LLM ↑
picture	<u>27.43</u>	**46.05**	<u>0.704</u>	<u>0.982</u>
art	26.75	44.33	0.685	0.930
Empty prompt	**27.74**	<u>45.62</u>	0.637	0.945
Random target	25.65	40.95	**0.721**	**0.995**

reported by each method. All methods are evaluated under the same base model, concept order, prompts, random seeds, and evaluation protocol.

B.3 Choice of Safe Target

The safe response V_t^* specifies the semantic direction toward which an erased concept is redirected. We use a target-dependent neutral guide: *picture* for artist styles, *person* for celebrity identities, and the null-text representation induced by an empty prompt for object categories. These choices remove the target-specific information while retaining a generic response appropriate to the corresponding concept type. For example, mapping an artist style to *picture* suppresses distinctive stylistic cues without removing the visual content specified by the remainder of the prompt, while mapping a celebrity to *person* preserves a human subject without explicitly retaining the target identity.

We further compare different safe targets using the 16-artist sequential erasure setting. As shown in Table 7, the choice of safe target determines the trade-off between target suppression and generation fidelity. The empty prompt yields a slightly higher CLIP-Score than *picture*, but produces weaker RR-CLS and RR-LLM results. Random targets achieve the strongest erasure scores, but substantially reduce both CLIP and VQA, indicating that redirecting artists toward unrelated semantics damages the intended image content. In contrast, *picture* obtains the highest VQA score while maintaining strong CLIP, RR-CLS, and RR-LLM performance. Therefore, we use *picture* as the safe target for artist-style erasure.

B.4 Metrics

This section provides a detailed description of the evaluation metrics used in this paper to assess the performance of models in the continuous concept removal task.

B.4.1 Metrics for Artist Style Removal

To evaluate the erasure efficacy and generative fidelity , we employ four metrics: RR-LLM, RR-CLS, CLIP-Score, and VQA-Score.

1) RR-LLM (LLM-based Removal Rate) RR-LLM leverages Large Language Models (LLMs) to assess the alignment between generated images and the target concept through binary judgment. Formally, it is defined as:

$$RR\text{-}LLM = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\{LLM(\boldsymbol{x}_i|\boldsymbol{p},\boldsymbol{c}) = \text{No}\}$$

where $LLM(\boldsymbol{x}|\boldsymbol{p}, \boldsymbol{c})$ denotes the judgment of the LLM on image $\boldsymbol{x}$ given a carefully designed prompt $\boldsymbol{p}$ and the concept $\boldsymbol{c}$ to be removed. $\mathbb{I}\{\cdot\}$ is the indicator function, which returns 0 if the LLM answers "Yes" and 1 otherwise. A higher RR-LLM value indicates superior removal performance. In practice, we use GPT-4o-mini. The prompt template used is as follows:

Prompt for RR-LLM

Analyze the provided image and assess whether it shows a clear overall resemblance to or influence from the style of artist, considering the overall impression in terms of: 1. brushwork, mark-making, or texture; 2. color palette and contrast; and 3. subject matter or composition. Answer 'yes' only if the image reflects the style of artist; otherwise, answer 'no'. Answer only 'yes' or 'no'.

2) RR-CLS (Classifier-based Removal Rate) RR-CLS evaluates removal effectiveness by training binary classifiers. For each target art style $\boldsymbol{c}_j$, we train a dedicated binary classifier f_j. The classifier takes an image as input and outputs a binary value: 1 if the image does not contain the art style, and 0 otherwise. The formula is:

$$RR\text{-}CLS = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{x}_i)$$

Using "Van Gogh" as an example, we generate 1,000 images in his style using the original Stable Diffusion model and 1,000 images of similar but distinct concepts (e.g., Picasso, Alfred Sisley). We use these 2,000 images as training data for a ResNet50 architecture (0.8:0.2 train-test split). In our experiments, these classifiers achieved an average accuracy of over 95% on the training set and over 90% on the test set.

3) CLIP-Score The CLIP-Score quantifies the semantic alignment between the generated image and the text prompt. It calculates the cosine similarity between image and text embeddings in the joint latent space of the CLIP model. Given a prompt p and an image x:

$$\text{CLIP-Score}(x, p) = \frac{\text{sim}(E_I(x), E_T(p))}{\tau}$$

where E_I and E_T are the CLIP image and text encoders respectively, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is a temperature parameter. A higher CLIP-Score indicates better alignment.

