YOUR UNLEARNING GIVES YOU AWAY: IDENTIFYING ERASED CONCEPTS IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing attacks on unlearned diffusion models assume that the erased concepts are known in advance and focus on recovering them. In practice, however, model providers may not disclose which concepts have been removed, and even with access to the original base model, an adversary may still lack a clear target to attack. In this paper, we aim to answer the following critical but overlooked questions: ***which concepts have been erased from the model, and how many have been erased in total?*** To this end, we present TRACER, a framework that rapidly and accurately identifies erased concepts and estimates their number. TRACER efficiently identifies erased concepts without generating and classifying images. By combining lightweight spectral analysis of weight footprints, it enables efficient search over large candidate vocabularies. To distinguish multiple erased concepts, we introduce a footprint coverage objective that guides sequential discovery. TRACER estimates the number of erased concepts by detecting a sharp decline in candidate confidence as the selected concepts account for the erasure footprint, without requiring labeled examples for calibration. The framework requires only lightweight linear algebra and limited forward probes, with no prior knowledge of the unlearning algorithm. Experiments across text-to-image and text-to-video backbones and diverse unlearning methods demonstrate that TRACER identifies erased concepts and estimates their number *in seconds*, achieving $\mathbf{150}$–$\mathbf{137{,}000}\times$ and $\mathbf{133}$–$\mathbf{20{,}000}\times$ speedups over MIA and brute-force search on image and video models, respectively, with substantially higher identification accuracy.

1 INTRODUCTION

Diffusion models (Rombach et al., 2022; Podell et al., 2023; Ding et al., 2022) exhibit powerful generative capabilities. Despite these advances, their large-scale deployment has raised concerns about copyright infringement and harmful or biased content generation. Concept unlearning (Bourtoule et al., 2021; Xu et al., 2024) aims to suppress designated concepts while preserving other generative capabilities. Most existing methods achieve this by modifying pretrained model weights, demonstrating promising results in unlearning individual concepts or sets of concepts.

Existing attacks exploit vulnerabilities in unlearned models to regenerate erased content, but they typically assume prior knowledge of the specific concepts to recover (Lu et al., 2026; Pham et al., 2023; Rusanovsky et al., 2025; Zhang et al., 2024c; Chin et al., 2023; Tsai et al., 2024; Yang et al., 2024; Petsiuk & Saenko, 2024; Liu et al., 2026; Xiang et al., 2026; Horwitz et al., 2024). In practice, model providers may not disclose what they have removed. Even access to the original base model does not directly reveal which of its many concepts the provider chose to erase. For example, a provider may erase a small set of individuals' identities for privacy reasons. Although the base model may still generate depictions of these individuals, access to this capability alone does not tell an adversary who was selected for protection. Identifying the erased concepts therefore provides information beyond access to the original model: it reveals the provider's undisclosed targets and allows an adversary to focus subsequent attacks on the content the provider sought to suppress. This raises two critical but overlooked questions: ***without any prior knowledge of the erasure targets, can an adversary know which concepts have been erased from the model, and how many have been erased in total?*** To the best of our knowledge, no existing method systematically addresses this identification problem in unlearned diffusion models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

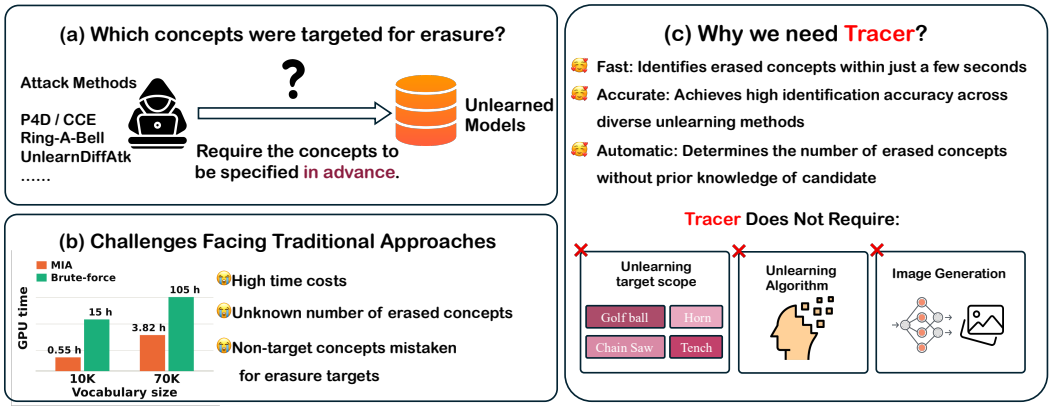

Figure 1: (a) Motivation: existing unlearning attacks assume known unlearning targets, leaving the identification of targets unexplored. (b) Current problem: adapted MIA and brute-force method incur high computational costs and struggle to determine target counts and distinguish targets from affected non-targets. (c) TRACER: identifies erased concepts and estimates their number in seconds without prior target knowledge, knowledge of the unlearning algorithm, or image generation.

Potential approaches include adapting traditional membership inference attacks (MIAs) (Fu et al., 2023; Kong et al., 2024; Duan et al., 2023; Wu et al., 2022; Matsumoto et al., 2023) or comparing images generated by the original and unlearned models for each candidate. However, extending these approaches to reliable and efficient identification of unknown erasure targets presents challenges. ***First***, unlearning can affect both target and non-target concepts. Changes in membership scores or generated content therefore cannot reliably distinguish directly targeted concepts from indirectly affected ones. ***Second***, the number of erased concepts is unknown. Identifying the target set requires determining which candidates to select and when to stop, without target labels for calibration. ***Third***, both approaches incur substantial computational costs even with a predefined candidate set. Without a narrowly specified set, the search may span a vocabulary containing tens of thousands of words, making repeated generation or MIA-based probing prohibitively expensive.

In this paper, we introduce a novel method, Tracing Concepts Erased by Unlearning (**TRACER**), for rapidly and accurately identifying erased concepts and estimates their number ***without any prior knowledge*** of the erasure targets. An intuitive but inconclusive approach is to compare the unlearned model weights with its original base model. However, these weight differences span distinct representation spaces, making it difficult to map them directly to erased concepts. *To turn these differences into usable evidence*, TRACER constructs Gram matrices from weight differences and uses a small set of probe words to identify compatible representation groups, then applies spectral reweighting to balance signals before aggregation, producing an erasure footprint for each group. These footprints provide the basis for identifying candidate concepts whose representations collectively explain the observed changes. *To identify multiple erased concepts*, we formulate identification as a subspace coverage problem, which measures how much of the erasure footprint a set of candidate concepts can collectively explain. Specifically, starting from an empty set, TRACER iteratively adds the candidate that best explains the remaining footprint energy, reducing redundant selections and directing the search toward other erased concepts. *To estimate the number of erased concepts*, we track candidate confidence throughout this coverage-based selection process. As the selected concepts account for the dominant footprint energy, a sharp decline in confidence indicates a cutoff in the selected sequence, allowing us to estimate the erased-set size without labeled calibration data. We summarize our contributions as follows:

- To the best of our knowledge, we are the *first* to systematically study erasure-target identification in unlearned diffusion models, exposing an overlooked vulnerability in diffusion unlearning.
- We propose TRACER, an efficient framework that accurately identifies multiple erased concepts from large candidate vocabularies without generating or classifying images, making broad searches for undisclosed erasure targets practical.
- TRACER automatically estimates how many concepts were erased, without labeled calibration examples, knowledge of the unlearning algorithm, or information about which weights were modified.

- Experiments across text-to-image and text-to-video backbones and diverse unlearning methods show that TRACER accurately identifies erased concepts and estimates their number from large candidate vocabularies in seconds, achieving **150–137,000**$\times$ and **133–20,000**$\times$ speedups over MIA and brute-force search on image and video models, respectively.

