Universal Cross-Prompt Adversarial Attacks on Promptable Concept Segmentation

Ziqi Zhou[1,2], **Yifan Hu**[3], **Yufei Song**[3], **Haowen Jiang**[4], **Xianlong Wang**[5], **Shengshan Hu**[3], **Dezhong Yao**[4], **Leo Yu Zhang**[6]
[1] College of Computer Science, Chongqing University
[2] Sichuan Engineering Technology Research Center for Data Security and Privacy Protection
[3] School of Cyber Science and Engineering, Huazhong University of Science and Technology
[4] School of Computer Science and Technology, Huazhong University of Science and Technology
[5] Department of Computer Science, City University of Hong Kong
[6] School of Information and Communication Technology, Griffith University
zhouziqi@cqu.edu.cn, {hyf1009,yufei17,jianghaowen,hushengshan,dyao}@hust.edu.cn
xianlong.wang@my.cityu.edu.hk, leo.zhang@griffith.edu.au

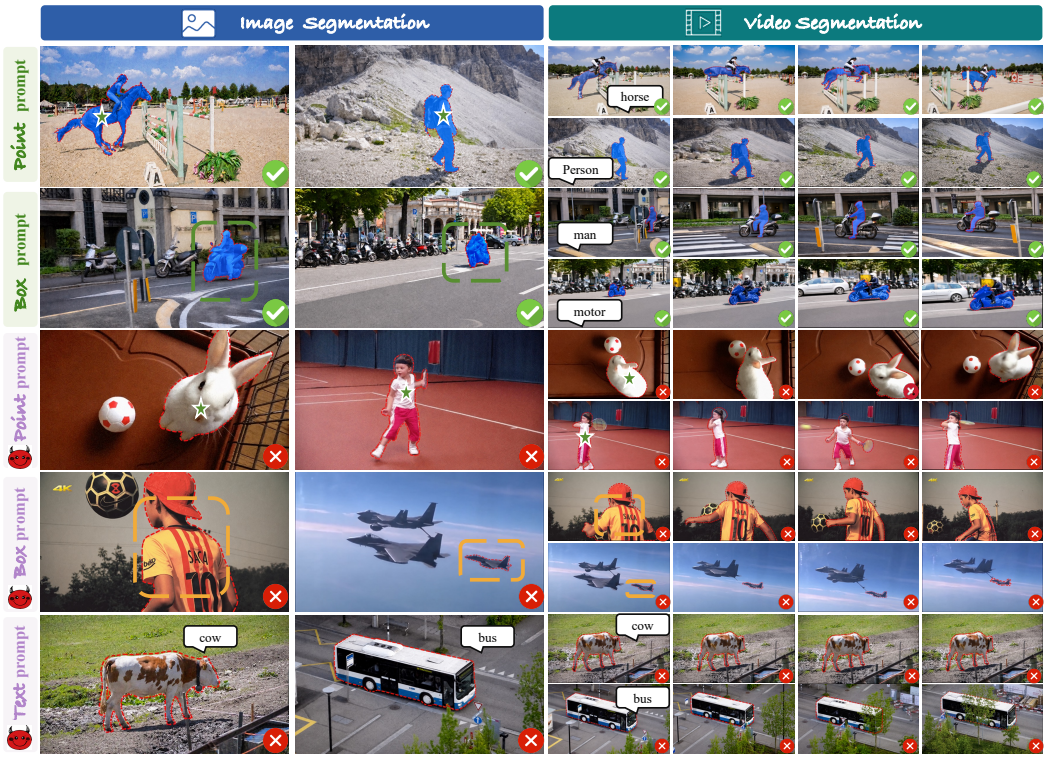


Figure 1: Visualization of SAM3 segmentation results on benign samples and adversarial examples generated by our method under point, box, and text prompts. The first and second rows correspond to benign samples, while the third to fifth rows correspond to adversarial examples. For better visualization, we use red dashed contours to indicate the ground-truth masks of the target objects, and blue highlights to indicate the regions segmented by SAM3.

Abstract

The ***Segment Anything Model*** (SAM) achieves remarkable performance in visual segmentation. The latest SAM3 extends promptable segmentation to concept-level prediction, broadening the scope of segmentation foundation models. While recent

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

works reveal that SAM and SAM2 are vulnerable to adversarial examples, the robustness of SAM3 under the concept segmentation paradigm remains unexplored. In addition, existing adversarial attacks on SAM-series models exhibit limited cross-prompt transferability. To this end, we propose AdvPCS, a universal cross-prompt adversarial attack for ***P**romptable **C**oncept **S**egmentation* (PCS), including a min-max prompt optimization strategy, a global–local perception deception attack, and a temporal transition deviation attack. Specifically, we first identify the hardest-to-attack prompts via the min-max bilevel optimization. In the inner maximization, we select the highest-response candidate within each prompt type under the current perturbation. In the outer minimization, we update the perturbation against the selected prompts. Given the selected prompts, we apply the perception deception attack to minimize both global and local existence probabilities under joint prompting, and employ the temporal transition deviation attack to maximize inter-frame semantic inconsistency and corrupt memory pointers. Extensive experiments on four benchmark datasets show that a single *universal adversarial perturbation* (UAP) generated by our method generalizes across frames from different videos and achieves strong attack performance under point, box, and text prompts. In particular, it reduces the average mIoU of various PCS models on the SA-CO dataset to below 5% under text prompts, demonstrating strong attack ability. The code is available at https://github.com/alphanull-cqu/AdvPCS.

1 Introduction

The Segment Anything Model (SAM) and its variants [2, 12, 21] have demonstrated strong generalization in visual segmentation. SAM [12] introduces the *promptable visual segmentation* (PVS) paradigm, enabling general-purpose image segmentation via prompts. Moreover, SAM2 [21] extends this framework to videos by incorporating memory mechanisms that ensure temporally consistent segmentation and tracking. To achieve *promptable concept segmentation* (PCS), the recent SAM3 [2] introduces a unified framework for detection, tracking, and segmentation, which takes simple noun phrases as input to understand, track, and precisely segment target objects. Compared with SAM and SAM2, SAM3 not only models texture and boundary cues but also identifies all instances in the scene that match the given concept, and outputs their segmentation masks with unique identities.