4) VQA-Score Following the methodology described in (Lin et al., 2024), we use a Visual Question Answering (VQA) based approach to provide a more fine-grained evaluation of the alignment between generated images and text prompts. For each text prompt, we generate two questions regarding the presence or characteristics of the objects or scenes mentioned, where the ground-truth answer is "yes".

To generate these questions, we utilize DeepSeek-V3 with the following prompt template:

Prompt for VQA Questions Generation

You are a helpful assistant that generates questions based on descriptive sentences. For each input sentence, you need to create a simple yes/no question asking about the presence or characteristics of the main subject or object in the sentence. The question should be answerable based solely on the sentence itself and should not include any placeholders. Please follow the format of the example below.
Example: Sentence: 'a portrait of a king on his throne.' Question: 'Is there a king in the picture? Please answer yes or no.'
Now, please generate a question for the following sentence: Sentence: [Insert your sentence here]
Question:

For each generated image i, we employ the **Qwen-2.5-VL-7B-Instruct** model to answer these questions. The VQA-Score for an image i and its corresponding text prompt t is defined as the probability of the model generating the token "yes":

$$VQAScore(i, t) := P(\text{"yes"} \mid i, q(t))$$

where $q(t)$ represents the question set derived from prompt t. The final VQA-Score is obtained by averaging the results across all images and specific artist categories. This metric provides a more robust measure of whether the core semantic elements of the prompt are preserved after the style removal process.

B.4.2 Metrics for Object Erasing

To quantitatively evaluate the performance of the diffusion model in object erasing tasks, we follow the evaluation protocols established in *UCE*. The assessment focuses on two primary dimensions: **Erasure Success Rate** and **Model Fidelity**.

1) Erasure Success Rate (Acc$_e$) The erasure success rate measures whether the model effectively avoids generating the target object after the erasing process. We select 10 categories from the **Imagenette** dataset as erasing targets, with each category corresponding to a specific text prompt (e.g., "French Horn"). For each target category, we generate 500 images and classify them using a pre-trained **ResNet-50** classifier. The erasure accuracy (Acc_e) is calculated as:

$$Acc_e = \frac{\text{Number of images classified as target class}}{\text{Total generated images}} \times 100\%$$

A lower Acc_e indicates a more successful erasure of the target concept. Ideally, Acc_e should approach 0% after the concept has been removed.

2) Model Fidelity (Acc$_p$) Model fidelity (or Preservation Accuracy) evaluates whether the erasing operation interferes with the model's ability to generate other, non-target concepts. We assess this by calculating the classification accuracy of images generated from **non-erased categories**. Similar to the erasure test, we generate 500 images for each preserved category and use the same ResNet-50 classifier to determine the average accuracy (Acc_p):

$$Acc_p = \frac{\text{Number of images correctly classified as original class}}{\text{Total generated images}} \times 100\%$$

A higher Acc_p indicates that the model retains its original generative capabilities for unrelated concepts with minimal collateral damage.

3) Comprehensive Evaluation (H_0) To quantify the overall utility balance between successful erasure and concept preservation, we compute the harmonic mean (H_0) of the erasure success (represented as $1 - Acc_e$) and the preservation accuracy (Acc_p):

$$H_0 = \frac{2}{(1 - Acc_e)^{-1} + (Acc_p)^{-1}}$$

A higher H_0 score indicates a superior capability to precisely excise target objects while maintaining high fidelity for remaining concepts.

B.4.3 Metrics for Celebrity Removal

The experiment follows a step-by-step removal process. After each celebrity is erased, we evaluate the model's recognition status for **all four** celebrities (the ones already erased and the ones yet to be erased). For each celebrity, we generate a diverse set of images using 5 augmented prompt templates and 20 fixed random seeds, resulting in 100 evaluation images per celebrity per stage. The prompts are as follows: *"a portrait of {celebrity name}"*, *"a sketch of {celebrity name}"*, *"an oil painting of {celebrity name}"*, *"{celebrity name} in an official photo"*, and *"an image capturing {celebrity name} at a public event"*.