2 RELATED WORK

2.1 CONCEPT UNLEARNING IN DIFFUSION MODELS.

Fine-tuning methods steer target-concept noise toward anchor distributions (Gandikota et al., 2023; Kumari et al., 2023; Zhang et al., 2024a; Heng & Soh, 2023; Fan et al., 2023; Lyu et al., 2024; Lu et al., 2024; Zhao et al., 2024; Huang et al., 2024; Bui et al., 2025; Thakral et al., 2025; Li et al., 2025a; Wu et al., 2025). Closed-form methods edit attention projections without retraining (Gandikota et al., 2024; Orgad et al., 2023; Basu et al., 2024; Chavhan et al., 2024; Li et al., 2025c; Wang et al., 2025; Deng et al., 2026b;a; Yao et al.; Biswas et al., 2026; Lee et al., 2025). Inference-time methods suppress undesirable concepts via guidance or activation intervention (Schramowski et al., 2023; Chen et al., 2025; Deja et al., 2025; Zhang et al., 2025). Recent work extends erasure to flow-matching transformers, including SD3 and FLUX (Gao et al., 2025a; Ranjan et al., 2026; Jiang et al., 2026; Lee et al., 2026). Parallel work targets text-to-video models through editing, low-rank refusal vectors, and training-free safeguards (Liu & Tan, 2024; Yoon et al., 2025; Ye et al., 2025; Xu et al., 2025; Facchiano et al., 2026). Robust methods resist adversarial prompts and subsequent fine-tuning (Zhang et al., 2024b; Kim et al., 2024; Gong et al., 2024; Srivatsan et al., 2025; Bui et al., 2024; Li et al., 2026b; Gao et al., 2025b; Li et al., 2026a).

2.2 ADVERSARIAL VULNERABILITIES OF UNLEARNED MODELS.

Recent analyses show erased concepts are often suppressed, not removed, resurfacing under subsequent benign fine-tuning or alternative sampling strategies (Rusanovsky et al., 2025; Suriyakumar et al., 2024; Lu et al., 2026; Yuan et al., 2025). Adversarial attacks recover erased concepts. White-box attacks exploit model gradients or internal signals (Yang et al., 2024; Chin et al., 2023; Zhang et al., 2024c; Pham et al., 2023), whereas black-box attacks search adversarial prompts with query-only model access (Tsai et al., 2024; Lyu et al., 2025; Yook et al., 2025). Non-prompt inference-time attacks recover concepts via concept arithmetic, image-guided inversion, or text-free inversion (Petsiuk & Saenko, 2024; Liu et al., 2026; Xiang et al., 2026), while weight-space attacks use fine-tuning or pre-erasure weights (George et al., 2025; Horwitz et al., 2024).

Beyond unlearning attacks, diffusion models themselves also face classical privacy attacks on their original training data. Membership inference attacks determine whether a given sample actually belongs to the training set (Matsumoto et al., 2023; Duan et al., 2023; Kong et al., 2024; Fu et al., 2023; Wu et al., 2022). Data extraction and memorization attacks reconstruct or replicate individual training images from the model (Carlini et al., 2023; Somepalli et al., 2023a;b).

2.3 POSITIONING OF OUR STUDY.

Our study aims to identify the concepts targeted for erasure in an unlearned model. Existing attacks (Sec 2.2) primarily recover content associated with known erased concepts. TRACER identifies unknown erasure targets and estimates their number without image generation, fine-tuning, or knowledge of the erasure algorithm. This setting shifts the focus from attacking a model to regenerate known erased concepts to identifying which concepts were erased.

3 METHOD

3.1 UNLEARN THREAT MODEL

What the adversary holds. A publisher starts from public base weights W, applies a published concept-unlearning method (Section 2.1) to erase an unknown set $S^\star$ of concepts, and releases the resulting edited weights W'. The adversary has access only to W, W', and a vocabulary $\mathcal{C}$ of ordinary words assumed to contain $S^\star$. In particular, the adversary knows neither $|S^\star|$ nor which

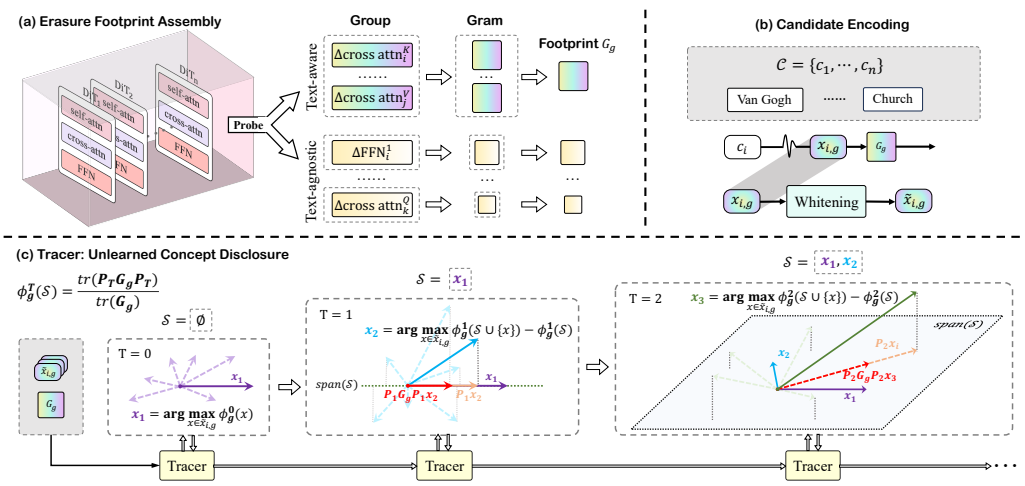

Figure 2: **Overview of TRACER. (a) Erasure footprint assembly.** Erasure imprints are grouped by their probed input representations. Spectrally reshaped and reweighted Gram matrices are fused within compatible groups into footprints G_g. **(b) Candidate encoding.** Candidates are encoded in each group's input space, whitened to suppress shared structure, and normalized. **(c) Unlearned concept disclosure.** TRACER iteratively selects concepts explaining the remaining footprint energy. selection combines standardized gains across groups with confidence-based weights. Label-free self-calibration fixes parameters and group weights. Confidence drops determine set size without image generation or algorithm knowledge.

erasure algorithm was used, has no labelled examples of erased or non-erased models, and therefore cannot construct shadow models for validation. The adversary may evaluate the full computation graph on text inputs and record the inputs to all changed weight matrices, which are identified directly by comparing W and W'. Text-encoder matrices are probed from the prompt alone, whereas generative-backbone matrices are probed by pairing the prompt with a single fixed latent at several noise levels. Both probes require only the available model weights and architecture. The sampling loop is never executed, no image is decoded, and no model output is observed.

What the attack recovers. The goal of the attack is to recover the identities of the erased concepts: a subset $\hat{S} \subseteq \mathcal{C}$ that should coincide with $S^\star$, together with an inferred count n intended to match $|S^\star|$. Because TRACER automatically searches all available aligned model matrices for weight changes and reads the representations they receive, it requires no knowledge of either the erasure algorithm or its edit locations. It therefore applies with minimal assumptions to weight-editing concept-unlearning methods across different model components and architectures.

3.2 ERASURE FOOTPRINT ASSEMBLY

Erasure locations depend on the unlearning method and architecture: closed-form editors typically modify cross-attention key and value projections, while fine-tuning may affect other attention and feed-forward layers, including joint or temporal attention in diffusion-transformer and video backbones. TRACER identifies changes from weight differences throughout the computation graph, without knowledge of edit locations. For each matrix ℓ, its *imprint* is $\Delta W_\ell = W'_\ell - W_\ell$; collectively, non-negligible imprints form the erasure ***footprint***. We retain ℓ in $\mathcal{L}$ when its relative Frobenius change $\|\Delta W_\ell\|_F / \|W_\ell\|_F$ exceeds $\tau = 10^{-8}$. This threshold excludes numerical noise. It is four orders of magnitude below the smallest relative imprint observed in our experiments.