SAMs are shown to be vulnerable to adversarial examples [15, 16, 24, 27, 33, 34, 36, 37], where attackers add carefully crafted and imperceptible perturbations to target frames to induce incorrect predictions. Existing attacks on SAM mainly focus on PVS. They either corrupt the image encoder features without relying on prompts [6, 32] or optimize adversarial perturbations with multiple random prompts [38] to achieve cross-prompt transferability. A recent work, UAP-SAM2 [35] disrupts temporal semantic consistency, significantly degrading video segmentation performance in SAM2. Although these methods perform well on SAM and SAM2 under a specific prompt type, their effectiveness drops substantially under diverse prompt types (See Tab. 2). Moreover, to the best of our knowledge, the robustness of emerging PCS models remains unexplored.

Different from SAM [12] and SAM2 [21], which rely on user-provided pixel-level prompts (*e.g.*, points or boxes), SAM3 [2] introduces a detector to determine whether a concept-level object exists in the current frame and then segments all regions associated with the queried concept. To study this new property, we conduct exploratory experiments in Sec. 3.2 from the perspectives of improving both attack effectiveness and transferability. These experiments reveal two key insights: (1) deceiving the perception mechanism in SAM3 improves attack effectiveness, and (2) mining the hardest-to-attack prompts for attack optimization enhances cross-prompt transferability. Moreover, for video tasks, effectively disrupting semantic consistency across consecutive frames remains a challenging issue.

In this paper, we propose AdvPCS, a brand-new cross-prompt universal adversarial attack for promptable concept segmentation models. AdvPCS integrates a min-max bilevel adversarial prompt optimization strategy, a global–local perception deception attack, and a temporal transition deviation attack. Our method targets both cross-prompt transferability and perception-level distortion, aiming to generate a *universal adversarial perturbation* (UAP) that generalizes across diverse prompts and frames. To enhance cross-prompt transferability, we design a min–max bilevel adversarial prompt optimization strategy that identifies the hardest-to-attack prompts for optimization. We first construct diverse candidates for each prompt type. The inner maximization selects the hardest-to-

attack candidate within each type, and the outer minimization updates the universal perturbation against the selected prompts. For perception-level distortion, we design a global–local perception deception attack that jointly minimizes the global and local presence probabilities under the three types of prompts, while constraining their pairwise differences to prevent the perturbation from overfitting to any single prompt. We then design a temporal transition deviation attack that disrupts semantic consistency across consecutive frames and deliberately misguides memory pointers, thereby enhancing cross-frame attack effectiveness.

To better evaluate the transferability, we train UAPs only on the SA-CO [2] dataset and evaluate them on other datasets under point, box, and text prompts. Fig. 1 shows visualizations of the adversarial examples generated by our method. The qualitative results in the last three rows show that these samples cannot be effectively segmented by SAM3 under all three types of prompts. The quantitative results in Sec. 4 demonstrate that our method achieves strong attack performance on both video and image segmentation tasks across four datasets. Our main contributions are summarized as follows:

- We propose AdvPCS, a novel cross-prompt universal adversarial attack for promptable concept segmentation models. We generate a single UAP only on the SA-CO dataset, which generalizes across different videos, frames, prompt types, and prompt instances, and achieves strong attack performance against PCS models.
- We design a min–max bilevel adversarial prompt optimization strategy that identifies the most robust prompt instances across text, point, and box modalities, thereby significantly improving cross-prompt transferability of the UAP.
- Extensive experiments on four datasets show that our method transfers across frames from different videos and achieves strong attack performance under point, box, and text prompts. Comparative and defense experiments further validate its effectiveness and superiority.

2 Related Work

2.1 Segment Anything Models

The Segment Anything Model (SAM) [12] first formulates image segmentation as an interactive prompt learning problem and enables general-purpose segmentation with prompts such as points and boxes. SAM2 [21] extends this framework to videos by introducing Transformer-based [25] temporal modeling and a memory mechanism, enabling temporally consistent segmentation and tracking across frames. It unifies image and video processing within a single architecture. SAM3 [2] introduces *Promptable Concept Segmentation* (PCS) with a decoupled detection and tracking architecture, where both components share a unified perception encoder. The detector performs open-vocabulary object detection, while the tracker inherits the memory mechanism of SAM2 for temporal identity tracking. The perception encoder processes each image once and produces prompt-independent features shared by both modules. Benefiting from their strong segmentation capabilities [3, 11, 13, 26, 31], they are widely adopted in applications such as medical image analysis [5, 22] and autonomous driving [1, 4].

2.2 Adversarial Attacks on SAMs

Recent works [8, 10, 15, 16, 20, 23, 27] study adversarial examples to mislead SAM and its downstream tasks. AttackSAM [30] exploits SAM's mask generation mechanism to craft targeted perturbations that mislead object segmentation. Although these methods effectively mislead SAM under fixed prompt settings, their effectiveness drops significantly as prompts vary. Consequently, recent studies [6, 9, 14, 35, 38] shift focus to cross-prompt attacks against SAMs. Zheng et al. [32] and Croce et al. [6] propose transferable attacks that operate solely on the image encoder, removing dependence on specific prompts. DarkSAM [38] further introduces a hybrid frequency–spatial adversarial attack to enhance cross-prompt attack effectiveness. Recently, UAP-SAM2 [35] shows that attacks designed for SAM1 fail on SAM2 due to semantic consistency across consecutive frames. In this work, we focus on the robustness of emerging PCS models and aim to generate a single UAP that effectively fools these models across different frames, prompt types, and prompt instances.