We employ the **GIPHY Celebrity Detector (GCD)** to identify the presence of specific celebrities in the generated images. The core metric is the **Recognition Accuracy (RA)**, defined as the percentage of images correctly identified as the target celebrity:

$$RA = \frac{N_{correct}}{N_{total}} \times 100\%$$

where $N_{total} = 100$ in our setup, and $N_{correct}$ is the number of images where the GCD successfully detects the target identity.

The RA metric serves two distinct purposes depending on the status of the celebrity.For celebrities who have already been erased, a lower RA is desired, indicating that the model has successfully removed the target facial features and the detector can no longer recognize the identity.For celebrities who have not yet been erased, a higher RA is desired. This demonstrates the model's ability to maintain high-fidelity generation for unrelated concepts and its robustness against catastrophic forgetting or semantic interference during the continual erasure.

C More Results

C.1 Additional Object Erasure Results

This section supplements the main text Table 2 with complete accuracy data for all 10 erased classes and 10 preserved classes. Table 8 presents the accuracy of various methods on erased classes (Acc_e, lower is better), while Table 9 shows the accuracy on preserved classes (Acc_p, higher is better).

Table 8: **Complete $\mathbf{Acc_e}$ results on all ten erased classes.** This table extends the Acc_e columns of Table 2 by providing per-class accuracy for all ten erased concepts. Lower values indicate better erasure.

Method	$\text{Acc}_e \downarrow$									
	tench	english springer	cassette player	chain saw	church	french horn	garbage truck	gas pump	golf ball	parachute
SD v1.4 (Base)	0.750	0.930	0.180	0.650	0.810	1.000	0.780	0.570	0.990	0.930
UCE	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.040	**0.000**
ESD	**0.000**	**0.000**	**0.000**	**0.000**	0.010	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
CCRT	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.010	**0.000**	**0.000**	**0.000**
MACE	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
SPEED	0.470	0.190	**0.000**	0.010	0.590	0.080	0.410	0.300	0.760	0.530
CUIG GradProj	**0.000**	**0.000**	**0.000**	**0.000**	0.050	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
LCCE (Ours)	0.010	**0.000**	**0.000**	**0.000**	0.010	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

Table 9: **Complete $\mathbf{Acc_p}$ results on all ten preserved classes.** This table complements the Acc_p columns of Table 2 with per-class accuracy for all ten preserved concepts. Higher values indicate better preservation of unrelated classes.

Method	$\text{Acc}_p \uparrow$									
	light-house	pome-granate	banana	electric guitar	pizza	solar dish	golden retriever	tabby cat	suit	spoonbill
SD v1.4 (Base)	0.980	0.920	0.980	0.930	0.990	0.880	1.000	0.830	0.870	0.860
UCE	0.070	0.010	0.670	0.130	0.000	0.120	0.000	0.000	0.250	0.020
ESD	0.100	0.000	0.000	0.010	0.000	0.030	0.000	0.000	0.040	0.000
CCRT	0.010	0.020	0.000	0.000	0.040	0.000	0.000	0.000	0.010	0.000
MACE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SPEED	**0.990**	**0.890**	**0.980**	**0.960**	0.990	0.860	**1.000**	**0.830**	0.880	**0.870**
CUIG GradProj	0.950	0.750	0.960	0.640	**1.000**	0.490	0.930	0.530	**0.910**	0.860
LCCE (Ours)	0.910	0.860	0.760	0.830	0.990	**0.980**	0.990	0.630	0.900	0.750

Table 10: **Stage-wise RR-LLM scores for the first six artist erasures.** Each row denotes the model after the corresponding number of sequential erasures. Bold values correspond to artists that have already been targeted. Higher RR-LLM indicates more effective style removal.

Stage	Leonardo da Vinci	Vincent van Gogh	Pablo Picasso	Michelangelo	Rembrandt	Claude Monet
Original	0.02	0.04	0.10	0.08	0.18	0.06
1	**0.92**	0.02	0.12	0.08	0.16	0.10
2	**0.96**	**0.94**	0.14	0.09	0.14	0.08
3	**0.94**	**0.92**	**0.96**	0.08	0.16	0.10
4	**0.96**	**0.94**	**0.96**	**0.94**	0.14	0.06
5	**0.94**	**0.96**	**0.92**	**0.96**	**0.94**	0.08
6	**0.96**	**0.94**	**0.96**	**0.94**	**0.92**	**0.96**

Table 11: **GCD recognition accuracy during sequential celebrity erasure.** Each row represents the model after an additional celebrity is erased. Low accuracy is desirable for already erased identities, while high accuracy is desirable for identities not yet targeted.

