# BREAKING THE PRESENCE GATE: UNIVERSAL ADVERSARIAL DISRUPTION OF SAM3 CONCEPT SEGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

SAM3 introduces open-vocabulary concept segmentation. Given a text query, it determine whether the concept is present and segment every matching instance. Its final candidate scores are multiplied by a shared presence probability, creating a single decision point through which many masks can be suppressed. We study the adversarial robustness of this design and propose Universal Concept Disruption (UCD), a single image-space perturbation trained across images and noun phrases. UCD attacks SAM3 at three coupled levels: it disrupts the text-conditioned visual representation, maximizes divergence in prompt-shared features, and jointly suppresses presence-gated scores and the spatial quality of surviving masks. Under a common universal-perturbation protocol, UCD is evaluated on SACo-Gold, LVIS, RefCOCO, PhraseCut, and OpenImages. With an $\ell_\infty$ budget of $8/255$, it reduces mean mask AP from 59.43 to 18.73 and mean concept-grounding $F_1$ from 50.32 to 20.49, consistently exceeding adapted attacks designed for SAM, SAM2, or generic classifiers. The same perturbation transfers to SAM3.1 and to video inference without re-optimization, exposing failures that persist through temporal memory. Prompt ensembling, head fine-tuning, and temporal filtering offer only partial recovery and can substantially reduce clean performance.

## 1 INTRODUCTION

Driven by their exceptional zero-shot generalization capabilities, the Segment Anything Model (SAM) series has emerged as a crucial bridge between visual foundation models and diverse downstream perception tasks. Specifically, SAM (Kirillov et al., 2023) provides a robust framework for prompt-driven image segmentation, and SAM2 (Ravi et al., 2025) advances this paradigm by incorporating temporal awareness for seamless video segmentation and object tracking. The recently released SAM3 (Carion et al., 2025) changes the operating point. It introduces open-vocabulary concept segmentation, and the model is expected to determine whether the concept exists and segment all matching instances.

Adversarial attacks introduce visually imperceptible perturbations to images, causing the model to produce incorrect predictions and fail at its intended task, thereby achieving the attack objective. The vulnerability of the SAM family to such attacks has been experimentally validated (Long et al., 2025; Qiao et al., 2023). Specifically, methods such as DarkSAM (Zhou et al., 2024) and Attack-SAM (Zhang et al., 2023) are designed to target the original SAM. For SAM2, UAP-SAM2 (Zhou et al., 2026) utilizes universal adversarial perturbations; in scenarios where the initial $k$ frames of a video are continuously attacked, it drastically degrades SAM2's video segmentation performance. However, the robustness of SAM3 against adversarial attacks remains unexplored.

We first verify whether existing attacks can be directly transferred to SAM3 image segmentation. To make the comparison fair, we adapt representative SAM, SAM2, and general adversarial baselines into the same universal-perturbation (Moosavi-Dezfooli et al., 2017) setting, using the same train-test separation, random seeds, optimization budget, and evaluation pipeline. Across five datasets, attacks that do not directly target SAM3's concept-scoring mechanism produce small or inconsistent degradation, while even the strongest adapted baseline remains below our SAM3-specific attack. Existing universal attacks against SAM aim to make the model segment nothing or segment incorrectly



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.

across prompts. SAM2 attacks further account for video memory and cross-frame prompt reuse. However, SAM3 differs significantly from SAM1 and SAM2, being specifically tailored for concept segmentation. To enable its concept segmentation capabilities, SAM3 features a highly integrated architecture design on both the input and output sides. A failure is no longer determined only by whether a mask is geometrically distorted, it can also occur when the queried concept is absent and suppresses otherwise plausible candidates.

On the input side, the original SAM conducted a proof-of-concept for text prompts by directly extracting embeddings via a frozen CLIP (Radford et al., 2021) text encoder. However, its performance was highly suboptimal, which led to subsequent decoupled designs that relied on cascading external CLIP-based detectors to provide spatial prompts. In contrast, SAM3 integrates Perception Encoder (PE)-style (Bolya et al., 2026) vision-language representations into a unified text-conditioned grounding architecture and is trained end-to-end on the SA-Co (Carion et al., 2025) dataset. On the output side, SAM3 introduces a presence token to address the challenges of open-vocabulary segmentation. This mechanism incorporates a presence token for concept existence and computes a final candidate score that combines query confidence with presence confidence. Functioning as an existence switch, the presence token is globally multiplied by the scores of all individual masks. While this design enhances the robustness of open-vocabulary segmentation, it inadvertently exposes a novel and critical universal attack surface.

Motivated by this observation, we propose Universal Concept Disruption (UCD), the first cross-concept universal adversarial attack tailored to SAM3 image segmentation. UCD treats SAM3 as a concept-grounding system rather than a promptable mask decoder, and disrupts it at three coupled levels. At the input level, UCD learns a single bounded perturbation from (image, noun-phrase (NP)) pairs, thereby activating the same text-conditioned concept path used at inference instead of relying on geometric prompts. At the feature level, UCD perturbs the prompt-shared visual representation, turning the common image backbone into a transferable attack carrier across concepts, datasets, and model variants. At the output level, UCD targets SAM3's presence-gated concept decision together with the spatial validity of retained masks, causing present concepts to disappear or become geometrically unreliable. We conduct extensive experiments on five datasets, showcasing superior attack efficacy of UCD. Our contributions are threefold:

- We propose the first universal cross-concept adversarial attack, termed UCD, specifically targeting the open-vocabulary image segmentation capabilities of SAM3.
- We develop a SAM3-specific concept disruption framework that couples text-conditioned input optimization, prompt-shared feature disruption, and presence output corruption.
- We comprehensively evaluate our method across five datasets under a unified universal-perturbation protocol. The results demonstrate that our approach consistently surpasses existing attacks.

## 2 METHODOLOGY

### 2.1 TASK DEFINITION AND OVERVIEW

**SAM3 Promptable Concept Segmentation.** We focus on the single-image, text-prompted setting. Let $x \in [0,1]^{3\times H\times W}$ be an image and let $t$ be a text prompt. SAM3 encodes the image and text, then runs a grounding stage that returns a set of candidate boxes, masks, query logits, and a prompt-level presence logit. We denote the image backbone by $E_\theta^{\text{img}}$, the text encoder by $E_\theta^{\text{text}}$, and the grounding head by $G_\theta$. The simplified image inference path is

$$F = E_\theta^{\text{img}}(x), \quad q = E_\theta^{\text{text}}(t), \quad (O, p) = G_\theta(F, q), \tag{1}$$

where $O = \{(\ell_i, b_i, m_i)\}_{i=1}^N$ and $p$ is the prompt-level presence logit. Here $\ell_i$ is the raw query logit for candidate $i$, $b_i$ is its box, and $m_i$ is its mask logit map. For each candidate, the final concept score is the product of the query probability and the presence probability:

$$s_i(x,t) = \sigma(\ell_i(x,t)) \cdot \sigma(p(x,t)). \tag{2}$$

Post-processing keeps candidates whose $s_i$ exceeds a threshold $\tau$ and converts retained mask logits into binary masks. A candidate with high mask quality can still be discarded if the model predicts the concept as absent, because the presence probability globally gates all candidate scores.

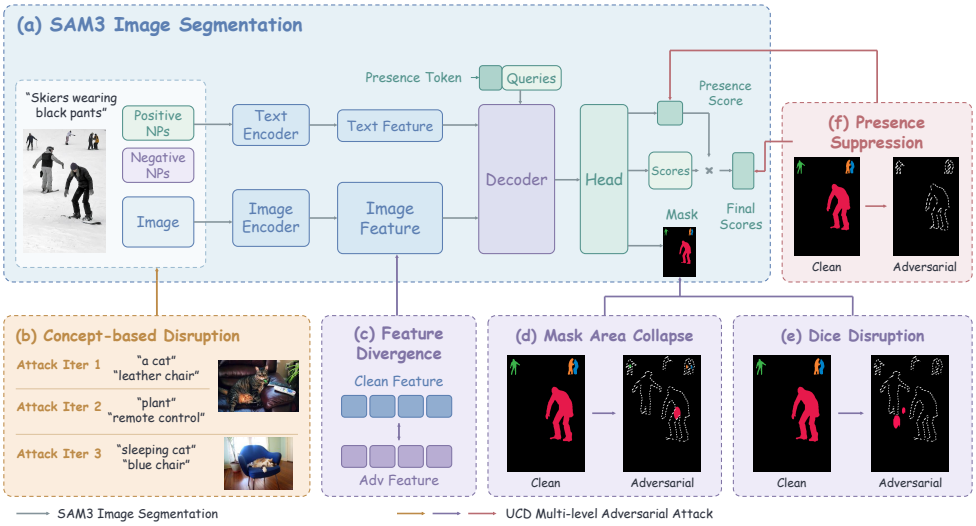



Figure 1: Universal Concept Disruption attacks SAM3 at input, feature, and output levels while optimizing a single image-space perturbation shared across prompts.

