SEGERASE: DUAL-OUTPUT CONCEPT UNLEARNING FOR TEXT-PROMPTED SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-promptable segmenters learn open-vocabulary grounding from web-scale data and can consequently be prompted to segment sensitive or proprietary concepts. Existing unlearning methods reshape a single global decision, such as a predicted label or a generated image. However, unlearning for segmenters is a dense prediction problem governed by two complementary outputs: pixel-wise mask logits determine *where* a concept is segmented, while an emission score determines *whether* its mask is returned. We therefore introduce SegErase, a concept-unlearning framework that jointly unlearns both outputs through two complementary components. First, Concept Attention-Guided Unlearning concentrates mask-side unlearning on pixels most associated with the target. Second, Teacher-Relative Score Unlearning applies a bounded reduction to target scores while anchoring retained and background predictions to a frozen teacher. Extensive experiments on SAM 3 and COSINE demonstrate that SegErase attains a more favorable unlearning-retention balance than existing unlearning methods across standard concepts, unseen prompt reformulations, and multi-target settings.

1 INTRODUCTION

Promptable and open-vocabulary segmentation models (Lüddecke & Ecker, 2022; Wang et al., 2022; Yang et al., 2022; Kirillov et al., 2023; Carion et al., 2026; Liu et al., 2025) have achieved remarkable open-vocabulary grounding, translating points, boxes, exemplars, or free-form language into object masks, powered by pretraining on vast, web-scale multimodal data. However, this training paradigm allows the models to localize an arbitrary concept on demand, including sensitive, private, or proprietary content that a deployed text interface should not return. To mitigate these risks, researchers have developed *machine unlearning* techniques (Cao & Yang, 2015; Yao et al., 2024; Gandikota et al., 2023; Fan et al., 2024), which remove a specified concept from a model while preserving its response to unrelated inputs.

However, existing unlearning methods are designed for classifiers and generators, whose output is a single label or token distribution. A promptable segmenter instead produces a dense mask together with an emission score that decides whether the mask is returned, so transplanting these methods to the segmentation interface faces the following limitations: **(i) Uniform target-negative supervision over dense mask logits.** A promptable segmenter represents an object through dense pixel-level logits rather than a single label, yet existing unlearning objectives apply a constant negative weight across all target pixels. This ignores that a concept is not uniformly expressed within its mask: interior pixels, discriminative parts, and boundary pixels contribute unequally to whether the model recognizes the concept at all. Dense-prediction losses have long exploited non-uniform pixel weighting (Ronneberger et al., 2015; Bulò et al., 2017; Lin et al., 2017; Kervadec et al., 2019), and ViT attention supplies a concept-responsive weighting signal that is already computed inside the decoder (Caron et al., 2021; Oquab et al., 2024). Uniform unlearning fails to exploit this signal, allocating supervision to pixels with little concept evidence rather than concentrating it on those with stronger evidence. **(ii) Unbounded target reduction on a shared score path.** Beyond the mask, a text-prompted segmenter uses an emission score to gate output. Existing methods unlearn a target by driving its score toward an absolute zero or fixed background. Because the scorer is reused by every retained prompt, such an unbounded objective keeps shifting the shared score distribution long after the target ceases to be emitted, thereby degrading retained prompts.

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.

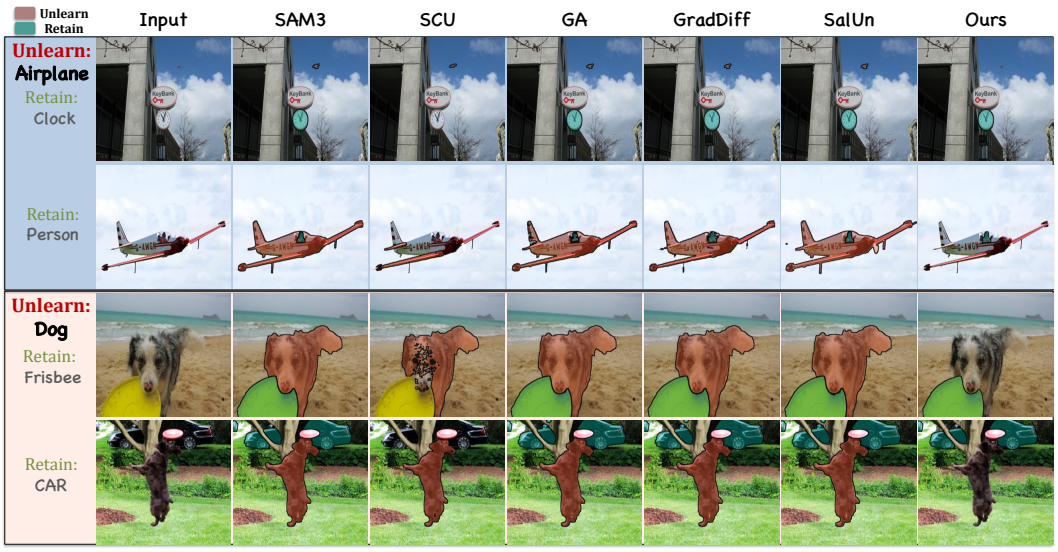

Figure 1: **Qualitative concept unlearning with retained-object preservation.** Brown denotes target masks and teal denotes retained-object masks. For the airplane and dog concepts, SegErase removes target predictions while preserving the clock, person, frisbee, and car; competing methods either leave target masks or alter retained predictions.

To address these limitations, we introduce SegErase, which unlearns the target across both outputs in a single adaptation process. We first propose Concept Attention-Guided Unlearning (CAGU), which converts assigned-query cross-attention into mean-preserving target-region weights. These weights redistribute target-negative supervision over the dense mask logits, so that the unlearning signal concentrates on the pixels that carry the prompted concept instead of spreading uniformly across the target mask. To unlearn the score output without pulling the shared scorer toward an absolute zero, we further introduce Teacher-Relative Score Unlearning (TRSU), which asks only for a bounded reduction of the target score relative to a frozen teacher and anchors the retained-positive and background scores to that same teacher.

In summary, our contributions are threefold. **(i)** We identify two limitations of existing unlearning methods for text-prompted segmentation. Uniform mask supervision overlooks spatial variations in concept evidence, while continued suppression of target scores can degrade retained predictions through the shared scorer. **(ii)** We propose SegErase, a joint mask-and-score unlearning method that combines attention-guided spatial unlearning with bounded teacher-relative score reduction. **(iii)** We empirically demonstrate that SegErase effectively unlearns target concepts while preserving non-target predictions across two segmenters and diverse concept settings.

2 Related Work

Promptable segmentation. Vision–language pretraining makes visual concepts addressable through natural language (Radford et al., 2021). Referring and prompted segmentation grounds these descriptions at pixel level (Lüddecke & Ecker, 2022; Wang et al., 2022; Yang et al., 2022), while open-vocabulary systems learn transferable text-aligned masks through grouping, mask classification, diffusion features, or generalized decoding (Xu et al., 2022; Ding et al., 2023; Xu et al., 2023; Zou et al., 2023; Liu et al., 2025). These models commonly build on query-based set prediction and masked attention (Carion et al., 2020; Cheng et al., 2022). SAM established promptable zero-shot segmentation from spatial prompts (Kirillov et al., 2023); subsequent variants improve efficiency (Zhao et al., 2023; Zhang et al., 2023), extend the interface to video (Ravi et al., 2025), and support noun phrases or visual exemplars (Carion et al., 2026). The resulting text interface requires unlearning the target prompt while preserving responses to other prompts and following the decoder query responsible for its mask.