3 Methodology

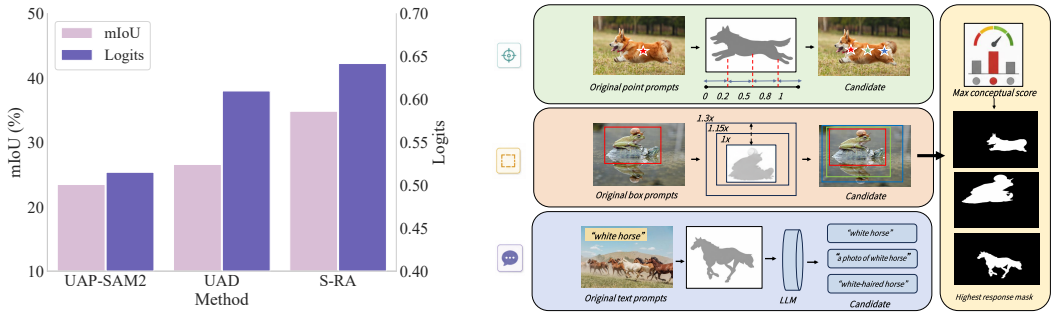
3.1 Problem Formulation

We consider a sequence of N frames, $\mathbf{X} = \{x_i\}_{i=1}^{N}$, with corresponding prompts $\mathbb{P} = \{p_i^j \mid i = 1, \ldots, N,\ j \in \{t, p, b\}\}$, where p_i^j denotes a prompt of type j for frame i, and t, p, and b represent text, point, and box prompts, respectively. Let f_θ denote SAM3 or its variants. f_θ predicts segmentation masks $\mathbf{Y} = \{y_i\}_{i=1}^{N}$ for each frame x_i. For a frame x_i, a pixel at coordinates (m, n), denoted as x_i^{mn}, is considered part of the masked region if its corresponding mask value y_i^{mn} exceeds zero. SAM3 consists of the following modules: an image encoder that extracts visual features from each input frame, a detector that predicts object presence based on the extracted features, a memory fusion module that integrates current-frame features with historical memory representations, and a tracker that generates candidate masks and selects the optimal mask as the final segmentation result. Following [35], we assume that the attacker has access to the SAM3 and can leverage public datasets to generate adversarial perturbations. The attacker aims to craft a UAP δ for each video frame such that the target object cannot be effectively segmented by SAM3. Meanwhile, δ remains imperceptible by enforcing an l_p-norm constraint bounded by a predefined perturbation budget ϵ.

3.2 Motivation

Motivated by the structural characteristics of SAM3, we design the attack from two perspectives: improving its effectiveness and transferability. Here are two key insights behind our method:

Insight I: Deceiving the perception mechanism improves attack effectiveness. Compared to previous models, SAM3 introduces a detector that jointly recognizes and localizes objects associated with open-vocabulary prompts, providing stronger semantic grounding. However, existing attacks do not explicitly target concept-level perception induced by textual prompts. Under open-vocabulary settings, ambiguities such as polysemy and subjective descriptions make it difficult for conventional feature-level perturbations [6, 32] or global semantic attacks [17] to consistently disrupt all candidate objects. We evaluate three state-of-the-art (SOTA) attacks for SAM and SAM2, including UAP-SAM2 [35], UAD [16], and S-RA [23], by generating adversarial examples on the SA-CO dataset to attack SAM3. As shown in Fig. 2a, the average results across the three types of prompts (*i.e.*, point, box, and text) reveal a clear correlation between the degradation in segmentation performance and the decrease in the detector's existence confidence. This observation motivates us to design attacks that directly target the perception mechanism to enhance attack effectiveness.



(a) Logit and mIoU comparison

(b) Hardest-to-attack prompts mining

Figure 2: Illustration of the key insights for designing effective attacks against PCS

Insight II: Hardest-to-attack prompt mining improves cross-prompt transferability. In addition to input variations, diverse prompt types and instances pose significant challenges for transferability. Existing attacks typically optimize for a single prompt type, and their effectiveness degrades substantially when the prompt type changes. Similarly, feature-space attacks become less effective when prompt information is rich. Inspired by adversarial training [17], we improve transferability from two perspectives: (i) jointly optimizing across point, box, and text prompts to enhance robustness under heterogeneous prompts; and (ii) identifying the hardest-to-attack prompt within each type to further strengthen transferability. As illustrated in Fig. 2b, we design three mining strategies for different prompt types. Specifically, ❶ for *text prompts*, we expand a given prompt into multiple semantically similar candidates using LLMs (*e.g.*, GPT5.4) and select the one that yields the highest

object existence score; ❷ for *point prompts*, we construct multiple candidates from the foreground mask, including left-biased, center, and right-biased points based on spatial quantiles, and select the point that maximizes the response; ❸ for *box prompts*, we generate a default bounding box from the foreground mask and further create scaled variants (*e.g.*, enlarged by 15% and 30%) while keeping the center fixed. The above mining strategy identifies the hardest-to-attack prompts for each type, which are then used for subsequent attack optimization.

Takeaway. *An effective attack on SAM3 requires deceiving its global–local perception mechanism while achieving strong transferability across videos, frames, prompt types, and prompt instances.*

3.3 AdvPCS: A Complete Illustration

In this section, we propose AdvPCS, a brand-new cross-prompt universal adversarial attack for PCS models. The pipeline of AdvPCS illustrated in Fig. 3 integrates a min–max bilevel adversarial prompt optimization strategy, a perception-deception attack, and a temporal transition deviation attack. The overall optimization objective of AdvPCS is as follows:

$$\mathcal{J}_{total} = \mathcal{J}_{pa} + \mathcal{J}_{ta}, \tag{1}$$

where $\mathcal{J}_{pa}$ is the global–local perception deception attack loss and $\mathcal{J}_{ta}$ denotes the temporal transition deviation attack loss.

Min–max bilevel adversarial prompt optimization. To improve cross-prompt transferability, we propose a min–max bilevel adversarial prompt optimization strategy that identifies the hardest-to-attack prompts for optimization. We first construct diverse candidates for point, box, and text prompts. In the inner maximization, we fix the current perturbation and select the highest-response candidate within each prompt type based on the detector's confidence scores. In the outer minimization, we fix the selected prompts and update the universal perturbation by minimizing the attack objective, thereby improving transferability across prompt types and instances. This optimization process can be formulated as:

$$\min_{\|\delta\|_p \le \epsilon} \max_{p_i^j \in \mathbb{P}} \sum_{i=1}^{N} \sum_{j \in \{t,p,b\}} \mathcal{L}\big(f_\theta(x_i + \delta,\ p_i^j),\ y_i\big), \tag{2}$$

where $\mathcal{L}(\cdot)$ is the segmentation loss.

Global–local perception deception attack. We aim to degrade SAM3's segmentation capability by deceiving its perception stage. From an analysis of its perception module, SAM3 employs a detector that produces a global perception logit and local perception logits to recognize and localize the target object specified by the input prompt. Specifically, the global perception logit captures the overall confidence of the target's presence in a frame, while the local perception logits are more sensitive to prompt-conditioned local features. Accordingly, we disrupt perception by suppressing both global and local confidence of the target across all the three prompt types. We further impose a consistency degradation constraint on prompt-dependent local responses to prevent the perturbation from overfitting to any single prompt. Hence, we define the perception-deception objective as follows:

$$\mathcal{J}_{pa} = \mathcal{J}_{global} + \mathcal{J}_{local} + \mathcal{J}_{reg}, \tag{3}$$

where $\mathcal{J}_{global}$ denotes the global existence weakening loss, $\mathcal{J}_{local}$ denotes the local existence weakening loss, and $\mathcal{J}_{reg}$ denotes the cross-prompt consistency regularization loss.