**Universal Concept Attack against Images.** Targeting images, we attack SAM3 with a single additive perturbation $\delta$ shared across all images and prompts:

$$x^{\text{adv}} = \Pi_{[0,1]}\left(x + \delta\right), \quad \|\delta\|_\infty \le \epsilon, \tag{3}$$

where $\epsilon$ denotes the bound on the perturbation magnitude. The attacker has white-box access to the released SAM3 model and optimizes $\delta$ over public training samples.

We consider an untargeted objective. In SAM3, concept segmentation prompts are categorized into positive prompts (where the concept is present in the image) and negative prompts (where the concept is absent from the image). For positive prompts, the attack aims to remove or corrupt the correctly predicted concept masks. For negative prompts, however, applying the same loss function as that used for positive prompts is ill-posed.

For a training set $\mathcal{D} = \{(x_j, t_j)\}_{j=1}^{M}$ of positive prompts, the ideal universal attack objective can be formulated as minimizing the retained clean-prediction utility:

$$\min_{\|\delta\|_\infty \le \epsilon} \mathbb{E}_{(x,t)\sim\mathcal{D}} \left[\mathcal{U}\left(f_\theta(x^{\text{adv}}, t), f_\theta(x, t)\right)\right], \tag{4}$$

where $\mathcal{U}$ measures the remaining agreement or utility of the adversarial prediction relative to the clean prediction. Minimizing $\mathcal{U}$ induces adversarial failure under the clipped input $x^{\text{adv}}$ from Eq. (3).

### 2.2 Universal Concept Disruption

In SAM3, a concept query passes through a hierarchy of decisions:

- The image must produce features that match the NPs.
- The grounding head must assign high candidate scores under the global presence gate.
- Retained candidates must have useful mask geometry.

A perturbation may disturb low-level image features without changing the final concept decision. It may suppress a score while leaving a surviving mask spatially correct, or it may alter a mask while the presence-gated score remains high enough for post-processing.

As shown in Figure 1, the attack has three architectural levels. First, at the input-concept level, UCD trains on actual positive image-NP pairs rather than geometric prompt grids. This activates SAM3's text-conditioned grounding path during optimization and avoids reducing the task to

prompt-agnostic mask failure. Second, at the feature level, UCD maximizes the discrepancy between clean and adversarial image features. Because the visual backbone is shared across text prompts, this term encourages transfer beyond the specific concepts observed during training. Third, at the output level, UCD directly attacks the presence-gated concept decision and the geometry of the retained masks. The score term targets the final score in Eq. (2), while the mask-area and clean-mask Dice terms make high-score survivors spatially useless. This hierarchy focuses UCD on the disappearance and spatial corruption of present concepts. Activating absent concepts is a complementary false-positive threat model, but it is not the objective studied in this paper.

For every sampled image, we first compute clean image features and clean top masks without gradient. We then apply $\delta$, run the same text prompts through SAM3, and back-propagate through the image backbone and grounding head into $\delta$. Let $\mathcal{X}_+$ be the set of training images and let $\mathcal{T}_+(x)$ be the positive noun phrases associated with image $x$. The full UCD objective is

$$\delta^\star = \arg\min_{\|\delta\|_\infty \le \epsilon} \mathbb{E}_{x\sim\mathcal{X}_+}\Bigg[\frac{1}{|\mathcal{T}_+(x)|}\sum_{t\in\mathcal{T}_+(x)}\Big(\lambda_s\mathcal{L}_{\text{score}}(x,t,\delta)+\lambda_a\mathcal{L}_{\text{area}}(x,t,\delta) + \lambda_d\mathcal{L}_{\text{dice}}(x,t,\delta)\Big) - \lambda_f\mathcal{D}_{\text{feat}}(x,\delta)\Bigg]. \tag{5}$$

We denote the expected objective being minimized in Eq. (5) by $\mathcal{J}(\delta)$. The three prompt-conditioned losses are averaged over the positive noun phrases for each image, while the prompt-shared feature term is applied once per image.

**Presence Suppression.** Let $S(x,t) = \{s_i(x,t)\}_{i=1}^N$ be the final SAM3 concept scores from Eq. (2). With $\mathcal{I}_s = \text{TopK}(S(x^{\text{adv}},t),k)$, we penalize high scores through

$$\mathcal{L}_{\text{score}}(x,t,\delta) = \frac{1}{k}\sum_{i\in\mathcal{I}_s}[s_i(x^{\text{adv}},t)-m_s]_+. \tag{6}$$

Here $[z]_+ = \max(z,0)$ and $m_s$ is a score margin. This term attacks the post-processing threshold rather than an intermediate query logit alone. As a result, it simultaneously accounts for the model's concept existence estimate and candidate confidence.

**Mask Area Collapse.** Score suppression alone can lower confidence while leaving plausible spatial masks. To attack segmentation quality, we add a mask-area loss on the top-scoring candidates. Let $M_i(x,t) = \sigma(m_i(x,t))$ be the soft mask probability for candidate $i$. With $\mathcal{I}_a = \text{TopK}(S(x^{\text{adv}},t),k_a)$, where the selected scores are treated as fixed for gradient computation, we minimize

$$\mathcal{L}_{\text{area}}(x,t,\delta) = \frac{1}{k_a}\sum_{i\in\mathcal{I}_a}\frac{1}{HW}\sum_{u,v}M_i^{uv}(x^{\text{adv}},t). \tag{7}$$

This term encourages high-confidence masks to shrink, which is useful when a candidate remains above threshold but no longer covers the target concept.

**Dice Disruption.** The area loss can over-favor tiny masks but does not explicitly push masks away from the clean spatial support of all high-score instances. Therefore, we store a clean high-score mask set

$$\mathcal{I}_c = \text{TopK}(S(x,t),k_d), \quad \mathcal{M}_c = \{M_j(x,t) : j\in\mathcal{I}_c\}, \tag{8}$$

and minimize the best Dice overlap between each top adversarial mask and this clean set. Let $\mathcal{I}_d = \text{TopK}(S(x^{\text{adv}},t),k_d)$ and

$$\text{Dice}(A,B) = \frac{2\langle A,B\rangle+\eta}{\|A\|_1+\|B\|_1+\eta}. \tag{9}$$

The Dice disruption loss is

$$\mathcal{L}_{\text{dice}}(x,t,\delta) = \frac{1}{k_d}\sum_{i\in\mathcal{I}_d}\max_{j\in\mathcal{I}_c}\text{Dice}\big(M_i(x^{\text{adv}},t),M_j(x,t)\big). \tag{10}$$

Although Eq. (10) is written as a loss to minimize, each term is the strongest Dice overlap with any clean high-score instance. Minimizing it lowers spatial agreement with the clean multi-instance prediction. The clean masks are detached so gradients update only the universal perturbation.

**Feature Divergence.** Text prompts vary widely across concept segmentation benchmarks. To reduce overfitting to individual prompts, we also attack the image representation shared across prompts. Let $F_\ell(x)$ denote the last feature map in SAM3's image-backbone FPN. We maximize normalized feature divergence:

$$\mathcal{D}_{\text{feat}}(x, \delta) = \frac{\|F_\ell(x^{\text{adv}}) - F_\ell(x)\|_2^2}{\|F_\ell(x)\|_2^2 + \eta}. \tag{11}$$

The negative sign in Eq. (5) maximizes this divergence while all other terms are minimized. Prompt-invariant image features are a natural attack surface for cross-prompt universal perturbations, but it is paired here with SAM3-specific score and mask objectives.