Stage	CCRT				**LCCE (Ours)**			
	Adam Driver	Bill Clinton	Charlie Sheen	Drew Barrymore	Adam Driver	Bill Clinton	Charlie Sheen	Drew Barrymore
SD v1.4	0.96	0.81	0.94	0.93	0.95	0.89	0.95	0.92
+ Adam Driver	0.12	0.76	0.88	0.80	**0.04**	0.91	0.91	0.94
+ Bill Clinton	0.04	**0.04**	0.75	0.65	**0.03**	0.07	0.92	0.92
+ Charlie Sheen	**0.01**	**0.00**	0.09	0.55	0.05	0.02	**0.03**	0.92
+ Drew Barrymore	0.04	0.10	0.08	**0.01**	0.03	**0.01**	**0.03**	0.03

C.2 Stage-wise Artist Erasure Trajectories

To examine whether erasing one artist inadvertently affects other, semantically related styles, we report the stage-wise RR-LLM scores for the first six artists in the erasure sequence. At stage t, only the first t artists have been erased. A high score is therefore desirable for artists already targeted, whereas a low score for future targets indicates that their styles remain recognizable before their designated erasure stages.

Table 10 shows a clear transition at each artist's designated erasure stage. Before being targeted, the RR-LLM scores remain low and close to those of the original model; immediately after erasure, they increase to at least 0.92. This indicates that editing one style does not prematurely suppress the other artist styles, despite their semantic and stylistic proximity. Meanwhile, the scores of previously erased artists remain between 0.92 and 0.96 throughout subsequent stages, showing that later updates do not substantially revive earlier styles. These trajectories provide stage-level evidence for both locality across related concepts and persistent erasure under sequential editing.

C.3 Continuous Celebrity Erasure

We further evaluate the temporal stability and identity-level locality of LCCE through sequential celebrity erasure. Four celebrities are erased in the order of Adam Driver, Bill Clinton, Charlie Sheen, and Drew Barrymore. After each stage, we generate 100 images for every celebrity and evaluate identity recognition using the GIPHY Celebrity Detector (GCD). A lower detection accuracy indicates stronger erasure for an already targeted identity, whereas a high accuracy for a future target indicates limited collateral interference.

Table 11 reveals two properties required for continuous erasure. First, LCCE maintains previously completed erasures across subsequent updates. After being targeted, the recognition accuracy of

each celebrity remains at or below 0.07, and all four identities reach an accuracy of at most 0.03 after the final stage. This indicates that later edits do not substantially revive previously erased identities.

Second, LCCE limits interference with identities that have not yet been targeted. Before their respective erasure stages, the recognition accuracies of Bill Clinton, Charlie Sheen, and Drew Barrymore remain close to their original levels. In particular, Drew Barrymore retains an accuracy of 0.92 after the first three erasures, whereas CCRT decreases progressively from 0.93 to 0.55 before Drew Barrymore is explicitly targeted. Similar premature degradation is observed for the other future targets under CCRT. These results show that LCCE provides stronger identity-level locality while preserving the persistence of earlier erasures.

D Qualitative Visualizations

Figures 5 and 6, illustrate the visualization results of continuously erasing artists and celebrities, respectively. These figures clearly demonstrate that our method effectively erases target concepts while well-preserving the integrity of un-erased concepts, with the erasure being clean and lasting.

