Forget What Has Seen: Selective Concept Unlearning in Segmentation Foundation Models

Miaozeng Du[1,3*], Jiaqi Li[2,3*], Sirui Pan[2], Yi Zhan[4], Guilin Qi[2,3†], Yuxin Zhang[2,3], Rihui Jin[2,3], Yinjia Shu[2], Qianshan Wei[2]

[1]College of Software Engineering, Southeast University, Nanjing, China
[2]School of Computer Science and Engineering, Southeast University, Nanjing, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[4]School of Computer Science, Peking University, Beijing, China
{miaozengdu,jqli,gqi}@seu.edu.cn

Abstract

Machine unlearning (MU) has emerged as a critical tool for removing sensitive or personal information from machine learning models, empowering individuals with the right to be forgotten. While MU has achieved success in classification and generative tasks, whether this technique can be effectively applied to segmentation foundation models remains uncertain. To address this issue, we propose an efficient method, Selective Concept Unlearning (SCU), to unlearn the segmentation capability of target concepts. SCU consists of several key aspects: (1) The Multi-level Forgetting Module, designed with a hierarchical three-level suppression strategy, including *(i)* distillation-level: Negative distillation steers model's output distribution away from teacher's correct outputs, erasing its learned concept recognition. *(ii)* attention-level: Attention suppression minimizes model's attention to target regions. *(iii)* output-level: Directly erases predictions for the target by relabeling as background. (2) The Preservation Module ensures maintaining segmentation quality for non-target concepts. Additionally, we introduce a set of metrics to evaluate segmentation unlearning methods. Experiments demonstrate that SCU consistently outperforms existing baselines.

1 Introduction

Image segmentation foundation models, such as the Segment Anything Model (SAM) (Kirillov et al. 2023), have achieved unprecedented success in recent years. SAM's ability to precisely segment diverse objects with minimal to no annotations has spurred its rapid adoption across different scenarios, including medical imaging (Ma et al. 2024; Wu et al. 2025), autonomous driving (Shan and Zhang 2023), robotics (Wang et al. 2023), and environmental monitoring (Huang et al. 2024; Osco et al. 2023). However, as these systems become increasingly integrated into sensitive and high-stakes environments, their powerful zero-shot capabilities and inherent ability to process vast quantities of visual data give rise to significant security, sensitivity, and privacy concerns. Concurrently, the concept of machine unlearning (MU) (Shintre, Roundy, and Dhaliwal 2019; Ginart et al. 2019) has emerged as a crucial paradigm for selectively removing or suppressing knowledge from foundation models while preserving the overall performance. Previous studies have investigated the effectiveness of machine unlearning in diverse tasks, ranging from image classification to generative scenarios (e.g. Large Language Models (Hu et al. 2024; Chen and Yang 2023), text-to-image models (Lu et al. 2024; Fan et al. 2023; Zhang et al. 2024), and Multimodal Large Language Models (Li et al. 2024)).

Figure 1: SCU selectively forgets sensitive concepts (fighter jet) while preserving non-sensitive segmentation (bicycles).

However, whether MU can be effectively applied to segmentation foundation models has not been thoroughly explored. Unlearning target concepts in segmentation foundation models presents significant challenges. A primary obstacle is **high-difficulty region-wise forgetting**. Unlike classification tasks that only require predicting one or a few labels and can achieve unlearning by substituting the target label, segmentation task requires dense pixel predictions, with each concept represented by a specific region composed of numerous pixels scattered throughout the image. This presents a far more complex and large-scale unlearning target. Another challenge is **model degradation**. In the Segmentation foundation model, target-specific features are tightly entangled with background content and non-target concepts, making it difficult to forget target features with-

*These authors contributed equally.
†Corresponding author.

out disrupting the surrounding structure. Consequently, suppressing the features for the target concept risks interfering with those of others. Fig 1 illustrates what an ideal solution should accomplish: nullifying the model's knowledge of a target concept, without inflicting collateral damage on its ability to segment other concepts.

To address the above challenges, we take the first step to explore MU in segmentation foundation models by proposing a novel method, SCU. We start by first introducing a new task, namely segmentation concept unlearning, in which a set of concepts can be safely removed from the pretrained segmentation model. To realize this goal, we design a framework comprising a Multi-level Forgetting Module and a Preservation Module. Multi-level Forgetting Module employs a hierarchical strategy across three levels: (i)Distillation level reverses supervision from a pretrained teacher model steers model's output distribution away from original predictions on the target. (ii)Attention level, attention suppression in the final attention layer reduces model's focus on the target concept, weakening attention responses to the target region. (iii)Output level relabels the target region as background to unlearn the original concept, further erasing the model's ability to segment the target. Furthermore, the Representation Preservation Module imposes explicit constraints on non-target regions to maintain the segmentation performance on unrelated concepts. Finally, to provide a thorough evaluation of MU on the segmentation task, we develop an evaluation schema including efficacy, generality, and specificity. Efficacy and generality assess the effectiveness of the unlearning methods, while specificity evaluates the utility of the segmentation foundation model post-unlearning. To validate the effectiveness of our method, we compare it against 5 representative MU methods. The experimental results reveal that our approach exceeds these comparative methods. We summarize the main contributions as follows:

- To the best of our knowledge, we are the first to investigate machine unlearning in segmentation foundation models.
- We propose a method consisting of Multi-level Forgetting Module and Preservation Module, where Multi-level Forgetting Module performs hierarchical unlearning at the distillation, attention, and output level, each targeting a different aspect of the unlearning process.
- We further design comprehensive evaluation metrics tailored for segmentation unlearning, which assess both forgetting effectiveness and performance retention. Experimental results on multiple baselines and segmentation backbones demonstrate our method consistently achieves the best trade-off between forgetting and preservation.