**Optimization.** We optimize $\delta$ with Adam (Kingma & Ba, 2014) and project it into the $\ell_\infty$ ball after every step:

$$\tilde{\delta} \leftarrow \text{Adam}(\delta, \nabla_\delta \mathcal{J}),\ \delta \leftarrow \text{clip}(\tilde{\delta}, -\epsilon, \epsilon). \tag{12}$$

# 3 Experiments

## 3.1 Experimental Setup

**Datasets and Metrics.** We evaluate UCD on five concept-segmentation datasets. SACo-Gold is released with SAM3 and evaluates promptable concept segmentation with both positive and negative NPs. LVIS (Gupta et al., 2019) evaluates open-vocabulary category segmentation with positive and negative category prompts. RefCOCO (Kazemzadeh et al., 2014), PhraseCut (Wu et al., 2020), and OpenImages (Kuznetsova et al., 2020) provide complementary referring-expression, phrase-grounding, and open-vocabulary segmentation evaluations. We report $\text{cgF}_1$ and COCO-style AP for masks and boxes. $\text{cgF}_1$ is the official SAM3 concept-grounding metric, but it is not applicable when a sampled set contains only positive prompts. For each image benchmark, we sample 100 images per dataset for evaluation. For randomly sampled evaluation subsets, we repeat the sampling and evaluation procedure three times and report the averaged metrics. Unless otherwise stated, seed 0 is used as the base sampling seed. We use the official pretrained SAM3 model as the target.

**Implementation Details.** The main perturbation is trained on 50 randomly sampled LVIS-train images and 50 randomly sampled SACo-Silver images. For each image, at most four positive NPs are used to form the positive-prompt training set $\mathcal{D}_+$. Training is grouped by image: we perform two updates for each sampled image and run one epoch over the sampled training set. The perturbation is randomly initialized and constrained by $\epsilon = 8/255$ in normalized pixel units. We use a learning rate of $0.03$. All images are resized to $1008 \times 1008$ for training and evaluation. During evaluation, candidate masks are kept with a final-score threshold of $0.5$ and binarized with a mask threshold of $0.5$. The loss weights are $\lambda_s = 0.5$, $\lambda_a = 0.15$, $\lambda_d = 0.5$, $\lambda_f = 1$. For the TopK, we use $k = 10$, $k_a = k_d = 3$. All experiments are conducted on a single NVIDIA RTX A6000 GPU.

## 3.2 Comparison Study

Table 1 compares UCD with four SAM3-adapted baseline attacks: DarkSAM (Zhou et al., 2024), S-RA (Shen et al., 2024), UAPGD (Deng & Karam, 2020), and UAP-SAM2 (Zhou et al., 2026). Each baseline is optimized as a universal image perturbation under the same perturbation budget, random seeds, and optimization schedule as UCD, and is evaluated on disjoint test samples.

UCD achieves the strongest attack among the compared methods. Averaged over the five benchmarks, UCD reduces mask AP from 59.43 to 18.73, while the strongest baseline on this metric, UAP-SAM2, remains at 56.82. The same trend holds for localization: UCD lowers the average box AP from 60.35 to 17.77, compared with 58.56 for UAPGD, the strongest baseline on this metric. On concept-aware recognition, UCD reduces the average $\text{cgF}_1$ from 50.32 to 20.49, compared with 46.00 for UAPGD, and achieves the lowest adversarial $\text{cgF}_1$ on both SACo-Gold and LVIS. The adapted SAM/SAM2-style and generic universal perturbation baselines produce limited or inconsistent degradation and sometimes fail to reduce AP. UCD consistently outperforms these baselines across all five datasets. Although the perturbation is trained on a small set of LVIS and SACo-Silver images, it remains effective across benchmarks with category prompts, referring expressions, phrase-level prompts, and open-vocabulary object annotations. These results support the

Table 1: Comparison with SAM3-adapted baseline attacks under the same fair universal-perturbation training budget. We report COCO-style mask AP (%), COCO-style bounding-box AP averaged over five datasets (%), and concept-aware F1 ($cgF_1$, %) on datasets with negative prompts. Clean denotes the unattacked SAM3 performance. Lower adversarial values indicate stronger attacks; the best attack values are shown in bold.

| | Mask AP ↓ (%) | | | | | | Box AP ↓ (%) | $cgF_1$ ↓ (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | SACo-Gold | LVIS | RefCOCO | Phrase Cut | Open Images | Avg. | Avg. | SACo-Gold | LVIS | Avg. |
| Clean | 65.68 | 64.00 | 57.30 | 34.34 | 75.85 | 59.43 | 60.35 | 56.43 | 44.21 | 50.32 |
| DarkSAM | 65.41 | 62.43 | 57.45 | 35.72 | 76.59 | 59.52 | 60.51 | 57.81 | 42.67 | 50.24 |
| S-RA | 61.54 | 61.85 | 57.71 | 34.97 | 76.22 | 58.46 | 59.53 | 51.69 | 42.95 | 47.32 |
| UAPGD | 57.00 | 61.43 | 58.09 | 35.12 | 77.09 | 57.75 | 58.56 | 49.57 | 42.42 | 46.00 |
| UAP-SAM2 | 57.30 | 61.07 | 56.63 | 34.58 | 74.50 | 56.82 | 58.63 | 49.23 | 43.18 | 46.20 |
| Ours | **18.69** | **30.51** | **13.77** | **9.44** | **21.23** | **18.73** | **17.77** | **17.40** | **23.57** | **20.49** |

Table 2: UCD ablations. All scores are adversarial percentages (↓); Mask AP and $cgF_1$ denote averages over the five image benchmarks and over SACo-Gold/LVIS, respectively. Bold marks the strongest attack in each panel. Query + presence denotes the presence-gated query score; mixed-data rows list LVIS/SACo image counts. Unless varied, the setting is Full UCD with $\epsilon = 8/255$, text-prompt training, and four aligned prompts per image. Evaluation always uses text prompts.

(a) Concept target.

| Variant | Mask AP | $cgF_1$ |
|---|---|---|
| Query only | 37.45 | 32.33 |
| Presence only | 57.34 | 45.03 |
| Query + presence | **18.73** | **20.49** |

(b) Training prompts.

| Variant | Mask AP | $cgF_1$ |
|---|---|---|
| Point grid ($4\times4$) | 27.94 | 29.93 |
| Point grid ($8\times8$) | 26.42 | 29.39 |
| Text prompts | **18.73** | **20.49** |

(c) Perturbation budget.

| Budget | Mask AP | $cgF_1$ |
|---|---|---|
| $4/255$ | 56.73 | 44.05 |
| $8/255$ | 18.73 | 20.49 |
| $16/255$ | **5.81** | **7.37** |

(d) Loss components.

| Variant | Mask AP | $cgF_1$ |
|---|---|---|
| w/o score | 36.85 | 33.77 |
| w/o mask area | 46.28 | 38.34 |
| w/o mask dice | 31.01 | 29.48 |
| w/o feature | 57.45 | 44.33 |
| Full UCD | **18.73** | **20.49** |

(e) Image-prompt pairing.

| Variant | Mask AP | $cgF_1$ |
|---|---|---|
| 1 prompt | 46.00 | 39.43 |
| 2 prompts | 37.15 | 31.16 |
| 3 prompts | 24.91 | 29.72 |
| 4 (shuffled) | 27.71 | 30.21 |
| 4 prompts | **18.73** | **20.49** |

(f) Training data.

| Variant | Mask AP | $cgF_1$ |
|---|---|---|
| Full UCD | **18.73** | **20.49** |
| SACo only | 35.29 | 34.26 |
| Mixed 25/25 | <u>29.80</u> | 29.80 |
| Mixed 100/100 | 32.51 | <u>29.57</u> |

cross-dataset generalization of UCD and suggest that jointly attacking feature representations, final concept scores, and mask geometry is more effective than the adapted baseline objectives.

### 3.3 Ablation

Table 2 summarizes all six ablation studies under the main evaluation protocol. Mask AP Avg is averaged over the five image benchmarks, and $cgF_1$ Avg is averaged over SACo-Gold and LVIS. Unless otherwise specified, Full UCD uses $\epsilon = 8/255$.