Combining these imprints introduces two challenges. **First**, imprints may receive different representations: some act on text encodings; others receive internal activations with different dimensions and semantics. We identify these representations using probe words rather than layer names. An imprint is *text-aware* if, for every probe, its input has greater cosine similarity to the matching text encoding than to any other probe encoding. This correspondence remains exact for imprints acting on text encodings but can disappear as textual information mixes with image features. Text-aware imprints

share one representation group. Each imprint failing the test is *text-agnostic* and forms a singleton group with location-specific input. Only imprints within the same group are aggregated.

Second, imprints within a group may have different output widths despite sharing an input representation, preventing direct summation. The Gram matrix $M_\ell = \Delta W_\ell^\top \Delta W_\ell$ removes the output dimension while preserving imprint strength along the shared input space, making these matrices summable. Direct summation may let strongly edited matrices obscure weaker but informative imprints. We therefore reshape each Gram matrix before aggregation. Writing $M_\ell = Q\Lambda Q^\top$, the eigenvectors in Q specify its principal directions, and each eigenvalue λ measures the corresponding strength. Reshaping preserves these directions, raises eigenvalues normalized by $\lambda_{\max}$ to a power ρ, and discards negligible directions. We additionally weight imprints by their inverse edit magnitude, with strength controlled by β:

$$G_g = \sum_{\ell \in g} \|\Delta W_\ell\|_F^{-2\beta} \left[\!\left[M_\ell\right]\!\right]_\rho, \qquad \left[\!\left[M_\ell\right]\!\right]_\rho = Q \operatorname{diag}\Big((\lambda/\lambda_{\max})^\rho \mathbf{1}[\lambda/\lambda_{\max} \geq \varepsilon_0]\Big) Q^\top. \tag{1}$$

Here, $\mathbf{1}[\cdot]$ denotes the indicator function, and ε_0 is the relative spectral cutoff. The self-calibrated exponents ρ and β control the relative strengths of directions within each imprint and contributions across imprints, respectively. This produces one *aggregated footprint* G_g per representation group.

3.3 Geometry-Aware Candidate Encoding

Each vocabulary entry in $\mathcal{C}$ is a *candidate*. We encode each candidate in the input space of its group's footprint. We insert the candidate into a set T of neutral prompt templates, pass the prompts through the network, and average the representations received by group g across templates to obtain $x_{c,g}$. One template turns *church* into "a photo of a church." For a text-aware group, we use the text encoding at the token position specified by offset o. For a text-agnostic group, we use the internal activation at that location, averaged over applicable token positions. Groups in the generative backbone additionally average over fixed noise levels. Template averaging reduces sensitivity to phrasing and emphasizes the candidate. Each candidate is represented separately in each group's input space.

Candidate representations are unevenly distributed within each group's input space: a few directions carry much of the vocabulary energy, giving many candidates large shared components. Candidates aligned with these directions may score highly against different footprints regardless of which concepts were erased. We suppress this shared structure to emphasize candidate-specific information. We construct a regularized second-moment matrix over the full vocabulary and normalize the transformed representations:

$$\Sigma_g = \mathbb{E}_{c \in \mathcal{C}}\left[x_{c,g} x_{c,g}^\top\right] + \varepsilon_w I, \qquad \tilde{x}_{c,g} = \frac{\Sigma_g^{-\gamma/2} x_{c,g}}{\left\|\Sigma_g^{-\gamma/2} x_{c,g}\right\|}, \tag{2}$$

where the expectation is the uncentered second moment and I is the identity. We set $\varepsilon_w = 10^{-3}\mathbb{E}_{c \in \mathcal{C}}\|x_{c,g}\|^2/d_g$, where d_g is the group's input dimension. Regularization ensures positive definiteness and limits amplification of weak directions. We compute $\Sigma_g^{-\gamma/2}$ through eigendecomposition. Directions carrying more vocabulary energy are attenuated more strongly. γ controls suppression from no transformation to full whitening. o and γ are selected by self-calibration in Section 3.4 before iterative candidate selection. Unit normalization prevents larger norms from inflating candidate scores in footprint comparisons using $\tilde{x}_{c,g}$.

3.4 Unlearned Concept Disclosure

Subspace Coverage Scoring. The footprint identifies a subspace whose eigenvectors may rotate within eigenspaces with equal or nearly equal eigenvalues. We therefore match candidate spans to the footprint rather than associate individual spectral directions with words. For $S \subseteq \mathcal{C}$, coverage measures the fraction of footprint energy captured by its candidate span:

$$\phi_g(S) = \frac{\operatorname{tr}\left(P_{S,g} G_g P_{S,g}\right)}{\operatorname{tr}(G_g)}, \tag{3}$$

where $P_{S,g}$ projects orthogonally onto $\operatorname{span}\{\tilde{x}_{c,g} : c \in S\}$ and $\operatorname{tr}(\cdot)$ denotes the trace. The singleton score $s_g(c) = \phi_g(\{c\})$ defines the group's initial field. Additional coverage measures contributions

beyond S, with zero gain for candidates within the span. To measure field separation, we define

$$\mathrm{conf}_\kappa(s) = \frac{\max_c s(c) - \mathrm{med}(s)}{q_{1-\kappa}(s) - \mathrm{med}(s) + \varepsilon_c}. \tag{4}$$

Here, s is the field over candidates, $\mathrm{med}(s)$ its median, and $q_{1-\kappa}(s)$ its $(1-\kappa)$-quantile. $\kappa \in (0, 1/2)$ specifies the upper-tail comparison range, while $\varepsilon_c > 0$ stabilizes the ratio. A constant field has zero confidence. The numerator measures the leader's separation from the median; the denominator normalizes it by the upper-tail spread. This reduces sensitivity to score scale while retaining dependence on the candidate vocabulary. Groups observe different representations, so we standardize their gains and weight them by the separation of leading candidates in their initial fields:

$$z_g(c, S) = \frac{\phi_g\big(S \cup \{c\}\big) - \phi_g(S) - \mu_g(S)}{\sigma_g(S) + \varepsilon_z}, \qquad w_g = \frac{\max\{\mathrm{conf}_\kappa(s_g), \varepsilon_g\}}{\sum_{g'} \max\{\mathrm{conf}_\kappa(s_{g'}), \varepsilon_g\}}, \tag{5}$$

where $\mu_g(S)$ and $\sigma_g(S)$ are the gain mean and standard deviation over $\mathcal{C}$. The constants $\varepsilon_z > 0$ and $\varepsilon_g > 0$ stabilize standardization and ensure positive group weights, respectively. For any parameter setting, the initial combined field is $s^{(1)}(c) = \sum_g w_g z_g(c, \varnothing)$.

Self-Calibration. Candidate fields depend on spectral shaping ρ, whitening γ, imprint-amplitude weighting β, and the token offset o. Without a labelled validation set, the adversary evaluates each setting through its initial combined field, before selecting a candidate. Writing $\theta = (\beta, \gamma, \rho, o)$, we evaluate a predefined finite set Θ and select $\theta^\star = \arg\max_{\theta \in \Theta} \mathrm{conf}_\kappa(s_\theta^{(1)})$. $s_\theta^{(1)}$ is constructed with the footprint operators, candidate representations, and group weights induced by θ. This selection requires no knowledge of the erased-set size. The setting and its group weights remain fixed throughout the search. Candidate gains and their standardization are updated as the span expands.

Progressive Concept Pursuit. With the calibrated setting, we greedily construct a candidate sequence by maximizing additional coverage across groups. Overlapping vocabulary expressions may remain prominent after a related candidate is selected because their representations need not lie in its span. We suppress these lexical overlaps after each selection. Let $\mathcal{F}(c)$ contain c and all vocabulary entries whose token sets are subsets or supersets of its token set. These entries are excluded from subsequent selection using only the vocabulary, without erased-concept labels or an assumption of semantic equivalence. Starting from $S_0 = D_0 = \varnothing$, let S_j and D_j denote the selected and excluded entries after step j. Each step combines the group gains into the global score $s^{(j)}(c) = \sum_g w_g z_g(c, S_{j-1})$. We select the candidate c_j with the highest score among entries not in D_{j-1}. We add c_j to the selected set and exclude all entries in $\mathcal{F}(c_j)$ from subsequent selection. Other suppressed entries are not added to the selected span. The search uses a fixed horizon J without prior knowledge of the erased-set size. We record confidence over eligible candidates initially and after each selection to determine how many selected candidates are retained in the output.