2 Related Work

2.1 Machine Unlearning

Machine unlearning enables selective removal of data from trained models to comply with privacy rights such as the "right to be forgotten" (Cao and Yang 2015), correct data errors (Goel et al. 2024), and mitigate bias (Steinhardt, Koh, and Liang 2017), without full retraining. Existing methods contain two main paradigms: exact unlearning (Bourtoule et al. 2021), which seeks models statistically equivalent to retraining from scratch via approaches such as SISA (Bourtoule et al. 2021) or differential privacy mechanisms (Gupta et al. 2021) but incurs substantial computational cost; and approximate unlearning (Golatkar, Achille, and Soatto 2020; Nguyen, Low, and Jaillet 2020), which improves efficiency using influence functions (Koh and Liang 2017), gradient-based updates (Neel, Roth, and Sharifi-Malvajerdi 2021), or knowledge distillation (Tarun et al. 2023), typically at the expense of formal guarantees of complete data removal (Chen et al. 2025).

2.2 Segmentation Foundation Models

The foundation model paradigm has revolutionized image segmentation, with Segment Anything Model (SAM) (Kirillov et al. 2023) as a cornerstone. Built on a Vision Transformer(Dosovitskiy et al. 2020), SAM provides remarkable zero-shot segmentation capabilities. Subsequent works focus on open-vocabulary recognition (Liang et al. 2023), improving efficiency (Zhao et al. 2023), and refining mask quality (Xie et al. 2024; Lin et al. 2025). However, the intersection of unlearning and large-scale models is underexplored. Their massive size makes exact unlearning methods computationally infeasible, while the effectiveness of approximate methods in such complex, high-dimensional parameter spaces is an open question. Our work addresses this crucial gap by proposing an efficient unlearning framework tailored for large segmentation models.

3 Problem Definition

Let $\mathcal{M}_T$ denote an image segmentation foundation model with parameters T, originally trained on a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{m}_i)\}_{i=1}^{N}$. $\mathbf{x}_i$ is an input image and $\mathbf{m}_i$ is the corresponding ground truth segmentation mask. We define the forgetting set $\mathcal{D}_f = \{(\mathbf{x}_j^C, \mathbf{m}_j^C)\}_{j=1}^{K}$ as a collection of K image-mask pairs associated with the targeted unlearning concepts C. Each image $\mathbf{x}_j^C$ contains the target concept C, and each $\mathbf{m}_j^C$ is the segmentation mask of the target concept C. To support the unlearning process and evaluate its effectiveness, we divide the forgetting set $\mathcal{D}_f$ into two disjoint subsets: a training subset $\mathcal{D}_f^{train}$ which contains a small number of samples used to train the unlearned model, and a testing subset $\mathcal{D}_f^{test}$, which is used to assess the generalization and completeness of the forgetting effect.

We define the goal of MU in image segmentation foundation models as follows:

> Machine unlearning in segmentation foundation models aims to eliminate the model's ability to segment a specific set of target concepts, while preserving its segmentation performance on all non-target contents.

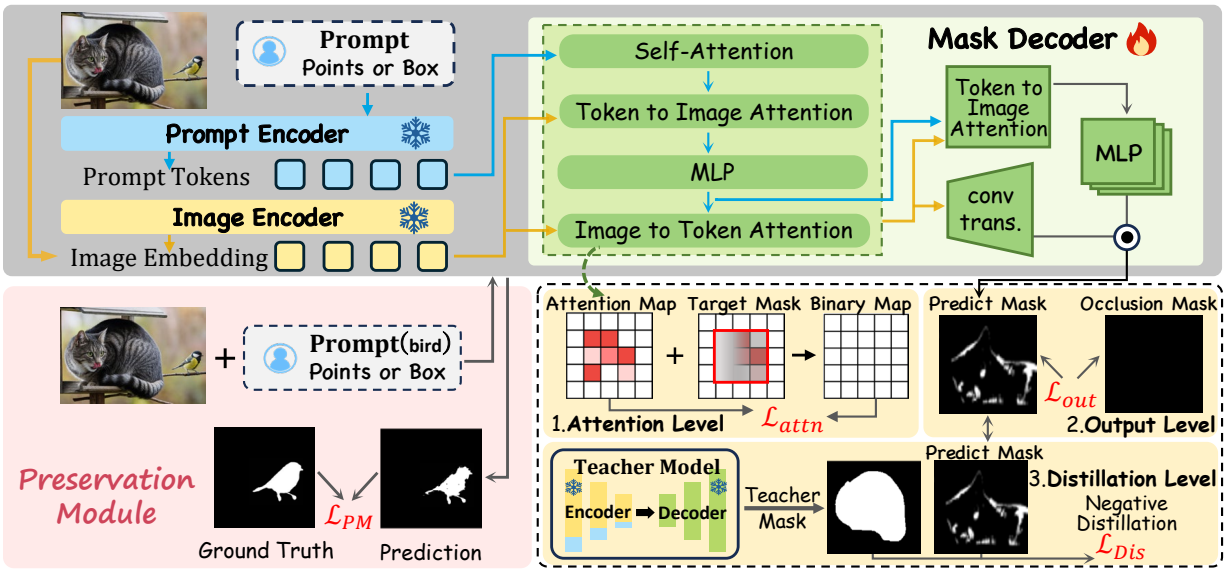

Figure 2: Overview of Unlearning Process in Segmentation Foundation Model using SCU. The yellow region shows Multi-level Forgetting Module ("cat" as target), while Preservation Module maintains segmentation for non-target concepts ("bird").