**Concept Target.** Table 2a shows that targeting either query confidence or presence alone is insufficient. Attacking their presence-gated combination gives the strongest concept-decision disruption.

**Training Prompt Form.** Table 2b compares text-prompt training with point-grid training. The sparse and dense grids use $4 \times 4$ and $8 \times 8$ points, respectively. Grid supervision still transfers, but text prompts are stronger, supporting optimization on SAM3's text-conditioned concept path.

**Perturbation Budget.** Table 2c shows the expected budget-robustness trade-off: $4/255$ is weak, $8/255$ is the main setting, and $16/255$ further strengthens the attack.

**UCD Loss Components.** Table 2d shows that all terms contribute to UCD. Removing feature divergence causes the largest drop in attack strength, while removing score or mask-geometry losses also weakens the result. Thus, feature disruption drives transfer, and the output losses align it with SAM3's final concept masks.

**Image-Prompt Pairing.** Table 2e shows that using more aligned prompts per image improves the attack. In the shuffled variant, prompt strings are globally permuted across the sampled positive training items after dataset sampling. Shuffling prompts remains effective but weaker than the aligned four-prompt setting, suggesting that both prompt diversity and image-text alignment help.

**Training Data Composition.** Table 2f shows that mixing LVIS and SACo training images is more effective than SACo-only training. The 25/25 mixture improves both metrics over SACo only, whereas increasing the mixture to 100/100 does not outperform Full UCD under the same per-image optimization schedule. Appendix D provides further discussion.

### 3.4 Attack Transferability

**Black-Box Transfer.** We evaluate black-box transfer by training the perturbation on the released SAM3 image model and applying it directly to SAM3.1 (Carion et al., 2025) without re-optimization, tuning, or access to SAM3.1 gradients. Because SAM3.1 keeps the same text-conditioned concept-segmentation interface and presence-gated prediction semantics, this setting tests whether UCD transfers beyond one checkpoint. The SAM3.1 evaluation uses the same settings as the main experiments. Table 3 reports mask AP and $\text{cgF}_1$, values in parentheses are clean-to-adversarial drops. $\text{cgF}_1$ is reported for SACo-Gold and LVIS.

The SAM3-trained perturbation transfers strongly to SAM3.1. Average mask AP drops from 45.75 to 13.66, and average $\text{cgF}_1$ over SACo-Gold and LVIS drops from 51.72 to 20.24. Since these drops occur without SAM3.1-specific optimization, they support the claim that prompt-shared image features, presence-gated scores, and mask geometry form a stable attack surface across the SAM3 family.

Table 3: Black-box transfer from SAM3 to SAM3.1. The perturbation is trained on SAM3 and evaluated on SAM3.1 without re-optimization. Parentheses in the adversarial columns denote the drop from clean performance.

| Dataset | Mask AP | | $\text{cgF}_1$ | |
|---|---|---|---|---|
| | Clean | Adv. | Clean | Adv. |
| SACo-Gold | 59.87 | 14.43 (-45.44) | 58.97 | 17.51 (-41.46) |
| LVIS | 51.05 | 22.17 (-28.88) | 44.47 | 22.97 (-21.50) |
| RefCOCO | 32.94 | 11.24 (-21.70) | - | - |
| PhraseCut | 26.07 | 7.37 (-18.70) | - | - |
| OpenImages | 58.81 | 13.10 (-45.71) | - | - |
| Average | 45.75 | 13.66 (-32.09) | 51.72 | 20.24 (-31.48) |

**Image-to-Video Transfer Attack.** We next test cross-task transfer to SAM3 video inference. The perturbation is trained only with the image objective in Eq. (5) and then injected into the video frame stream. No video frames or temporal losses are used during attack training.

Let $v = (x_1, \ldots, x_L)$ denote a video clip and let $t$ be a noun phrase prompt. SAM3 receives the prompt on the first frame and propagates the resulting concept tracks. For prefix length $k$, we perturb only the first $k$ frames:

$$\tilde{x}_i^{(k)} = \begin{cases} \Pi_{[0,1]}(x_i + \delta), & 1 \le i \le k, \\ x_i, & k < i \le L. \end{cases} \tag{13}$$

All later frames are clean. To isolate temporal transfer, we evaluate only the tail

$$\mathcal{E}_k = \{k+1, \ldots, L\}. \tag{14}$$

For any video metric $M$ where larger is better, we measure the clean-to-adversarial drop on this tail:

$$\Delta_M(k) = M(\Phi_\theta(v,t)|_{\mathcal{E}_k}, g|_{\mathcal{E}_k}) - M\Big(\Phi_\theta(\tilde{v}^{(k)},t)|_{\mathcal{E}_k}, g|_{\mathcal{E}_k}\Big), \tag{15}$$

where $\Phi_\theta$ is the SAM3 video predictor and $g$ is the ground-truth masklet annotation.

We evaluate on the validation split of SACo-VEval using 15-frame clips and $k \in \{1, 7, 14\}$, corresponding to attack/evaluation lengths of $1/14$, $7/8$, and $14/1$. After filtering short clips, the evaluation contains 50 videos and 200 video-prompt pairs, with at most four positive prompts involved in optimization per video. Figure 2(a) plots official SACo-VEval drops on the tail frames.

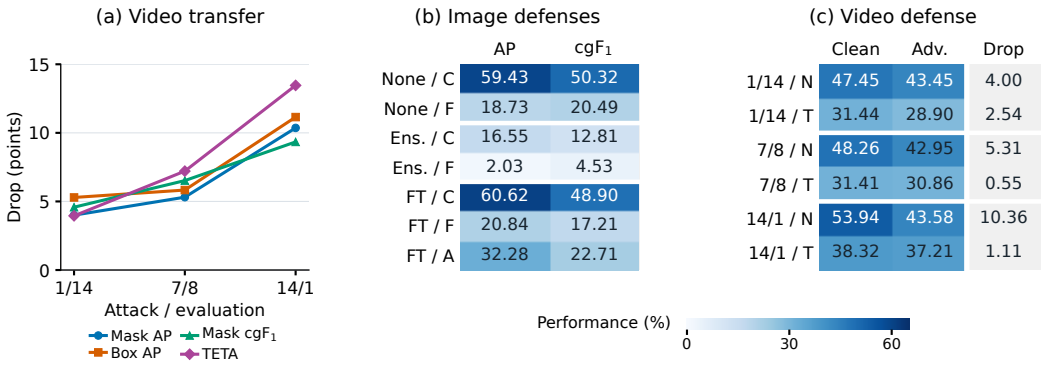



Figure 2: Video transfer and defense evaluation. (a) Clean-to-adversarial drops on clean tail frames of 15-frame SACo-VEval clips; larger drops indicate stronger transfer. (b) Dataset-averaged Mask AP and cgF$_1$ (%). Ens.: prompt ensemble; FT: head fine-tuning; C/F/A: clean/fixed/adaptive UCD. (c) Tail-frame Mask AP (%) by attack/evaluation length; N/T: no filter/temporal filter. Blue encodes performance; gray Drop entries are clean minus adversarial (points). Full defense results appear in Appendix E.

The image-trained perturbation transfers to video without video-specific optimization. Attacking only the first frame already reduces tail performance by 4.00 mask AP, 5.29 box AP, 4.58 mask cgF$_1$, and 3.95 TETA points. Longer attacked prefixes generally increase the effect: in the $14/1$ setting, the evaluated frame is clean, yet the drops reach 10.36 mask AP, 11.15 box AP, 9.34 mask cgF$_1$, and 13.46 TETA points. This shows that prefix corruption can be carried forward by SAM3 video memory to later clean frames.

### 3.5 Defense-Oriented Evaluation

We evaluate three task-relevant mitigation families: prompt ensembling, adversarial head fine-tuning, and temporal consistency filtering. For each defense, we report clean performance, fixed-attack performance, and, when the defense is differentiable or can be approximated, adaptive-attack performance. We avoid using image corruptions as defenses, because they can reduce clean performance enough to obscure whether the defense is meaningful.