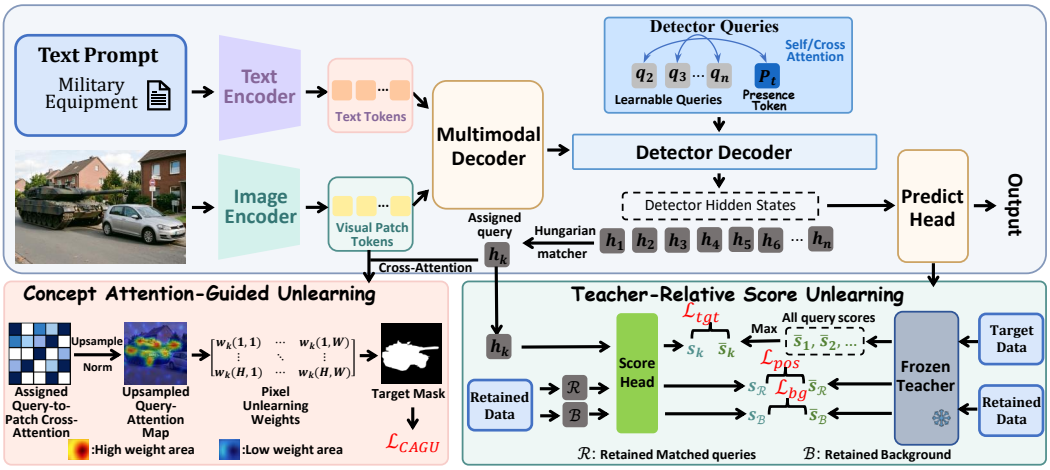

Figure 2: **The SegErase framework.** Native query assignment identifies the candidate responsible for the target region. CAGU converts that candidate's current image-patch attention into the pixel weights w_k used for mask-side unlearning; TRSU applies a bounded reduction to its emission score while anchoring retained and background scores to a frozen teacher.

Machine unlearning. Machine unlearning removes specified training influence without full retraining (Cao & Yang, 2015). Exact, certified, and partitioned schemes provide strong deletion semantics for particular learners or training pipelines (Ginart et al., 2019; Guo et al., 2020; Bourtoule et al., 2021; Brophy & Lowd, 2021; Ghazi et al., 2023). Approximate neural methods instead balance target unlearning and retained utility through optimization, parameter dampening, or restricted updates (Golatkar et al., 2020; Thudi et al., 2022; Liu et al., 2022; Kurmanji et al., 2023; Foster et al., 2024; Fan et al., 2024). This trade-off has also been studied in language and multimodal foundation models (Yao et al., 2024; Maini et al., 2024; Ma et al., 2025; Fan et al., 2025; Shen et al., 2025; Li et al., 2025; Min et al., 2025) and in diffusion models through negative guidance, attention re-steering, and projection edits (Gandikota et al., 2023; Zhang et al., 2024; Gandikota et al., 2024). SCU brings concept unlearning to segmentation foundation models through feature, attention, and output level objectives (Du et al., 2026). In contrast, SegErase explicitly follows the dense mask and scalar emission-score paths of a text-prompted segmenter and preserves non-target outputs.

3 Task Definition

A text-prompted segmenter f_θ maps an image I and a text prompt p to N candidate predictions,

$$f_\theta(I,p) = \{(Z_i,\, s_i)\}_{i=1}^{N}, \qquad Z_i \in \mathbb{R}^{H\times W},\ s_i \in \mathbb{R}, \tag{1}$$

where Z_i determines *where* a concept lies and s_i determines *whether* its mask is emitted. This interface covers query-based decoders ($N{>}1$) and single-output decoders ($N{=}1$).

For a target concept $c^\star$, let $\mathcal{D}_G$ contain target examples and $\mathcal{D}_S$ retained examples. Concept unlearning adapts f_θ to $f_{\theta'}$ so that target masks are no longer returned on $\mathcal{D}_G$, while predictions on $\mathcal{D}_S$ remain close to those of f_θ. We express this method-agnostic objective as

$$\min_{\theta'}\ \mathcal{R}_G(f_{\theta'};\mathcal{D}_G) \quad \text{s.t.} \quad \mathcal{R}_S(f_{\theta'}, f_\theta;\mathcal{D}_S) \le \epsilon, \tag{2}$$

where $\mathcal{R}_G$ measures target response and $\mathcal{R}_S$ measures deviation from the original model on retained examples.

4 METHOD

4.1 OVERVIEW

Figure 2 summarizes SegErase. CAGU uses the assigned query's image attention to weight mask-side unlearning, while TRSU reduces its emission score relative to a frozen teacher and anchors retained and background scores.

For each target sample (I_k, p_k, Y_k), where $Y_k \in \{0,1\}^{H\times W}$ is the target mask, both losses act on the responsible candidate $\omega(k)$, and we write $Z_k := Z_{\omega(k)}$ and $s_k := s_{\omega(k)}$. For a set-prediction decoder, ω is the model's native training-time assignment,

$$\omega = \arg\min_{\pi\in\Pi} \sum_k \mathcal{C}\big(\pi(k), Y_k\big), \tag{3}$$

where $\mathcal{C}$ is its matching cost and Π the one-to-one assignments. We recompute ω at each forward pass and supervised layer so the losses follow the current target query.

4.2 CONCEPT ATTENTION-GUIDED UNLEARNING

With the responsible candidate identified, we first unlearn its dense mask output. Rather than treating all target pixels equally, CAGU uses the candidate's cross-attention over image patches to weight target-negative supervision on the mask logits. This spatial weighting distinguishes regions with stronger concept evidence from less informative interior or boundary pixels.

Let $\Omega = [H] \times [W]$ denote the mask grid and $x = (u, v) \in \Omega$ a spatial location. At every student forward pass, we extract the current head-averaged attention map A_k; it is recomputed as the student changes rather than read from a precomputed cache. We bilinearly resize it to mask resolution and min–max normalize it to $\bar{A}_k \in [0,1]^{H\times W}$. We define the pixel weights directly as

$$w_k(x) = b + a\left[\gamma \bar{A}_k(x) + (1-\gamma)\frac{\sum_{x'\in\Omega} Y_k(x')\bar{A}_k(x')}{\sum_{x'\in\Omega} Y_k(x')}\right], \tag{4}$$

where $a > 0$ controls the attention scale, $b \geq 0$ is a base weight, and $\gamma \in [0,1]$ controls spatial contrast. The target-region mean of w_k is $b + a\sum_x Y_k(x)\bar{A}_k(x)/\sum_x Y_k(x)$ for every γ. Thus, γ redistributes supervision across pixels without changing its total amount over the target. Setting γ=0 gives uniform weighting, whereas γ=1 uses the full attention contrast, so any gain is attributable to allocation rather than loss magnitude.