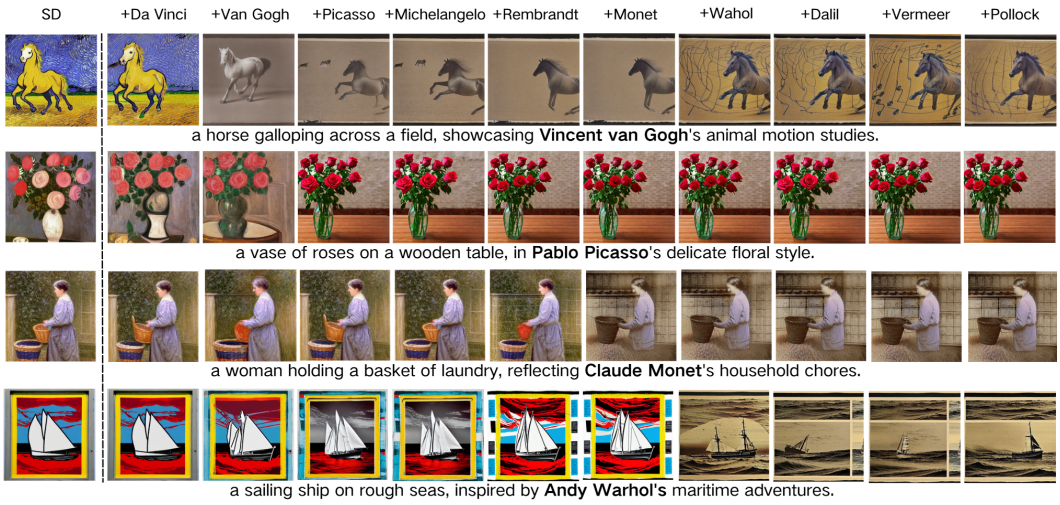


Figure 5: Continual erasure of visualization results from 10 artists.

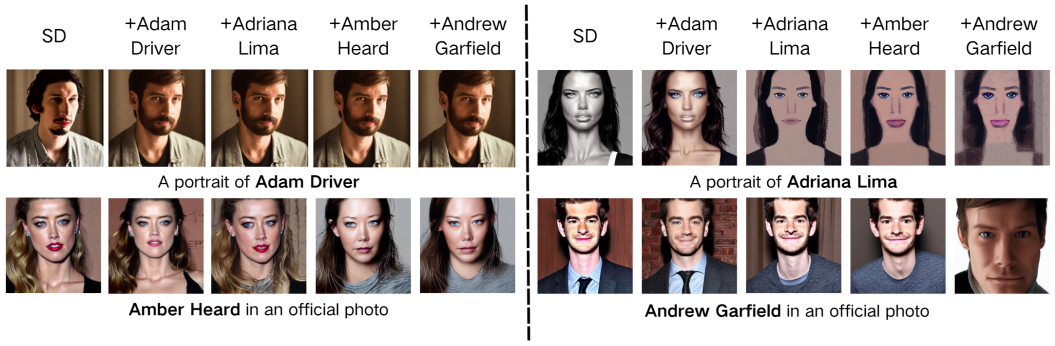
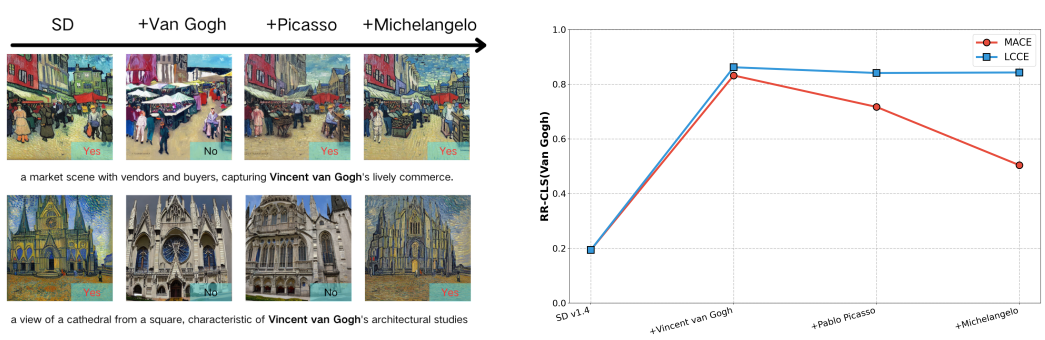


Figure 6: Continual erasure of visualization results from 4 celebrities.

E Concept Revival Discussion

When addressing continuous concept erasure tasks, we observe a critical challenge that is uncommon in single-step erasure methods but becomes pivotal: Concept Revival. When these methods are



(a) (b)

Figure 7: **(a)Visualization of Concept Revival in MACE.** In the bottom right corner of each image is the RR-LLM result, where "No" indicates successful erasure and "Yes" indicates style residue. **(b) Visualization of RR-CLS Trend.** The figure shows the RR-CLS for the Van Gogh style, where MACE's RR-CLS for Van Gogh gradually decreases during subsequent erasures, indicating that the erased style has been partially revived.

applied in multi-round iterative erasure scenarios, concepts that were successfully erased in earlier steps may unexpectedly re-emerge during subsequent erasure operations, thereby undermining the effectiveness of prior erasures.

