Transferable Attacks on Open-Vocabulary Video Instance Segmentation via Dual-Objective Triggers

Minghao Shou[1], **Kesen Wang**[2], **Tong Zhang**[1], **Han Bao**[1] and **Zonghui Wang**[1]
[1]Zhejiang University
[2]King Abdullah University of Science and Technology
shouminghao@zju.edu.cn, kesen.wang@alumni.kaust.edu.sa, {tz_zju, baohan21, zhwang}@zju.edu.cn

Abstract

Open-vocabulary video instance segmentation (OV-VIS) couples spatial-temporal reasoning with language grounding, yet its adversarial robustness has remained largely unexplored. We present the Dual-Objective Triggers (DOT), the first transferable attack on OV-VIS that simultaneously exploits the vision–language coupling and temporal coherence. DOT deploys a Dual Semantic Perturbation Module that combines two complementary triggers: a Semantic Suppression Trigger disrupts the alignment between the true object and the query, while a Plausible Replacement Trigger steers the tracker toward a phantom trajectory that is visually plausible and text-consistent. To amplify cross-model transferability without sacrificing perceptual fidelity, we introduce Phase-Guided Adversarial Training, which injects perturbations primarily in the phase spectrum while blending amplitudes with clean references. Extensive experiments on four state-of-the-art OV-VIS implementations demonstrate that DOT reduces mAP by up to 69.3% and achieves attack success rate of up to 98%, outperforming the strongest baseline by a factor of 1.6× on average, while maintaining a PSNR of 52.4 dB, thus exposing critical security vulnerabilities and laying a foundation for future research on robust and trustworthy vision–language systems.

1 Introduction

Video instance segmentation (VIS) is a fundamental task in video understanding, aiming to segment and track object instances across a video sequence. Deep learning models have greatly improved VIS performance [Zhang *et al.*, 2023b; Zheng *et al.*, 2024a; Lee *et al.*, 2024]. Advancing this paradigm, Open-vocabulary video instance segmentation (OV-VIS) extends this by incorporating language guidance, allowing segmentation of arbitrary categories defined by text. Despite these achievements, the adversarial robustness of OV-VIS remains underexplored. In particular, its vulnerability to transfer-based adversarial attacks has not been systematically studied. While prior work [Gu *et al.*, 2022; Jia *et al.*, 2023] has demonstrated the susceptibility of traditional image-based segmentation models to adversarial perturbations, existing methods perform poorly on OV-VIS, failing to reveal its inherent risks. This gap is not merely a theoretical gap but a practical security concern; it leaves a clear opening for adversaries.

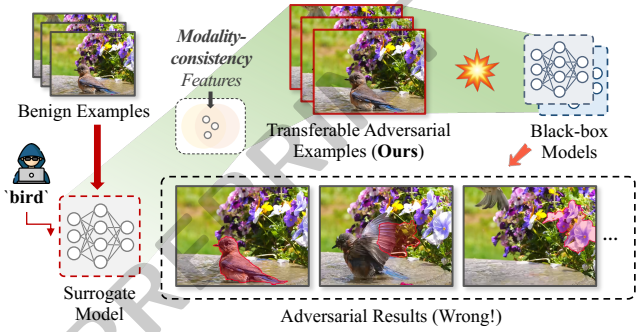

Figure 1: Given a user-specified text query and a surrogate OV-VIS model, our Dual-Objective Triggers (DOT) craft transferable adversarial videos that fool unseen OV-VIS models, causing their predicted masks to drift spatially (erroneous regions highlighted in red).

OV-VIS systems now underpin safety-critical applications, *e.g.*, autonomous driving and video surveillance. However, adversaries can manipulate textual queries or exploit model vulnerabilities to suppress or misdirect the segmentation of critical objects, thereby causing negative economic and societal impacts. These concerns underscore the urgency of systematically investigating adversarial attacks on OV-VIS. The interplay between open-vocabulary queries and the requirement for spatio-temporal consistency of instance masks presents distinctive challenges for attack design. Concretely, in a typical attack scenario, the adversary designs to mislead the model's segmentation mask, causing it to deviate from the intended target. Simultaneously, the attacker must preserve spatio-temporal coherence among instances; otherwise, excessive disorder would render the attack conspicuous and easily detected. The difficulty is further exacerbated in real-world scenarios, where OV-VIS models vary significantly in terms of architecture, training data, and vision-language fusion strategies. Such heterogeneity introduces substantial uncertainty, preventing the adversary from reliably inferring the

model in use. Consequently, any practical attack must therefore generalize across architectures, data distributions, and fusion strategies to achieve high transferability.

Transferable adversarial attacks [Goodfellow *et al.*, 2014; Wei *et al.*, 2018; Yin *et al.*, 2023; Chen *et al.*, 2025a] generally craft input perturbations that generalize across a set of held-out models; however, design assumptions inherited from static single-modality image regimes substantially limit their effectiveness in open-vocabulary video instance segmentation. An effective perturbation must first disrupt the fine-grained alignment between free-form textual queries and visual content, yet existing methods often neglect the linguistic modality or replace it with sparse geometric cues such as points or bounding boxes [Zheng *et al.*, 2024b; Zhang *et al.*, 2023a], thereby failing to exert direct influence over the language-conditioned mask prediction. Attacks that manipulate only a subset of frames fail to preserve temporal coherence, inducing perceptible flickering artifacts, and allow recurrent inference to recover the correct target in later frames [Pony *et al.*, 2021; Li *et al.*, 2023]. Moreover, the pronounced heterogeneity of OV-VIS architectures, training corpora, and fusion strategies leads to a sharp decline in transfer performance once a perturbation is evaluated outside its development family. These observations indicate that a practical transferable attack for OV-VIS must jointly optimize over vision and language and remain robust to architectural diversity. These qualities current approaches do not yet exhibit.