CAGU then drives the selected logits to background inside the target mask,

$$\mathcal{L}_{\text{CAGU}} = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_x Y_k(x)\, w_k(x)\ \text{BCE}\big(Z_k(x), 0\big)}{\sum_x Y_k(x)}, \tag{5}$$

leaving logits outside Y_k untouched.

4.3 TEACHER-RELATIVE SCORE UNLEARNING

Mask-side unlearning alone does not determine whether the resulting mask is returned. We therefore next adapt the emission score with TRSU. Because this score path is shared across prompts, forcing the target score to an absolute zero continues updating it after the target has stopped being emitted and can reduce retained scores. We instead define the score terms relative to a frozen teacher f_θ, limiting each change to what is required.

Let $\bar{s}$ denote teacher scores and s student scores on the same input, with $\mathcal{R}$ the retained assigned candidates and $\mathcal{B}$ the unassigned candidates. For the target we require a finite reduction m below the teacher's own response; for $r \in \mathcal{R}$ we forbid a drop of more than δ but leave increases free:

$$\mathcal{L}_{\text{tgt}} = \frac{1}{K}\sum_{k=1}^{K}\big[\, s_k - (\bar{s}_k - m)\,\big]_+, \tag{6}$$

$$\mathcal{L}_{\text{pos}} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}}\big[\,(\bar{s}_r - \delta) - s_r\,\big]_+. \tag{7}$$

The two hinges have complementary roles. The target term opens a bounded gap, whereas the retained term prevents an excessive drop. Once the target has fallen m below the teacher, Equation 6 contributes no gradient.

The unassigned candidates $\mathcal{B}$ carry the background objectness distribution, which should not move in either direction, so we anchor them symmetrically:

$$\mathcal{L}_{\text{bg}} = \frac{1}{|\mathcal{B}|} \sum_{q \in \mathcal{B}} D_{\text{KL}}\Big(\text{Bern}\big(\sigma(\bar{s}_q)\big) \,\|\, \text{Bern}\big(\sigma(s_q)\big)\Big). \tag{8}$$

Since the teacher and student are two forward passes of the same architecture, $\bar{s}_k$ is read off without assuming that candidate indices correspond across optimization steps. We use the teacher's largest score for the image–prompt pair. The three terms combine as

$$\mathcal{L}_{\text{TRSU}} = \lambda_q \mathcal{L}_{\text{tgt}} + \lambda_+ \mathcal{L}_{\text{pos}} + \lambda_b \mathcal{L}_{\text{bg}}. \tag{9}$$

4.4 Joint Objective

Decoders refine their predictions layer by layer, so we apply both objectives at each of the L supervised layers, reusing that layer's own selector output:

$$\mathcal{L} = \frac{1}{L} \sum_{\ell=1}^{L} \Big(\lambda_p \, \mathcal{L}_{\text{CAGU}}^{(\ell)} + \mathcal{L}_{\text{TRSU}}^{(\ell)}\Big). \tag{10}$$

5 Experiments

5.1 Experimental Setup

Datasets and splits. The object benchmark contains four COCO concepts (Lin et al., 2014): airplane, bicycle, dog, and cat. For each concept, we use 15 images for unlearning, 50 for validation, and 40 for testing. The PACO benchmark (Ramanathan et al., 2023) contains two parts, two attributes, and four materials, with 30/20/20 images per target for unlearning, validation, and testing, respectively. All three sets are image-disjoint, and the test set is used only for final evaluation. We additionally evaluate unseen prompt reformulations and joint 1/2/4/8-object unlearning using the corresponding object splits. For the data-scaling study, nested unlearning sets contain 15, 30, 60, and 120 images per concept, retain the original 15 images, and add disjoint COCO images. Each setting uses two data passes and are evaluated on the same 80 test images per concept.

Models and baselines. We evaluate SAM 3 (Carion et al., 2026), whose detector decoder exposes native query assignments, and COSINE (Liu et al., 2025), which has a single prompt-conditioned output. We compare with Gradient Ascent, Gradient Difference, SalUn (Fan et al., 2024), and SCU (Du et al., 2026). On SAM 3, every method uses rank-16 LoRA (Hu et al., 2022) on the same trainable modules, prompts, and splits; mechanism and scaling analyses use the same fixed 30-update SegErase recipe. On COSINE, all methods update the transformer decoder while DINOv2 (Oquab et al., 2024), OpenCLIP (Cherti et al., 2023), and the merge module remain frozen.

Metrics and evaluation setting. We evaluate concept unlearning through each model's deployed text interface on image-disjoint test sets. Writing $\text{IoU}(f(I,p), Y)$ for mask IoU after thresholding and NMS, we report

$$G_{\text{IoU}} = \frac{1}{|\mathcal{D}_G|} \sum_{(I,p,Y) \in \mathcal{D}_G} \text{IoU}(f(I,p), Y), \quad S_{\text{IoU}} = \frac{1}{|\mathcal{D}_S|} \sum_{(I,p,Y) \in \mathcal{D}_S} \text{IoU}(f(I,p), Y), \tag{11}$$

with lower target Generality G_{IoU} and higher retained Specificity S_{IoU} preferred. Following referring-segmentation protocols (Wang et al., 2022; Yang et al., 2022), we also report $\text{P@}\tau = |\mathcal{D}|^{-1} \sum \mathbb{1}[\text{IoU} > \tau]$, a sample-level mask precision distinct from detection AP. We use $\tau = .5$ for the object benchmark and $\tau = .3$ for PACO and prompt variants. Every method is evaluated through its native segmentation outputs. All test images share a fixed score threshold $\tau_s = .5$ and the same NMS configuration within each architecture.

Table 1: **Object unlearning across two text-prompted segmenters.** Macro results on the same four image-disjoint object splits, reported as mean ± standard deviation across independent runs. P@.5 is the fraction of image–prompt samples with mask IoU above .5.

Model	Method	G IoU ↓	G P@.5 ↓	S IoU ↑	S P@.5 ↑
SAM 3	Original	.794±.000	.912±.000	.670±.000	.781±.000
	Gradient Ascent	.277±.052	.304±.069	.653±.017	.762±.024
	Gradient Difference	.491±.156	.556±.173	**.664±.009**	**.768±.015**
	SalUn	.293±.122	.321±.139	.659±.003	.759±.003
	SCU	.368±.061	.385±.067	.661±.004	.761±.015
	SegErase	**.029±.028**	**.033±.037**	.635±.009	.727±.023
COSINE	Original	.705±.000	.831±.000	.517±.000	.576±.000
	Gradient Ascent	.323±.078	.369±.077	.485±.005	.543±.008
	Gradient Difference	.323±.016	.358±.025	.492±.006	.548±.013
	SalUn	.405±.014	.462±.017	**.499±.001**	**.561±.004**
	SCU	.309±.018	.340±.036	.489±.006	.542±.014
	SegErase	**.110±.081**	**.117±.100**	.425±.013	.444±.025

5.2 Main Results

Table 1 compares the same four object concepts across both architectures at the reported operating points. On SAM 3, the best baseline reaches .277 G IoU and .304 G P@.5. SegErase lowers these values to .029 and .033, relative reductions of 89.5% and 89.1%, respectively. The P@.5 result means that only 3.3% of target image–prompt pairs retain a mask with IoU above .5, showing that the low mean IoU is consistent across samples. At the same operating point, SegErase retains .635 S IoU and .727 S P@.5, corresponding to 94.8% and 93.1% of the Original model's retained performance.