Erased-Set Size Detection. We determine the candidate sequence's cutoff by tracking how clearly the remaining field separates its leader from the background. A pronounced decline suggests exhaustion of the dominant erasure signal. After selecting j candidates, we record $a_j = \mathrm{conf}_\kappa(s^{(j+1)})$ for $j = 0, \ldots, J$. Confidence is computed only over candidates that have neither been selected nor excluded by lexical suppression. Thus, a_0 measures initial separation, while a_J measures separation after the final selection. We measure the confidence decline following the i-th selection as

$$r_i = \log \max\{a_{i-1}, \varepsilon_a\} - \log \max\{a_i, \varepsilon_a\}, \qquad i = 1, \ldots, J, \tag{6}$$

where $\varepsilon_a > 0$ keeps logarithms defined. A positive r_i indicates a decline, with its magnitude measuring the log confidence ratio. The largest decline provides the default size estimate. Some profiles rise as selecting dominant candidates makes remaining targets more distinguishable, then decline in stages because related vocabulary entries remain prominent. The largest decline may therefore occur after the erased set has been selected. For these profiles, we identify the first sufficiently large decline. The initial-rise condition holds when confidence increases in at least $B - 1$ of the first B transitions, where $3 \le B \le J$ is a fixed window length. Let $r_{\max}$ denote the largest log confidence drop and $\alpha \in (0, 1)$ a relative threshold. If the initial-rise condition holds and $r_{\max} > 0$, we estimate $n = \min\{i : r_i \ge \alpha r_{\max}\}$ over $i = 1, \ldots, J$, selecting the first drop reaching a fraction α of the largest drop. Otherwise, n is the index of the largest drop, with ties resolved by choosing the

Table 1: Identification accuracy (%) of erased object concepts across three text-to-image backbones on Imagenette-20. Shaded cells report evaluated results. "–" indicates an unevaluated setting.

Method	Venue	Erased-set size K					Avg. (↑)
		1	3	5	10	20	
Stable Diffusion v1.5							
ESD	ICCV 2023	100	100	94	–	–	98.0
UCE	WACV 2024	100	100	100	100	100	100.0
MACE	CVPR 2024	100	100	100	100	–	100.0
FMN	CVPR 2024a	100	100	100	–	–	100.0
ConceptPrune	ICLR 2024	100	100	100	94	–	98.5
Sculpting	ICCV 2025b	100	100	100	100	–	100.0
SPEED	ICLR 2025c	100	100	–	–	–	100.0
ScaPre	ICLR 2026b	100	100	100	100	100	100.0
FIA	ICML 2026a	100	100	100	100	–	100.0
AdvUnlearn	NeurIPS 2024b	100	–	–	–	–	100.0
STEREO	CVPR 2025	100	100	–	–	–	100.0
Meta-Unlearning	ICCV 2025b	100	–	–	–	–	100.0
Stable Diffusion 3							
UCE	WACV 2024	100	100	94	–	–	98.0
EraseFlow	NeurIPS 2026	100	100	100	–	–	100.0
FLUX.1							
EraseAnything	ICML 2025a	100	100	–	–	–	100.0
ScaPre	ICLR 2026b	100	100	100	100	–	100.0

earliest selection. Both rules use only the confidence sequence. The estimate n determines the set $\hat{S} = S_n$, comprising the first n candidates. The horizon J bounds the search without prescribing the output size. Implementation settings are detailed in Appendix A.1 and Appendix A.2.

4 Experiment

4.1 Experimental Settings

We evaluate TRACER on three text-to-image backbones: Stable Diffusion v1.5 (Rombach et al., 2022), Stable Diffusion 3 (Esser et al., 2024), and FLUX.1 (Labs, 2024). The text-to-video backbones are CogVideoX-5B (Yang et al., 2025) and Wan2.1-T2V-1.3B (Wan et al., 2025). We use official implementations and released checkpoints without modification. Object identification is evaluated on the ten Imagenette classes (Howard et al., 2019). Explicit-content evaluation uses I2P prompts (Schramowski et al., 2023). Artistic-style identification covers the artists evaluated by ESD (Gandikota et al., 2023) and UCE (Gandikota et al., 2024). Generation quality is evaluated on MS COCO (Lin et al., 2014). We construct **Imagenette-20** by augmenting Imagenette's ten classes (Howard et al., 2019) with 10 non-overlapping common object categories from ImageNet-1k (Russakovsky et al., 2015) to evaluate identification across varying erased-set sizes. All experiments run on NVIDIA RTX A6000 GPUs.

4.2 Disclosing Erased Concept in Unlearned Models

Reproducing unlearning methods with their official implementations reveals limits on stable erasure capacity. Beyond these limits, models exhibit degraded forgetting or damaged generative capabilities. Shaded cells in Table 1 indicate evaluated settings within these limits; excluded settings are marked "–". We evaluate varying erased-set sizes, where K denotes the number of concepts jointly erased from each checkpoint. For each size, we sample concept sets from Imagenette-50, shared across methods and backbones. Each set is jointly erased to produce a checkpoint. We define *recovery accuracy* as the percentage of checkpoints whose recovered set exactly matches the ground-truth erased set. Correct recovery identifies all erased concepts without extras, jointly testing concept identities and inferred set size. Partial recovery counts as failure. The Avg. column averages recovery accuracy over evaluated set sizes within each method's stable erasure range, excluding "–". Higher values indicate better performance (↑). Table 4 extends the same protocol to text-to-video models. Evaluation is restricted to individual concepts and small groups, where the tested methods maintain stable erasure. TRACER achieves perfect recovery for both methods on CogVideoX-5B and for Refusal on Wan2.1-T2V-1.3B. For VideoEraser on Wan2.1-T2V-1.3B, recovery remains perfect for individual concepts, with a modest decline under joint erasure. These results demonstrate that erasure

Table 2: Exact estimation accuracy (%) of the erased-set size K across evaluated image and video unlearning methods.

Method	Ground-truth size K				
	1	3	5	10	20
text-to-image (Stable Diffusion v1.5)					
ESD	100	100	100	–	–
UCE	100	100	100	100	100
MACE	100	100	100	100	–
ConceptPrune	100	100	98	–	–
SPEED	100	100	92	–	–
ScaPre	100	100	100	100	100
text-to-image (FLUX.1)					
EraseAnything	100	100	–	–	–
EraseFlow	100	100	100	–	–
text-to-video (CogVideoX-5B)					
Refusal	100	100	100	–	–
VideoEraser	100	100	–	–	–

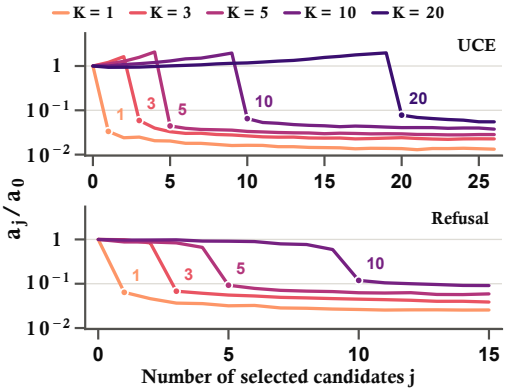

Figure 3: Normalized confidence a_j/a_0 drops at the true erased-set size K for UCE and Refusal.

Table 3: Recall@K (%) with the true K provided, evaluated over 50 checkpoints per setting.