To tackle the above challenges, we propose Dual-Objective Triggers (DOT) in Figure 1, the first transfer-oriented adversarial framework for OV-VIS. Specifically, DOT introduces two modules, namely a **Dual Semantic Perturbation Module** (DSPM) and a **Phase-Guided Adversarial Training strategy** (PGAT). DSPM is designed to produce transferable perturbations with two complementary adversarial triggers. These triggers break the association between the text query and the genuine object while keeping spatio-temporal consistency. The first trigger, termed the Semantic Suppression Trigger, aims to weaken or erase the semantic features of the original target object in the video. The second trigger, named Plausible Replacement Trigger, leads the model to generate a misleading yet coherent instance trajectory and align the trajectory semantically with the corresponding class. To balance stealth and efficacy, we further devise a Phase-Guided Adversarial Training strategy that blends clean and adversarial amplitudes while keeping the adversarial phase intact. This suppresses visible artifacts and lets the phase carry the semantic structure that drives transferability.

In brief, our contributions can be summarized as follows:

1. To our knowledge, this is the first to systematically investigate how open-vocabulary video instance segmentation behaves under cross-model adversarial settings, revealing that current state-of-the-art systems remain highly vulnerable despite their strong spatio-temporal and language–vision coupling.
2. We introduce a novel framework that couples a Semantic Suppression Trigger with a Plausible Replacement Trigger to break cross-modal alignment and preserve realistic video dynamics, and we further enhance its transferability and imperceptibility through a Phase-Guided Adversarial Training strategy.
3. Extensive experiments on four state-of-the-art OV-VIS architectures show that our method DOT cuts mAP by up to 69.3% while retaining a perceptual quality of 52.4 dB PSNR, outperforming the strongest baseline by an average factor of 1.6×.

2 Related Works

Open–Vocabulary Video Instance Segmentation. OV-VIS treats classification, tracking, and segmentation under a single language-conditioned framework. The task was formalized by OV2Seg [Wang *et al.*, 2023], which introduced the large-scale LV-VIS benchmark and utilized a memory-induced transformer to propagate masks to unseen classes. Subsequent works refined the query design: InstFormer [Guo *et al.*, 2025] enriches instance tokens with CLIP text embeddings and re-trains only lightweight category heads, markedly improving cross-category generalization; CLIP-VIS [Zhu *et al.*, 2024] repurposes a frozen CLIP image encoder by adding a lightweight video head that propagates CLIP-based queries across frames, preserving zero-shot semantics with minimal extra computation; and OVFormer [Fang *et al.*, 2024] jointly encodes video frames and text prompts via a transformer backbone with a small alignment layer to bridge instance queries and vision–language features. Recent studies emphasize efficiency, such as GLEE [Wu *et al.*, 2024], which handles multiple tasks in a unified foundation model, and TrOY-VIS [Yan *et al.*, 2025], which introduces a lightweight tracking-by-query paradigm to reduce per-frame computation.

Adversarial Attacks. Adversarial attacks craft imperceptible input perturbations to mislead deep models [Goodfellow *et al.*, 2014; Szegedy *et al.*, 2013]. For videos, Flickering-Noise (FN) offers cheap, physically deployable temporal flashes [Lu *et al.*, 2017], while Frame-by-Frame (F&F) per-frame optimization extends classic image attacks and inspires sparser variants [Zhou *et al.*, 2023b; Mu *et al.*, 2021; Wang *et al.*, 2021]. SVASTIN explores the full spatio-temporal volume with transferable patches resilient to motion and deformation [Pan *et al.*, 2024]. Universal schemes such as BTC and US-Attack learn a single perturbation that generalizes across clips [Kim *et al.*, 2023; Li and Wang, 2021], whereas ARA limits gradient computation to Regions of Interest for interpretable, task-oriented noise [Li *et al.*, 2023]. The rise of open-vocabulary detectors has further revealed vulnerabilities in cross-modal alignment [Chhipa *et al.*, 2024; Schlarmann *et al.*, 2024; Zhang *et al.*, 2025; Nguyen Tiet Nguyen *et al.*, 2024]. However, existing video attacks are ill-suited to OV-VIS. FN and F&F perturb early detections or association graphs but leave the language branch untouched. SVASTIN targets per-frame semantic segmentation without enforcing temporal coherence. ARA assumes fixed categorical priors and thus fails to mislead open-set queries.

3 Preliminary

In this section, we first briefly introduce open-vocabulary video instance segmentation. Then, we formally define the threat model for a transfer attack on OV-VIS.

3.1 Open-Vocabulary Video Instance Segmentation

Given a training dataset $\mathcal{D}_{\text{train}}$ containing pixel-level annotations for a set of training categories $\mathcal{C}_{\text{train}}$, traditional VIS aims to train a model $\mathcal{F}(\cdot)$ to be tested on a dataset $\mathcal{D}_{\text{test}} = \{V_i\}$, where $V_i \in \mathbb{R}^{T\times H\times W\times 3}$ refers to a video clip with spatial shape (H, W) and T frames. $\mathcal{F}(\cdot)$ is supposed to predict a sequence of segmentation masks $\mathcal{M}_{pred} = \{m_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W}$ and a category label $c \in \mathcal{C}_{\text{train}}$ for each object in videos from the training categories, while the objects from novel categories $\mathcal{C}_{\text{novel}}$ are ignored.