The same pattern holds on COSINE. Relative to SCU, the strongest target-side baseline, SegErase reduces G IoU from .309 to .110 and G P@.5 from .340 to .117, improvements of 64.4% and 65.6%. It retains .425 S IoU and .444 S P@.5, or 82.2% and 77.1% of the Original model. Thus, SegErase is the only method with the lowest target response under both IoU and P@.5 on both architectures. The higher retention ratios on SAM 3 are consistent with its assignment-specific mask and score outputs, whereas COSINE performs unlearning through a shared prompt-conditioned output. Per-concept, 78-category, and multi-target results appear in Appendices C, D, and E.

5.3 Scaling the Unlearning Set

Figure 3 scales the unlearning set by 8× while holding the number of data passes fixed. The 15-image setting obtains .023 G IoU and .963 S-IoU retention. With 30 images, these metrics improve to .005 and .979, respectively. Increasing the set to 60 images further reduces G IoU to .002 and raises S-IoU retention to .994. At 120 images, G IoU remains .002 and S-IoU retention remains .991. The gains therefore persist across the full range of adaptation-set sizes.

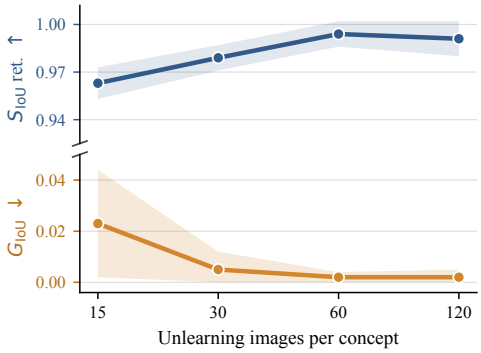

Figure 3: **Data scaling on COCO.** G_{IoU} and S_{IoU} retention on 80 test images per concept.

The largest improvement occurs in the first doubling: from 15 to 30 images, target IoU falls by 78.3% while retained IoU recovery rises by 1.6 percentage points. At 60 images, the standard deviation of G IoU decreases from .021 to .002, alongside the highest observed retention. The nearly unchanged means at 60 and 120 images suggest that performance reaches a stable regime around 60 unlearning images per

Table 2: **PACO cross-type unlearning.** Test-set results over two parts, two attributes, and four materials.

Method	G IoU ↓	G P@.3 ↓	S IoU ↑	S P@.3 ↑
Original SAM 3	0.279±0.000	0.388±0.000	0.737±0.000	0.925±0.000
Gradient Ascent	0.205±0.009	0.277±0.016	0.742±0.000	0.932±0.003
Gradient Difference	0.146±0.010	0.212±0.017	0.741±0.002	0.934±0.002
SalUn	0.171±0.005	0.231±0.006	**0.742±0.005**	**0.935±0.004**
SCU	0.081±0.038	0.108±0.059	0.729±0.003	0.917±0.005
SegErase	**0.002±0.004**	**0.004±0.007**	0.734±0.004	0.915±0.010

Table 3: **Unseen prompt-variant attacks across architectures.** Jailbreak reformulations and adaptive paraphrases are excluded from training. We retain only image–prompt pairs recognized by the corresponding original model at IoU $> .3$.

Model	Method	Jailbreak ↓		Adaptive ↓	
		IoU	P@.3	IoU	P@.3
SAM 3	Original	.806±.000	1.000±.000	.826±.000	1.000±.000
	Gradient Ascent	.646±.040	.800±.039	.645±.038	.781±.038
	Gradient Difference	.687±.021	.864±.027	.708±.030	.870±.019
	SalUn	.652±.039	.810±.043	.665±.037	.807±.041
	SCU	.607±.017	.762±.019	.637±.013	.787±.021
	SegErase	**.344±.016**	**.425±.017**	**.407±.026**	**.492±.030**
COSINE	Original	.737±.000	1.000±.000	.705±.000	1.000±.000
	Gradient Ascent	.175±.059	.267±.077	.116±.035	.188±.047
	Gradient Difference	.152±.006	.230±.007	.092±.004	.145±.005
	SalUn	.199±.011	.302±.018	.123±.007	.198±.008
	SCU	.152±.008	.231±.010	.095±.003	.150±.004
	SegErase	**.040±.029**	**.058±.040**	**.017±.009**	**.023±.015**

concept. The same trend on the expanded 80-image test set shows that the result extends beyond the original 40 test images.

5.4 Generalization Across Target Types

Table 2 extends the evaluation from object nouns to a matched PACO matrix of parts, attributes, and materials. SegErase reduces macro G IoU from SCU's .081 to .002 while retaining 99.6% of the Original model's non-target IoU (.734 versus .737). The result indicates that the learned operating point extends beyond salient object categories to compositional targets. Unlike object nouns, these targets may occupy only a small region, describe appearance shared by multiple objects, or depend on the relationship between a part and its parent object. Strong performance across all three types shows that SegErase does not rely on the target being a standalone object with a single contiguous mask.

5.5 Robustness to Unseen Prompts

Table 3 evaluates jailbreak reformulations and adaptive descriptions that are never used during adaptation. On SAM 3, SegErase lowers jailbreak/adaptive G IoU from SCU's .607/.637 to .344/.407; on COSINE, it lowers the strongest baseline values from .152/.092 to .040/.017. The remaining SAM 3 response shows that reformulation is the hardest evaluated setting, but SegErase is strongest on both IoU and P@.3 across the two suites and architectures. This experiment tests whether the learned unlearning is tied to the exact wording used during adaptation. The consistent ranking across reformulations indicates that the method acts on the target concept rather than memorizing a single prompt string.

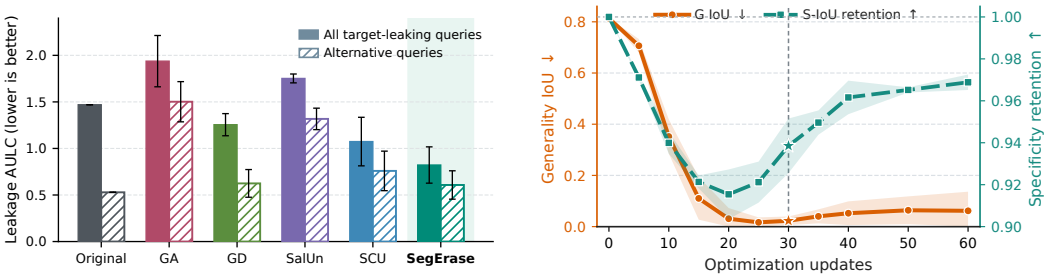

Figure 4: **Output-space and optimization diagnostics. Left:** Pre-NMS target-query leakage for the reported SAM 3 models; AULC integrates target-overlapping queries over all score thresholds, and hatched bars exclude the currently assigned target query. **Right:** validation-set trajectory through 60 updates; the star marks update 30 in this separate extended schedule.