Baseline Method		Recall@K (↑)			Recall@10 (↑)
		$K=1$	$K=3$	$K=5$	
ESD	MIA	42.0	6.7	4.8	27.1
	Brute-force	82.0	33.3	16.2	53.3
	TRACER	**100.0**	**100.0**	**100.0**	**100.0**
UCE	MIA	60.0	24.0	14.0	39.6
	Brute-force	92.0	68.7	36.6	72.2
	TRACER	**100.0**	**100.0**	**100.0**	**100.0**
SPEED	MIA	78.0	63.3	42.4	68.9
	Brute-force	76.0	60.0	38.8	66.5
	TRACER	**100.0**	**100.0**	**100.0**	**100.0**
ScaPre	MIA	80.0	72.0	48.2	74.7
	Brute-force	90.0	78.7	52.0	80.4
	TRACER	**100.0**	**100.0**	**100.0**	**100.0**

Table 4: Identification accuracy (%) of erased object concepts across two text-to-video backbones with $K \in \{1, 3\}$, following the protocol of Table 1.

Method	Venue	Size K		Avg. (↑)
		1	3	
CogVideoX-5B				
VideoEraser	EMNLP 2025	100	100	100.0
Refusal	ICLR 2026	100	100	100.0
Wan2.1-T2V-1.3B				
VideoEraser	EMNLP 2025	100	90	95.0
Refusal	ICLR 2026	100	100	100.0

footprints remain informative across video architectures with different temporal modeling, enabling accurate identification of both individual targets and jointly erased concept sets.

To assess size estimation separately from concept identification, Table 2 reports the percentage of checkpoints whose inferred count matches the true erased-set size. TRACER achieves perfect size estimation in most settings across image and video models, with occasional errors for ConceptPrune and SPEED under joint erasure. Figure 3 illustrates the confidence dynamics for UCE and REFUSAL. Normalized confidence remains high during early selections, then drops sharply once the selected sequence reaches the true erased-set size. This decline reflects reduced separation between the leading remaining candidate and the background, suggesting exhaustion of the dominant erasure signal. The confidence-drop rule uses this transition to estimate the target count without prior size information. Artistic-style identification results are reported in Appendix B.

4.3 COMPARISON WITH TRADITIONAL ADVERSARIAL METHODS

We compare TRACER with two approaches. Adapted MIA ranks concepts by aggregated changes in sample-level membership scores. Brute-force probing ranks candidates by comparing images generated before and after unlearning. Implementation details are provided in Appendix A.3. We evaluate ESD, UCE, SPEED, and ScaPre on Stable Diffusion v1.5, with 50 checkpoints per erased-set size $K \in \{1, 3, 5\}$. All approaches share checkpoints and 1,000-name vocabulary. To accommodate conventional methods, all receive the true K; TRACER uses its iterative selection order without size estimation. Unlike exact-set accuracy, this comparison credits partial recovery. Let $T_m^{(L)}$ contain the first L candidates for checkpoint m. We compute $\text{Recall@}L = \frac{100}{50} \sum_{m=1}^{50} |T_m^{(L)} \cap S_m^\star|/K$, where $S_m^\star$ is the erased set. Table 3 reports Recall@K for each K; recovering three of five targets scores 60%. Recall@10 uses $L = 10$ with denominator K. Its reported value equally averages the three K settings. TRACER maintains high recall across methods and set sizes, while baseline recall

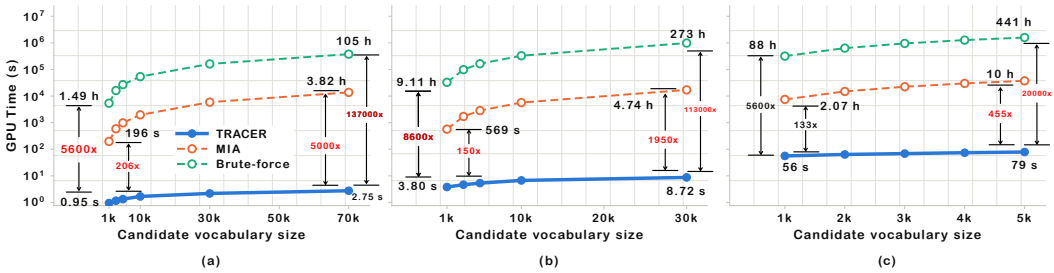

Figure 4: Runtime across three methods, showing the scalability of TRACER to large candidate sets.

Table 5: Component ablations in Recall@5 (%) with the true number of erasure targets $K = 5$ provided.

Variant	ESD	UCE	SPEED	ScaPre
Full TRACER	**100.0**	**100.0**	**100.0**	**100.0**
Single imprint per group	84.0	94.8	88.4	91.6
w/o background suppression	76.8	92.4	90.8	88.0
w/o progressive selection	56.4	78.4	82.8	81.2

Table 6: Calibration and size-detection ablations in exact-set recovery accuracy (%).

Variant	ESD	UCE	SPEED	ScaPre
Full TRACER	**100.0**	**100.0**	**100.0**	**100.0**
w/o self-calibration	78	84	92	86
Largest-drop only	72	64	78	80

generally declines as K grows. Allowing ten candidates improves baseline coverage but leaves targets unidentified, supporting the effectiveness of weight footprints beyond behavioral changes alone.

Figure 4 compares runtime across candidate vocabulary sizes on (a) Stable Diffusion v1.5, (b) FLUX.1, and (c) CogVideoX-5B. TRACER exhibits modest runtime growth as the vocabulary expands across all backbones. Across the evaluated vocabulary sizes, TRACER achieves 150–137,000× and 133–20,000× speedups over MIA and brute-force probing on image and video models, respectively. Weight analysis with limited forward probes avoids repeated image or video generation and sample-level membership evaluation, enabling efficient concept identification over large candidate vocabularies.

4.4 Ablation Studies

We evaluate components on 50 checkpoints per unlearning method at $K = 5$. Table 5 provides K to isolate concept identification from size estimation. The single imprint variant retains the imprint with the highest initial confidence in each group. Removing background suppression sets $\gamma = 0$ while preserving unit normalization. Removing progressive selection uses the initial ranking without updating coverage gains, retaining lexical suppression. TRACER achieves 100% Recall@5 across four methods. Using a single imprint reduces recall by 5.2 to 16.0 percentage points, supporting evidence integration across weight changes. Removing background suppression causes declines of 7.6 to 23.2 points. Replacing progressive selection with static ranking produces the largest reduction across methods, reaching 43.6 points on ESD. These results support updating candidate scores to capture information beyond the selected span. Table 6 evaluates exact recovery of the erased set without providing K. Replacing self-calibration with fixed parameters reduces accuracy to between 78% and 92%. Always selecting the largest confidence drop, without accounting for an initial rise, yields accuracy between 64% and 80%. The full method achieves 100% across four methods, supporting both calibration for each checkpoint and size estimation adapted to the confidence profile.

5 Conclusion

We presented TRACER, which accurately identifies erasure targets and their number from publicly released model weights in seconds without image generation. Experiments demonstrate its effectiveness across image and video backbones and diverse unlearning methods. It enables label-free search over large candidate vocabularies, achieving substantial speedups over membership inference attacks and brute-force generation-based probing methods. Our work establishes erasure-target identification as a distinct task and provides a practical and computationally efficient framework for assessing what model weights reveal about erasure targets despite suppressed generation.

AI USE STATEMENT

In this work, generative AI tools were used solely to correct grammatical errors and improve the clarity of the English writing, without changing the scientific content. They were not used to generate research ideas, develop the methodology, conduct experiments, write or analyze code, interpret results, formulate claims, or create figures and other research artifacts. All scientific content and contributions were developed independently by the authors.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, private data, or sensitive user information. All models and datasets used in this study are publicly available and are employed solely for academic research. We are not aware of any ethical concerns, conflicts of interest, or issues related to discrimination, privacy, security, or legal compliance arising from this work. The authors follow established responsible research practices.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of this work. The complete methodological pipeline and all components of the proposed approach are described in detail in the main paper, with additional implementation details provided in the appendix. These descriptions are intended to enable independent reproduction of our method and experimental results. Upon acceptance, we will publicly release the complete source code and accompanying configurations.