**Prompt ensembling.** SAM3's concept path is text-conditioned, so a natural test-time mitigation is to query multiple paraphrases of the same noun phrase and aggregate the predictions. For a prompt $t$, we construct a small deterministic template set

$$\mathcal{E}(t) = \{t,\ \text{“a photo of } t\text{”},\ \text{“all } t\text{”},\ \text{“the visible } t\text{”}\}. \tag{16}$$

We merge predictions with mask non-maximum suppression and retain candidates whose mean final concept score across matched ensemble members exceeds the standard threshold. This defense tests whether UCD overfits to one lexical realization of the concept and whether voting over prompt variants can recover missed detections.

**Adversarial head fine-tuning.** We evaluate a defense that freezes the image backbone and text encoder, and fine-tunes only the grounding head and presence-score layers on a mixture of clean and UCD-perturbed training samples. For every positive image-prompt pair, the model is trained to recover the clean concept score and mask assignment under both $x$ and $\Pi_{[0,1]}(x+\delta)$. Negative prompts are kept in the batch to discourage a trivial increase of all presence scores. This defense directly targets the component attacked by UCD while limiting clean-performance drift. To avoid overstating robustness, we also retrain an adaptive UCD perturbation against the fine-tuned model and evaluate whether the recovered robustness persists under a white-box attacker.

**Temporal consistency filtering.** For video inference, we test whether the image-trained perturbation can be mitigated by rejecting temporally unstable concept tracks. Given frame-level track scores $s_i$ and masks $M_i$, the filter maintains a track only when its median score over a sliding window of length $w$ remains above the standard threshold and its mask remains spatially consistent with

the previous accepted mask:

$$\mathrm{keep}(i) = \mathbb{1}\big[\mathrm{median}(s_{i-w+1:i}) \geq \tau \wedge \mathrm{IoU}(M_i, M_{i-1}) \geq \tau_{\mathrm{temp}}\big]. \tag{17}$$

We report metrics on the clean tail frames, as in Fig. 2(a). This isolates whether temporal post-processing can prevent early corrupted concept states from propagating into later frames.

**Image defenses.** For image evaluation, we report the same summary metrics used throughout the ablation study. Complete numerical results for both image and video defenses are provided in Appendix E. Figure 2(b) shows that prompt ensembling lowers both clean and attacked performance, rather than preserving clean utility. Although it lowers the fixed-UCD drop, it also reduces clean Mask AP Avg from 59.43 to 16.55. The apparent robustness therefore comes mainly from suppressing predictions on clean images. Lightweight head fine-tuning preserves clean utility, with 60.62 Mask AP Avg and 48.90 $\mathrm{cgF}_1$ Avg, but fixed UCD still reduces performance to 20.84 Mask AP Avg and 17.21 $\mathrm{cgF}_1$ Avg. Retraining UCD gives 32.28 Mask AP Avg and 22.71 $\mathrm{cgF}_1$ Avg, which is higher than the fixed-attack score but still below clean performance.

**Video defense.** Figure 2(c) evaluates temporal filtering in the same image-to-video transfer setting as Fig. 2(a), using only tail-frame Mask AP. The filter reduces the clean-to-adversarial drop in all prefix settings, as shown in the gray Drop column, and nearly eliminates it when the attacked prefix is long. However, this reduction comes with a large clean-performance cost: for example, at $1/14$, clean Mask AP decreases from 47.45 without filtering to 31.44 with filtering. The filter therefore behaves as an aggressive post-processing rule rather than a robust recovery mechanism.

Overall, these defenses do not invalidate the main UCD finding. Test-time prompt ensembling and temporal filtering can reduce the measured attack drop, but only by lowering clean concept-segmentation quality. Head fine-tuning is less destructive, yet UCD remains effective against both the original and fine-tuned models. Robust defenses for SAM3 therefore need to preserve the presence-gated concept decision rather than simply pruning uncertain predictions.

## 4 Related Work

**Segment Anything Models.** SAM introduced class-agnostic promptable segmentation from points, boxes, masks, and automatic prompts (Kirillov et al., 2023), and has often served as the segmentation prior in detector- or grounding-driven open-vocabulary pipelines (Li et al., 2021; Ren et al., 2024). SAM2 (Ravi et al., 2025) keeps this geometric-prompt interface while extending it to video with streaming memory. SAM3 (Carion et al., 2025) instead moves to Promptable Concept Segmentation (PCS). For UCD, the change is SAM3's presence-token gate, which makes concept existence a shared post-processing decision and exposes the attack surface targeted in this work.

**Adversarial Attacks on SAM Series.** Adversarial attacks (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2015) add small perturbations to induce incorrect predictions, with universal adversarial perturbations learning one reusable perturbation across inputs (Moosavi-Dezfooli et al., 2017). Several studies (Zhou et al., 2024; Xia et al., 2024; Lu et al., 2024; Liu et al., 2026; Huang et al., 2024; Zheng et al., 2025; Croce & Hein, 2024; Han et al., 2024) have analyzed or attacked SAM-series models. UAP-SAM2 (Zhou et al., 2026) learns cross-prompt universal perturbations under video memory with target scanning and dual semantic deviation. Unlike these attacks, UCD trains on positive (image, NP) pairs and directly targets SAM3.

## 5 Conclusion

This paper studies the adversarial robustness of SAM3 image segmentation and identifies the presence-gated concept decision as a new universal attack surface. We proposed UCD, a cross-concept universal adversarial perturbation. Experiments across five image benchmarks show that UCD consistently outperforms attack baselines, and ablations verify the contribution of each component. Our defense-oriented evaluation further suggests that simple prompt ensembling, head fine-tuning, and temporal filtering are insufficient to restore robust concept segmentation without cost.

## AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing this manuscript. The authors are responsible for verifying the implementation, data statistics, mathematical definitions, citations, and all claims.

## REFERENCES


Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems*, 38:60884–60937, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *IEEE*, 2017.

Francesco Croce and Matthias Hein. Segment (almost) nothing: Prompt-agnostic adversarial attacks on segmentation models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 425–442, 2024. doi: 10.1109/SaTML59370.2024.00028.

Yingpeng Deng and Lina J. Karam. Universal adversarial attack via enhanced projected gradient descent. *IEEE*, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Computer Science*, 2014.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Dongshen Han, Chaoning Zhang, Sheng Zheng, Chang Lu, Yang Yang, and Heng Tao Shen. Sam meets uap: Attacking segment anything model with universal adversarial perturbation, 2024. URL https://arxiv.org/abs/2310.12431.

Shize Huang, Qianhui Fan, Zhaoxin Zhang, Xiaowen Liu, Guanqun Song, and Jinzhe Qin. Segment shards: Cross-prompt adversarial attacks against the segment anything model. *Applied Sciences (2076-3417)*, 14(8), 2024.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, and Jenq Neng Hwang. Grounded language-image pre-training. 2021.

Xiaoliang Liu, Furao Shen, and Jian Zhao. Region-guided attack on the segment anything model. *Neural Networks*, 193(c):108058, 2026.

Jiahuan Long, Zhengqin Xu, Tingsong Jiang, Wen Yao, Shuai Jia, Chao Ma, and Xiaoqian Chen. Robust sam: on the adversarial robustness of vision foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5775–5783, 2025.

Jiahao Lu, Xingyi Yang, and Xinchao Wang. Unsegment anything by simulating deformation. *IEEE*, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2017.

Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *IEEE*, 2015.

Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. *IEEE*, 2017.

Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, and Feng Yan. Grounded sam: Assembling open-world models for diverse visual tasks. 2024.

Yifan Shen, Zhengyuan Li, and Gang Wang. Practical region-level attack against segment anything models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 194–203, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *Computer Science*, 2013.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10216–10225, 2020.

Song Xia, Wenhan Yang, Yi Yu, Xun Lin, Henghui Ding, Ling Yu Duan, and Xudong Jiang. Transferable adversarial attacks on sam and its downstream models. 2024.

Chenshuang Zhang, Chaoning Zhang, Taegoo Kang, Donghun Kim, Sung-Ho Bae, and In So Kweon. Attack-sam: Towards evaluating adversarial robustness of segment anything model. *arXiv preprint arXiv:2305.00866*, 1(3):5, 2023.