In contrast, Open-Vocabulary VIS aims to train a model $\mathcal{F}(\cdot)$ on $\mathcal{D}_{\text{train}}$, and then test on $\mathcal{D}_{\text{test}}$ for both training categories $\mathcal{C}_{\text{train}}$ and novel categories $\mathcal{C}_{\text{novel}}$. Let $\mathcal{C}$ denote the complete category set, defined as the union of training and novel classes: $\mathcal{C} \coloneqq \mathcal{C}_{\text{train}} \cup \mathcal{C}_{\text{novel}}$. Specifically, given a test video clip $V_i \in \mathbb{R}^{T\times H\times W\times 3}$ during inference, the trained model $F_s(\cdot)$ is required to predict the segmentation mask sequence $\mathcal{M}_{pred}$ and the category label $c \in \mathcal{C}$ for each object p in V_i:

$$\{m_1, m_2, \cdots, m_T\} = \mathcal{F}_\theta(V_i, c), \tag{1}$$

where $m_t \in \mathbb{R}^{H\times W}$ is the segmentation mask for each frame t. In the experiments, we term the training categories as *base categories*, while the categories disjoint from base categories are referred to as *novel categories*.

3.2 Threat Model

Attacker's Goals: Given a clean video V and the corresponding label c, we seek to generate a video-domain adversarial patch $\mathcal{P} = \{p_t\}_{t=1}^{T}$ to mislead the surrogate model $\mathcal{F}_s$ into producing a mask sequence that deviates from the ground truth and instead aligns with a predefined target mask sequence $\mathcal{M}_{\text{target}} = \{\widetilde{m}_t\}_{t=1}^{T}$:

$$\mathcal{M}_{\text{pred}} = \mathcal{F}_s(V \oplus \mathcal{P}, c) \approx \mathcal{M}_{\text{target}}, \tag{2}$$

where the operator $\oplus$ represents an addition of a video adversarial patch $\mathcal{P}$. To satisfy the imperceptibility constraint, the perturbation must remain nearly imperceptible to human observers and downstream preprocessing. We therefore impose a perceptual budget with $\varepsilon \ll 1$:

$$\| \mathcal{P} \|_\infty \leq \varepsilon. \tag{3}$$

Within these constraints, the attacker optimises $\mathcal{P}$ on a single surrogate model, yet expects it to transfer across heterogeneous unseen OV-VIS models encountered in the wild. By generating adversarial examples under these objectives and directly applying them to the target OV-VIS model, the attacker can influence the model's output.

Attacker's Capabilities: We assume a strict limited-access threat model that mimics realistic deployments: the attacker can access or train one surrogate OV-VIS network but has no knowledge of the architecture, parameters, or training data of the production model. The perturbation is crafted on this surrogate and then directly injected into the target stream—no gradient queries, score queries, or test-time fine-tuning are allowed. Consequently, the attack must rely purely on cross-model transferability to succeed against heterogeneous OV-VIS systems in the wild.

4 Transferable OV-VIS Attack

In this section, we present the proposed Dual-Objective Triggers (DOT) framework, a transfer-oriented adversarial attack framework, as illustrated in Figure 2. The framework comprises two key components: the Dual Semantic Perturbation Module (DSPM) and the Phase-Guided Adversarial Training (PGAT) strategy. Specifically, DSPM is designed to suppress the semantic features of the source object, while forcing the model to produce trajectories that are deceptive yet temporally coherent. Furthermore, PGAT exploits phase information in the frequency domain to enhance both the robustness and the cross-model transferability of the generated adversarial examples.

4.1 Dual Semantic Perturbation Module

OV-VIS tackles video instance segmentation conditioned on open-vocabulary text queries. In contrast to fixed-category benchmarks, the task requires models to jointly align free-form language with visual content and to preserve consistent instance identities across the video timeline. To effectively attack OV-VIS models, we aim to break the association between the target object and the given text description, while ensuring that the generated segmentation remains temporally coherent. Since both triggers are injected only at the input level and do not rely on model-specific gradients or internal features, DSPM is inherently architecture-agnostic and thus highly transferable.

Given a clean video $V = \{v_t\}_{t=1}^{T}$ and a corresponding textual query c, we utilize a surrogate OV-VIS model $\mathcal{F}_s$ to obtain a sequence of predicted instance masks $\mathcal{M} = \{m_1, m_2, \cdots, m_T\}$ over T frames of the target. Let the superscript s mark the source object and p mark the distractor (phantom). To induce predictions that are spatially and temporally consistent yet deviated from the original object location, we introduce two adversarial patches: Semantic Suppression Trigger $\mathcal{P}^s = \{p_t^s\}_{t=1}^{T}$ and Plausible Replacement Trigger $\mathcal{P}^p = \{p_t^p\}_{t=1}^{T}$. The purpose of $\mathcal{P}^s$ is to suppress or eliminate the semantic features of the original target object, thereby hindering the model's ability to recognize it. In contrast, $\mathcal{P}^p$ is designed to guide the model toward generating a misleading instance trajectory that remains coherent across space and time. The patch p_t^s is centered at the centroid $r_t^s \in \{0,1\}^{H\times W}$ of the predicted instance mask m_t, and $r_t^p \in \{0,1\}^{H\times W}$ is obtained by spatially shifting r_t^s with an offset $(\Delta x, \Delta y)$ to r_{t-1}^s. For the t-th frame, an adversarial frame is formulated as:

$$v_t^{\text{adv}} = v_t + r_t^s \odot p_t^s + r_t^p \odot p_t^p, \tag{4}$$

where $v_t \in \mathbb{R}^{H\times W\times 3}$ denotes the clean input frame and v_t^{adv} denotes its adversarial counterpart. p_t^s and p_t^p are applied to disjoint spatial regions defined by binary masks r_t^s and r_t^p.

After obtaining the adversarial video, we employ the surrogate model to infer the corresponding instance predictions.

