Localized-Attention-Guided Concept Erasure for Text-to-Image Diffusion Models

Zhuan Shi[*]
McGill University and Mila
zhuan.shi@mila.quebec

Alireza Dehghanpour Farashah[*]
McGill University and Mila
alireza.farashah@mila.quebec

Rik de Vries[*]
EPFL - Swiss Federal Technology Institute of Lausanne
rik.de.vries23@gmail.com

Golnoosh Farnadi
McGill University and Mila
farnadig@mila.quebec

Abstract

Concept erasure has become a fundamental safety requirement for text-to-image diffusion models, enabling removal of objectionable or copyrighted content without costly retraining. To preserve generative capacity, localized concept erasure is proposed which confines edits to the region occupied by the target concept and leaves the remainder of the scene untouched. However, existing localized concept erasure still suffer from a Concept Neighborhood gap: suppressing the target often attenuates neighboring, semantically related concepts, diminishing overall fidelity and limiting practical utility. To bridge this gap, we present Localized-Attention-Guided Concept Erasure (LACE), a training-free framework whose three stages progress from coarse to fine control: (1) Representation-space projection, which suppresses the target concept subspace while reinforcing semantic neighbors; (2) Attention-guided spatial gate, which derives a spatial mask identifying regions of residual concept activation and conduct attention suppression; (3) Gated Feature Clean-up, which performs a hard scrub on gated feature activations. This three-stage pipeline enables precise and localized removal of visual concepts while retaining semantic structure and expressiveness. Experiments show that LACE effectively removes targeted concepts, preserves semantically related neighbors, and maintains overall image composition.

1 Introduction

Recently, text-to-image (T2I) diffusion models have been widely adopted in creative and industrial domains for generating high-quality visuals from a wide range of prompts Song et al. [2020], Nichol et al. [2021], Rombach et al. [2022], Ramesh et al. [2022], Saharia et al. [2022], Yang et al. [2023]. However, their training on large-scale, uncurated datasets Schuhmann et al. [2022], Carlini et al. [2019] poses risks of reproducing copyrighted artistic style Jiang et al. [2023], Setty [2023], Shi et al. [2024a] or harmful content Mirsky and Lee [2021], Schramowski et al. [2023]. To enable safe and compliant deployment, concept erasure, which refers to the removal of specific visual concepts from the model's generative capacity, has become a critical requirement.

Recent studies on concept erasure in T2I diffusion models mainly follow two lines. Training-based approaches fine-tune model components, modify prompt embeddings, or apply gradient-driven edits to suppress target concepts Li et al. [2024a], Liu et al. [2024], Zhang et al. [2024], Liang et al. [2024]. While often effective, they demand substantial computation, weight access, and carefully curated

[*]Equal Contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: GenProCC.

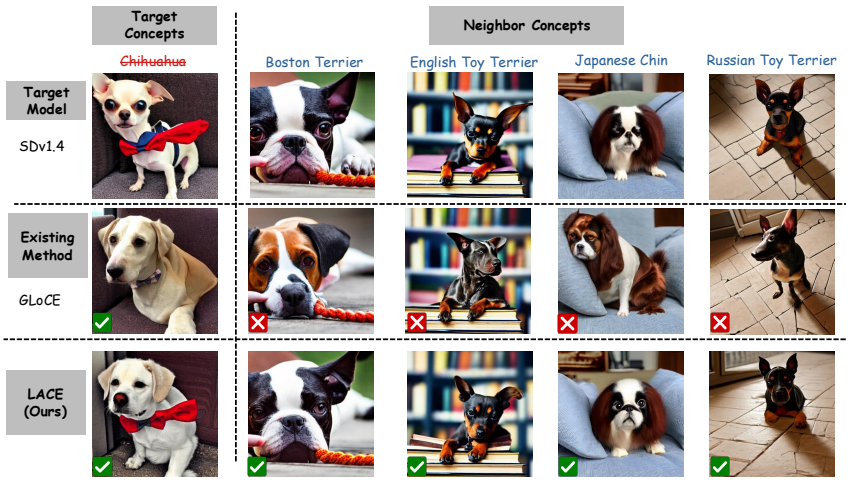


Figure 1: Erasure Effectiveness and Neighbor Retention: GLoCE vs. LACE(Ours)

data to prevent forgetting or unintended shifts. Training-free approaches, in contrast, operate in the input or latent space, such as projection into orthogonal subspaces Gandikota et al. [2023, 2024], Gong et al. [2024], Biswas et al. [2025] or attention suppression Orgad et al. [2023], which remove concepts without retraining. Despite their efficiency, these global operations can still distort unrelated content or cause semantic drift, motivating the need for more localized and targeted erasure strategies.

To maintain fidelity and spatial precision, **Localized Concept Erasure** has recently been proposed Lee et al. [2025]. Instead of suppressing a concept globally across all latent features, localized erasure seeks to remove the target concept only from the region of the image where it visibly appears, leaving the remainder of the scene untouched. In detail, GLoCE introduces a gated low-rank adapter that attenuates the influence of the target token during denoising. However, we identify a critical limitation in this formulation: the *neighbor gap*-a phenomenon where semantically adjacent concepts are unintentionally suppressed alongside the target. To illustrate this, we conduct an experiment in which a specific dog breed is erased. As shown in Fig. 1, while both methods succeed in suppressing the target, GLoCE also degrades the generation quality of other dog breeds, indicating a lack of semantic precision in preserving neighboring concepts.