Given an input frame x_i, we add a UAP δ to obtain the adversarial frame $\tilde{x}_i = x_i + \delta$. Under text, point, and box prompts, the detector produces the corresponding global perception logits r_i^t, r_i^p, and r_i^b, as well as local perception logits z_i^t, z_i^p, and z_i^b, where the superscripts t, p, and b denote the text, point, and box prompt settings, respectively. We first suppress the global existence confidence of the target object under three types of prompts. $\mathcal{J}_{global}$ is expressed as:

$$\mathcal{J}_{global} = \frac{1}{N} \sum_{i=1}^{N} \left(\left\| r_i^t - \tau \right\|_2^2 + \left\| r_i^p - \tau \right\|_2^2 + \left\| r_i^b - \tau \right\|_2^2 \right), \tag{4}$$

where $\tau < 0$ is a predefined low target logit value. By uniformly suppressing global perception logits toward a shared low-confidence target, we weaken SAM3's ability to recognize the target object globally. We then further suppress the local perception logits under the three prompt types. Our

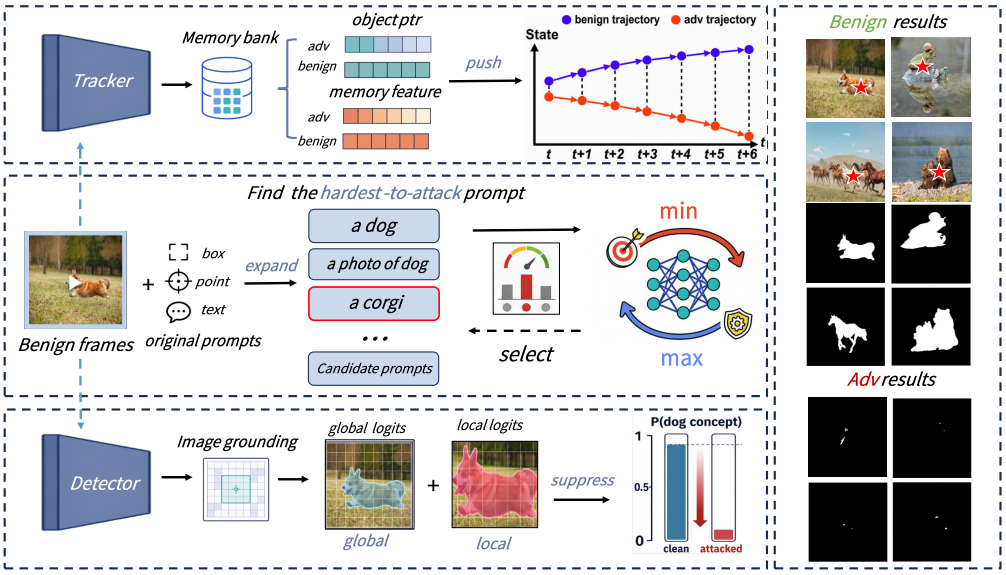


Figure 3: The framework of AdvPCS

empirical results show that box prompts are more difficult to attack. Therefore, we place greater emphasis on box prompts during the optimization process. $\mathcal{J}_{local}$ is formulated as follows:

$$\mathcal{J}_{local} = \frac{1}{N}\sum_{i=1}^{N}\left(\left\|z_i^t - \tau\right\|_2^2 + \left\|z_i^p - \tau\right\|_2^2 + \lambda\left\|z_i^b - \tau\right\|_2^2\right), \tag{5}$$

where λ is a hyperparameter that controls the influence of box prompts during the adversarial optimization process. To avoid overfitting to a single prompt type, we enforce consistent degradation of local responses across prompts in the probability space. Let $p_i^t = \sigma(z_i^t)$, $p_i^p = \sigma(z_i^p)$, and $p_i^b = \sigma(z_i^b)$ denote the prediction probabilities under the text, point, and box prompts, respectively, where $\sigma(\cdot)$ is the sigmoid function. We define $\mathcal{J}_{reg}$ as follows:

$$\mathcal{J}_{reg} = \frac{1}{3N}\sum_{i=1}^{N}\left(\left\|p_i^t - p_i^p\right\|_2^2 + \left\|p_i^t - p_i^b\right\|_2^2 + \left\|p_i^p - p_i^b\right\|_2^2\right). \tag{6}$$

Temporal transition deviation attack. Motivated by [35], which shows that memory weakens adversarial attacks in SAM2, we propose a temporal transition deviation attack that improves cross-frame transferability by disrupting temporal semantic consistency and misleading memory propagation. Starting from the second frame, we feed both benign and adversarial frames into the tracker and extract the corresponding object pointers and memory features. Let $\mathbf{o}_i$ denote the object pointer at the i-th frame, and let $\mathbf{m}_i$ denote the corresponding features. For simplicity, we define the adjacent-frame pointer transition and memory transition as follows:

$$\Delta\mathbf{o}_i = \mathbf{o}_i - \mathbf{o}_{i-1}, \qquad \Delta\mathbf{m}_i = \mathbf{m}_i - \mathbf{m}_{i-1}. \tag{7}$$

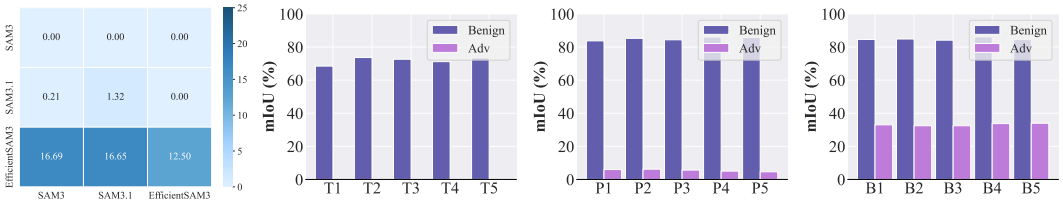
Accordingly, we use $\Delta\mathbf{o}_i^{\text{adv}}$ and $\Delta\mathbf{o}_i^{\text{benign}}$ to denote the pointer transitions of the adversarial and benign sequences, respectively, and $\Delta\mathbf{m}_i^{\text{adv}}$ and $\Delta\mathbf{m}_i^{\text{benign}}$ are defined in the same way for the memory features. Unlike [35], which focuses on maximizing feature discrepancies between adjacent adversarial frames, we instead emphasize *the deviation between adversarial and benign memory trajectories* at each time step. For each frame, we compute the temporal transition deviation π_i between the adversarial and benign samples:

$$\pi_i = \underbrace{\cos(\Delta\mathbf{o}_i^{\text{adv}}, \Delta\mathbf{o}_i^{\text{benign}})}_{\textbf{Pointer Deviation}} + \underbrace{\cos(\Delta\mathbf{m}_i^{\text{adv}}, \Delta\mathbf{m}_i^{\text{benign}})}_{\textbf{Feature Deviation}} \tag{8}$$

where $\cos(\cdot)$ denotes the cosine similarity. Finally, we obtain the objective of the temporal transition deviation attack $\mathcal{J}_{ta}$ as follows:

Table 1: The mIoU (%) of AdvPCS under different settings.

Prompt	Dataset	Video Segmentation								Image Segmentation							
		SAM3		SAM3.1		E-SAM3		Avg		SAM3		SAM3.1		E-SAM3		Avg	
		BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU
Point	YouTube	85.29	6.29	85.48	17.58	56.58	28.56	75.78	17.48	74.17	0.09	53.59	1.74	53.38	36.60	60.38	12.81
	MOSE	69.12	8.17	76.82	23.69	42.79	15.40	62.91	15.75	78.90	0.84	43.22	2.59	48.86	32.83	56.99	12.09
	DAVIS	69.73	7.89	81.59	39.13	51.55	30.69	67.62	25.90	68.21	0.32	41.95	0.61	44.98	33.81	51.71	11.58
Box	YouTube	88.28	39.62	87.86	52.93	57.51	19.25	77.88	37.27	88.19	4.18	87.48	12.61	71.04	59.43	82.24	25.41
	MOSE	75.15	25.20	82.73	43.20	39.74	13.37	65.87	27.26	86.54	7.95	79.15	19.08	64.48	54.80	76.72	27.28
	DAVIS	88.14	41.13	88.90	61.14	53.88	23.77	76.97	42.01	87.98	3.95	88.49	14.91	65.95	52.89	80.81	23.92
Text	SA-CO	68.48	0.00	68.73	1.32	73.52	12.50	70.24	4.61	87.54	0.07	72.65	0.61	75.06	11.19	78.42	3.96



(a) Text transfer (b) Text prompt (c) Point prompt (d) Box prompt

Figure 4: Transferability study results. (a) Cross-model attack transfer results; (b)–(d) cross-prompt instance attack transfer results under different prompt types, respectively.

$$\mathcal{J}_{ta} = \frac{1}{N-1} \sum_{i=2}^{N} \pi_i \tag{9}$$

By directly reducing the cosine similarity between the adversarial and benign temporal transitions, the adversarial sequence is progressively driven away from the benign temporal trajectory, so that the corrupted temporal cues accumulated from previous frames can further misguide subsequent predictions and amplify the attack effect over time.

4 Experiments

4.1 Experimental Setup

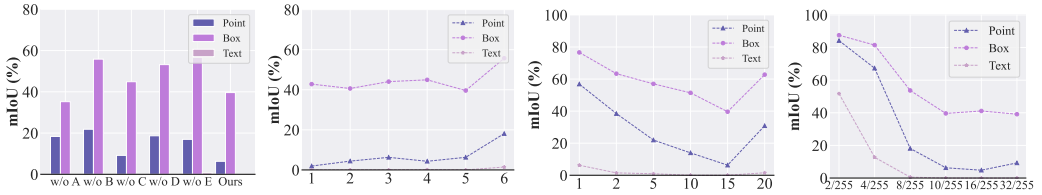
Datasets and models. We evaluate the proposed attack on four public video segmentation benchmarks, including YouTube [28], DAVIS [18], MOSE [7], and SA-CO [2]. To further evaluate attack performance on image segmentation, following UAP-SAM2 [35], we randomly sample 100 videos from YouTube [28], DAVIS [18], MOSE [7], and SA-CO [2], and extract 15 frames per video to construct image-level test sets. We resize all frames to $3 \times 1008 \times 1008$. We use official SAM3 [2], SAM3.1 [2], and EfficientSAM3 [29] as victim models for evaluation.

Attack settings. Following [35, 38], we set the perturbation upper bound for AdvPCS to $10/255$, with a batch size of 1 and 15 training epochs. We fix the random seed to 30 in all experiments for reproducibility. We set λ to 5 and τ to $\log(0.01/0.99)$. To fully evaluate the attack performance of AdvPCS under different prompts, we test it on SAM3 and its variants with point, box, and text prompt modes. To better assess transferability, we train all methods on the SA-CO dataset and test them on other datasets. For video segmentation, we optimize on 100 randomly sampled videos with 15 frames each. For image segmentation, we use 1500 frames for optimization.

Evaluation metrics. We use the *mean Intersection over Union* (mIoU) to evaluate performance. It measures the average overlap between predicted masks and ground-truth masks. A lower mIoU indicates a stronger attack. We denote the mIoU computed on adversarial inputs as *Adversarial mIoU* (AoU) and the mIoU computed on benign inputs as *benign mIoU* (BoU).

Table 2: The mIoU (%) of comparison study. The best results are highlighted in bold.

Method	Video Segmentation								Image Segmentation							
	SA-CO	YouTube		MOSE		DAVIS		Avg	SA-CO	YouTube		MOSE		DAVIS		Avg
	Text	Point	Box	Point	Box	Point	Box		Text	Point	Box	Point	Box	Point	Box	
Benign	68.48	85.29	88.28	69.12	75.15	69.73	88.14	77.74	87.54	74.17	88.19	78.90	86.54	68.21	87.98	81.65
AttackSAM [30]	53.12	80.39	85.69	65.60	70.38	68.03	87.10	72.90	63.54	61.21	83.24	58.79	81.77	57.93	85.20	70.24
DarkSAM [38]	56.16	83.29	86.97	66.20	73.57	62.04	87.44	73.67	70.22	64.79	86.68	65.88	83.81	61.97	86.78	74.30
S-RA [23]	16.74	32.23	51.18	34.43	44.16	34.03	67.40	40.02	21.63	11.28	22.59	21.15	39.74	17.43	27.22	23.01
UAD [16]	6.55	25.79	50.26	24.14	33.14	33.12	53.15	32.31	7.47	12.43	48.87	14.08	46.68	12.67	47.84	27.15
UAP-SAM2 [35]	1.18	22.37	53.96	13.44	36.62	24.23	57.01	29.83	2.86	4.42	19.21	3.23	26.41	4.11	20.84	11.58
Ours	**0.00**	**6.29**	**39.62**	**8.17**	**25.20**	**7.89**	**41.13**	**18.33**	**0.07**	**0.09**	**4.18**	**0.84**	**7.95**	**0.32**	**3.95**	**2.49**