REFERENCES

Samyadeep Basu, Nanxuan Zhao, Vlad Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *International Conference on Learning Representations*, volume 2024, pp. 17592–17603, 2024.

Shristi Das Biswas, Arani Roy, and Kaushik Roy. Cure: Concept unlearning via orthogonal representation editing in diffusion models. *Advances in Neural Information Processing Systems*, 38: 82539–82563, 2026.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

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

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

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.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024.

Die Chen, Zhiwen Li, Mingyuan Fan, Cen Chen, Wenmeng Zhou, Yanhao Wang, and Yaliang Li. Growth inhibitors for suppressing inappropriate image concepts in diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 79164–79184, 2025.

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.

Kamil Deja et al. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025.

Kaiyuan Deng, Bo Hui, Gen Li, Jie Ji, Minghai Qin, Geng Yuan, and Xiaolong Ma. Forget-it-all: Multi-concept machine unlearning via concept-aware neuron masking. *arXiv preprint arXiv:2601.06163*, 2026a.

Kaiyuan Deng, Gen Li, Yang Xiao, Bo Hui, and Xiaolong Ma. Forget many, forget right: Scalable and precise concept unlearning in diffusion models. *arXiv preprint arXiv:2601.06162*, 2026b.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902, 2022.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Simone Facchiano, Stefano Saravalle, Matteo Migliarini, Edoardo De Matteis, Alessio Sampieri, Andrea Pilzer, Emanuele Rodolà, Indro Spinelli, Luca Franco, and Fabio Galasso. Video unlearning via low-rank refusal vector. In *International Conference on Learning Representations*, volume 2026, pp. 132049–132078, 2026.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. A probabilistic fluctuation based membership inference attack for diffusion models. *arXiv preprint arXiv:2308.12143*, 2023.

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

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

Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025a.

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

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

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.

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.

Eliahu Horwitz, Jonathan Kahana, and Yedid Hoshen. Recovering the pre-fine-tuning weights of generative models. *arXiv preprint arXiv:2402.10208*, 2024.

Jeremy Howard et al. Imagenette: A smaller subset of imagenet for quick experiments. https://github.com/fastai/imagenette, 2019.

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

Nanxiang Jiang, Zhaoxin Fan, Baisen Wang, Daiheng Gao, Junhang Cheng, Jifeng Guo, Yalan Qin, Yeying Jin, Hongwei Zheng, Faguo Wu, et al. Z-erase: Enabling concept erasure in single-stream diffusion transformers. *arXiv preprint arXiv:2603.25074*, 2026.

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*, pp. 461–478. Springer, 2024.

Fei Kong, Jinhao Duan, Heng Tao Shen, Xiaoshuang Shi, Xiaofeng Zhu, Kaidi Xu, et al. An efficient membership inference attack for the diffusion model by proximal initialization. In *International Conference on Learning Representations*, volume 2024, pp. 57876–57896, 2024.

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

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. *Advances in Neural Information Processing Systems*, 38:70876–70911, 2026.

Black Forest Labs. Flux: Advancing generative models for creative workflows. https://huggingface.co/black-forest-labs/FLUX.1, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18596–18606. IEEE, 2025.

Uichan Lee, Jeonghyeon Kim, and Sangheum Hwang. Localized concept erasure in text-to-image diffusion models via high-level representation misdirection. In *The Fourteenth International Conference on Learning Representations*, 2026.

Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. *Advances in Neural Information Processing Systems*, 38:17754–17789, 2026a.

Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. *arXiv preprint arXiv:2505.11131*, 2025a.

Fengpeng Li, Kemou Li, Qizhou Wang, Bo Han, and Jiantao Zhou. Aegis: Adversarial target-guided retention-data-free robust concept erasure from diffusion models. *arXiv preprint arXiv:2602.06771*, 2026b.

Gen Li, Yang Xiao, Jie Ji, Kaiyuan Deng, Bo Hui, Linke Guo, and Xiaolong Ma. Sculpting memory: Multi-concept forgetting in diffusion models via dynamic mask and concept-aware optimization. *arXiv preprint arXiv:2504.09039*, 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Renyang Liu, Guanlin Li, Tianwei Zhang, and See-Kiong Ng. Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 143251–143275, 2026.

Shiqi Liu and Yihua Tan. Unlearning concepts from text-to-video diffusion models. *arXiv preprint arXiv:2407.14209*, 2024.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? *Advances in Neural Information Processing Systems*, 38:14751–14775, 2026.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Xinqi Lyu, Yihao Liu, Yanjie Li, and Bin Xiao. Pla: Prompt learning attack against text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16851–16860, 2025.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pp. 77–83. IEEE, 2023.

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

Vitali Petsiuk and Kate Saenko. Concept arithmetics for circumventing concept inhibition in diffusion models. In *European Conference on Computer Vision*, pp. 309–325. Springer, 2024.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*, 2023.

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. *arXiv preprint arXiv:2307.01952*, 2023.

Ravi Ranjan, Utkarsh Grover, Xiaomin Lin, and Agoritsa Polyzou. Razor: Ratio-aware layer editing for targeted unlearning in vision transformers and diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7998–8008, 2026.

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

Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2966–2975. IEEE, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

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*, pp. 22522–22531, 2023.

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, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

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

Vinith Menon Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9121–9130, 2025.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

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. IEEE, 2025.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28263–28273. IEEE, 2025.

Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models. *arXiv preprint arXiv:2210.00968*, 2022.

Qianlong Xiang, Miao Zhang, Haoyu Zhang, Kun Wang, Junhui Hou, and Liqiang Nie. Tina: Text-free inversion attack for unlearned text-to-image diffusion models. *arXiv preprint arXiv:2603.17828*, 2026.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. Videoeraser: Concept erasure in text-to-video diffusion models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5965–5994, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7737–7746, 2024.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Quanming Yao, Yang Liu, Zhen Wang, Yatao Bian, et al. Erasing concept combination from text-to-image diffusion model. In *The Thirteenth International Conference on Learning Representations*.

Xiaoyu Ye, Songjie Cheng, Yongtao Wang, Yajiao Xiong, and Yishen Li. T2vunlearning: A concept erasing method for text-to-video diffusion models. *arXiv preprint arXiv:2505.17550*, 2025.

Hyun Jun Yook, Ga San Jhun, Jae Hyun Cho, Min Jeon, Donghyun Kim, Tae Hyung Kim, and Youn Kyu Lee. Zium: Zero-shot intent-aware adversarial attack on unlearned models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3926–3935, 2025.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Xiaoyong Yuan, Xiaolong Ma, Linke Guo, and Lan Zhang. What lurks within? concept auditing for shared diffusion models at scale. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4154–4168, 2025.

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

Lingyun Zhang, Yu Xie, Yanwei Fu, and Ping Chen. Concept replacer: Replacing sensitive concepts in diffusion models via precision localization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8172–8181. IEEE, 2025.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024b.

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

Mengnan Zhao, Lihe Zhang, Tianhang Zheng, Yuqiu Kong, and Baocai Yin. Separable multi-concept erasure from diffusion models. *arXiv preprint arXiv:2402.05947*, 2024.

A Implementation Details and Evaluation Protocols

A.1 Hyperparameter Settings and Implementation Details

Self-calibration grid. We select $\theta = (\beta, \gamma, \rho, o)$ by maximizing the confidence of the initial combined field, as described in Section 3.4. Table 7 lists the candidate values. The predefined finite set Θ is their full Cartesian product, yielding $2 \times 2 \times 3 \times 2 = 24$ settings for the main configuration. The selected setting and its group weights remain fixed throughout the search. Candidate gains and their standardization are updated as the selected span expands.

Table 7: Self-calibration search grid for the main configuration.

Parameter	Candidate values	Role
β	$\{0, 1\}$	Controls imprint-amplitude weighting through $\|\Delta W_\ell\|_F^{-2\beta}$. Setting $\beta = 0$ removes amplitude weighting.
γ	$\{0, 1\}$	Controls background suppression, from no transformation to full regularized whitening. Unit normalization is retained in both cases.
ρ	$\{0, 0.5, 1\}$	Controls spectral shaping within each imprint.
o	$\{2, 1\}$	Specifies the token position relative to the non-padding sequence length for token-based text encodings.