To bridge this gap, we introduce Localized-Attention-Guided Concept Erasure(**LACE**), a training-free framework that performs localized concept erasure while explicitly preserving the semantics of neighboring concepts and global image quality. LACE follows a three-stage pipeline that progresses:

- **Stage 1: Representation-Space Projection.** we perform spectrum-aware projection in the token embedding space to suppress the semantic subspace of the target concept while preserving nearby semantics. We apply this projection to the Key and Value matrices in the UNet cross-attention layers, ensuring consistent erasure across both the input prompt and the model's internal attention activations.
- **Stage 2: Attention-Guided Spatial Gating.** We execute a forward pass to extract attention maps from an early cross-attention layer. Then we identify live target tokens based on their projection magnitude and construct a spatial gate that highlights regions where residual concept attention remains. This gate is reused across layers to suppress attention toward target tokens in a second forward pass.
- **Stage 3: Gated Feature Clean-up.** For regions identified by the attention gate, we apply a hard scrub operation that projects UNet hidden features away from the target concept subspace. This ensures complete elimination of residual traces without affecting the rest of the image.

Extensive experiments on 3 datasets across diverse prompts and erasure scenarios demonstrate that LACE effectively removes targeted concepts, preserves neighbor concepts and maintains the quality of the generated images.

Our contributions are as follows:

- We identify and formalize the Concept Neighborhood gap in current localized, training-free concept erasure methods.
- We propose LACE, a principled, training-free pipeline that explicitly preserves neighbor concepts while achieving complete and localized erasure of the target.
- Experiments show that LACE achieves high erasure precision, preserves semantic neighbors, and maintains overall generation quality across different datasets and settings.

2 Related Work

Concept Erasure in Text-to-Image Diffusion Models. Training-based interventions are the dominant approach to concept erasure, enabling control through parameter updates. Typical methods include retraining or finetuning on filtered datasets with negative guidance Li et al. [2024b], Gandikota et al. [2023], Zhang et al. [2024], Chin et al. [2023], or minimizing divergence between harmful and safe concepts Shi et al. [2024b]. Other strategies involve adversarial training Kim et al. [2024], preference optimization Park et al. [2024], Das et al. [2024], and self-supervised latent manipulation Li et al. [2024a]. Partial finetuning targets specific layers to forget undesired knowledge Lu et al. [2024], Heng and Soh [2023]. While effective, these methods require substantial computation and risk overfitting or forgetting Chang et al. [2024]. Recent efforts shift towards training-free strategies. Some methods directly mask latent features correlated with the concept Orgad et al. [2023], others apply projection into the null space of semantic embeddings Gandikota et al. [2024], Gong et al. [2024], and spectral methods like CURE Biswas et al. [2025] decompose activations to suppress concept-aligned directions. Despite their practicality, these techniques often erase concepts too aggressively or too imprecisely, harming surrounding visual fidelity.

Localized Concept Erasure and its Limitations. To balance safety and generative quality, *Localized Concept Erasure* (LCE) was proposed in GLoCE Lee et al. [2025], which limits erasure to the spatial and temporal regions where the target concept appears. It uses gated LoRA adapters applied only to attention-predicted areas. While LCE improves visual fidelity and avoids global degradation, it only attenuates rather than fully removes the concept, allowing it to resurface during denoising. The learned gating mechanism adds complexity and lacks flexibility. Crucially, it does not preserve semantically related neighbor concepts, which may also be unintentionally suppressed. Our method, LACE, builds upon LCE by applying explicit projection and residual feature removal, while preserving neighbor semantics to maintain visual and structural consistency.

Concept Neighborhood Preservation. The *Concept Neighborhood problem* refers to the unintended removal of semantically related but valid concepts during erasure. FADE Thakral et al. [2025] highlights this issue, showing that removing “cat" can unintentionally affect “lion" or “tiger" due to shared embedding components. Although FADE mitigates this via disentangled attention filters, it relies on finetuning and cannot fully prevent leakage. Inspired by FADE, we address this challenge in a training-free setup by explicitly constructing a neighbor concept subspace. We identify the target concept's top-k semantic neighbors using cosine similarity in CLIP embedding space, followed by concreteness and popularity filtering. These neighbors are then used to guide selective projection that removes the target direction while preserving related semantics.

3 Preliminaries and Problem Formulation

Let $\mathcal{M}_\theta$ denote a pretrained text-to-image diffusion model parameterized by θ, which synthesizes image samples $\mathbf{x}_{0:T}$ conditioned on a text prompt $\mathcal{P}$ through a denoising process. Given a discrete vocabulary $\mathcal{V}$ and its associated token embeddings $E \in \mathbb{R}^{|\mathcal{V}| \times d}$, the prompt $\mathcal{P}$ is tokenized as a sequence $[w_1, \ldots, w_n]$, with corresponding CLIP embeddings $[x_1, \ldots, x_n]$.

Let $F \subset \mathcal{V}$ be a set of target tokens representing the visual concept to be erased. Let $N \subset \mathcal{V}$ denote the neighborhood of F—tokens semantically close to the target concept but not to be removed. Our goal is to construct a modified model $\tilde{\mathcal{M}}$ that satisfies:

- **Erasure completeness:** For any prompt containing tokens in F, images sampled from $\tilde{\mathcal{M}}$ should not exhibit recognizable visual traces of the target concept.