$$\arg\min_{T} \left\{ \mathbb{E}_{\left(\mathbf{x}_j^C, \mathbf{m}_j^C\right) \in \mathcal{D}_f} \left[\sum_{p \in \Omega\left(\mathbf{m}_j^C\right)} \hat{\mathbf{m}}_T\left(\mathbf{x}_j^C\right)[p] \right] + \lambda \mathbb{E}_{(\mathbf{x}_i, \mathbf{m}_i) \in \mathcal{D} \setminus \mathcal{D}_f} \mathcal{L}_{\text{seg}}\left(\hat{\mathbf{m}}_T\left(\mathbf{x}_i\right), \mathbf{m}_i\right) \right\} \tag{1}$$

4 Methodology

4.1 Overview

In this paper, we introduce a novel unlearning framework SCU designed to selectively forget specific concepts in pretrained segmentation models, without degrading the segmentation performance on other concepts. As illustrated in Fig. 2, our framework consists of several components: Multi-level Forgetting Module, and Preservation Module. Multi-level Forgetting Module implements hierarchical unlearning. At distillation level, negative distillation uses teacher model guiding the segmentation predictions for target concepts via a negative distillation loss, forcing them to diverge from the original segmentation, thus facilitating targeted unlearning. At attention level, attention suppression reduces the decoder's focus on the target concept, weakening the model's feature response in the target region. At output level, we alter the supervision for the target region, guiding model to suppress and eventually unlearn recognition of the original concept region at the output stage. The Preservation Module imposes segmentation constraints on non-target regions, helping prevent the model from performance degradation and maintaining model's segmentation effectiveness.

4.2 Multi-level Forgetting Module

Distillation-level: We introduce a negative distillation strategy at the distillation level to enable effective unlearning. Unlike conventional distillation, which encourages the student model to mimic the teacher, negative distillation explicitly drives the student's predictions for the target concept to diverge from those of the teacher model. This approach encourages the model to respond differently from the original model for the target concept, thereby achieving the desired forgetting effect.

We designate the original pretrained SAM models as the 'bad teacher' $\mathcal{M}_T$, which provides an initial segmentation baseline for target concepts. The goal is to make the student model (the unlearning SAM model) $\mathcal{M}_S$ to generate predictions that diverge from $\mathcal{M}_T$'s output, thereby effectively forgetting the target concept's segmentation capability. For a given input image $\mathbf{x}$ containing the target concept C, denote the $\mathcal{M}_T$'s output as $\mathbf{y}_T = \mathcal{M}_T(x)$ and the $\mathcal{M}_S$'s output as $\mathbf{y}_S = \mathcal{M}_S(x)$. To rigorously quantify the discrepancy between $\mathbf{y}_T$ and $\mathbf{y}_S$, we introduce a negative distillation loss that inverts the conventional MSE objective. Specifically, the negative distillation loss is defined as:

$$\mathcal{L}_{Dis} = -\lambda \,\text{MSE}\left(\mathbf{y}_T, \mathbf{y}_S\right) = -\lambda \frac{1}{N} \sum_{i=1}^{N} \left(y_{T,i} - y_{S,i}\right)^2 \tag{2}$$

where λ is a hyperparameter that scales the unlearning force and N is the number of elements in the output vectors. The negative sign reverses the typical distillation objective, minimizing $\mathcal{L}_{Dis}$ actually increases the discrepancy between $\mathbf{y}_T$ and $\mathbf{y}_S$. Furthermore, the gradient for the student output is computed as:

$$\nabla_{\mathbf{y}_S} \mathcal{L}_{Dis} = -\frac{2\lambda}{N} \left(\mathbf{y}_T - \mathbf{y}_S\right), \tag{3}$$

which explicitly drives the update in $\mathbf{y}_S$ away from $\mathbf{y}_T$, quantifies how a small change in $\mathbf{y}_S$ affects the negative distillation loss.

Attention-level: In transformer-based segmentation foundation models, the decoder tends to retain high attention on previously learned regions due to deep fitting during pretraining. This persistent attention bias often leads to continued activation of the target region even after label modification, thus hindering effective unlearning (Chefer, Gur, and Wolf 2021; Li et al. 2018). To address this issue, we introduce an attention strategy that explicitly reduces the attention weights assigned to the target region. By minimizing the attention from the mask token to image tokens associated with the target concept, the model is guided to reduce its internal reliance on target-related features, thereby promoting effective unlearning.