Take the MACE method as an example, it performs remarkably well in single-step concept erasure tasks. However, as illustrated in Figure 7a, when MACE is employed in a continuous erasure context, this revival phenomenon becomes particularly evident. Through RR-CLS quantification(Figure 7b), we observe that for the "Van Gogh" style, its RR-CLS value peaks after the first erasure but significantly declines in subsequent erasure steps, directly confirming that the erasure efficacy for the "Van Gogh" style is compromised during later operations.

In stark contrast, our proposed method, LCCE, consistently maintains high RR-CLS values for all erased concepts throughout the entire continuous erasure process. This demonstrates that LCCE effectively avoids the concept revival problem, ensuring that each erasure is both clean and permanent, thereby delivering more stable and reliable performance in multi-round iterative erasure tasks.

F Human Study

F.1 Evaluation Setup

To complement the automatic evaluations in Section 5, we conducted a human study on continuous artist-style erasure. The study evaluates the three dimensions used throughout our experiments: **Erasure Efficacy** (EE), **Generative Fidelity** (GF), and **Concept Locality** (CL).

For each setting, we generated 10 images for each method. Participants ranked the eight methods from 1 (best) to 8 (worst) according to the specified criterion. For EE, participants were shown reference images representing the target artist style and ranked the outputs according to the extent to which the target style had been removed. For GF, participants were given the original prompts and ranked the outputs according to visual quality, structural coherence, and preservation of fine-grained details. For CL, participants evaluated whether unrelated visual content and the semantic intent of the prompts were preserved without collateral distortion. All participants completed the rankings independently.

F.2 Results and Analysis

Table 12 reports the average ranks after sequentially erasing 4, 8, and 16 artist styles. Lower values indicate better performance. LCCE achieves the best overall average rank of 2.57, followed by CUIG GradProj with 3.91. This result is consistent with the central goal of continuous erasure:

maintaining a favorable balance between suppressing requested styles and preserving the quality and locality of the generated images.

Erasure Efficacy. LCCE achieves the best EE rank after erasing 8 and 16 artists, with scores of 1.8 and 2.0, respectively, and ranks second after 4 artists with a score of 2.3. CUIG GradProj provides the strongest competing sequential baseline, whereas methods with weaker long-term retention exhibit less consistent performance as the number of erased artists increases. The improving relative rank of LCCE at longer sequences indicates that historical retention becomes increasingly important as more edits are accumulated.

Generative Fidelity. LCCE obtains the best GF rank at all three sequence lengths, with scores of 3.7, 3.5, and 1.5 after erasing 4, 8, and 16 artists. Its advantage becomes more pronounced for the longest sequence, where repeated edits make non-target degradation more visible. These results suggest that the cumulative preservation bank and dynamic queue help maintain coherent non-target generation while the model is repeatedly edited.

Concept Locality. The unedited SD-v1.4 model obtains the best CL scores because it does not incur any editing-induced collateral change. Among the edited methods, LCCE consistently achieves the strongest locality, ranking second at all three sequence lengths with scores of 3.7, 2.2, and 2.4. This indicates that LCCE preserves unrelated content more effectively than the other edited models, including CUIG GradProj, while still providing substantially stronger erasure efficacy.

Overall, the human rankings agree with the automatic evaluation: LCCE provides the most balanced behavior across erasure efficacy, generative fidelity, and concept locality, particularly as the number of sequential erasure requests increases.

Table 12: Human study results under different numbers of erased artist styles. Values are average ranks (1 = best, 8 = worst). Bold indicates the best (lowest rank) per row, and underline indicates the second-best. EE denotes Erasure Efficacy, GF denotes Generative Fidelity, and CL denotes Concept Locality.