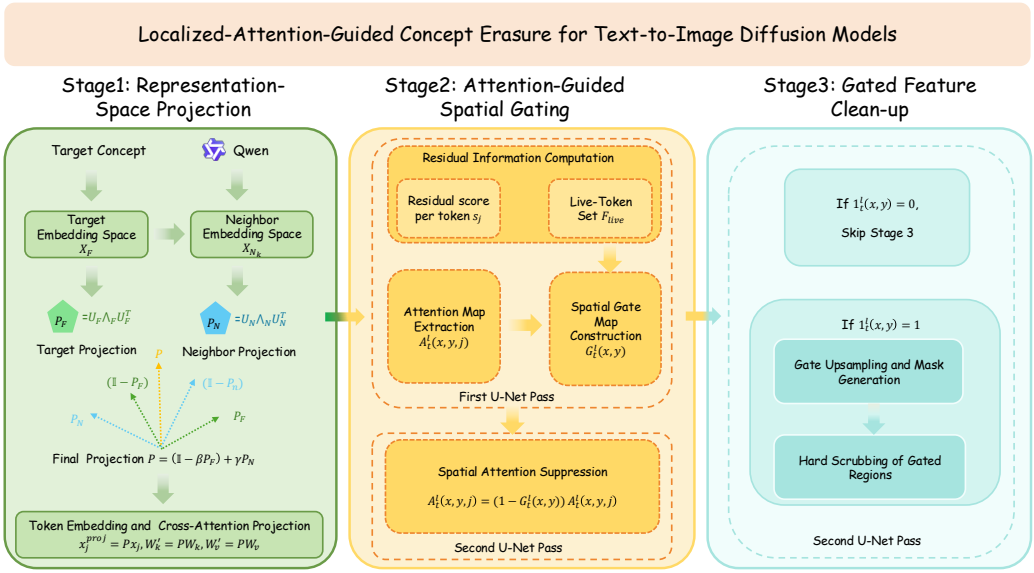


Figure 2: Workflow of LACE.

- **Neighbor preservation:** For prompts containing tokens in $N \setminus F$, the output distribution of $\tilde{\mathcal{M}}$ should remain close to the baseline model $\mathcal{M}_\theta$.
- **Global quality retention:** For prompts not mentioning F or N, the image quality and diversity should remain nearly unchanged.
- **Training-free practicality:** The transition $\mathcal{M}_\theta \to \tilde{\mathcal{M}}$ must avoid backpropagation, auxiliary datasets, or network modifications.

4 Main Approach

We introduce Localized-Attention-Guided Concept Erasure(**LACE**), a training-free, three-stage pipeline that removes the target concept while preserving semantically related neighbor concepts and maintaining generation fidelity. As Fig. 2 shows, the three stages progressively refine the erasure: from representation-space projection, to spatial localization, to gated cleanup.

4.1 Stage 1: Representation-Space Projection

Stage 1 is to operate in the token-level representation space to selectively suppress the semantics of the target concept while restoring coherent generation capacity by reinforcing semantically related neighbor concepts. This is achieved by explicitly projecting embeddings away from the target concept subspace and re-injecting a controlled, weighted subspace spanned by neighbor tokens. This dual operation ensures that target concepts are erased at the semantic embedding level, while the expressiveness and plausibility of the output are preserved.

Identifying the Target Concept Subspace. Let $F \subset \mathcal{V}$ be the set of target tokens to be erased, and $X_F = \{x_j \mid w_j \in F\} \subset \mathbb{R}^d$ be their corresponding embeddings. We compute the low-rank representation of the target concept via Singular Value Decomposition(SVD)Wall et al. [2003], Baker [2005], obtaining a rank-r orthonormal basis:

$$X_{\mathcal{F}} = U_{\mathcal{F}} \Sigma_{\mathcal{F}} V_{\mathcal{F}}^{\top}, \quad U_{\mathcal{F}} \in \mathbb{R}^{d \times r} \tag{1}$$

where $\Sigma_{\mathcal{F}} = \mathrm{diag}(s_1, \ldots, s_r)$.

To reflect the relative contribution of each direction, inspired by Biswas et al. [2025], we adopt Spectral Expansion mechanism for spectral regularization which selectively modulates singular

vectors based on their relative significance to control the strength of forgetting. Specifically, we define the spectral expansion function as:

$$\lambda_i^{(\mathcal{F})} = \frac{\alpha_{\text{target}} \cdot r_i^{(\mathcal{F})}}{(\alpha_{\text{target}} - 1) \cdot r_i^{(\mathcal{F})} + 1}, \quad \text{where } r_i^{(\mathcal{F})} = \frac{s_i^2}{\sum_j s_j^2}, \tag{2}$$

Then we compute the corresponding projection operator that captures the target concept subspace as follows:

$$P_{\mathcal{F}} = U_{\mathcal{F}} \Lambda_{\mathcal{F}} U_{\mathcal{F}}^\top, \quad \text{where } \Lambda_{\mathcal{F}} = \text{diag}(\lambda_1^{(\mathcal{F})}, \ldots, \lambda_r^{(\mathcal{F})}), \tag{3}$$

This formulation enables fine-grained control over suppression strength along each semantic axis of the target concept.