SAM's mask decoder receives a sequence of input tokens composed of a prompt embedding $\{p_i\}_{i=1}^{N_p}$ and a learnable mask token q. This sequence attends to the image patch embedding $\{\mathbf{v}_j\}_{j=1}^{M}$ through prompt-to-image cross-attention, where each query token (prompt or mask) retrieves information from the image tokens. The cross-attention matrix $\mathbf{A} \in \mathbb{R}^{(N_p+1)\times M}$ represents the attention from each input token(prompt tokens and a learnable mask token) to the image patch tokens. We extract the attention vector corresponding to the mask token and reshape it into a spatial map $\mathbf{A}^{\text{s}} \in \mathbb{R}^{H_p \times W_p}$, where each element indicates the attention assigned to a specific image patch region. We construct a binary mask $\mathbf{m}_{\text{attn}} \in \{0,1\}^{H_p \times W_p}$, where patches overlapping with the target region are assigned 0 and the others 1. This mask is generated by downsampling the original ground truth mask to the patch resolution (H_p, W_p). The attention-level suppression loss is then defined as:

$$\mathcal{L}_{\text{attn}} = \frac{1}{H_p W_p} \sum_{i=1}^{H_p} \sum_{j=1}^{W_p} \mathbf{m}_{\text{attn}}(i,j) \cdot \left(\mathbf{A}^{\text{s}}(i,j)\right)^2, \quad (4)$$

This formulation penalizes attention in spatial regions corresponding to the target concept and encourages the model to reallocate focus to non-target regions, thereby facilitating selective unlearning at the attention level.

Output-level: Although distillation and attention level reduce model's reliance on target information, they may not fully remove the model's ability to segment the target concept. Thus, an explicit output-level intervention is necessary to directly constrain segmentation results and ensure thorough erasure of the target from predictions. To address this, we introduce an output-level supervision mechanism. This approach directly constrains the predicted segmentation mask for the target region, guiding the model toward an occlusion pattern and further weakening its segmentation of the target concept. Let the ground truth mask for the target concept C in the input image be $\mathbf{m}_o$. Then construct an occlusion mask $\mathbf{m}_c$ of the same dimensions, where each element is defined as

$$\mathbf{m}_c(i) = 0, \quad \forall i \in \{1, \ldots, N\}, \quad (5)$$

where N represents the total number of pixels in the region corresponding to C.

The unlearning (student) model $\mathcal{M}_S$ produces a predicted segmentation mask $\mathbf{m_s}$ for the target concept. We enforce the output align with $\mathbf{m_c}$ via a loss function based on binary cross-entropy between $\mathbf{m_s}$ and $\mathbf{m_c}$:

$$\begin{aligned}\mathcal{L}_{\text{out}} = -\frac{1}{N} \sum_{i=1}^{N} &\left[\mathbf{m}_c(i) \log\left(\mathbf{m}_s(i)\right)\right. \\ &\left. + \left(1 - \mathbf{m}_c(i)\right) \log\left(1 - \mathbf{m}_s(i)\right)\right],\end{aligned} \quad (6)$$

This loss penalizes any positive prediction in the masked region, despite the presence of a prompt, thereby forcing the model to unlearn association with the target concept.

4.3 Preservation Module

To preserve segmentation performance on non-target objects during selective unlearning, we introduce the Preservation Module (PM). The main challenge arises from the shared network backbone, which encodes features common to both target and non-target regions. Suppressing target-specific features may inadvertently disturb the representations of non-target concepts, leading to degraded segmentation. To address this, the PM enforces feature invariance in non-target regions using a dual loss strategy. Given an input image $\mathbf{x}$, the ground truth mask $\mathbf{m}_{gt}^{non}$ for non-target objects, along with the corresponding prediction $\mathbf{m}_{pred}^{non}$ is used to compute an output-level preservation loss. In the non-target regions, BCE loss is computed as:

$$\begin{aligned}\mathcal{L}_{BCE}^{non} = -\frac{1}{N} \sum_{i=1}^{N} &\left[\mathbf{m}_{gt}^{non}(i) \log\left(\mathbf{m}_{pred}^{non}(i)\right)\right. \\ &\left. + \left(1 - \mathbf{m}_{gt}^{non}(i)\right) \log\left(1 - \mathbf{m}_{pred}^{non}(i)\right)\right],\end{aligned} \quad (7)$$

Alongside $\mathcal{L}_{BCE}^{non}$ for pixel-level precision, $\mathcal{L}_{Dice}^{non}$ enhances region-level consistency and class imbalance by maximizing the overlap between predicted and true masks. $\mathcal{L}_{Dice}^{non}$ is defined as:

$$\mathcal{L}_{\text{Dice}}^{\text{non}} = 1 - \frac{2\sum_{i=1}^{N} \mathbf{m}_{gt}^{\text{non}}(i)\mathbf{m}_{\text{pred}}^{\text{non}}(i) + \epsilon}{\sum_{i=1}^{N} \mathbf{m}_{gt}^{\text{non}}(i) + \sum_{i=1}^{N} \mathbf{m}_{\text{pred}}^{\text{non}}(i) + \epsilon}, \quad (8)$$

where ϵ is a small constant to ensure numerical stability.