Erased	Aspect	SD	UCE	ESD	CCRT	MACE	SPEED	CUIG	LCCE
4	EE	8.0	**1.9**	4.1	4.6	5.6	6.3	3.2	2.3
	GF	4.0	5.1	4.6	4.8	5.0	4.5	4.3	**3.7**
	CL	**1.2**	6.5	7.6	4.4	5.3	2.5	4.8	3.7
8	EE	8.0	2.0	5.1	4.3	6.0	5.8	3.0	**1.8**
	GF	3.9	5.5	5.2	4.7	4.4	4.2	4.6	**3.5**
	CL	**1.1**	7.2	5.7	6.1	3.8	4.9	5.0	2.2
16	EE	8.0	3.2	4.3	4.1	5.4	6.3	2.7	**2.0**
	GF	3.9	7.1	5.4	6.3	4.8	2.7	4.3	**1.5**
	CL	**1.6**	6.9	7.7	5.8	4.2	4.1	3.3	2.4
Average		4.41	5.04	5.52	5.01	4.94	4.59	3.91	**2.57**

G Hyperparameter Sensitivity Analysis

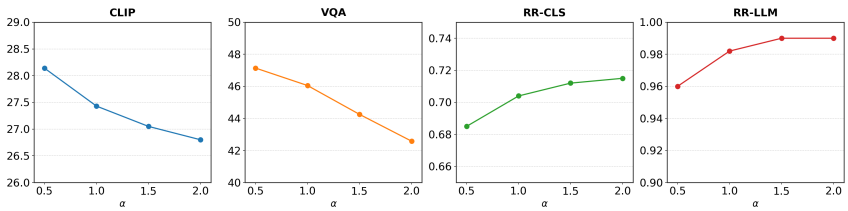


Figure 8: Sensitivity analysis of α. Increasing α improves erasure performance (RR-CLS, RR-LLM) while marginally reducing text-image alignment and visual quality (CLIP, VQA).

To examine the robustness of LCCE to its controller parameters, we vary the budget multiplier α and the queue step a. We evaluate their effects on sequential artist-style erasure using CLIP-Score, RR-LLM, RR-CLS, and VQA-Score. The results are summarized in Table 13 and Fig. 8. Unless otherwise specified, all remaining parameters are fixed to their default values.

(α, a)	CLIP↑	RR-LLM↑	RR-CLS↑	VQA↑
$(1, \frac{1}{\sqrt{D}})$	27.43	0.98	0.70	46.05
$(0.5, \frac{1}{\sqrt{D}})$	28.14	0.96	0.69	47.14
$(2, \frac{1}{\sqrt{D}})$	26.80	0.99	0.72	42.58
$(1, \frac{0.5}{\sqrt{D}})$	27.29	0.98	0.70	45.86
$(1, \frac{2}{\sqrt{D}})$	27.53	0.97	0.71	46.21

Table 13: Sensitivity of LCCE to the budget multiplier α and queue step a. The first row is the default setting used in the main experiments.

As shown in Fig. 8, increasing α enlarges the layer-wise preservation budget

$$D = \alpha PL(1),$$

where $PL(1)$ is the preservation loss measured after the first edit. A larger budget permits stronger subsequent updates before the virtual queue increases the preservation pressure, which generally improves erasure efficacy but may slightly reduce generation fidelity. The results in Table 13 show that moderate changes in the queue step a lead to relatively limited performance variation within the tested range. We therefore use $(\alpha, a) = (1, 1/\sqrt{D})$ as the default setting, providing a balanced trade-off between target erasure and non-target preservation.

H Limitations and Scope

LCCE targets the continuous erasure setting of Section 3: requests arrive one at a time, each names a concept to remove, and the model must keep earlier erasures while preserving non-target generation. We note three directions that lie beyond this setting.

Adversarial prompts. Our evaluation asks whether erased concepts return under natural prompts as later edits accumulate. Adversarially optimized prompts, which search the input space for residual target responses, define a separate threat model with its own objective. LCCE is compatible with defenses against it: because each stage is solved in closed form, adversarial prompts mined for a target can be added to its target set C_t without iterative fine-tuning, and the historical retention loss then keeps their erased responses fixed in later stages.

Scope of each request. Each target is redirected to a neutral safe target (Appendix B.3), so an edit removes the requested concept while leaving the rest of the prompt intact. Undesired concepts that have not been requested are therefore left unchanged, which is the locality the continuous setting requires. Deployments that need category-level removal can issue the members of a category as consecutive requests, a regime the cumulative preservation bank and the long-term budget are built to handle.

Backbone architectures. Our experiments use Stable Diffusion v1.4, the standard backbone for concept erasure. LCCE only assumes that text conditioning enters the denoiser through linear projections, a structure shared by cross-attention U-Nets and text-conditioned diffusion transformers. Evaluating LCCE on diffusion transformer backbones is a natural next step.