Mining Neighbor Concept. Given the target concepts $X_{\mathcal{F}}$, we use the following steps to obtain Neighbor concepts:

- **Embedding-Based Retrieval.** Let $\mathcal{C}_{\text{all}}$ denote a large external concept pool (e.g., Wikipedia titles). Using a pretrained sentence embedding model (e.g., Qwen-embedding), we compute cosine similarities between each $x_f \in X_{\mathcal{F}}$ and all $x_i \in \mathcal{C}_{\text{all}}$:

$$\cos(x_f, x_i) = \frac{x_f^\top x_i}{\|x_f\| \cdot \|x_i\|}. \tag{4}$$

  We select the top-k most similar concepts to form an initial candidate set $\mathcal{C}_k$.
- **Concreteness Filtering.** We use a pretrained RoBERTa-based SVR model Wartena [2024] to estimate a concreteness score $s_i \in [1, 5]$ for each candidate $c_i \in \mathcal{C}_k$. Only concepts with $s_i \geq \tau$ (e.g., $\tau = 3.5$) are retained.
- **Popularity Filtering.** To remove obscure concepts, we enforce a minimum popularity threshold $\text{Pop}(c_i) \geq P_{\text{thresh}}$. We employ page view statistics as a surrogate for popularity.
- **CLIP-based Final Reranking.** Remaining candidates are re-ranked by their CLIP similarity to the original target embeddings. The final top-k embeddings $\{x_j\}_{j \in \mathcal{N}_k}$ form the neighbor concept set $\mathcal{N}_k$.

Neighbor Subspace Construction. Let $X_{\mathcal{N}_k}$ be the stacked embeddings of the selected neighbors. We perform SVD to extract a low-rank basis:

$$X_{\mathcal{N}_k} = U_{\mathcal{N}} \Sigma_{\mathcal{N}} V_{\mathcal{N}}^\top, \quad U_{\mathcal{N}} \in \mathbb{R}^{d \times r}. \tag{5}$$

where $\Sigma_{\mathcal{F}} = \text{diag}(\sigma_1, \ldots, \sigma_r)$.

Similar to Target Concept Subspace, we firstly define the spectral expansion function:

$$\lambda_i^{(\mathcal{N})} = \frac{\alpha_{\text{neighbor}} \cdot r_i^{(\mathcal{F})}}{(\alpha_{\text{neighbor}} - 1) \cdot r_i^{(\mathcal{F})} + 1}, \quad \text{where } r_i^{(\mathcal{F})} = \frac{\sigma_i^2}{\sum_j \sigma_j^2}, \tag{6}$$

Then we construct neighbor space as follows:

$$P_{\mathcal{N}} = U_{\mathcal{N}} \Lambda_{\mathcal{N}} U_{\mathcal{N}}^\top, \quad \text{where } \Lambda_{\mathcal{N}} = \text{diag}(\lambda_1^{(\mathcal{N})}, \ldots, \lambda_r^{(\mathcal{N})}), \tag{7}$$

This projection operator $P_{\mathcal{N}}$ reinforces directions aligned with semantically related neighbor concepts while maintaining a data-driven structure.

Final Projection Operator. We define a composite projection operator that simultaneously removes the target concept and injects neighbor semantics:

$$P = (I - \beta P_{\mathcal{F}}) + \gamma P_{\mathcal{N}}, \tag{8}$$

where $\beta, \gamma \in [0, 1]$ are user-defined hyperparameters that control the erasure and reinjection strength, respectively.

Prompt Embedding Rewriting. For each token embedding in the prompt:

$$x_j^{\text{proj}} = \begin{cases} Px_j, & w_j \in \mathcal{F} \\ x_j, & \text{otherwise} \end{cases} \tag{9}$$

UNet Cross-Attention Rewriting Let $W_K, W_V \in \mathbb{R}^{d \times d}$ be the original Key and Value projection matrices. Apply:

$$W_K' = PW_K, \quad W_V' = PW_V \tag{10}$$

This ensures that the attention mechanism no longer attends to the erased concept subspace, while reinforcing semantically coherent neighbor features.

4.2 Stage 2: Attention-Guided Spatial Gating

While Stage 1 neutralizes target semantics, some residual influence may persist in the network's attention flow. Stage 2 introduces a spatial attention gate to locate and suppress these signals. In detail, This stage modifies the model's cross-attention layers using a two-pass mechanism for each denoising timestep t.

Attention Map Extraction at First Pass. We run a dry forward pass using the modified embeddings and extract attention maps $A_t^\ell(x, y, j)$ from the DownBlock-2 of the UNet, where each A_t^ℓ reflects the attention at pixel (x, y) to token j at timestep t.

Residual Influence Detection. For each token x_j, we compute its activation under the erased concept subspace:

$$s_j = \|P_F x_j\|_2 \tag{11}$$

Then we build the Live-token Set F_{live} as follows:

$$\{j \mid s_j > \delta_{\text{token}}\} \tag{12}$$

Gate Map Construction. We derive a spatial gate $G_t(x, y)$ by summing attention over live target tokens:

$$G_t(x, y) = \sum_{j \in F_{\text{live}}} A_t^\ell(x, y, j) \tag{13}$$

This gate identifies the pixels where the residual presence of the target concept is detected.