(a) Module (b) Weight (c) Epoch (d) Epsilon

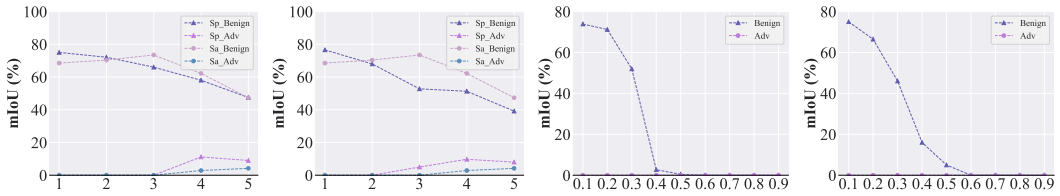
Figure 5: Ablation results on the effect of different factors on the attack performance of AdvPCS

4.2 Attack Performance

In this section, we evaluate the attack performance of AdvPCS on both video and image segmentation tasks, analyzing its effectiveness across datasets, models, and prompt types. We report quantitative results in Table 1. All UAPs are trained only on SA-CO and directly transferred to other datasets and settings. From the overall results, adversarial inputs consistently lead to a significant drop in mIoU across all datasets, models, and prompt types. This trend holds in both video and image settings, demonstrating the strong effectiveness of the proposed method. For instance, in video tasks, AdvPCS reduces the average mIoU from 70.24% to 4.61% on SA-CO under text prompts, and achieves low average mIoU values of 17.48% on YouTube and 25.90% on DAVIS under point prompts. A similar trend is observed in image settings, confirming strong cross-dataset transferability. The attack is effective across all prompt types, including point, box, and text. Text prompts exhibit the most pronounced degradation, likely due to their higher reliance on high-level semantic reasoning. On the SA-CO dataset, the image mIoU drops substantially from 78.42% to 3.96%, indicating that the model nearly fails to segment the target under adversarial perturbations. Point and box prompts also show clear performance degradation, but their AoU values remain relatively higher. In addition to cross-dataset transferability, as illustrated in Fig. 4, we further evaluate the transferability of AdvPCS across different SAM-series models and prompt instances. Fig. 4(a) shows the text transfer grid from SA-CO to other models, indicating that adversarial examples generated for one model can also degrade the segmentation performance of other models, demonstrating notable cross-model transferability. Fig. 4(b)-(d) show the attack performance on five selected groups under text, point, and box prompts, respectively. For each prompt type, all prompt instances experience consistently strong attack effects, indicating that AdvPCS maintains high cross-prompt transferability across different input prompts.

4.3 Comparison Study

In this section, we compare AdvPCS with SOTA methods, including Attack-SAM [30], Dark-SAM [38], S-RA [23], UAD [16], and UAP-SAM2 [35]. All baselines are adapted to a universal adversarial attack framework with the same optimization settings to ensure fair comparison. Table 2 reports the quantitative results on both video and image segmentation tasks. Across all datasets, AdvPCS consistently achieves the lowest mIoU values, indicating the strongest attack performance. For instance, in video segmentation, AdvPCS reduces the average mIoU to 18.33%, whereas the second-best method, UAP-SAM2, achieves 29.83%. On the SA-CO dataset under text prompts, AdvPCS completely eliminates segmentation performance, achieving 0.00% mIoU compared with 1.18% for UAP-SAM2. For image segmentation, AdvPCS also outperforms all baselines, reducing



(a) SAM3-Corruption (b) SAM3.1-Corruption (c) SAM3-Pruning (d) SAM3.1-Pruning

Figure 6: Results of defense evaluation. (a)-(b): Data pre-processing, (c)-(d): Model pruning

the average mIoU to 2.49%, far below the next best result of 11.58% by UAP-SAM2. We also provide visual comparisons of adversarial examples generated by these attacks in Sec. E.

4.4 Ablation Study

In this section, we explore the effect of four key factors on the attack performance of AdvPCS. The victim model is SAM3, and all experiments are conducted on the SA-CO dataset.

The effect of modules. We explore the contribution of each component within AdvPCS to its overall attack performance. We remove each component individually while retaining all other components to evaluate its individual impact. For simplicity, we denote the $\mathcal{J}_{local}$, $\mathcal{J}_{global}$, $\mathcal{J}_{ta}$, $\mathcal{J}_{reg}$, and min–max bilevel adversarial prompt optimization strategy as A–E, respectively. The results shown in Fig. 5(a) indicate that removing any single module significantly reduces the attack efficacy compared with the complete AdvPCS. This demonstrates that all modules are essential for achieving optimal performance. The complete AdvPCS consistently achieves the lowest mIoU across point, box, and text prompts.

The effect of hyperparameter λ. We study the effect of the weight λ assigned to box prompts on attack performance. Experiments are conducted with box prompt weights ranging from 1 to 6. As shown in Fig. 5(b), attack performance increases with higher weights and reaches a peak at a weight of 5, beyond which further increases yield minimal improvements. Based on these results, a weight of 5 is adopted as the default configuration.

The effect of training epochs. We examine the effect of varying the number of training epochs, ranging from 1 to 20, on AdvPCS performance. Fig. 5(c) shows that attack performance rises rapidly in the early epochs and stabilizes after 15 epochs. Since additional epochs offer only marginal improvements, 15 epochs are selected as the default setting.

The effect of perturbation budget ϵ. We evaluate AdvPCS with perturbation budgets ranging from $2/255$ to $32/255$. As shown in Fig. 5(d), larger ϵ generally results in stronger attacks. Even with a small budget of $4/255$, AdvPCS significantly reduces the mIoU for all prompt types, demonstrating the method's effectiveness under limited perturbations.

4.5 Defense Study

In this section, we evaluate the robustness of AdvPCS against common defense strategies applied to SAM3-series models, including data pre-processing [19] and model pruning [39].