Thus, the combined loss for maintaining accurate segmentation on non-target regions in PM is expressed as:

$$\mathcal{L}_{\text{PM}} = \mathcal{L}_{BCE}^{\text{non}} + \mathcal{L}_{\text{Dice}}^{\text{non}} \quad (9)$$

5 Experiments

5.1 Experimental Settings

Baselines. In our experiments, we compare our method with five representative unlearning approaches. These include three established baselines: Random Labeling (RL) (Golatkar, Achille, and Soatto 2020), Gradient Ascent (GA) (Thudi et al. 2022), and L1-Sparse (L1-Sparse) (Jia et al. 2023). Additionally, we introduce two new boundary-based baselines derived from decision boundary manipulation

Model	Method	Efficacy (↓)		Generality (↓)		Specificity (↑)		Normalized Score (↑)
		IOU	AP	IOU	AP	IOU	AP	
SAM1	Gradient Ascent	18.6	21.8	22.8	25.4	16.1	18.2	0.637
	Boundary Shrink	68.0	72.6	63.5	66.1	**75.5**	**76.4**	0.333
	Random Labeling	63.4	68.5	57.5	59.2	71.5	72.4	0.389
	Boundary Expand	64.1	65.4	57.9	60.7	71.3	73.9	0.393
	L1-sparse	21.0	26.9	24.1	29.5	37.4	42.9	0.723
	SCU(ours)	**16.3**	**18.8**	**21.5**	**23.4**	61.3	63.5	**0.923**
SAM2	Gradient Ascent	23.2	25.1	27.3	30.6	23.9	26.7	0.617
	Boundary Shrink	72.8	73.0	65.8	69.6	**78.4**	**84.9**	0.347
	Random Labeling	69.5	71.8	67.5	70.1	76.6	79.8	0.332
	Boundary Expand	71.9	74.5	64.7	67.3	76.5	81.9	0.343
	L1-sparse	25.3	28.1	27.4	31.5	40.9	46.8	0.706
	SCU(ours)	**19.4**	**21.0**	**24.6**	**26.7**	68.6	70.6	**0.929**

Table 1: Comparison with existing machine unlearning methods. Complete per-concept results are in the appendix.

techniques (Chen et al. 2023), namely Boundary Shrink (BS) and Boundary Expand (BE).

Model and Training. We conduct our experiments on two widely adopted versions of the Segment Anything Model: SAM1 (ViT-L) and SAM2 (sam2_hiera_large). During training, the image encoder is frozen, and only the mask decoder is trained. All models are trained for 50 steps using the proposed unlearning framework. We adopt the AdamW optimizer with a learning rate 1e-4, a batch size of 4. All experiments are conducted on NVIDIA A100 GPUs (40 GB).

Dataset. To evaluate machine unlearning within segmentation foundation models, we curate a dataset by collecting image-mask pairs from the PASCAL VOC and COCO segmentation datasets. The PASCAL VOC dataset contains annotations for 20 object categories, from which we select 8 representative classes: aeroplane, monitor, bicycle, bottle, cat, sheep, bird, and chair, as target concepts for unlearning experiments. Due to limited number of annotated concepts per class in VOC, we supplement each category with additional samples from the COCO dataset. To enable effective unlearning, we filter and retain only those samples where target concepts can be successfully segmented by SAM. For each selected category, we collect 15 image-mask pairs as the training subset seen during unlearning. The remaining images containing the same concept, drawn from VOC and COCO, form the unseen testing subset $\mathcal{D}_f^{test}$, which is used to evaluate the generalization of forgetting. Across 8 categories, the test set contains 773 images, yielding a total of 893 segmentation pairs used in experiments.

5.2 Evaluation Metrics

To comprehensively evaluate the forgetting behavior of segmentation foundation models, we adopt four metrics from different dimensions:

Efficacy examines how effectively the unlearned model $\mathcal{M}_S$ has unlearned the seen examples. This metric is evaluated on forgotten training set $\mathcal{D}_f^{train}$ and unlearned model's predictions are consistent unlearning objective.

Generality assesses generalization of forgetting unseen instances of target concept. We evaluate unlearning student model $\mathcal{M}_S$ on $\mathcal{D}_f^{test}$, which contains diverse samples of target concept not encountered during unlearning.

Specificity measures the preservation of segmentation performance on non-target content.

All three evaluation dimensions above are quantified using two metrics: Intersection-over-Union (IoU) and Average Precision (AP). In our setting, AP is computed under a single IoU threshold of 0.3. Specifically, for each prediction, we consider it correct if its IoU with the ground truth mask exceeds 0.3. The final AP score reflects the proportion of such correct predictions across the dataset, thereby evaluating the model's ability to localize and delineate the target concept under a relaxed matching criterion.

Normalized Score balances forgetting efficacy, generalization, and specificity. Since lower values are preferable for Efficacy and Generality, we first reverse their direction via $x' = x_{\max} - x$, and then apply Min-Max normalization to scale all metrics to $[0, 1]$. The final score is obtained by computing the average of the three normalized values, offering a holistic view of each method's overall performance.

Method	Effi. (↓)		Gene. (↓)		Spec. (↑)		Normalize Score (↑)
	IOU	AP	IOU	AP	IOU	AP	
w/o PM	17.6	28.2	**17.1**	**23.2**	21.4	29.5	0.538
w/o Attn	26.5	25.4	26.7	27.1	36.5	38.7	0.421
w/o Dis	32.8	32.4	32.5	33.1	36.8	38.5	0.108
w/o Out	27.9	28.7	30.2	32.6	32.5	33.8	0.379
SCU(ours)	**16.3**	**18.8**	21.5	23.4	**61.3**	**63.5**	**0.949**

Table 2: Ablation study of different modules.