Attention Suppression at the Second Pass. For each layer ℓ, we apply:

$$A^\ell(x, y, j) \leftarrow (1 - S_t(x, y)) \cdot A^\ell(x, y, j), \quad \text{if } j \in F_{\text{live}} \tag{14}$$

This suppresses target concept attention in gated spatial regions while preserving unaffected ones.

4.3 Stage 3: Gated Feature Clean-up

In the final stage of LACE, we remove any residual traces of the target concept that persist despite earlier representation-space and attention-level modifications. This stage acts only on pixels identified by the attention gate as retaining residual target activation and is completely bypassed otherwise, ensuring minimal interference.

Step 3.1 Gate Upsampling and Mask Generation
Let $S_t \in \mathbb{R}^{32 \times 32}$ be the spatial attention gate derived from Stage 2, indicating the cumulative attention mass over target tokens in DownBlock-2. For each scrubbed UNet layer ℓ operating at spatial resolution $H_\ell \times W_\ell$, we first upsample the gate using bilinear interpolation:

$$G_t^\ell = \text{Upsample}(G_t) \in \mathbb{R}^{H_\ell \times W_\ell} \tag{15}$$

We then compute a binary mask:

$$1_t^\ell(x, y) = \begin{cases} 1, & \text{if } G_t^\ell(x, y) \geq \delta_{\text{scrub}} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Table 1: Quantitative Comparison on Oxford Flowers Dataset.

	Camellia					Anthurium					Alpine Sea Holly				
	Acc_t (↓)	**Acc_r** (↑)	**Hcc** (↑)	**CLIP** (↑)	**KID** (↓)	**Acc_t** (↓)	**Acc_r** (↑)	**Hcc** (↑)	**CLIP** (↑)	**KID** (↓)	**Acc_t** (↓)	**Acc_r** (↑)	**Hcc** (↑)	**CLIP** (↑)	**KID** (↓)
SD	100.00	100.00	0.00	32.55	-	100.00	100.00	0.00	32.56	-	100.00	100.00	0.00	32.52	-
UCE	34.78	63.58	64.39	31.56	0.62	**0.00**	61.27	75.98	31.75	0.53	**0.00**	66.84	80.12	32.00	0.47
RECE	**0.00**	76.69	86.81	31.77	0.10	**0.00**	64.59	78.49	31.98	0.15	**0.00**	70.39	82.62	31.25	0.14
GLoCE	100.00	85.83	0.00	32.17	0.17	83.33	**85.15**	27.88	32.22	0.17	44.44	85.66	67.40	32.17	0.22
Ours	4.35	**95.36**	**95.50**	**32.41**	**0.07**	16.67	83.69	**83.51**	**32.68**	**0.08**	**0.00**	**87.11**	**93.11**	**32.55**	**0.11**

Table 2: Quantitative Comparison on Stanford Dogs Dataset: Concept Erasure for Different Dog Types

	Bluetick					Chesapeake Bay Retriever				
	Acc_t (↓)	**Acc_r** (↑)	**Hcc** (↑)	**CLIP** (↑)	**KID** (↓)	**Acc_t** (↓)	**Acc_r** (↑)	**Hcc** (↑)	**CLIP** (↑)	**KID** (↓)
SD	100.00	100.00	0.00	34.98	-	100.00	100.00	0.00	34.97	-
UCE	**0.00**	59.57	74.66	34.57	0.14	4.00	51.81	67.30	34.14	0.30
RECE	**0.00**	77.11	**87.08**	34.51	**0.04**	**0.00**	64.84	78.67	34.40	0.11
GLoCE	38.89	77.39	68.29	34.39	0.172	84.00	78.58	26.59	34.10	0.17
Ours	16.67	**78.43**	80.81	**34.70**	0.06	16.00	**79.13**	**81.49**	**34.70**	**0.07**

where $\delta_{\text{scrub}} \in [0, 1]$ is a fixed threshold. In practice, we activate Stage 3 only when any location in the gate satisfies $1_t^\ell(x, y) = 1$, indicating strong target presence.

Step 3.2 Hard Scrubbing of Gated Regions
At each activated scrub layer ℓ, we directly zero out the latent features at gated positions. Let $h_t^\ell(x, y) \in \mathbb{R}^d$ be the feature at position (x, y) and timestep t. We apply the following update:

$$h_t^\ell(x, y) \leftarrow \begin{cases} \mathbf{0}, & \text{if } 1_t^\ell(x, y) = 1 \\ h_t^\ell(x, y), & \text{otherwise} \end{cases} \tag{17}$$

This aggressive hard scrubber ensures that residual activations corresponding to the target concept are completely eliminated in identified regions. Unlike projection-based soft suppression, zeroing is non-reversible, and used only when prior stages are insufficient.