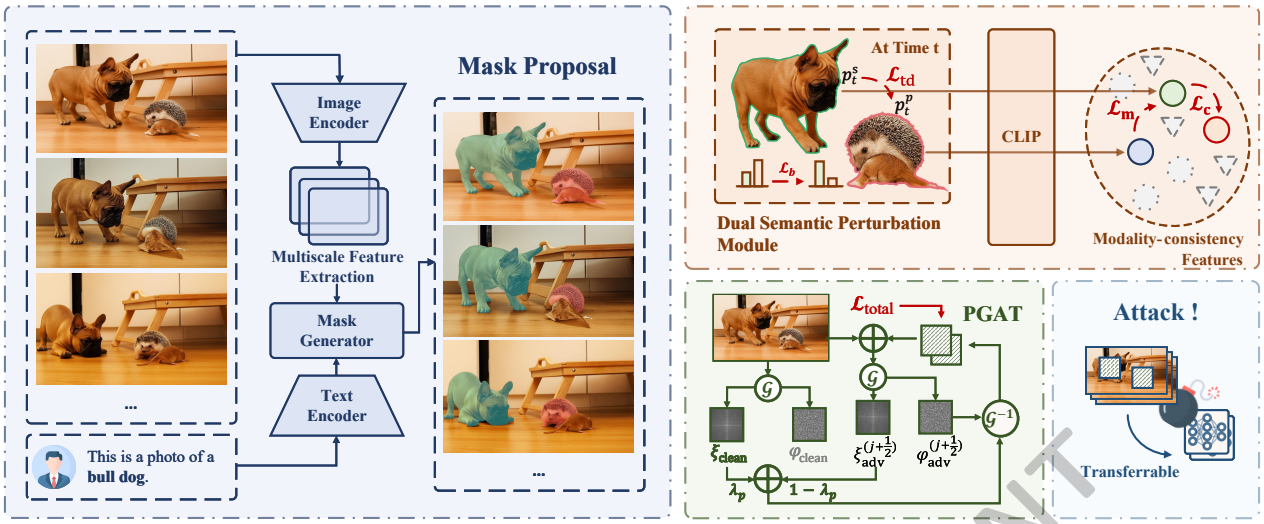

Figure 2: Overview of DOT. (a) A surrogate OV-VIS model first extracts candidate masks. (b) The Dual Semantic Perturbation Module (DSPM) applies two learnable triggers, including a Semantic Suppression Trigger (SST) and a Plausible Replacement Trigger (PRT), whose optimization is guided by CLIP-based supervision. (c) Phase-Guided Adversarial Training (PGAT) refines these triggers in the frequency domain by blending clean amplitude with adversarial phase. (d) The resulting imperceptible perturbations transfer to unseen OV-VIS models.

For a given adversarial video, in order to ensure that the predicted instances remain spatially and temporally aligned with the adversarial target while exhibiting controlled deviations, we introduce a trajectory deviation loss, which constrains the extent of spatiotemporal displacement between the predicted trajectory and the target trajectory.

$$\mathcal{L}_{\text{td}} = \frac{1}{T}\sum_{t=1}^{T}\Big(1 - \frac{2\,\|m_t^s \cdot m_t^p\|_1}{\|m_t^s\|_1 + \|m_t^p\|_1}\Big). \tag{5}$$

For a given adversarial video, to enforce consistency between the masks predicted by the model and the target masks $\mathcal{M}_{\text{target}} = \{\widetilde{m}_t\}_{t=1}^{T}$, we employ a pixel-wise binary cross-entropy (BCE) loss, formulated as

$$\mathcal{L}_{\text{b}} = -\frac{1}{T}\sum_{t=1}^{T}\frac{1}{H\times W}\sum_{i=1}^{H\times W}\Big[q_t^i \log(y_t^i) + (1-q_t^i)\,\log(1-y_t^i)\Big], \tag{6}$$

where $y_t^i \in [0,1]$ denotes the predicted foreground probability at pixel i in frame t, and $q_t^i \in \{0,1\}$ represents the corresponding binary value of pixel i in the target mask $\widetilde{m}_t$.

In our task, the optimization objective targets two adversarial components: the Semantic Suppression Trigger and the Plausible Replacement Trigger. To ensure these adversarial patches function as intended, we process the adversarial video on a frame-by-frame basis. Specifically, for the t-th frame, we randomly sample K regions centered around and encompassing the Semantic Suppression Trigger, denoted as $\{v_{t(i)}^s\}_{i=1}^K$, and crop K regions centered around the Plausible Replacement Triggers, denoted as $\{v_{t(i)}^p\}_{i=1}^K$. Each of the cropped regions is then resized to a uniform resolution and passed through the CLIP model to extract its corresponding feature representation as follows:

$$\begin{aligned}(x_{t(1)}^s, x_{t(2)}^s, \cdots, x_{t(K)}^s) &= f_{\text{CLIP}}(v_{t(1)}^s, v_{t(2)}^s, \cdots, v_{t(K)}^s),\\ (x_{t(1)}^p, x_{t(2)}^p, \cdots, x_{t(K)}^p) &= f_{\text{CLIP}}(v_{t(1)}^p, v_{t(2)}^p, \cdots, v_{t(K)}^p).\end{aligned} \tag{7}$$

After obtaining the cropped feature representations of the Semantic Suppression Trigger $\{\{x_{t(i)}^s\}_{i=1}^K\}_{t=1}^T$ and the Plausible Replacement Trigger $\{\{x_{t(i)}^p\}_{i=1}^K\}_{t=1}^T$, we introduce a contrastive loss to disrupt the original alignment between the source object's visual trajectory and its corresponding textual query, while simultaneously binding the Plausible Replacement Trigger region to the same textual query.