Sheng Zheng, Chaoning Zhang, and Xinhong Hao. Black-box targeted adversarial attack on segment anything (sam). *Trans. Multi.*, 27:1901–1913, January 2025. ISSN 1520-9210. doi: 10.1109/TMM.2024.3521769. URL https://doi.org/10.1109/TMM.2024.3521769.

Ziqi Zhou, Yufei Song, Minghui Li, Shengshan Hu, Xianlong Wang, Leo Yu Zhang, Dezhong Yao, and Hai Jin. Darksam: Fooling segment anything model to segment nothing. *Advances in Neural Information Processing Systems*, 37:49859–49880, 2024.

Ziqi Zhou, Yifan Hu, Yufei Song, Zijing Li, Shengshan Hu, Leo Yu Zhang, Dezhong Yao, Long Zheng, and Hai Jin. Vanish into thin air: Cross-prompt universal adversarial attacks for sam2. *Advances in Neural Information Processing Systems*, 38:142839–142861, 2026.

## A Dataset Introduction

We evaluate UCD on five complementary image benchmarks. Together, these datasets test whether a universal perturbation transfers across promptable concept segmentation, long-tailed category segmentation, referring expressions, phrase grounding, and open-vocabulary object annotations.

**SACo-Gold.** SACo-Gold is part of the SA-Co benchmark released with SAM3 (Carion et al., 2025). It pairs images with noun phrases and provides instance masks for all objects that match each phrase. It also includes negative noun phrases with no matching object, making it suitable for evaluating both mask AP and concept-grounding $\mathrm{F}_1$.

**LVIS.** LVIS is a large-vocabulary instance segmentation benchmark with a long-tailed category distribution (Gupta et al., 2019). We convert category names into text prompts and evaluate whether SAM3 can retain masks for rare and common object categories under attack. Like SACo-Gold, LVIS supports both positive and negative category prompts in our protocol, so it contributes to both AP and $\mathrm{cgF}_1$ evaluation.

**RefCOCO.** RefCOCO is derived from the ReferItGame collection of natural-language referring expressions (Kazemzadeh et al., 2014). Each prompt describes a specific target object in an image, often using attributes, location, or relations to distinguish it from nearby objects. This benchmark tests whether UCD transfers from category-style prompts to instance-level referring expressions.

**PhraseCut.** PhraseCut is a phrase-based segmentation benchmark built for language-guided region segmentation in natural scenes (Wu et al., 2020). Its phrases can refer to objects, stuff regions, attributes, parts, and relationships. We use it to evaluate whether the perturbation disrupts more compositional phrase queries, beyond single noun categories.

**OpenImages.** OpenImages provides large-scale image annotations covering diverse visual concepts and object categories (Kuznetsova et al., 2020). In our evaluation, it serves as an additional open-vocabulary segmentation benchmark with broad image diversity. It complements LVIS by testing category prompts on a different image source and annotation distribution.

## B Optimization Algorithm

Algorithm 1 summarizes the optimization procedure for Universal Concept Disruption (UCD).

**Algorithm 1** Universal Concept Disruption

**Require:** SAM3 model $f_\theta$, training pairs $\mathcal{D}$, budget $\epsilon$
Initialize $\delta \sim \mathcal{U}(-\epsilon, \epsilon)$
**for** each optimization step **do**
  Sample an image and its positive text prompts
  Compute clean features and clean top masks without gradient
  Apply $x^{\mathrm{adv}} = \Pi_{[0,1]}(x + \delta)$
  Accumulate score, area, and Dice losses over prompts
  Add feature-divergence
  Update $\delta$ with Adam and project to $[-\epsilon, \epsilon]$
**end for**
**return** universal perturbation $\delta$

## C Evaluation Metrics

The main paper reports COCO-style Average Precision (AP) for masks and boxes, and concept-grounding $\mathrm{F}_1$ ($\mathrm{cgF}_1$) for datasets that contain both positive and negative prompt-image pairs. Mask AP and box AP use the same definition, with mask IoU and box IoU respectively.

**Average Precision.** For an IoU threshold $\tau$, predictions are sorted by confidence and matched one-to-one to ground-truth instances. A prediction is a true positive if its IoU with an unmatched ground-truth instance is at least $\tau$; otherwise it is a false positive. Let $p_\tau(r)$ denote the interpolated precision at recall level $r$. COCO-style AP is computed over the 101 recall levels $\mathcal{R} = \{0, 0.01, \ldots, 1\}$ and then averaged over the ten IoU thresholds $\mathcal{T} = \{0.50, 0.55, \ldots, 0.95\}$:

$$AP^\tau = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} p_\tau(r), \qquad AP = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} AP^\tau. \tag{18}$$

$\text{AP}_{50}$ and $\text{AP}_{75}$ correspond to $AP^\tau$ at $\tau = 0.50$ and $\tau = 0.75$, respectively. We report AP values as percentages.

**Concept-Grounding $\text{F}_1$.** We follow the official SAM3 $\text{cgF}_1$ definition (Carion et al., 2025). The metric combines a positive-pair mask $\text{F}_1$ term with an image-level presence term. For positive prompt-image pairs, let $TP_\tau$, $FP_\tau$, and $FN_\tau$ be the total counts after one-to-one matching at IoU threshold $\tau$. The positive micro $\text{F}_1$ is

$$pmF_1^\tau = \frac{2TP_\tau}{2TP_\tau + FP_\tau + FN_\tau}, \qquad pmF_1 = \frac{1}{10} \sum_{\tau \in \mathcal{T}} pmF_1^\tau. \tag{19}$$

For the image-level presence classification over positive and negative prompt-image pairs, the Matthews correlation coefficient is

$$\text{IL_MCC} = \frac{\text{IL_TP} \cdot \text{IL_TN} - \text{IL_FP} \cdot \text{IL_FN}}{\sqrt{D}}, \tag{20}$$

where

$$D = (\text{IL_TP} + \text{IL_FP})(\text{IL_TP} + \text{IL_FN}) \cdot (\text{IL_TN} + \text{IL_FP})(\text{IL_TN} + \text{IL_FN}). \tag{21}$$

The reported percentage-form $\text{cgF}_1$ is

$$\text{cgF}_1 = 100 \cdot pmF_1 \cdot \text{IL_MCC}. \tag{22}$$

Because IL_MCC requires both positive and negative prompt-image pairs, $\text{cgF}_1$ is not reported for positive-only sampled subsets.

## D Training Data

The training-data panel of Table 2 (Table 2f) evaluates how training data composition affects the universal perturbation. SACo-only training is substantially weaker than Full UCD, but it remains effective, reaching 35.29 Mask AP Avg and 34.26 $\text{cgF}_1$ Avg. Using a smaller balanced mixture of LVIS and SACo improves the result to 29.80 Mask AP Avg and 29.80 $\text{cgF}_1$ Avg, showing that category diversity and concept-segmentation examples are complementary. Increasing the mixture to 100/100 images does not further improve the attack under the same per-image optimization schedule.

## E Complete Defense Evaluation Results

Tables 4 and 5 provide the complete numerical results visualized in Figure 2(b,c). The image evaluation uses the same metric averages as the main experiments. The video evaluation reports tail-frame Mask AP on SACo-VEval for each attack/evaluation length.

Table 4: Complete image defense results (%). Fixed UCD uses the perturbation from Table 1; adaptive UCD is retrained against the fine-tuned head.

| Defense | Setting | Mask AP Avg | $cgF_1$ Avg |
|---|---|---|---|
| None | Clean | 59.43 | 50.32 |
| None | Fixed UCD | 18.73 | 20.49 |
| Prompt ensemble | Clean | 16.55 | 12.81 |
| Prompt ensemble | Fixed UCD | 2.03 | 4.53 |
| Head fine-tuning | Clean | 60.62 | 48.90 |
| Head fine-tuning | Fixed UCD | 20.84 | 17.21 |
| Head fine-tuning | Adaptive UCD | 32.28 | 22.71 |

Table 5: Complete video defense results. Scores are tail-frame Mask AP (%) on SACo-VEval. Drop is clean minus adversarial, in percentage points.

| | No filter | | | Temporal filter | | |
|---|---|---|---|---|---|---|
| Attack/eval | Clean | Adv. | Drop | Clean | Adv. | Drop |
| 1/14 | 47.45 | 43.45 | 4.00 | 31.44 | 28.90 | 2.54 |
| 7/8 | 48.26 | 42.95 | 5.31 | 31.41 | 30.86 | 0.55 |
| 14/1 | 53.94 | 43.58 | 10.36 | 38.32 | 37.21 | 1.11 |

## F Qualitative Visualization

Figures 3–7 visualize representative results from the main UCD setting. Each figure contains four rows of examples. The left margin lists the text prompt, the clean-to-adversarial top score change, and the number of retained masks. The two image columns compare the clean SAM3 overlay with the adversarial overlay under the same prompt.