5 Experiments

5.1 Experimental Setup

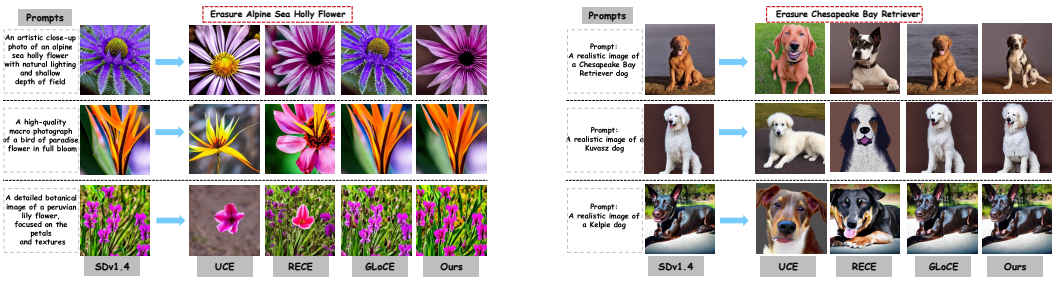
Dataset: (1) fine-grained datasets including Oxford Flowers Nilsback and Zisserman [2008] and Stanford Dogs Khosla et al. [2011] to assess concept erasure in high-similarity settings by removing one class and measuring retention on others; (2) localized erasure using the Celebrity dataset from GLoCE Lee et al. [2025], to evaluates fine-grained, identity-specific erasure while preserving co-occurring concepts.

Metrics: Our evaluation focuses on two main objectives: (i) effectively removing target concept and preserve retain concept, and (ii) preserving the model's ability to generate high-quality images for unrelated prompts.

To measure erasure, we use (1) **Target Accuracy ($\mathbf{Acc_t}$)**: the percentage of generated images that still contain the target concept after unlearning. Lower Acc_t indicates better forgetting. (2) **Retain Accuracy ($\mathbf{Acc_r}$)**: the percentage of unrelated or neighbor prompts that produce semantically correct outputs, Higher Acc_r indicates better retention. (3) **Harmonic mean of $(1 - \mathbf{Acc_t})$ and $\mathbf{Acc_r}$ (Hcc)**, following Lu et al. [2024], defined as:

$$Hcc = 2 \times \frac{(1 - Acc_t) \times Acc_r}{(1 - Acc_t) + Acc_r}$$

For generation quality, we use (1) **CLIP**: between the generated image and the input prompt, and (2) **Kernel Inception Distance (KID)** Bińkowski et al. [2018] over outputs from retained prompts to



(a) (b)

Figure 3: Qualitative comparison results on (a) Oxford Flowers (Removing *Alpine Sea Holly Flower*), (b) Stanford Dogs (Removing *Chesapeake Bay Retriever*).

evaluate visual fidelity after unlearning. KID computes the squared Maximum Mean Discrepancy (MMD) between feature representations of generated images by both original and unlearned model:

$$\begin{aligned}\mathrm{MMD}(p,q) = \mathbb{E}_{x,x'\sim p}\left[K(x,x')\right] + \mathbb{E}_{y,y'\sim q}\left[K(y,y')\right] \\ - 2\mathbb{E}_{x\sim p, y\sim q}\left[K(x,y')\right],\end{aligned} \tag{18}$$

Baselines: We evaluate our method against a selected set of state-of-the-art unlearning approaches, including UCE Gandikota et al. [2024], RECE Gong et al. [2024], and GLoCE Lee et al. [2025]. These baselines are chosen because they represent the most recent and effective strategies designed for concept erasure. In particular, GLoCE is the current SOTA for localized concept erasure, while UCE and RECE are among the strongest methods for non-localized erasure. We do not include older baselines, as our focus is on benchmarking against the most competitive and relevant methods in their respective settings.

The detail of **Environmental Setup and Hyper Parameters** is provided in Appendix A.

5.2 Results on Oxford Flowers

We first perform targeted unlearning of three flower types: Camellia, Anthurium, and Alpine Sea Holly, with results in Table 1 and Figure 3a. While UCE and RECE achieve low target class accuracy ($\mathrm{Acc_t}$), they markedly degrade non-target accuracy ($\mathrm{Acc_r}$). In contrast, our method attains similarly low $\mathrm{Acc_t}$ while preserving much higher $\mathrm{Acc_r}$, yielding the highest Hcc scores across all flowers and demonstrating a superior balance between erasure and retention. It also consistently achieves the highest CLIP and lowest KID scores, showing that erased samples remain photorealistic and semantically aligned with their prompts.

5.3 Results on Stanford Dog

We also evaluate on Stanford Dogs, a fine-grained benchmark with high intra-class variability. As shown in Table 2 and Figure 3b, RECE attains the lowest $\mathrm{Acc_t}$ for single-concept removal, but our method offers a better balance between $\mathrm{Acc_t}$ and $\mathrm{Acc_r}$, achieving competitive erasure while preserving generalization. In terms of quality, our approach ranks among the best across breeds and often surpasses others. Although GLoCE yields strong CLIP and KID scores, it fails to erase effectively, as indicated by high $\mathrm{Acc_t}$.

5.4 Results on Localized Celebrity Erasure

We evaluate localized concept erasure using the GLoCE benchmark Lee et al. [2025], which measures the ability to remove a target individual while preserving co-occurring identities. The benchmark covers four celebrities—Anna Kendrick, Elon Musk, Anne Hathaway, and Bill Clinton—with prompts in the form "an image of [target] and [retained]" (e.g., "an image of Elon Musk and Amanda Seyfried"). Accuracy is computed via a pretrained celebrity classifier Hasty et al. [2024]. Using the same dataset and prompt structure as GLoCE ensures direct comparability and highlights our method's precision in identity-specific erasure. Qualitative results are shown in Figure 4, where each row corresponds to a target concept to be unlearned (e.g., *Bill Clinton* in the first row, *Elon Musk* in the second), with quantitative results in Table.6 in Appendix B.