To achieve this, we propose a contrastive loss to push the Semantic Suppression Trigger features $x_{t(i)}^s$ away from the genuine victim class description y_{src} and closer to a fabricated class description y_{tgt}. At the same time, it pulls the Plausible Replacement Trigger features $x_{t(i)}^p$ toward y_{src}. The loss is defined as:

$$\mathcal{L}_{\text{c}} = -\frac{1}{TK}\sum_{t,i}\Big[\log\frac{e^{x_{t(i)}^s \cdot y_{\text{tgt}}}}{\sum_y e^{x_{t(i)}^s \cdot y}} + \log\frac{e^{x_{t(i)}^p \cdot y_{\text{src}}}}{\sum_y e^{x_{t(i)}^p \cdot y}}\Big], \tag{8}$$

where $\{y_j\}_{j=1}^{\mathcal{C}}$ denote the set of text embeddings for all $\mathcal{C}$ classes.

To further guide the model toward predicting the instance mask in the region of the Plausible Replacement Trigger, we introduce an additional margin loss to enforce that the similarity between the Plausible Replacement Trigger features $x_{t(i)}^p$ and the source class description y_{src} exceeds that between the Semantic Suppression Trigger features $x_{t(i)}^s$ and y_{src} by at least a margin γ:

$$\mathcal{L}_m = \frac{1}{TK}\sum_{t,i}\max(0, \gamma - [(x_{t(i)}^p \cdot y_{\text{src}}) - (x_{t(i)}^s \cdot y_{\text{src}})]), \tag{9}$$

where γ is a positive margin hyperparameter that controls the minimum separation between the two similarity scores. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{td}}\mathcal{L}_{\text{td}} + \lambda_{\text{b}}\mathcal{L}_{\text{b}} + \lambda_{\text{c}}\mathcal{L}_{\text{c}} + \lambda_m\mathcal{L}_m. \tag{10}$$

Note that every loss term is defined purely on surrogate outputs; no target-model feedback is required, which is critical for transfer attacks.

4.2 Phase-Guided Adversarial Training Strategy

Recent studies on frequency-domain representations of images have revealed that the phase spectrum encodes most of the structural and semantic information, while the amplitude spectrum mainly governs low-level appearance such as color and contrast [Chen *et al.*, 2025b; Li *et al.*, 2024; Zhou *et al.*, 2023a; Rahaman *et al.*, 2019]. Inspired by these findings, we ask: can adversarial optimization be guided to alter phase components primarily, so as to maximize feature-level impact while preserving visual plausibility? This question motivates our Phase-Guided Adversarial Training strategy, which explicitly regularizes perturbations in the frequency domain to encode semantics in phase and preserve amplitude statistics for natural appearance.

We implement this idea inside a PGD-style optimization that jointly updates the two patches. Let $\mathcal{L}_{\text{total}}$ denote the adversarial objective (contrastive / task losses + regularizers). At PGD iteration j the intermediate (pre-regularization) updates for the t-th frame are

$$\begin{aligned} p_t^{s(j+\frac{1}{2})} &= p_t^{s(j)} - \alpha \cdot \operatorname{sign}\big(\nabla_{p_t^s}\mathcal{L}_{\text{total}}\big), \\ p_t^{p(j+\frac{1}{2})} &= p_t^{p(j)} - \alpha \cdot \operatorname{sign}\big(\nabla_{p_t^p}\mathcal{L}_{\text{total}}\big), \end{aligned} \tag{11}$$

where α is the step size.

Following the gradient step, we apply a frequency-domain regularization to improve visual quality and to bias perturbations toward phase changes. Let $\mathcal{G}(\cdot)$ be the Discrete Fourier Transform (DFT) that returns amplitude and phase, and $\mathcal{G}^{-1}(\cdot)$ its inverse. For a patch $p_t^* \in \{p_t^s, p_t^p\}$, denote

$$\begin{aligned} (\xi_{\text{adv}}^{*(j+\frac{1}{2})},\ \varphi_{\text{adv}}^{*(j+\frac{1}{2})}) &= \mathcal{G}\big(p_t^{*(j+\frac{1}{2})}\big), \\ (\xi_{\text{clean}}^{*},\ \varphi_{\text{clean}}^{*}) &= \mathcal{G}\big(p_{t,\text{clean}}^{*}\big), \end{aligned} \tag{12}$$

where $p_{t,\text{clean}}^*$ is a clean reference patch (e.g., the initial patch or a local image crop without adversarial modification). We construct a regularized amplitude by linear interpolation:

$$\xi_{\text{r}}^{*(j+\frac{1}{2})} = \lambda_p \cdot \xi_{\text{clean}}^{*} + (1-\lambda_p)\cdot \xi_{\text{adv}}^{*(j+\frac{1}{2})},\ \lambda_p \in [0,1]. \tag{13}$$

We then reconstruct the spatial patch using the regularized amplitude and the adversarial phase, and enforce valid pixel range by clipping:

$$\begin{aligned} \tilde{p}_t^{*(j+1)} &= \mathcal{G}^{-1}\big(\xi_{\text{r}}^{*(j+\frac{1}{2})},\ \varphi_{\text{adv}}^{*(j+\frac{1}{2})}\big), \\ p_t^{*(j+1)} &= \text{Clip}\big(\tilde{p}_t^{*(j+1)},\ 0,\ 1\big). \end{aligned} \tag{14}$$

Intuitively, Eq. (13) preserves coarse appearance via the clean amplitude while allowing the phase φ_{adv} to encode adversarial structure that more strongly perturbs model features. Note we do not enforce a hard zeroing of amplitude changes; λ_p controls the trade-off between visual fidelity and attack flexibility.