Data pre-processing. We apply the Spatter (Sp) and Saturate (Sa) corruptions at severity levels from 0 to 5 on both SAM3 and SAM3.1 on SA-CO. As shown in Fig. 6(a) and Fig. 6(b), increasing the corruption severity causes the mIoU of benign samples to drop from approximately 70% to 40%, while the mIoU of adversarial examples remains consistently below 15%. This indicates that AdvPCS can still effectively degrade model performance even under strong pre-processing defenses.

Model pruning. We evaluate AdvPCS under varying pruning ratios on SAM3 and SAM3.1 using the DAVIS dataset. Fig. 6(c) and Fig. 6(d) show that as the pruning ratio increases from 0 to 0.6, the mIoU of benign samples drops sharply from approximately 70% to below 5%, whereas the mIoU of adversarial examples remains close to 0%. These results suggest that model pruning provides very limited effectiveness against AdvPCS.

5 Conclusions, Limitations, and Broader Impact

In this paper, we present AdvPCS, a novel cross-prompt universal adversarial attack against PCS models. By explicitly targeting both cross-prompt transferability and perception-level distortion, Ad-

vPCS generates a highly transferable universal adversarial perturbation that generalizes across diverse prompts and frames. AdvPCS integrates a min–max bilevel adversarial prompt optimization strategy, a global–local perception deception attack that suppresses global and local presence probabilities, and a temporal transition deviation attack that disrupts semantic consistency across frames. Extensive experiments on four datasets and three models demonstrate that AdvPCS consistently outperforms SOTA attacks, revealing critical vulnerabilities in emerging PCS models.

Despite its effectiveness, our method has two limitations. First, the min–max bilevel optimization incurs additional computational overhead due to prompt exploration and selection. Second, AdvPCS specifically targets concept-level presence prediction in PCS models, and may not generalize to conventional segmentation models or earlier frameworks such as SAM and SAM2. This work exposes security risks in emerging PCS models, particularly adversarial examples with cross-prompt and cross-frame transferability, and highlights the potential misuse of adversarial attacks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62472057 and 62372196. This work was also supported by the Open Fund of Sichuan Engineering Technology Research Center for Data Security and Privacy Protection (Grant No. SCDSPP-2026-4). Yufei Song is the corresponding author.

References

[1] Mario Camarena, Het Patel, Fatemeh Nazari, Evangelos Papalexakis, Mohamadhossein Noruzoliaee, and Jia Chen. Ad-sam: Fine-tuning the segment anything vision foundation model for autonomous driving perception. *arXiv preprint arXiv:2510.27047*, 2025.

[2] 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. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR'26)*, 2026.

[3] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 3367–3375, 2023.

[4] Jian Cheng. Ur-sam: Urban road segmentation in autonomous driving scenes based on sam. In *Proceedings of the 3rd International Conference on Algorithm, Image Processing and Machine Vision (AIPMV'25)*, pages 126–130. IEEE, 2025.

[5] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[6] Francesco Croce and Matthias Hein. Segment (almost) nothing: Prompt-agnostic adversarial attacks on segmentation models. In *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML'24)*, pages 425–442. IEEE, 2024.

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H. S. Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 20224–20234. IEEE, 2023.

[8] 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. *arXiv preprint arXiv:2310.12431*, 2023.

[9] 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*, 14(8):3312, 2024.

[10] Yan Jiang, Guisheng Yin, Ye Yuan, Jingjing Chen, and Zhipeng Wei. Cross-point adversarial attack based on feature neighborhood disruption against segment anything model. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'24)*, pages 1–6. IEEE, 2024.

[11] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS'23)*, volume 36, pages 29914–29934, 2023.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 3992–4003. IEEE, 2023.

[13] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 19499–19509, 2024.

[14] Yuchen Liu and Pengxu Wei. Cross-prompt adversarial attack on segment anything model. In *Proceedings of the 12th International Conference on Communications and Broadband Networking (ICCBN'24)*, pages 34–39, 2024.

[15] 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 (AAAI'25)*, volume 39, pages 5775–5783, 2025.

[16] Jiahao Lu, Xingyi Yang, and Xinchao Wang. Unsegment anything by simulating deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, pages 24294–24304, 2024.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations (ICLR'18)*, 2018.

[18] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[19] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'18)*, pages 8571–8580, 2018.

[20] 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.

[21] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Alwala Kalyan Vasudev, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR'25)*, 2025.

[22] Sourya Sengupta, Satrajit Chakrabarty, and Ravi Soni. Is sam 2 better than sam in medical image segmentation? In *Proceedings of the Medical Imaging 2025: Image Processing (SPIE'25)*, volume 13406, pages 666–672. SPIE, 2025.

[23] 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 (CVPR'24)*, pages 194–203, 2024.

[24] Yufei Song, Ziqi Zhou, Qi Lu, Hangtao Zhang, Yifan Hu, Lulu Xue, Shengshan Hu, Minghui Li, and Leo Yu Zhang. Segtrans: Transferable adversarial examples for segmentation models. *IEEE Transactions on Multimedia*, 2025.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS'17)*, volume 30, 2017.

[26] Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, and Xiaodan Liang. X-sam: From segment anything to any segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'26)*, volume 40, pages 26187–26196, 2026.

[27] Song Xia, Wenhan Yang, Yi Yu, Xun Lin, Henghui Ding, Lingyu Duan, and Xudong Jiang. Transferable adversarial attacks on sam and its downstream models. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS'24)*, pages 87545–87568, 2024.

[28] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[29] Chengxi Zeng, Yuxuan Jiang, and Aaron Zhang. Efficientsam3: Progressive hierarchical distillation for video concept segmentation from sam1, 2, and 3. *arXiv preprint arXiv:2511.15833*, 2025.

[30] 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.

[31] Jing Zhang, Zhikai Li, Xuewen Liu, and Qingyi Gu. Efficient-sam2: Accelerating sam2 with object-aware visual encoding and memory retrieval. In *Proceedings of the fourteenth International Conference on Learning Representations (ICLR'26)*, 2026.

[32] Sheng Zheng, Chaoning Zhang, and Xinhong Hao. Black-box targeted adversarial attack on segment anything (sam). *IEEE Transactions on Multimedia*, 27:1901–1913, 2024.

[33] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM'23)*, pages 6311–6320, 2023.