5.6 Ablation Study

Table 4 evaluates target masks before score filtering. Without TRSU, G IoU remains .742, showing that mask-side supervision alone leaves a strong target-shaped response. Removing CAGU increases G IoU from .029 to .639 and lowers S IoU from .635 to .629. The mean-matched comparison in Appendix B.1 isolates this spatial-allocation effect more directly. Replacing the teacher-relative margin with an absolute target reduces G IoU to .359 but also reduces S IoU to .557; the same retention trend appears on COSINE in Appendix B. Appendix B.2 further examines sensitivity to CAGU spatial strength, TRSU margin, and retained-positive weight. The results separate the roles of the two output paths: CAGU removes the target-shaped mask response, whereas TRSU prevents the score update from damaging retained predictions.

Table 4: **Component ablation on SAM 3.** G IoU uses the highest-IoU query before score filtering.

Variant	CAGU	G IoU ↓	S IoU ↑
w/o CAGU	×	.639±.056	.629±.012
w/o TRSU	√	.742±.062	.652±.012
Absolute score target	√	.359±.079	.557±.019
SegErase	√	**.029±.028**	**.635±.009**

5.7 Residual Query Leakage

Final-output IoU can hide target-shaped responses that remain below the score threshold or are removed by NMS. Figure 4 (left) therefore measures every pre-NMS query whose mask overlaps the target by more than .3. The figure uses the models reported in the object benchmark. Across four concepts, SegErase gives the lowest total leakage AULC among baselines (.822), compared with 1.074 for SCU and 1.470 for the Original model. Its nonzero alternative-query AULC (.608) also shows that final emitted IoU does not imply complete removal of every decoder response. Table 5 complements emitted output with oracle-mask and pre-NMS measurements on airplane. These views distinguish deployed unlearning from residual decoder response. These findings confirm that SegErase uppresses target-shaped responses more completely and more consistently than every baseline, at every stage of the decoding pipeline.

Table 5: **Pre-NMS query leakage on airplane.**

Method	AULC ↓	$Q_{.3}$ ↓	$Q_{.5}$ ↓	$U_{.3}$ (%) ↓
Original SAM 3	30.37	26	22	23.1
CAGU (mask only)	58.43	75	26	74.7
Gradient Ascent	108.36	115	23	81.7
Gradient Difference	24.29	21	20	14.3
SalUn	36.37	26	23	23.1
SCU	52.41	81	22	76.5
SegErase	**19.59**	**9**	**0**	**11.1**

5.8 Optimization Dynamics

Figure 4 (right) shows that target unlearning and retained utility evolve in opposite phases. Macro G IoU reaches .017±.019 at 25 updates and is .023±.017 at update 30, then rises to .052±.046 at 40 and .062±.075 at 60. Over the same interval, S-IoU retention recovers from .921±.010 to

Table 6: **Public-figure unlearning across two segmenters.** SAM 3 covers four identities; COSINE covers the two identities recognized by its Original model on the test set.

Model	Method	G IoU ↓	G P@.3 ↓	S IoU ↑	S P@.3 ↑
SAM 3	Original	.990±.000	1.000±.000	.866±.000	.876±.000
	Gradient Ascent	.739±.045	.753±.049	.828±.000	.841±.000
	Gradient Difference	.434±.273	.444±.277	.854±.013	.872±.015
	SalUn	.677±.071	.694±.076	.827±.008	.840±.008
	SCU	.336±.176	.345±.175	.854±.012	.872±.013
	SegErase	**.127±.000**	**.128±.000**	**.866±.000**	**.876±.000**
COSINE	Original	.614±.000	1.000±.000	.060±.000	.091±.000
	Gradient Ascent	**.000±.000**	**.000±.000**	.000±.000	.000±.000
	Gradient Difference	.258±.252	.356±.310	**.129±.073**	**.176±.099**
	SalUn	.252±.253	.356±.310	.114±.062	.144±.074
	SCU	.190±.231	.250±.250	.071±.063	.087±.077
	SegErase	**.000±.000**	**.000±.000**	.125±.069	.167±.095

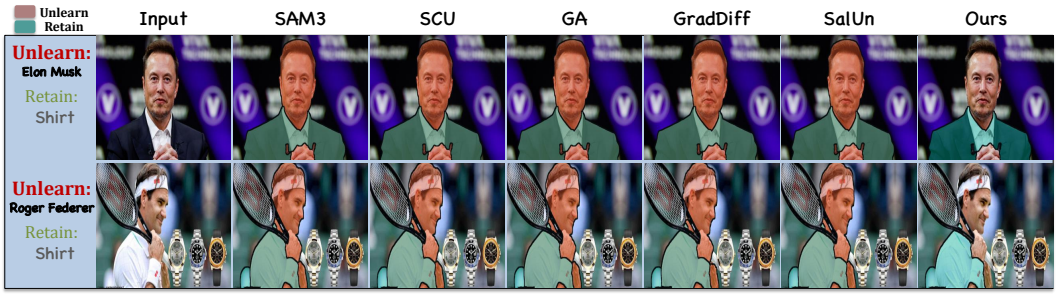

Figure 5: **Qualitative public-figure unlearning.** Brown denotes identity masks and teal denotes retained shirt masks. Across Elon Musk and Roger Federer examples, SegErase removes the identity mask while preserving the retained clothing region more faithfully than the competing methods.

.969±.004. Extending optimization therefore shifts the solution toward retention and increases target-side variance, revealing sensitivity to optimization length.

5.9 Public-Figure Unlearning

Table 6 extends the evaluation from category-level concepts to named identities. On SAM 3, SegErase reduces G IoU from SCU's .336 to .127 while matching the Original model's .866 S IoU. On COSINE, it reaches zero G IoU and retains .125 S IoU; Gradient Difference retains slightly more (.129) but leaves .258 target IoU. The qualitative examples in Figure 5 show the same distinction between removing the named identity and preserving the retained clothing region. These results show that SegErase transfers directly from category nouns to named identities, achieving the strongest suppression–retention trade-off on both architectures.

6 Conclusion

We introduced SegErase, a dual-output framework for concept unlearning in text-prompted segmentation. By jointly suppressing target masks and emission scores, SegErase removes target responses while preserving retained segmentation behavior across architectures and concept types. These results show that effective unlearning requires controlling both spatial mask representations and the scores that govern their emission. Future work could study larger concept collections, continual unlearning, and stronger adaptive attacks against the remaining internal representations.

AI USE STATEMENT

Generative AI tools assisted language, figure, and table refinement. The authors reviewed all AI-assisted text and take responsibility for the final content.

ETHICS STATEMENT

The main experiments use public COCO images and category annotations. The sensitive-region experiment uses pseudo-mask annotations.

REPRODUCIBILITY STATEMENT

Configurations, fixed operating points, and per-sample test predictions accompany the submission; Appendix A gives implementation details.