5.3 Main Results

Table 1 presents a comprehensive comparison of our SCU method with multiple machine unlearning baselines on both SAM1 and SAM2. The main observations are as follows:

Superior Unlearning Performance: SCU consistently achieves the best unlearning performance in terms of both Efficacy and Generality, where lower scores are better. On SAM2, SCU achieves an Efficacy IOU of 19.4, and 16.3 on SAM1, significantly outperforming weak baselines like

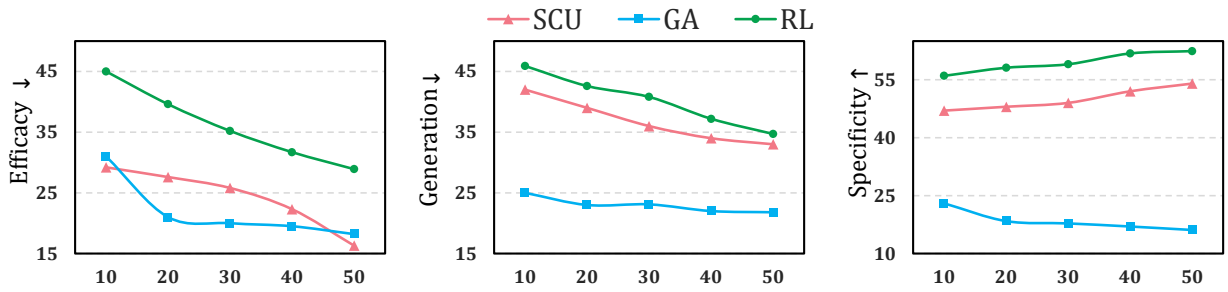

Figure 3: Results of unlearning 8 objects simultaneously using SAM1. The height of the bar chart denotes the mean of the two metrics(IOU and AP)

BE (74.5) and BS (73.0), which largely fail to unlearn the concept. This indicates that SCU most effectively unlearns model's knowledge of the target concept and reliably generalizes forgetting to unseen samples.

Optimal balance between Unlearning and utility: An ideal unlearning method should forget the target concept while maintaining performance on non-target concepts, as reflected by high Specificity. Baselines show a clear trade-off between forgetting and model utility: conservative methods such as RL, BE, BS maintain high Specificity but fail to forget the target concept (BE's Efficacy AP remains at 74.5 on SAM2, with Normalized Score at 0.343, making its high Specificity meaningless).In contrast, methods like GA and L1-sparse achieve moderate forgetting but cause a drastic drop in specificity (GA's Specificity AP dropping to 18.2 on SAM1 and 26.7 on SAM2).In comparison, SCU maintains high Specificity (AP: 63.5 on SAM1, 70.6 on SAM2,and the highest Normalized Score) while effectively erasing the target concept, achieving unlearning without compromising overall segmentation performance.

Robustness and scalability: Comparing SAM1 and SAM2 demonstrates each method's stability under architectural upgrades. Many baselines, like BE, show degraded unlearning on more powerful models (Efficacy AP increases from 65.4 to 74.5). In contrast, SCU maintains consistently high performance, with its Normalized Score improving from 0.923 to 0.929 and Efficacy, Generality, and Specificity nearly the same, highlighting SCU's robustness and adaptability to advanced model architectures.

In summary, SCU achieves thorough unlearning while maximally preserving general segmentation capabilities, showing remarkable stability on advanced architectures. The highest Normalized Scores (0.923 and 0.929) show effectiveness and superiority of SCU for selective concept unlearning in segmentation foundation models.

5.4 Ablation Study

To evaluate the contribution of each component, we perform ablation studies by removing individual levels of multi-level forgetting module and preservation module. All experiments are conducted on SAM1, and results are shown in Table 2. For direct comparison, all metrics use IOU as the representative measure. For each ablation, all other modules are retained to isolate the effect of the removed component.

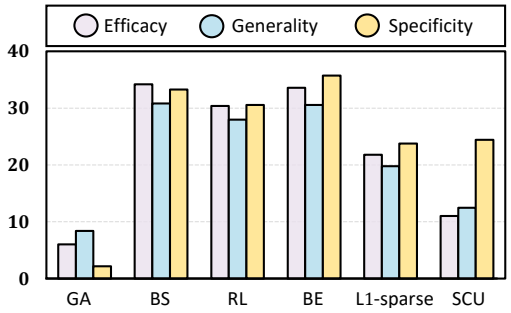

Figure 4: Results of unlearning 8 objects simultaneously using SAM. The height of the bar chart denotes the mean of the two metrics(IOU and AP)

Impact of Distillation-Level: Ablating Distillation-Level leads to a significant reduction in forgetting efficacy, with Efficacy increasing from 16.3 to 32.8. Generality and Specificity also decline, indicating that Distillation-Level is crucial for suppressing original segmentation behaviors and achieving effective unlearning.
Impact of Attention-Level: Removing the Attention-Level weakens forgetting and generalization (Efficacy and Generality rise to 26.5 and 26.7), underscoring its role in reducing the model's sensitivity to the target concept.
Impact of Output-Level: Excluding the Output-Level decreases specificity (IOU drops to 32.5), highlighting its importance for maintaining non-target segmentation accuracy.
Impact of Preservation Module: Removing the Preservation Module causes Specificity to drop sharply (61.3 to 21.4), highlighting its crucial role in maintaining non-target segmentation performance during unlearning.
Overall, the full method achieves best performance, demonstrating its effectiveness and reliability for unlearning.