Table 3: Multiple Concept Erasure on Oxford Flowers and Stanford Dog

Method	10 Flowers					10 Dogs				
	Acc_t (↓)	**Acc_r (↑)**	**Hcc (↑)**	**CLIP (↑)**	**KID (↓)**	**Acc_t (↓)**	**Acc_r (↑)**	**Hcc (↑)**	**CLIP (↑)**	**KID (↓)**
SD	100.00	100.00	0.00	32.70	-	100.00	100.00	0.00	34.99	-
UCE	11.35	25.50	39.61	29.45	1.39	3.75	10.81	19.44	26.74	5.31
RECE	**2.18**	12.39	21.99	26.17	3.40	**2.08**	20.61	34.05	30.05	2.37
GLoCE	87.34	**86.52**	22.09	**32.27**	**0.08**	81.25	**82.70**	30.57	**34.54**	**0.04**
Ours	**2.18**	53.73	**69.36**	31.89	0.43	32.92	72.00	**69.45**	34.24	0.17

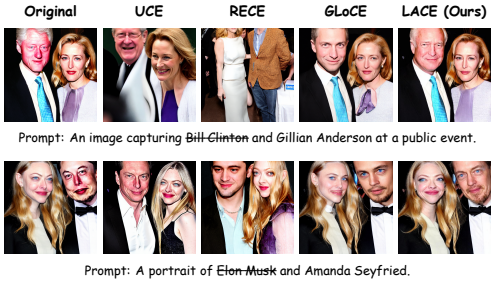


Figure 4: Qualitative Comparison on Celebrity.

Table 4: Ablation Study on Oxford Flowers

Variant	**Acc_t (↓)**	**Acc_r (↑)**	**Hcc (↑)**
Stage 1	28.08	**91.94**	77.76
Stage 1 + 2	23.91	**91.94**	81.62
Stage 1 + 2 + 3	**7.01**	88.72	**90.71**

5.5 Results on Multiple Concepts

We evaluate our method in the setting of multiple concept erasure by removing ten categories of flowers or dogs simultaneously. As shown in Table 3, our approach achieves effective erasure (low Acc_t) while preserving high Acc_r, Hcc, and CLIP scores, indicating strong retention and generation quality. These results highlight the robustness and scalability of our method compared to existing baselines. Although GLoCE achieves the lowest KID, this is mainly because it does not effectively erase the target concepts. In contrast, our method achieves good performance in both single and multiple concepts.

5.6 Ablation Study

We conduct ablation study on the Oxford Flowers dataset by incrementally adding each component of our framework and measuring its impact on unlearning and retention performance. Stage 1 corresponds to textual embedding projection, Stage 2 introduces attention suppression, and Stage 3 applies hard scrubbing at feature level. As shown in Table 4, the full model (Stage 1 + 2 + 3) significantly enhances unlearning performance and achieves the highest Hcc, indicating a best trade-off between erasure and preservation.

6 Conclusion

In this work, we introduce LACE, a training-free framework for localized concept erasure in text-to-image diffusion models. By combining projection-based representation editing, attention-guided spatial localization, and gated feature clean-up, LACE effectively suppresses targeted concepts while preserving semantically adjacent content and maintaining high visual fidelity. Extensive experiments across various datasets, prompts, and erasure settings demonstrate that LACE achieves precise and robust concept forgetting without retraining or compromising the model's generative capacity.

References

Kirk Baker. Singular value decomposition tutorial. *The Ohio State University*, 24:22, 2005.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. *arXiv preprint arXiv:2505.12677*, 2025.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=r1lUOzWCW`.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

Zhiyuan Chang, Mingyang Li, Junjie Wang, Yi Liu, Qing Wang, and Yang Liu. Repairing catastrophic-neglect in text-to-image diffusion models via attention-guided feature enhancement. *arXiv preprint arXiv:2406.16272*, 2024.

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

Anudeep Das, Vasisht Duddu, Rui Zhang, and N Asokan. Espresso: Robust concept filtering in text-to-image models. In *Proceedings of the Fifteenth ACM Conference on Data and Application Security and Privacy*, pages 305–316, 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*, pages 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*, pages 5111–5120, 2024.

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*, pages 73–88. Springer, 2024.

Nick Hasty, Ihor Kroosh, Dmitry Voitekh, and Dmytro Korduban. Giphy celebrity detector. `https://github.com/Giphy/celeb-detection-oss`, 2024. Accessed: 2025-08-02.

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

Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 363–374, 2023.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pages 461–478. Springer, 2024.

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*, pages 18596–18606, 2025.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024a.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating unsafe content generation in text-to-image models. *CoRR*, 2024b.

Siyuan Liang, Kuanrong Liu, Jiajun Gong, Jiawei Liang, Yuan Xun, Ee-Chien Chang, and Xiaochun Cao. Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning. *arXiv preprint arXiv:2403.16257*, 2024.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pages 93–109. 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*, pages 6430–6440, 2024.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

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.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

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

Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. *Advances in Neural Information Processing Systems*, 37:80244–80267, 2024.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 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*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

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

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

Riddhi Setty. Ai art generators hit with copyright suit over artists' images. *Bloomberg Law. Accessed on February*, 1:2023, 2023.