REFERENCES

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

Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1092–1104. PMLR, 2021.

Samuel Rota Bulò, Gerhard Neuhold, and Peter Kontschieder. Loss max-pooling for semantic image segmentation. In *CVPR*, pp. 7082–7091. IEEE Computer Society, 2017.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE Computer Society, 2015.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pp. 213–229. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9630–9640. IEEE, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pp. 1280–1289. IEEE, 2022.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pp. 2818–2829. IEEE, 2023.

Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8090–8102. PMLR, 2023.

Miaozeng Du, Jiaqi Li, Sirui Pan, Yi Zhan, Guilin Qi, Yuxin Zhang, Rihui Jin, Yinjia Shu, and Qianshan Wei. Forget what has seen: Selective concept unlearning in segmentation foundation models. In *AAAI*, pp. 20923–20931. AAAI Press, 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. In *ICLR*. OpenReview.net, 2024.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning. In *NeurIPS*, 2025.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI*, pp. 12043–12051. AAAI Press, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pp. 2426–2436. IEEE, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pp. 5099–5108. IEEE, 2024.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Pasin Manurangsi, Ayush Sekhari, and Chiyuan Zhang. Ticketed learning-unlearning schemes. In *COLT*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5110–5139. PMLR, 2023.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *NeurIPS*, pp. 3513–3526, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, pp. 9301–9309. Computer Vision Foundation / IEEE, 2020.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3832–3842. PMLR, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *MIDL*, volume 102 of *Proceedings of Machine Learning Research*, pp. 285–296. PMLR, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *ICCV*, pp. 3992–4003. IEEE, 2023.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *NeurIPS*, 2023.

Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Hui Zhang, Yongrui Chen, Qianshan Wei, Junfeng Fang, Ruipeng Wang, Sheng Bi, and Guilin Qi. Forget the token and pixel: Rethinking gradient ascent for concept unlearning in multimodal generative models. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pp. 12179–12200. Association for Computational Linguistics, 2025.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pp. 2999–3007. IEEE Computer Society, 2017.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *CoLLAs*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254. PMLR, 2022.

Yang Liu, Yufei Yin, Chenchen Jing, Muzhi Zhu, Hao Chen, Yuling Xi, Bo Feng, Hao Wang, Shiyu Li, and Chunhua Shen. Unified open-world segmentation with multi-modal prompts. In *ICCV*, pp. 21557–21567. IEEE, 2025.

Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts. In *CVPR*, pp. 7076–7086. IEEE, 2022.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *ICLR*. OpenReview.net, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for llms. *CoRR*, abs/2401.06121, 2024.

Thomas De Min, Subhankar Roy, Stéphane Lathuilière, Elisa Ricci, and Massimiliano Mancini. Group-robust machine unlearning. *Trans. Mach. Learn. Res.*, 2025, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: parts and attributes of common objects. In *CVPR*, pp. 7141–7151. IEEE, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*. OpenReview.net, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume abs/1505.04597, 2015.

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM unlearning via neural activation redirection. In *NeurIPS*, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *EuroS&P*, pp. 303–319. IEEE, 2022.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: clip-driven referring image segmentation. In *CVPR*, pp. 11676–11685. IEEE, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas M. Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pp. 18113–18123. IEEE, 2022.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, pp. 2955–2966. IEEE, 2023.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: language-aware vision transformer for referring image segmentation. In *CVPR*, pp. 18134–18144. IEEE, 2022.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In *NeurIPS*, 2024.

Chaoning Zhang, Dongsheng Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight SAM for mobile applications. *CoRR*, abs/2306.14289, 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 *CVPR Workshops*, pp. 1755–1764. IEEE, 2024.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *CoRR*, abs/2306.12156, 2023.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *CVPR*, pp. 15116–15127. IEEE, 2023.

A Implementation Details

All SAM 3 experiments use rank-16 LoRA ($\alpha_{\mathrm{L}} = 32$, dropout .1) on the vision encoder, text encoder, geometry encoder, DETR encoder and decoder, and mask decoder; pre-trained weights remain frozen. Unless otherwise stated, we optimize for 30 updates with batch size one, learning rate 5×10^{-5}, weight decay .01, three warmup updates, cosine decay, BF16 autocast, and gradient clipping at one. The loss coefficients are $\lambda_p = 1$, $\lambda_q = 20$, $\lambda_+ = .2$, and $\lambda_b = .05$; the score margin is $m = 3$, the retained-positive tolerance is $\delta = .1$, and CAGU uses $(a, b, \gamma) = (1, .1, .5)$. Candidate expansion is disabled. Every decoder layer uses its own native matching result.

Table 7: **Deployment cost across architectures.** Latency is model-forward time over 50 runs on one A100. Methods sharing a row have the same deployed computation graph and trainable surface.

Model	Deployed method	Added	Total	Latency (ms)	Mem. (GiB)
SAM 3	Original	0	840.5M	300.2±49.8	3.96
	GA/GD/SalUn/SCU/SegErase	18.5M	859.0M	449.3±55.2	5.30
COSINE	Original	0	681.3M	128.1±2.6	3.39
	GA/GD/SalUn/SCU/SegErase	0	681.3M	131.3±13.1	3.39

We use SegErase's final checkpoint and select baseline checkpoints on the validation set. Test images are not used for checkpoint, threshold, or hyperparameter selection.

For PACO, baselines run for up to 60 updates and their checkpoints are selected on the validation set; SegErase uses its final 30-update checkpoint.

The COSINE experiments use the same object splits and top-5 retained-category specificity protocol. CAGU derives spatial weights from semantic-prompt–visual-feature relevance, and TRSU applies a teacher-relative finite margin to the prompt-conditioned score map. The configuration uses 12 updates, learning rate 10^{-5}, CAGU strength $\gamma = .5$, score margin 6, retained replay of two samples per update, and retention weight 6. DINOv2, OpenCLIP, and the merge module remain frozen while the transformer decoder is updated.

Table 7 reports deployment cost for each architecture. On SAM 3, all rank-16 parameter methods add 18.5M LoRA parameters and measure 449.3 ± 55.2 ms per forward pass, compared with 300.2 ± 49.8 ms for the Original model. COSINE updates 17.9M existing decoder parameters in place and adds no parameters; Original and updated COSINE measure 128.1 ± 2.6 and 131.3 ± 13.1 ms with the same 3.39 GiB peak memory.

B Additional Objective Analyses

B.1 Mean-Matched Attention Weighting

Table 8: **Mean-matched spatial weighting comparison.** Test IoU from one run when the target-mask loss uses either a uniform weight or CAGU attention weights with the same target-region mean.

Target	G IoU ↓		S IoU ↑	
	Uniform	Attention	Uniform	Attention
Airplane	.565	**.494**	.562	**.569**
Bicycle	.318	**.249**	.677	**.679**
Dog	.263	**.159**	.738	**.747**
Cat	.153	**.035**	.736	**.737**
Macro	.325	**.234**	.678	**.683**

The matched-mean comparison changes only the spatial allocation of the mask loss. Attention weighting lowers G IoU for all four targets and reduces macro G IoU from .325 to .234, while macro S IoU increases slightly from .678 to .683. Because the target-region mean weight is identical, the unlearning gain comes from spatial allocation rather than a larger average loss magnitude. This matched comparison supports CAGU's spatial-allocation mechanism.