Figure 3: Qualitative examples on SACo-Gold.

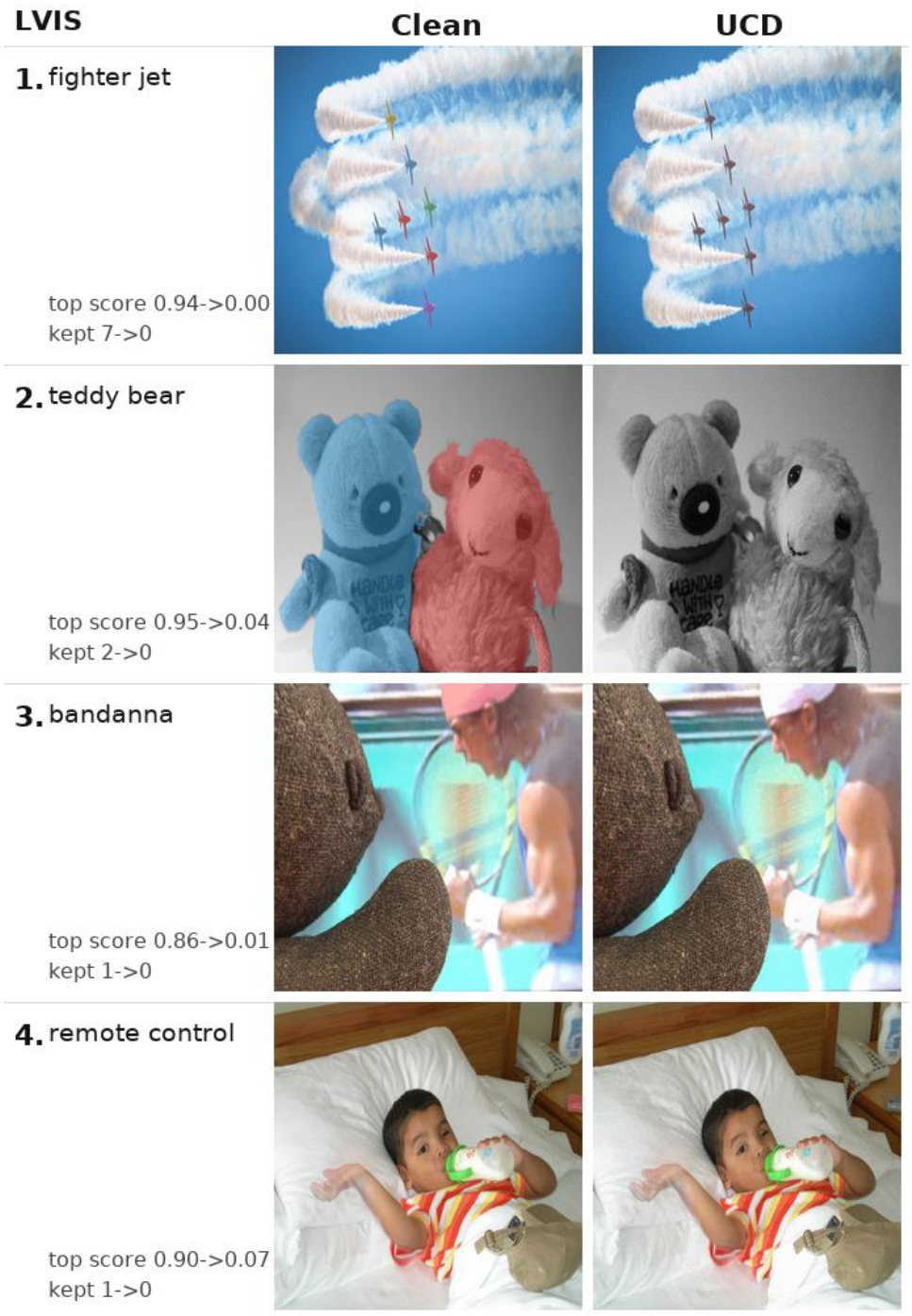



Figure 4: Qualitative examples on LVIS.

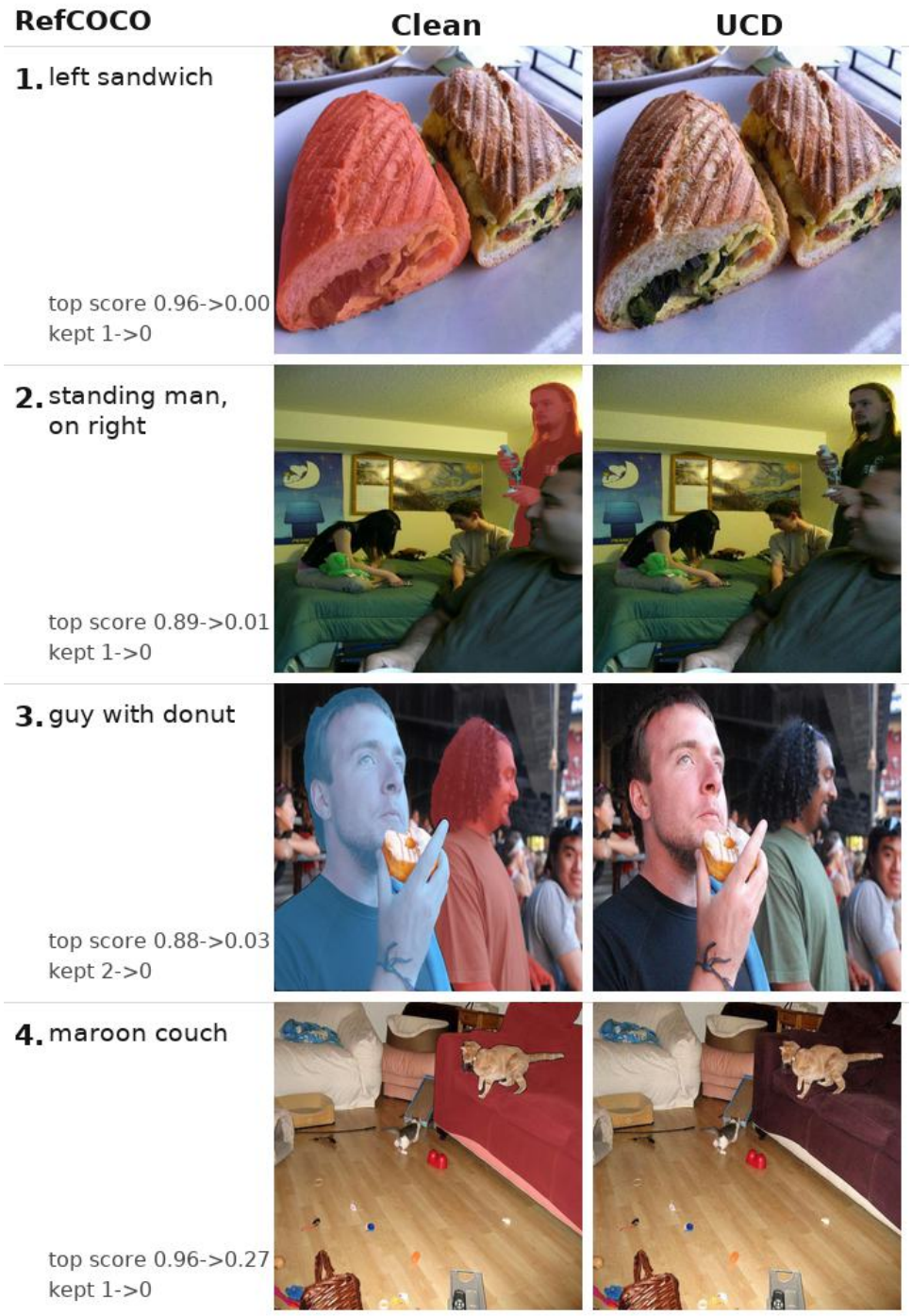



Figure 5: Qualitative examples on RefCOCO.