Zhuan Shi, Yifei Song, Xiaoli Tang, Lingjuan Lyu, and Boi Faltings. Copyright-aware incentive scheme for generative art models using hierarchical reinforcement learning. *arXiv preprint arXiv:2410.20180*, 2024a.

Zhuan Shi, Jing Yan, Xiaoli Tang, Lingjuan Lyu, and Boi Faltings. Rlcp: A reinforcement learning-based copyright protection method for text-to-image diffusion model. *arXiv preprint arXiv:2408.16634*, 2024b.

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

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9121–9130, 2025.

Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

Christian Wartena. Estimating word concreteness from contextualized embeddings. In *Proceedings of the 20th Conference on Natural Language Processing (Konvens 2024)*, pages 81–88, Vienna, Austria, 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.konvens-main.9/`.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024.

A Environmental Setup and Hyperparameters.

All experiments were implemented using PyTorch and based on the Stable Diffusion v1.4 architecture. Training and evaluation were performed on high-performance NVIDIA A100 GPUs with 80 GB of memory, running on a Linux-based system.

To define the neighborhood projection $P_{\mathcal{N}}$, we selected the top 10 nearest neighbor concepts in the embedding space. For the neighbor attention regularization term, we set $\alpha_{\text{neighbor}} = 100$.

Unless stated otherwise, we fixed the main regularization hyperparameters to $\beta = \gamma = 1.0$, which control the trade-off between forgetting the target concept and retaining unrelated content. These hyperparameters directly influence the optimization objective by adjusting the relative importance of target concept suppression (β) and retention of non-target concepts (γ).

To evaluate the sensitivity of our method to these hyperparameters, we performed ablation studies by varying β and γ on the Celebrity dataset. The following table reports the mean target and retain accuracies averaged over four identities.

Table 5: Mean Target and Retain Accuracy of Four Celebrities under Different β and γ Values

β	γ	**Mean Acc_t**	**Mean Acc_r**
0.75	1.25	28.33	71.68
0.80	1.20	11.84	81.34
0.90	1.10	8.34	88.51
1.00	1.00	**0.67**	**91.35**

As shown in Table 5, the combination of $\beta = \gamma = 1.0$ achieves best balance between minimizing target accuracy and maximizing retain accuracy (preserving unrelated concepts).

B Quantitative Results on the Celebrity Dataset.

Table 6: Quantitative Comparison on Celebrity Dataset

	Anna Kendrick			**Elon Musk**			**Anne Hathaway**			**Bill Clinton**			**Mean**		
	Acc_t (↓)	Acc_r (↑)	Hcc (↑)	Acc_t (↓)	Acc_r (↑)	Hcc (↑)	Acc_t (↓)	Acc_r (↑)	Hcc (↑)	Acc_t (↓)	Acc_r (↑)	Hcc (↑)	Acc_t (↓)	Acc_r (↑)	Hcc (↑)
UCE	**0.00**	58.00	73.42	2.00	56.67	71.81	**0.00**	64.00	78.05	**0.00**	58.67	73.95	0.50	59.83	74.31
RECE	**0.00**	46.67	63.64	0.67	24.67	39.52	**0.00**	34.00	50.75	**0.00**	20.00	33.33	**0.17**	31.83	46.81
GLoCE	1.34	**94.64**	**96.61**	0.67	**93.33**	**96.24**	2.00	**96.67**	**97.33**	0.00	**95.33**	**97.61**	1.00	**94.99**	**96.95**
Ours	**0.00**	88.00	93.62	**0.00**	90.67	95.11	0.67	95.33	97.30	2.00	91.34	94.55	0.67	91.35	95.17

In this section, we present a detailed comparison of our method against prior approaches on the Celebrity dataset. This benchmark consists of prompts that mention both a *target* identity (to be erased) and a *retain* identity (to be preserved), allowing for evaluation of both aspects simultaneously. We adopt the evaluation pipeline introduced by GLoCE. In this pipeline Giphy pretrained celebrity classifier is used to analyze the generated images and determine whether each identity (target or retained) is detected. The goal is to minimize the appearance of the target while maximizing the retention of unrelated identities.

Table 6 shows that while UCE and RECE achieve the lowest target accuracies (i.e., strongest forgetting), they suffer from significantly degraded retain accuracy, indicating poor preservation of non-target concepts. In contrast, both our method and GLoCE demonstrate strong overall performance by balancing effective concept forgetting with high retain accuracy. Notably, our method achieves lower average target accuracy in compare with GLoCE indicating superior erasure performance, while maintaining competitive retain accuracy.

It is important to note that we do not report CLIP score or KID for this dataset. Since each prompt contains both a target and a retained identity, global metrics like CLIP and KID, which measure overall semantic alignment or image quality, are not suitable for isolating the effect of targeted concept erasure.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim our methods could remove targeted concepts, preserves semantically related neighbors and and maintains overall image composition

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In experiments, we show that erasing the target concept and preserving adjacent concepts are greatly affected by parameters, which is a trade-off.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The answer NA means that the paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In our experiment and appendix, we provide implementation details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We plan to open source the code after paper get accepted

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We discuss those in the experiment session.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the standard protocol and only report the averaged results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe those in the appendix

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the code.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed social impact in the introduction and related work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: the paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: the paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: the paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer:[NA]

Justification: core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.