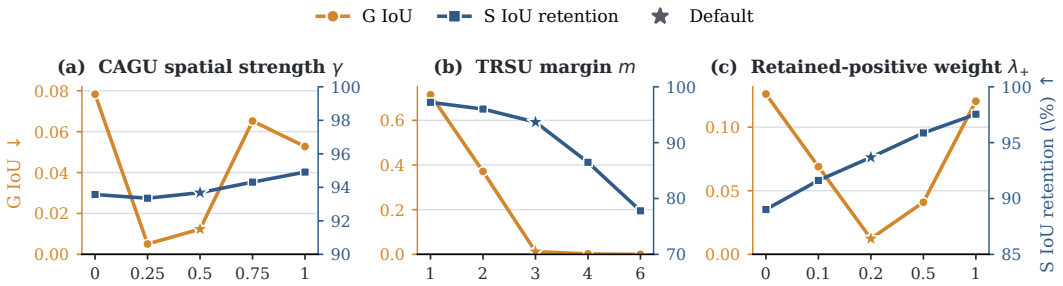

Figure 6: **Hyperparameter sensitivity on SAM 3.** Four-concept validation-set macro at the fixed 30-update endpoint of one run. Each panel varies one hyperparameter while fixing the others; stars mark the reported configuration. Lower G IoU and higher S IoU retention are better.

B.2 Hyperparameter Sensitivity

Figure 6 separates the roles of the three method-specific hyperparameters. For CAGU, $\gamma = .25$ and .5 both produce low target IoU (.005 and .012) with approximately 93–94% retained IoU; the reported $\gamma = .5$ does not rely on the best point of this sweep. The TRSU margin exposes the expected trade-off: increasing m from 1 to 6 lowers G IoU from .714 to zero while S IoU retention falls from 97.2% to 77.8%. The selected $m = 3$ retains 93.7% while reducing G IoU to .012. Finally, $\lambda_+ = .2$ gives the lowest G IoU in its sweep; larger retained-positive weights recover up to 97.5% S IoU but allow more target response. These sweeps support the reported setting as a balanced operating point rather than a test-selected optimum.

B.3 Architecture-Matched Score Objective

Table 9: **Bounded versus absolute score unlearning.** Four-concept test-set macro under the common evaluation protocol.

Model	Target score loss	G IoU ↓	G P@.5 ↓	S IoU ↑	S P@.5 ↑
SAM 3	BCE-to-background	**.001±.001**	**.000±.000**	.557±.019	.632±.027
	Teacher-relative margin	.029±.028	.033±.037	**.635±.009**	**.727±.023**
COSINE	BCE-to-background	**.012±.002**	**.000±.000**	.411±.015	.434±.012
	Teacher-relative margin	.110±.081	.117±.100	**.425±.013**	**.444±.025**

The architecture-matched comparison in Table 9 changes only the score target while holding all other components fixed. On COSINE, the margin exchanges G/S IoU .012/.411 for .110/.425, increasing retained IoU while keeping target IoU far below the original model. This is the same teacher-relative operating point reported in Table 1.

B.4 COSINE Retention Frontier

Figure 7 shows the effect of retained-positive pressure on the validation set. Increasing it moves G IoU from .074 at $\lambda_+ = 0$ to .338 at .20, while macro specificity retention rises from .918 to approximately .95. The response is not monotone per concept: the shared dense map lets protection of one retained category alter target and retained pixels elsewhere, and the dual multiplier reaches its clip value on the harder dog and cat runs. This pattern differs from SAM 3, where TRSU acts on assignment-selected scalar scores and retained constraints can protect other queries more directly.

C Per-Concept Test Results

Table 10 breaks down the main results by target. As shown in the SegErase's SAM 3 results, it gives the lowest G IoU for every object and lowers macro G IoU from the best comparison row's .277 to .029, an 89.5% reduction. Dog accounts for most of the remaining SAM 3 variance. On COSINE,

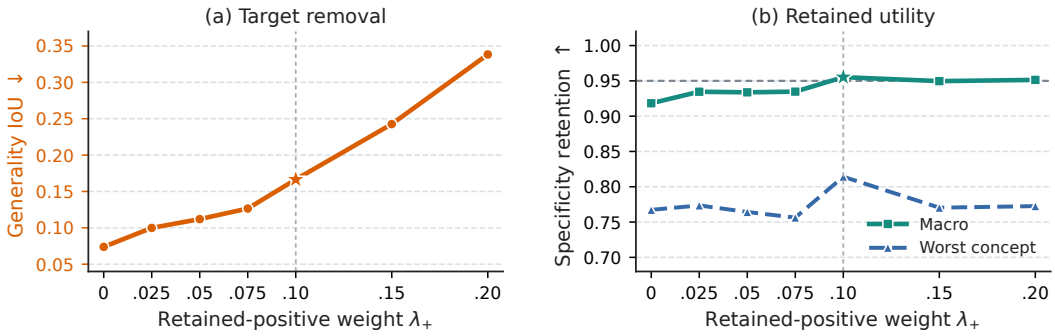

Figure 7: **Seven-point COSINE validation frontier.** All points use four-concept macro averaging and a 15-update budget. The star marks $\lambda_+ = .1$, the lowest-G point satisfying the 95% macro specificity-retention constraint; $\lambda_+ = .2$ also satisfies the macro constraint but leaves a higher target response. The worst-concept curve shows that bicycle remains below 90% retention, exposing dense-output coupling hidden by the macro average.

Table 10: **Per-concept object unlearning across architectures.**

Model	Method	Airplane	Bicycle	Dog	Cat	Macro G ↓	Macro S ↑
SAM 3	Original	.834±.000	.630±.000	.843±.000	.868±.000	.794±.000	.670±.000
	Gradient Ascent	.296±.210	.199±.086	.232±.071	.380±.064	.277±.052	.653±.017
	Gradient Difference	.655±.084	.335±.302	.344±.439	.629±.287	.491±.156	**.664±.009**
	SalUn	.245±.186	.105±.019	.300±.439	.521±.146	.293±.122	.659±.003
	SCU	.142±.069	.086±.016	.406±.258	.839±.013	.368±.061	.661±.004
	SegErase	**.032±.029**	**.026±.006**	**.056±.096**	**.001±.001**	**.029±.028**	.635±.009
COSINE	Original	.740±.000	.469±.000	.767±.000	.847±.000	.705±.000	.517±.000
	Gradient Ascent	.193±.107	.322±.012	.383±.278	.396±.184	.323±.078	.485±.005
	Gradient Difference	.167±.090	.306±.041	.235±.043	.586±.090	.323±.016	.492±.006
	SalUn	.258±.136	.367±.018	.347±.066	.647±.041	.405±.014	**.499±.001**
	SCU	.153±.077	.311±.042	.228±.025	.545±.093	.309±.018	.489±.006
	SegErase	**.027±.011**	**.043±.013**	**.000±.000**	**.369±.314**	**.110±.081**	.425±.013

SegErase is strongest on airplane and bicycle and gives the best macro G IoU; dog is statistically tied with SCU, while cat remains the hardest transfer target.