4.3 Universal DOT (U-DOT)

To improve transferability across unseen videos, we further explore a universal variant (U-DOT), in which a shared pair of triggers $(\mathcal{P}_u^s, \mathcal{P}_u^p)$ are optimized jointly over a batch of videos rather than per instance:

$$\min_{\mathcal{P}_u^s, \mathcal{P}_u^p} \mathbb{E}_{V\sim\mathcal{D}}\big[\mathcal{L}_{\text{total}}(V, \mathcal{P}_u^s, \mathcal{P}_u^p)\big]. \tag{15}$$

This allows the learned triggers to generalize across diverse scenes with negligible additional cost.

5 Experiments

5.1 Experiment Setup

Datasets and Metrics In this study, we benchmark our method using the LV-VIS and YTVIS-21 dataset, both are public corpus for OV-VIS. We report the Performance Drop (Δ %) in mean Average Precision (mAP) and mean Average Recall (mAR). To quantify the spatial deviation of the predicted masks caused by perturbations, we adopt metrics mIoU-GT and ASR_{IoU}@30 (ASR in short) following the work of [Chen *et al.*, 2025b]. mIoU-GT measures the IoU between ground truth and predictions, while ASR evaluates the success rate of adversarial videos in distorting under a 30% IoU threshold.

Baselines We port four strong segmentation attacks from Multiple Object Tracking and Video Instance Segmentation: Flickering-Noise (FN), Frame-by-Frame (F&F), SVASTIN, and Adversarial Region Attack (ARA). We generate pseudo boxes using a frozen Mask2Former for FN and F&F. We evaluate two OV-VIS pipelines, CLIP-VIS and OVFormer, both combining a CLIP-style image–text encoder with a masked-attention decoder (Mask2Former-style). We utilize the officially released pre-trained checkpoints and evaluate attacks under a strict black-box protocol setting without any direct access to model weights or gradients.

5.2 Transfer Settings and Implementations

We use InstFormer [Guo *et al.*, 2025] as the surrogate OV-VIS model. Unless otherwise stated, all adversarial examples are generated once on InstFormer (surrogate) and evaluated without gradient access on every target model, faithfully reflecting a transfer-attack scenario. Both the SST and PRT are square patches whose side length is set to $0.8\times$ the shorter edge of the detected bounding box, constrained to $[48, 256]$ pixels to maintain a reasonable scale. The patches are initialized with zeros, and $K = 4$ crops per frame are randomly sampled. Perturbations are optimized for $J = 60$ iterations with step size $\alpha = 2/255$ under an ℓ_∞ budget of $\varepsilon = 16/255$. The displacements Δx and Δy are randomly constrained to 10%–20% of the source object's shorter edge. We set $\lambda_{\text{td}} = 1.0$, $\lambda_{\text{b}} = 1.0$, $\lambda_{\text{c}} = 1.0$, $\lambda_{\text{m}} = 1.0$, and $\lambda_{\text{p}} = 0.5$. All experiments are performed on an NVIDIA A100-80 GB GPU with seed 42. To select the phantom object description y_{tgt}, we adopt a simple yet effective strategy: we choose the category that has the lowest CLIP similarity to the source category y_{src}.

Model	Method	LV-VIS				YTVIS-21			
		mAP Δ ↑	mAR Δ ↑	IoU-GT ↓	ASR ↑	mAP Δ ↑	mAR Δ ↑	IoU-GT ↓	ASR ↑
CLIP-VIS [R-50]	FN Attack	9.1	7.8	13.8	81.8	3.7	13.2	14.2	84.4
	F&F Attack	19.2	**33.4**	9.4	88.2	25.4	10.2	18.8	87.0
	SVASTIN	17.0	18.0	9.6	86.1	24.7	12.0	13.6	90.4
	ARA	21.8	31.0	8.8	91.1	48.7	22.4	12.1	93.8
	DOT (Ours)	**36.4**	30.2	**3.9**	**98.9**	**55.2**	**38.9**	**7.8**	**97.2**
CLIP-VIS [Conv-B]	FN Attack	3.2	4.7	12.5	85.3	32.3	17.4	21.1	73.8
	F&F Attack	13.3	21.4	11.4	89.8	28.8	45.0	13.6	84.6
	SVASTIN	9.6	3.7	10.3	87.8	23.5	12.8	13.9	82.7
	ARA	16.7	8.7	10.0	92.0	20.8	39.9	12.7	87.5
	DOT (Ours)	**28.5**	**23.6**	**9.3**	**95.0**	**91.3**	**88.0**	**5.4**	**98.4**
OVFormer [R-50]	FN Attack	24.9	16.8	11.1	81.5	24.5	17.5	18.0	89.7
	F&F Attack	36.4	35.9	11.1	87.1	89.0	21.4	16.0	87.5
	SVASTIN	26.3	16.1	12.6	79.0	83.5	**86.0**	21.2	82.2
	ARA	42.1	31.0	10.5	92.4	85.5	49.7	16.1	95.4
	DOT (Ours)	**69.3**	**61.2**	**9.2**	**98.0**	**91.5**	64.1	**5.6**	**98.6**
OVFormer [Swin-B]	FN Attack	28.4	20.1	10.6	84.4	38.4	33.6	14.7	86.7
	F&F Attack	41.7	46.8	**10.4**	84.1	85.0	84.3	31.3	88.9
	SVASTIN	16.7	7.0	13.7	79.0	84.1	82.8	25.6	81.4
	ARA	41.2	32.1	10.4	91.7	88.0	87.7	15.3	94.2
	DOT (Ours)	**62.8**	**52.8**	11.2	**97.9**	**94.6**	**92.4**	**6.0**	**98.2**

Table 1: Quantitative comparison of four different adversarial attack methods on the dataset LV-VIS and YTVIS-21 with surrogate model InstFormer. ↑ indicates higher is better, and ↓ indicates lower is better. Best results are **bold**, second best are underlined.