[34] Ziqi Zhou, Shengshan Hu, Ruizhi Zhao, Qian Wang, Leo Yu Zhang, Junhui Hou, and Hai Jin. Downstream-agnostic adversarial examples. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV'23)*, pages 4345–4355, 2023.

[35] 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. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS'25)*, 2025.

[36] Ziqi Zhou, Bowen Li, Yufei Song, Shengshan Hu, Wei Wan, Leo Yu Zhang, Dezhong Yao, and Hai Jin. Numbod: A spatial-frequency fusion attack against object detectors. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence (AAAI'25)*, 2025.

[37] Ziqi Zhou, Minghui Li, Wei Liu, Shengshan Hu, Yechao Zhang, Wei Wan, Lulu Xue, Leo Yu Zhang, Dezhong Yao, and Hai Jin. Securely fine-tuning pre-trained encoders against adversarial examples. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP'24)*, 2024.

[38] 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. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS'24)*, 2024.

[39] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

A Contents

- Sec. B: Experimental settings including datasets, evaluation metrics, and platform.
- Sec. C: Stability analysis under different random seeds.
- Sec. D: Multi-point evaluation study of the proposed method under multi-prompt settings.
- Sec. E: Supplementary visualization results of the comparison study.

B Experimental Setting

In this section, we provide details of our experimental settings.

B.1 Datasets

We evaluate our method on four benchmark datasets: YouTube-VOS [28], DAVIS [18], MOSE [7], and SA-CO [2].

- **YouTube:** YouTube [28] is a large-scale benchmark for video object segmentation. It contains 3,471 training videos, 474 validation videos, and 508 test videos, covering 94 object categories with pixel-level segmentation annotations. The dataset is widely used for video object segmentation, semi-supervised segmentation, and multi-object tracking.
- **DAVIS:** DAVIS [18] is a standard benchmark for video object segmentation. The DAVIS 2017 split contains 60 training videos, 30 validation videos, and 30 test-dev videos with high-quality pixel-level annotations. It is commonly used to evaluate precise object boundary segmentation and temporal consistency in multi-object video segmentation.
- **MOSE:** MOSE [7] is a large-scale benchmark for video object segmentation in complex scenes. It contains 1,507 training videos, 311 validation videos, and 331 test videos with pixel-level segmentation annotations. The dataset focuses on challenging scenarios such as heavy occlusion, object disappearance and reappearance, and frequent object interactions.
- **SA-CO:** SA-CO [2] is a large-scale dataset for open-vocabulary video segmentation. Its video component contains 52.5K videos and 467K spatiotemporal mask sequences with phrase-level segmentation annotations. The dataset supports research on text-prompted video segmentation, tracking, and open-vocabulary understanding.

B.2 Evaluation Metrics

We use *mean Intersection over Union* (mIoU) to evaluate segmentation performance. mIoU computes the average *Intersection over Union* (IoU) across all categories, where IoU measures the overlap between the predicted mask and the ground-truth mask:

$$\text{IoU} = \frac{\text{Predicted Region} \cap \text{Ground Truth Region}}{\text{Predicted Region} \cup \text{Ground Truth Region}}.$$

mIoU is the average of the IoUs for all categories, reflecting the model's overall performance in the segmentation task. A higher mIoU indicates better segmentation performance across categories, particularly in tasks with class imbalance or fine-grained segmentation.

B.3 Platform

We conduct experiments on a server running a 64-bit Ubuntu 20.04.1 system with an Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz processor, 125GB memory, and two NVIDIA A100-SXM4 GPUs, each with 80GB memory. The experiments are performed using the Python language and PyTorch library version 2.1.0.

C Stability Analysis

We examine whether seed-dependent data sampling affects evaluation results by replicating the AdvPCS attack under different random seed configurations. We further introduce three additional seeds to assess sensitivity to sampling variations. We use SAM3 as the victim model and evaluate on four video segmentation benchmarks: DAVIS, YouTube, MOSE, and SA-CO. Results in Fig. A1 summarize the variation in attack performance across seeds. The consistently small error bars indicate that AdvPCS maintains stable performance across all datasets. These results show that AdvPCS is largely insensitive to seed initialization, confirming its robustness under different random sampling conditions.

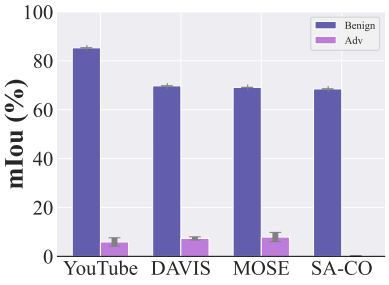


Figure A1: Stability analysis

D Multi-point Evaluation Study

We further analyze the effect of multi-prompt evaluation settings on AdvPCS. Using SAM3 as the victim model, we evaluate adversarial segmentation with 1–5 input points on three datasets: YouTube, DAVIS, and MOSE. Results in Tab. A1 show that increasing the number of prompt points provides richer spatial cues, which partially reduces the effect of adversarial perturbations. Nevertheless, even under multiple point prompts, AdvPCS maintains strong attack effectiveness and consistently disrupts SAM3's ability to produce coherent and semantically consistent segmentation. These results demonstrate that AdvPCS generalizes well across diverse prompting configurations, confirming its stability and effectiveness in multi-prompt settings.

Table A1: The mIoU (%) of AdvPCS under the multi-point evaluation mode.

Prompt	Prompt Number	Video Segmentation							
		YouTube		MOSE		DAVIS		Avg	
		BoU	AoU	BoU	AoU	BoU	AoU	BoU	AoU
Point	1	85.29	6.29	69.12	8.17	69.73	7.89	74.71	7.45
	2	82.64	17.61	71.92	12.67	82.08	16.84	78.88	15.71
	3	87.89	19.22	73.30	14.37	86.36	22.01	82.52	18.53
	4	87.62	22.12	75.14	16.61	86.52	23.02	83.09	20.58
	5	88.16	20.83	75.66	15.44	86.22	22.85	83.35	19.71

E Visualization Results of Comparison Study

We provide visualization results of comparative studies against Attack-SAM [30], DarkSAM [38], S-RA [23], UAD [16], and UAP-SAM2 [35]. We select SAM3 as the victim model and evaluate all methods on video segmentation tasks across the DAVIS dataset using point prompts. As shown in Fig. A2, AdvPCS outperforms all existing attack methods in video segmentation tasks.



Figure A2: Visualization of adversarial examples generated by the proposed method and existing methods on the DAVIS dataset for SAM3.