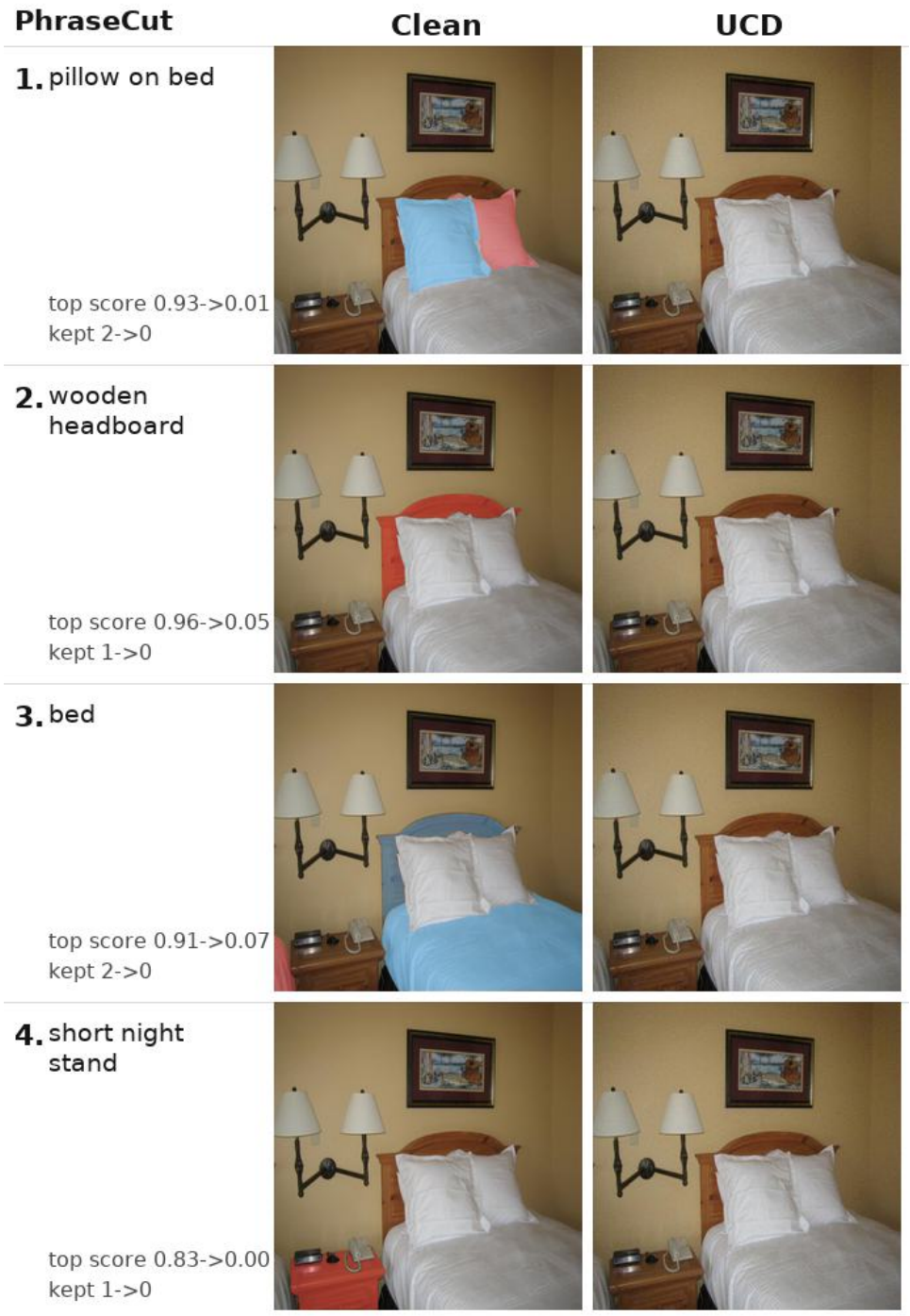



Figure 6: Qualitative examples on PhraseCut.

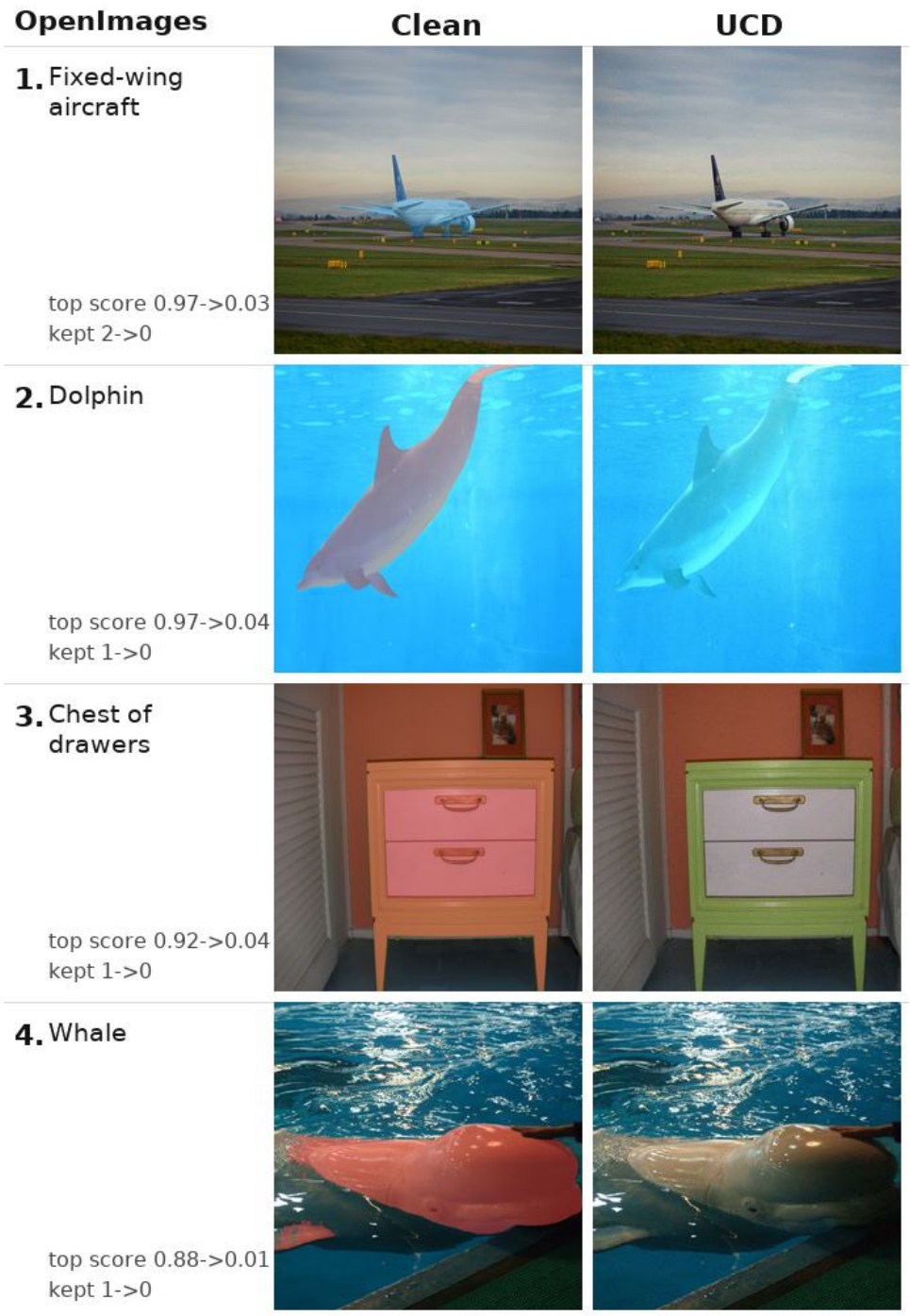



Figure 7: Qualitative examples on OpenImages.