	InstFormer				CLIP-VIS				OVFormer			
	mAP	mAR	IoU-GT	ASR	mAP	mAR	IoU-GT	ASR	mAP	mAR	IoU-GT	ASR
InstFormer	–	–	–	–	36.4	30.2	3.9	98.9	69.3	61.2	9.2	98.0
CLIP-VIS	43.1	36.4	8.4	96.7	–	–	–	–	37.2	30.4	9.1	96.2
OVFormer	69.3	62.9	7.3	97.2	49.8	41.5	7.9	98.9	–	–	–	–

Table 2: Cross-model transfer attack performance. Rows indicate the surrogate model used to craft adversarial examples; columns indicate the black-box target model evaluated. Values represent the drop (Δ) in mAP, mAR, mIoU-GT, and the Attack Success Rate (ASR) on LV-VIS. Diagonal entries (–) are omitted as they correspond to the white-box setting rather than transfer.

5.3 Experimental Results

Table 1 summarises DOT's performance on the LV-VIS and YTVIS-21 benchmarks across two OV-VIS architectures and two backbones. DOT consistently ranks first or second across all metrics and shows clear advantages on precision-oriented indicators.

Among the strongest baselines, ARA and F&F operate in feature space. On CLIP-VIS[ResNet-50], ARA achieves slightly higher recall gain, while F&F nearly matches DOT on OVFormer[Swin-B]. In contrast, DOT prioritises mislocalization through aggressive semantic suppression, resulting in comparable recall degradation but stronger localisation errors, reflected by higher IoU-GT under strong backbones.

Table 2 further demonstrates DOT's cross-architecture transferability. DOT induces substantial performance drops across models, with mAP reductions up to 69.3% and consistently high ASR (often above 96%), indicating robust, high-confidence mislocalisation even under transfer settings.

These results validate the effectiveness of DOT's dual-trigger design: semantic suppression disrupts correct grounding, while plausible replacement enforces coherent false trajectories, yielding strong architectural generality. Standard defenses such as JPEG compression and Bit-Red show only marginal mitigation effects.

5.4 Ablation Study

We perform an ablation study to assess the contributions of the DSPM and PGAT modules. Additional ablations on loss weights, patch size, and crop count show that DOT is robust to hyperparameter variations.

Effectiveness of the DSPM Module

Table 3 indicates that the trajectory-deviation loss $\mathcal{L}_{\text{td}}$ is essential for recall degradation, while the cross-modal contrastive loss $\mathcal{L}_{\text{c}}$ mainly governs precision by decoupling visual and textual semantics.

Trigger ablations reveal complementary effects: removing Plausible Replacement Trigger causes the largest precision drop, whereas removing Semantic Suppression Trigger mainly harms recall and leads to the lowest ASR, showing that a coherent phantom trajectory is crucial for effective attacks.

	mAP Δ ↑	mAR Δ ↑	ASR ↑
Ours	36.4%	30.2%	98.9%
w/o $\mathcal{L}_{\text{td}}$	20.5%	6.9%	86.7%
w/o $\mathcal{L}_{\text{b}}$	14.9%	19.8%	84.0%
w/o $\mathcal{L}_{\text{c}}$	11.9%	17.7%	86.0%
w/o $\mathcal{L}_{\text{m}}$	24.7%	20.8%	85.7%
w/o PRT	31.1%	17.1%	80.1%
w/o SST	25.0%	21.2%	65.0%

Table 3: Effect of removing each loss term or trigger.

	ASR ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	98.9%	52.4 dB	99.3%	.014
w/o PGAT	99.2%	49.3 dB	98.8%	.031

Table 4: Balance between ASR_{IoU} and imperceptibility by adopting PGAT strategy.

Balance of Effectiveness and Imperceptibility

We conducted ablations to isolate the contribution of individual components in DOT. Imperceptibility is evaluated using three metrics: Peak Signal-to-Noise Ratio (PSNR, dB), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

Table 4 shows that the PGAT-enhanced attack preserves comparable effectiveness while significantly improving perceptual fidelity. It boosts PSNR by 3.1 dB, increases SSIM to 99.3%, and reduces LPIPS by more than half.

Gap between Universal and Video-Specific Solution

Figure 3 compares a Universal DOT (UDOT) with the video-specific DOT under the same ℓ_∞ budget. UDOT achieves moderate degradation across detectors but is substantially weaker than per-video DOT. Its higher IoU-GT indicates more global, less object-aligned perturbations.

These results suggest that UDOT captures dataset-level priors with limited temporal adaptation, while per-video DOT exploits clip-specific gradients to induce stronger semantic disruption. In practice, UDOT offers a lightweight alternative under resource constraints, whereas per-video DOT remains optimal for maximal attack effectiveness.

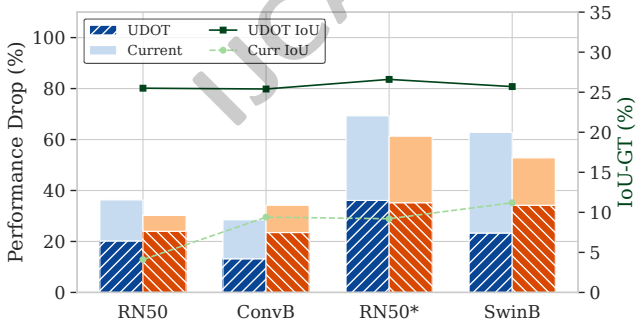

Figure 3: Comparison of attack effectiveness between the Universal DOT (UDOT) and our Current solution. Light and dark shades represent the Current and UDOT solutions, respectively. Similarly, light and dark line tracks IoU statistics.