For $\rho = 0$, spectral reshaping produces the orthogonal projector onto the eigenspace retained by the relative spectral cutoff ε_0 in Eq. equation 1. For $\rho = 1$, it retains the normalized eigenvalues $\lambda/\lambda_{\max}$ above this cutoff. The intermediate setting $\rho = 0.5$ reduces the disparity between the strengths of retained spectral directions.

Token Position Selection. For a text-aware group using token-based text encodings, the offset o specifies the token position used to represent each prompt, as described in Section 3.3. Let $L(p)$ denote the number of non-padding tokens in prompt p, including special tokens. The zero-based token index is

$$i_o(p) = \max\{L(p) - o, 0\}. \tag{7}$$

The position is determined separately for each prompt; larger offsets select earlier positions. For a CLIP sequence consisting of a beginning-of-text token, content tokens, and an end-of-text token, $o = 1$ selects the end-of-text token and $o = 2$ selects the last content token. Other tokenizers use the same index rule with their respective non-padding sequence lengths. Groups using pooled text representations are unaffected by o.

Confidence and Erased-Set Size Detection. Table 8 lists the fixed parameters for confidence computation, erased-set size detection, and numerical stabilization. We use $\kappa = 0.25$ for group weighting, self-calibration, and size estimation, so $q_{1-\kappa}$ in Eq. equation 4 is the 75th percentile. The search uses a fixed horizon $J = 96$. For size detection, we set $B = 8$ and $\alpha = 0.30$: the initial-rise condition requires increases in at least seven of the first eight transitions, and the first sufficiently large decline must reach 30% of the largest log confidence drop. The cutoff follows the rule in Section 3.4.

Datasets and Candidate Vocabularies. We construct **Imagenette-20** by extending the ten Imagenette classes Howard et al. (2019) with ten non-overlapping categories from ImageNet-1k Russakovsky et al. (2015): *zebra*, *giant panda*, *pizza*, *banana*, *airliner*, *school bus*, *violin*, *laptop*, *submarine*, and *umbrella*. This expanded set enables evaluation across a wider range of erased-set sizes. Artistic-style identification covers the artists evaluated by ESD Gandikota et al. (2023) and UCE Gandikota et al. (2024).

The erased concepts are drawn from these evaluation sets, while identification searches a larger candidate vocabulary $\mathcal{C}$. For object recovery, the standard vocabulary contains the 1,000 ImageNet-1k class names Russakovsky et al. (2015). We construct an expanded vocabulary of 73,406 entries by combining these class names with an English word list from GNU Aspell[1], including its spelling-variant component. This vocabulary covers all 1,000 ImageNet class names under case-insensitive

[1] https://aspell.net/

Table 8: Fixed parameters for confidence computation, erased-set size detection, and numerical stabilization in the main configuration.

Parameter	Value	Role
κ	0.25	Specifies the upper-tail comparison range in the confidence statistic.
J	96	Sets the fixed search horizon.
B	8	Sets the window length for the initial-rise condition.
α	0.30	Sets the relative threshold for the first sufficiently large decline.
ε_0	10^{-3}	Sets the relative spectral cutoff in Eq. equation 1.
ε_c	10^{-12}	Stabilizes the confidence denominator in Eq. equation 4.
ε_z	10^{-9}	Stabilizes gain standardization in Eq. equation 5.
ε_g	10^{-6}	Ensures positive group weights in Eq. equation 5.
ε_a	10^{-30}	Keeps logarithms defined in Eq. equation 6.

Table 9: Prompt templates for probe matching and candidate encoding. Each placeholder is replaced with the complete vocabulary entry.

Index	Object templates	Artistic-style templates
1	`a photo of a {}`	`art by {}`
2	`{}`	`{}`
3	`an image of a {}`	`painting by {}`
4	`a picture of a {}`	`artwork by {}`
5	`a photo of the {}`	`style of {}`

matching and retains multiword names and capitalized proper names. The 70,000-entry vocabulary used in the runtime comparison is a subset of this expanded pool. We will release the complete candidate vocabulary with our code.

Probe Selection and Prompt Templates. We use probe words for the representation-matching test in Section 3.2. Each probe is inserted into the applicable prompt templates in T, and its representations are averaged across templates. The test compares the inputs to edited matrices with the corresponding text encodings to distinguish text-aware from text-agnostic imprints.

Table 9 lists the template sets used for probe matching and candidate encoding. Each set contains five templates, including one that uses the vocabulary entry without additional text. Each placeholder is replaced with the complete entry. Template averaging reduces sensitivity to phrasing, as described in Section 3.3. Artistic-style experiments are conducted separately with both sets to assess sensitivity to the choice of templates.

A.2 Extension to Multiword Concept Identification

TRACER treats each entry in $\mathcal{C}$ as a complete candidate, which may contain multiple words. Identifying a multiword concept already present in $\mathcal{C}$ requires no change to the main pipeline. This section describes an iterative search extension that constructs candidate phrases from individual words. The extension uses the text-aware footprint to guide expansion and selects a phrase length without knowing the target length in advance.

Candidate Initialization and Expansion. We construct a candidate word vocabulary $\mathcal{V}$ from alphabetic CLIP tokenizer entries ending with `</w>`. Each entry contains at least two characters after removing the marker and is represented by a single token. The initial candidate set $\mathcal{C}^{(1)}$ contains all single-word entries in $\mathcal{V}$. Candidates at expansion round j contain exactly j words. This construction covers phrases whose constituent words are each represented by a retained token. At round j, we retain the b highest-scoring candidates in a set $\mathcal{B}^{(j)}$. Each retained candidate is expanded by inserting a word from $\mathcal{V}$ at every possible position:

$$\mathcal{C}^{(j+1)} = \left\{ \mathrm{Insert}(c, v, i) \,\middle|\, c \in \mathcal{B}^{(j)},\ v \in \mathcal{V},\ i = 0, \ldots, j \right\}. \tag{8}$$

Here, $\mathrm{Insert}(c, v, i)$ inserts v after the first i words of c, including before the first word and after the last. Duplicate sequences are removed, while distinct word orders remain separate candidates. Expansion draws from the full vocabulary $\mathcal{V}$, allowing words with low individual scores to contribute to a phrase.

Phrase Encoding and Footprint Scoring. At expansion round j, each candidate $c = (v_1, \ldots, v_j) \in \mathcal{C}^{(j)}$ is encoded as $[\mathrm{BOS}, v_1, \ldots, v_j, \mathrm{EOS}]$, where BOS and EOS mark the beginning and end of the sequence. We extract the CLIP hidden state at the final content-token position, corresponding to v_j. Causal attention allows this representation to incorporate all preceding words. The extension uses the bare phrase with a fixed extraction position, without template averaging or token-offset search. We center the candidate representations, apply regularized covariance whitening with exponent γ, and perform ℓ_2 normalization. The resulting unit vector for candidate c at round j is denoted by $\widetilde{x}_c^{(j)}$. The mean and covariance are recomputed over $\mathcal{C}^{(j)}$ at each expansion round. The initial single-word representations can be reused across checkpoints sharing the same base text encoder and preprocessing settings.

Following Section 3.2, we fuse spectrally reshaped weight footprints within the compatible text-aware group to construct $G_{g_{\mathrm{txt}}}$. This extension uses the 768-dimensional CLIP input space. Each candidate is scored by

$$s^{(j)}(c) = \left(\widetilde{x}_c^{(j)}\right)^\top G_{g_{\mathrm{txt}}} \widetilde{x}_c^{(j)}. \tag{9}$$

The quadratic form measures the footprint energy along the candidate's normalized representation. Dividing it by $\mathrm{tr}(G_{g_{\mathrm{txt}}})$ gives the singleton subspace coverage in this group. We use the unnormalized score and standardize it within each round as described below.