D Unlearning Across 78 COCO Categories

Table 11: **Concept unlearning across 78 COCO categories on SAM 3.** Values are category means ± standard deviations. G is averaged over all 78 target categories; S retention is averaged over the 76 categories with non-target examples. Each category uses 15 images for unlearning, 10 for validation, and 10 for testing.

Method	G IoU ↓	G P@.3 ↓	S IoU ret. ↑	S P@.3 ret. ↑
Original SAM 3	.735±.108	.929±.101	1.000±.000	1.000±.000
Gradient Ascent	.092±.011	.211±.011	.874±.140	.885±.148
Gradient Difference	.151±.177	.167±.231	.982±.082	.994±.065
SalUn	.101±.208	.170±.106	.890±.128	.909±.133
SCU	.539±.292	.694±.359	**1.003±.049**	**1.000±.054**
SegErase	**.028±.066**	**.033±.099**	.974±.066	.963±.070

Table 11 evaluates the same SAM 3 methods across 78 COCO categories. SegErase achieves the lowest target response on both metrics, reducing G IoU from .735 to .028 and G P@.3 from .929 to .033, while retaining 97.4% of the Original model's non-target IoU. Gradient Ascent and SalUn leave higher target responses and retain only 87.4% and 89.0% of non-target IoU, respectively. Gradient Difference retains 98.2% but leaves .151 target IoU, while SCU preserves non-target

Table 12: **Joint multi-target unlearning across two segmenters.** Mean ± standard deviation across independent runs for jointly forgotten concept sets of increasing size.

Model	Method	1 target		2 targets		4 targets		8 targets	
		G IoU ↓	S Ret. ↑	G IoU ↓	S Ret. ↑	G IoU ↓	S Ret. ↑	G IoU ↓	S Ret. ↑
SAM 3	Gradient Ascent	.464±.290	**.968±.020**	.840±.048	.991±.015	.485±.266	.949±.017	.786±.000	**1.000±.000**
	Gradient Difference	.695±.249	.944±.059	.635±.210	.999±.013	.638±.177	.990±.006	.454±.322	.970±.032
	SalUn	.601±.146	.930±.026	.582±.435	.971±.023	.379±.234	.953±.015	.786±.000	**1.000±.000**
	SCU	.559±.311	.921±.023	.421±.020	**1.020±.026**	.506±.203	.966±.022	.228±.162	.976±.010
	SegErase	**.000±.000**	.797±.006	**.076±.131**	.975±.057	**.004±.000**	**.995±.006**	**.000±.000**	.977±.011
COSINE	Gradient Ascent	**.000±.000**	.499±.009	**.000±.000**	.122±.008	**.000±.000**	.007±.008	**.000±.000**	.000±.000
	Gradient Difference	**.000±.000**	.846±.222	**.000±.000**	.692±.095	.003±.005	**.914±.019**	.017±.006	.768±.147
	SalUn	**.000±.000**	**.972±.094**	**.000±.000**	.646±.092	.005±.008	.899±.003	.030±.013	.707±.179
	SCU	.006±.010	.819±.222	**.000±.000**	.535±.143	.002±.004	.899±.013	.003±.002	.667±.163
	SegErase	.107±.084	.921±.125	**.000±.000**	**.816±.193**	.050±.040	.891±.025	.085±.012	**.826±.165**

Table 13: **Sensitive-region unlearning across two segmenters.** The COSINE target subset contains two test frames recognized by its Original model. Values are mean ± sample standard deviation across independent runs; lower G and higher S are better. Bold marks the best adapted result within each model.

Model	Method	G IoU ↓	G P@.3 ↓	S IoU ↑	S P@.3 ↑
SAM 3	Original	.965±.000	1.000±.000	.964±.000	1.000±.000
	Gradient Ascent	.391±.201	.451±.245	**.961±.002**	**1.000±.000**
	Gradient Difference	.322±.205	.353±.235	.960±.001	**1.000±.000**
	SalUn	.312±.290	.412±.408	.957±.005	**1.000±.000**
	SCU	.246±.221	.333±.296	.960±.000	**1.000±.000**
	SegErase	**.163±.166**	**.216±.223**	.956±.007	**1.000±.000**
COSINE	Original	.487±.000	1.000±.000	.162±.000	.212±.000
	Gradient Ascent	**.000±.000**	**.000±.000**	.022±.009	.033±.011
	Gradient Difference	**.000±.000**	**.000±.000**	.109±.071	.158±.107
	SalUn	**.000±.000**	**.000±.000**	.099±.075	.143±.117
	SCU	**.000±.000**	**.000±.000**	.096±.077	.136±.121
	SegErase	**.000±.000**	**.000±.000**	**.122±.076**	**.179±.117**

predictions but leaves substantial target response. These results show that SegErase maintains a strong unlearning–retention balance across a substantially broader set of concepts.

E Multi-Target Scaling

Table 12 evaluates the same 1/2/4/8-concept scaling protocol on both architectures with all parameter baselines rerun under the same protocol. On SAM 3, SegErase gives the lowest G IoU at every target count and reaches .975/.995/.977 S retention for 2/4/8 targets. On COSINE, SegErase maintains .816–.921 S retention across the multi-target settings and reaches zero G IoU at two targets.

F Sensitive-Region Evaluation

We additionally evaluate sensitive regions outside the COCO object benchmark. The COSINE subset contains two recognized test frames.

On SAM 3, SegErase gives the lowest target G IoU (.163) while retaining .956 S IoU. All COSINE baselines eliminate the target response on its two-frame subset; SegErase gives the highest S IoU among them (.122). Figure 8 shows representative target and retained prompts.

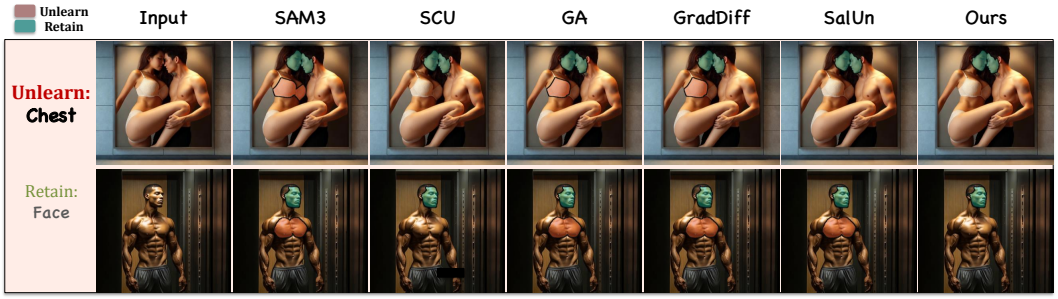

Figure 8: **Qualitative sensitive-region unlearning.** Brown denotes the target region and teal denotes the retained region. The examples compare chest unlearning with face retention across the Original SAM 3 model, four baselines, and SegErase.