5.5 Impacts of Training Steps

In this section, we analyze impact of training steps as shown in Figure 3. SCU remains stable across all metrics as number of training steps increases, indicating low sensitivity to training length. By contrast, GA and RL suffer a substantial drop in specificity with more steps, reflecting increasing disruption to non-target segmentation; their generality and efficacy also degrade over time, revealing unstable unlearn-

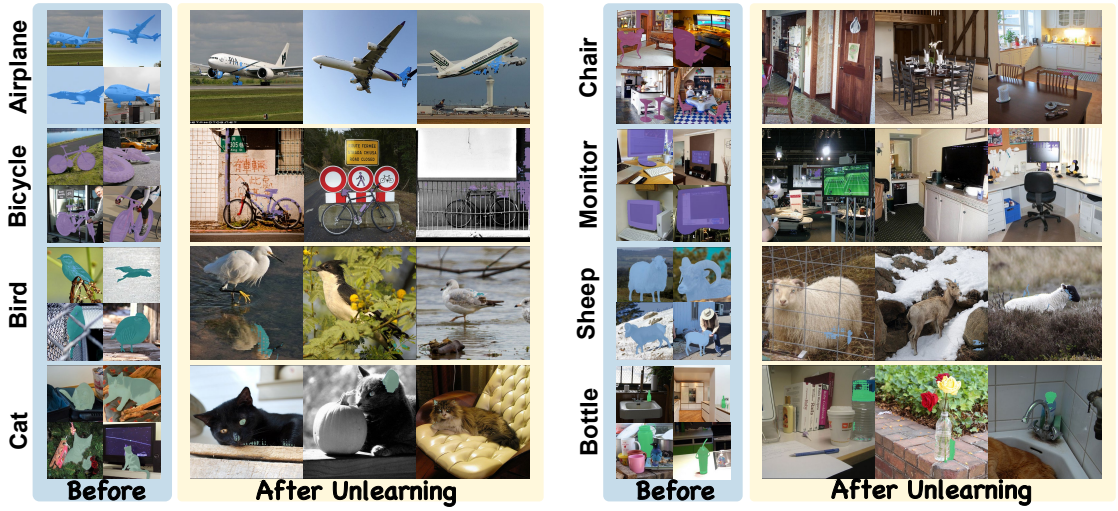

Figure 5: Results of forgetting 8 types of concepts using SCU on SAM. For each concept, the first 2×2 grid displays the original segmentation results from the pretrained SAM. The following three images show the segmentation outcomes after the concept has been selectively forgotten by SCU.

ing behavior and no performance gain from longer training.

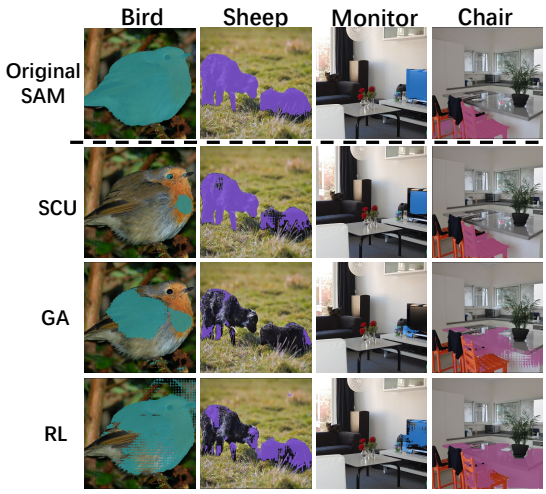

Figure 6: Visualization results of unlearning 'Bird' in SAM.

5.6 Results of Unlearning Multiple Objects

Figure 4 presents results from unlearning all 8 objects synchronously. We concatenate all the forgetting sets of these objects as fine-tuning data and train the model for 50 epochs. Results show that fine-tuning with the GA method severely degrades the SAM model's performance, rendering it largely unable to segment objects. Models fine-tuned with BS, RL, and BE methods exhibit reduced performance across all metrics but fail to achieve the desired unlearning effect. In contrast, SCU enables SAM to maintain robust retention performance while unlearning the target concepts.

5.7 Case Study on Unlearning

Figure 5 shows qualitative results of forgetting eight concept categories using SCU on SAM1. After unlearning, SAM fails to segment target concepts, with masks disappearing or reassigned to background regions. For Airplane and Monitor, masks are nearly absent, indicating successful forgetting, while for Bird and Cat—often blending with background—the model also loses attention to relevant regions. SCU shows particularly strong forgetting on small or less salient objects, likely due to weaker contextual cues and fewer supporting features, which makes these categories more vulnerable to targeted suppression. Overall, these results demonstrate that SCU effectively suppresses model's ability to segment forgotten concepts in visual segmentation.

Figure 6 illustrates unlearning of 'Bird' concept in SAM. GA and RL fail to forget the target concept. In particular, RL retains almost entire bird region with accurate boundaries, indicating minimal forgetting, while GA suppresses the bird but severely degrades segmentation of non-target concepts. In contrast, SCU produces nearly no segmentation for the bird, while simultaneously preserving the segmentation quality for unrelated concepts. This illustrates it's ability to strike a balance between forgetting target concepts and retaining segmentation utility for non-target categories.