Cross-Length Comparison and Stopping. We standardize candidate scores within each expansion round and apply a length penalty:

$$u^{(j)}(c) = \frac{s^{(j)}(c) - \mu_s^{(j)}}{\sigma_s^{(j)} + \varepsilon_z} - \lambda(j-1), \tag{10}$$

where j denotes the current expansion round and the number of words in each candidate. The quantities $\mu_s^{(j)}$ and $\sigma_s^{(j)}$ are the mean and standard deviation of scores over $\mathcal{C}^{(j)}$, and ε_z stabilizes the denominator. The standardized term measures a candidate's prominence relative to other candidates of the same word length. The coefficient λ penalizes each additional word beyond the first. Within a round, this transformation preserves candidate rankings and therefore leaves beam selection unchanged.

Let $c_j^\star = \arg\max_{c \in \mathcal{C}^{(j)}} u^{(j)}(c)$ denote the highest-scoring candidate at round j. We initialize the best phrase with $c_1^\star$. For each subsequent round $j \geq 2$, we update the best phrase only when

$$u^{(j)}(c_j^\star) > \max_{1 \leq q < j} u^{(q)}(c_q^\star), \tag{11}$$

where q indexes preceding expansion rounds and $c_q^\star$ is the highest-scoring candidate at round q. This condition requires the current round's best penalized score to exceed the best score from all preceding rounds. If it holds, we retain $c_j^\star$ and continue expansion unless the maximum permitted length has been reached. Otherwise, the search stops and returns the best phrase from the preceding rounds. We use a beam size of $b = 8$, a length penalty of $\lambda = 0.05$, and a maximum of three words.

This stopping rule selects the word length of an individual candidate. It is separate from erased-set size detection, which estimates the number of erased concepts. Cross-length comparison is heuristic because both the representation statistics and score distributions change across rounds. The search may therefore return a shorter phrase even when the complete target has been generated.

A.3 Baseline Implementations

We compare TRACER with brute-force generation probing and an adapted membership inference attack (Kong et al., 2024). Both baselines assign a score to each candidate $c \in \mathcal{C}$ based on changes between the public base model W and the unlearned model W'. Candidates are ranked in descending order of these scores.

Brute-Force Generation Probing. This baseline evaluates whether a candidate concept becomes less evident in generated images after unlearning. For each candidate c, we construct prompts containing the complete candidate name and generate images using both W and W'. Let $\mathcal{I}_W(c)$ and $\mathcal{I}_{W'}(c)$ denote the corresponding image sets. We score the candidate by the decrease in average concept evidence:

$$s_{\mathrm{BF}}(c) = \frac{1}{|\mathcal{I}_W(c)|} \sum_{I \in \mathcal{I}_W(c)} a(I, c) - \frac{1}{|\mathcal{I}_{W'}(c)|} \sum_{I \in \mathcal{I}_{W'}(c)} a(I, c), \tag{12}$$

where $a(I, c)$ measures the evidence for concept c in image I. A larger score indicates a greater decline in the model's ability to generate the candidate concept. The same evaluation function is used for both models.

This procedure requires image generation for every candidate in $\mathcal{C}$. Changes in generated content may also reflect effects on related non-target concepts, which can receive high scores despite not being directly targeted.

Adapted Membership Inference Attack. We adapt PIA Kong et al. (2024) from sample-level membership inference to concept-level identification. For each candidate c, let $\mathcal{X}_c$ denote its associated evaluation samples. We apply the PIA error statistic to the same samples under W and W' and aggregate the differences:

$$s_{\mathrm{PIA}}(c) = \frac{1}{|\mathcal{X}_c|} \sum_{x \in \mathcal{X}_c} \left[r_{W'}(x, c) - r_W(x, c) \right], \tag{13}$$

where $r_W(x, c)$ denotes the sample-level PIA error computed using model W with concept conditioning c. A larger score indicates a greater increase in error after unlearning. Averaging over $\mathcal{X}_c$ converts the sample-level changes into a single candidate score.

The adaptation ranks concepts by changes in the PIA statistic rather than making binary membership decisions for individual samples. Its concept scores may capture both direct erasure effects and changes affecting non-target concepts.

Table 10: Identification accuracy (%) of erased artistic styles across three text-to-image backbones. K denotes the number of jointly erased styles. Shaded cells report evaluated results. "–" indicates an unevaluated setting. Avg. is the arithmetic mean over evaluated settings within each row.

Method	Venue	Erased-set size K					Avg. ↑
		1	3	5	10	20	
Stable Diffusion v1.5							
ESD	ICCV 2023	100	100	100	92	–	98.0
UCE	WACV 2024	100	100	100	100	100	100.0
MACE	CVPR 2024	100	100	100	100	–	100.0
FMN	CVPR 2024a	100	100	100	–	–	100.0
ConceptPrune	ICLR 2024	100	100	100	100	–	100.0
Sculpting	ICCV 2025b	100	100	100	100	–	100.0
SPEED	ICLR 2025c	100	100	–	–	–	100.0
ScaPre	ICLR 2026b	100	100	100	100	100	100.0
FIA	ICML 2026a	100	100	100	100	100	100.0
AdvUnlearn	NeurIPS 2024b	100	–	–	–	–	100.0
STEREO	CVPR 2025	100	100	–	–	–	100.0
Meta-Unlearning	ICCV 2025b	100	–	–	–	–	100.0
Stable Diffusion 3							
UCE	WACV 2024	100	100	100	–	–	100.0
EraseFlow	NeurIPS 2026	100	100	100	–	–	100.0
FLUX.1							
EraseAnything	ICML 2025a	100	100	–	–	–	100.0
ScaPre	ICLR 2026b	100	100	100	100	–	100.0

B Additional Experimental Results

Table 10 reports the identification accuracy of erased artistic styles across three text-to-image backbones. TRACER achieves perfect identification in nearly all evaluated settings. On Stable Diffusion v1.5, ESD reaches 92% when ten styles are erased jointly, yielding an average accuracy of 98.0%. All other evaluated configurations achieve 100% accuracy. UCE, ScaPre, and FIA maintain perfect identification with twenty jointly erased styles. These results demonstrate that weight footprints reveal erasure targets beyond object concepts, extending to artistic styles across different backbones and unlearning methods.

We exclude facial-identity unlearning from this evaluation. In our preliminary checks, the tested methods suppress target identities but also degrade the generation of non-target content. These checkpoints therefore provide a limited basis for evaluating target identification under selective unlearning, as their weight changes also reflect broader losses in generative capability.

C Code Sources for Unlearning and Attack Methods

We list the public code repositories for the unlearning methods and the MIA baseline used in our experiments.

- ESD: https://github.com/rohitgandikota/erasing
- UCE: https://github.com/rohitgandikota/unified-concept-editing
- MACE: https://github.com/Shilin-LU/MACE
- FMN: https://github.com/SHI-Labs/Forget-Me-Not
- SPM: https://github.com/THU-MIG/SPM
- ConceptPrune: https://github.com/ruchikachavhan/concept-prune
- Sculpting: https://github.com/coulsonlee/Sculpting-Memory-ICCV-2025
- SPEED: https://github.com/Ouxiang-Li/SPEED
- ScaPre: https://github.com/kaiyuan02415/ScaPre
- FIA: https://github.com/kaiyuan02415/Forget-It-All
- AdvUnlearn: https://github.com/OPTML-Group/AdvUnlearn
- STEREO: https://github.com/koushiksrivats/robust-concept-erasing
- Meta-Unlearning: https://github.com/sail-sg/Meta-Unlearning
- EraseFlow: https://github.com/Abhiramkns/EraseFlow
- EraseAnything: https://github.com/tomguluson92/EraseAnything
- Refusal: https://github.com/4idrossifenil-etanammide/Video-Unlearning
- VideoEraser: https://github.com/bluedream02/VideoEraser
- SAFREE: https://github.com/jaehong31/SAFREE
- MIA baseline (PIA): https://github.com/kong13661/PIA