6 Conclusion

We propose SCU, an efficient selective concept unlearning method for segmentation foundation models. With a Multi-level forgetting and a preservation module, SCU erases target concepts while maintaining non-target segmentation quality. We further introduce new unlearning metrics, and experiments demonstrate that SCU achieves state-of-the-art forgetting and preservation performance.

Acknowledgments

This work is partially supported by National Nature Science Foundation of China under No. U21A20488. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine Unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, 141–159. IEEE.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer Interpretability Beyond Attention Visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 782–791. Computer Vision Foundation / IEEE.

Chen, A.; Li, Y.; Zhao, C.; and Huai, M. 2025. A survey of security and privacy issues of machine unlearning. *AI Mag.*, 46(1).

Chen, J.; and Yang, D. 2023. Unlearn What You Want to Forget: Efficient Unlearning for LLMs. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 12041–12052. Association for Computational Linguistics.

Chen, M.; Gao, W.; Liu, G.; Peng, K.; and Wang, C. 2023. Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 7766–7775. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Ginart, A.; Guan, M.; Valiant, G.; and Zou, J. Y. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Goel, S.; Prabhu, A.; Torr, P.; Kumaraguru, P.; and Sanyal, A. 2024. Corrective Machine Unlearning. *CoRR*, abs/2402.14015.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 9301–9309. Computer Vision Foundation / IEEE.

Gupta, V.; Jung, C.; Neel, S.; Roth, A.; Sharifi-Malvajerdi, S.; and Waites, C. 2021. Adaptive Machine Unlearning. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 16319–16330.

Hu, X.; Li, D.; Hu, B.; Zheng, Z.; Liu, Z.; and Zhang, M. 2024. Separate the Wheat from the Chaff: Model Deficiency Unlearning via Parameter-Efficient Module Operation. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 18252–18260. AAAI Press.

Huang, Z.; Jing, H.; Liu, Y.; Yang, X.; Wang, Z.; Liu, X.; Gao, K.; and Luo, H. 2024. Segment Anything Model Combined with Multi-Scale Segmentation for Extracting Complex Cultivated Land Parcels in High-Resolution Remote Sensing Images. *Remote Sensing*, 16(18): 3489.

Jia, J.; Liu, J.; Ram, P.; Yao, Y.; Liu, G.; Liu, Y.; Sharma, P.; and Liu, S. 2023. Model Sparsification Can Simplify Machine Unlearning. *CoRR*, abs/2304.04934.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 3992–4003. IEEE.

Koh, P. W.; and Liang, P. 2017. Understanding Black-box Predictions via Influence Functions. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. PMLR.

Li, J.; Wei, Q.; Zhang, C.; Qi, G.; Du, M.; Chen, Y.; Bi, S.; and Liu, F. 2024. Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Li, K.; Wu, Z.; Peng, K.; Ernst, J.; and Fu, Y. 2018. Tell Me Where to Look: Guided Attention Inference Network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 9215–9223. Computer Vision Foundation / IEEE Computer Society.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7061–7070.

Lin, Y.; Li, H.; Shao, W.; Yang, Z.; Zhao, J.; He, X.; Luo, P.; and Zhang, K. 2025. SAMRefiner: Taming Segment Anything Model for Universal Mask Refinement. *arXiv preprint arXiv:2502.06756*.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W. 2024. MACE: Mass Concept Erasure in Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 6430–6440. IEEE.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

Neel, S.; Roth, A.; and Sharifi-Malvajerdi, S. 2021. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning. In Feldman, V.; Ligett, K.; and Sabato, S., eds., *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, 931–962. PMLR.

Nguyen, Q. P.; Low, B. K. H.; and Jaillet, P. 2020. Variational Bayesian Unlearning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Osco, L. P.; Wu, Q.; De Lemos, E. L.; Gonçalves, W. N.; Ramos, A. P. M.; Li, J.; and Junior, J. M. 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation*, 124: 103540.

Shan, X.; and Zhang, C. 2023. Robustness of segment anything model (sam) for autonomous driving in adverse weather conditions. *arXiv preprint arXiv:2306.13290*.

Shintre, S.; Roundy, K. A.; and Dhaliwal, J. 2019. Making machine learning forget. In *Privacy Technologies and Policy: 7th Annual Privacy Forum, APF 2019, Rome, Italy, June 13–14, 2019, Proceedings 7*, 72–83. Springer.

Steinhardt, J.; Koh, P. W.; and Liang, P. 2017. Certified Defenses for Data Poisoning Attacks. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 3517–3529.

Tarun, A. K.; Chundawat, V. S.; Mandal, M.; and Kankanhalli, M. 2023. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling SGD: Understanding Factors Influencing Machine Unlearning. In *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*, 303–319. IEEE.

Wang, A.; Islam, M.; Xu, M.; Zhang, Y.; and Ren, H. 2023. Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–244. Springer.

Wu, J.; Wang, Z.; Hong, M.; Ji, W.; Fu, H.; Xu, Y.; Xu, M.; and Jin, Y. 2025. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 103547.

Xie, Z.; Guan, B.; Jiang, W.; Yi, M.; Ding, Y.; Lu, H.; and Zhang, L. 2024. Pa-sam: Prompt adapter sam for high-quality image segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, 1755–1764. IEEE.

Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; and Wang, J. 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156*.