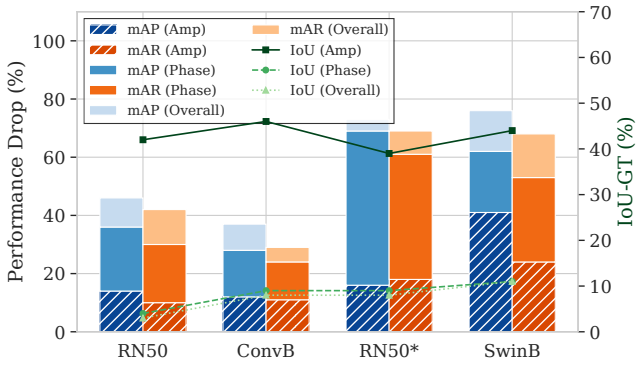

Figure 4: Comparison of Amplitude-only, Phase-only, and Mixed variants. Dark, medium, and light shades represent the three variants, respectively. Blue bars denote the drop in mAP, orange bars indicate the drop in mAR, and the line tracks IoU statistics. While phase modulation excels in transferability and stealth, the mixed strategy strikes the optimal balance between semantic impact and perceptual fidelity.

Amplitude vs. Phase vs. Mixed (PGAT Ablation)

To examine the contribution of the phase and amplitude components in the proposed PGAT, we conduct a controlled ablation where adversarial updates are applied to only the amplitude (Amplitude-guided), only the phase (Phase-guided), or both components with amplitude blending (Overall). All variants share identical optimization settings.

As shown in Figure 4, phase-guided perturbations consistently outperform amplitude-only variants in mAP/mAR degradation while maintaining much lower IoU-GT, indicating superior stealth and mask misalignment. The mixed strategy achieves the strongest overall attack with IoU-GT comparable to the phase-guided variant, offering the best trade-off between effectiveness and imperceptibility.

These results suggest that phase modulation captures semantically meaningful structure critical to model perception, while amplitude-only perturbations mainly affect low-level appearance. Blending both enables PGAT to preserve visual naturalness while enhancing cross-model transferability.

6 Conclusion

In this paper, we propose DOT, a transferable adversarial framework for open-vocabulary video instance segmentation that explicitly accounts for temporal coherence. DOT jointly suppresses genuine object semantics and induces coherent phantom trajectories through frequency-guided optimization. Its Dual Semantic Perturbation Module and Phase-Guided Adversarial Training enable imperceptible yet highly transferable phase-based perturbations. Extensive experiments across two architectures and four backbones show that DOT substantially outperforms existing attacks, achieving up to 69% ΔmAP degradation with near-perfect attack success rates. Beyond exposing vulnerabilities, DOT also yields challenging adversarial videos that can be used as hard negatives for improving robustness, highlighting critical security gaps in current vision-language pipelines.

Ethical Statement

This work studies transferable adversarial attacks on open-vocabulary video instance segmentation for the sole purpose of revealing and ultimately reducing real-world vulnerabilities. No private or proprietary data were used: all experiments rely exclusively on the publicly available OV-VIS benchmark, whose licenses permit research use. The method never interacts with human subjects, nor does it process personally identifiable information; accordingly, institutional review board (IRB) oversight is not required. We adhere to the IJCAI Code of Professional Ethics: the work advances scientific understanding of model robustness, the presentation is honest and complete, and the responsible-release protocol balances scientific openness with social responsibility. Ultimately, we believe the benefits of exposing OV-VIS vulnerabilities outweigh the residual risks, yet we urge practitioners to incorporate robust defense mechanisms before deploying vision–language models in safety-critical settings.

Contribution Statement

Minghao Shou and Kesen Wang contributed equally to this work. Zonghui Wang is the corresponding author. Minghao Shou and Kesen Wang conceived and designed the method, conducted the experiments, and wrote the manuscript. Tong Zhang and Han Bao provided feedback and assisted with experiments. Zonghui Wang supervised the project and revised the manuscript.

References

[Chen *et al.*, 2025a] Long Chen, Yuling Chen, Zhi Ouyang, Hui Dou, Yangwen Zhang, and Haiwei Sang. Boosting adversarial transferability in vision-language models via multimodal feature heterogeneity. *Scientific Reports*, 15(1):7366, 2025.

[Chen *et al.*, 2025b] Xingbai Chen, Tingchao Fu, Renyang Liu, Wei Zhou, and Chao Yi. Mba: Multimodal bidirectional attack for referring expression segmentation models. *arXiv preprint arXiv:2506.16157*, 2025.

[Chhipa *et al.*, 2024] Prakash Chandra Chhipa, Kanjar De, Meenakshi Subhash Chippa, Rajkumar Saini, and Marcus Liwicki. Open-vocabulary object detectors: Robustness challenges under distribution shifts. In *European Conference on Computer Vision*, pages 62–79. Springer, 2024.

[Fang *et al.*, 2024] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *European Conference on Computer Vision*, pages 225–241. Springer, 2024.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Gu *et al.*, 2022] Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer, 2022.

[Guo *et al.*, 2025] Pinxue Guo, Hao Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, and Wenqiang Zhang. Openvis: Open-vocabulary video instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3275–3283, 2025.

[Jia *et al.*, 2023] Xiaojun Jia, Jindong Gu, Yihao Huang, Simeng Qin, Qing Guo, Yang Liu, and Xiaochun Cao. Transegpgd: Improving transferability of adversarial examples on semantic segmentation. *arXiv preprint arXiv:2312.02207*, 2023.

[Kim *et al.*, 2023] Hee-Seon Kim, Minji Son, Minbeom Kim, Myung-Joon Kwon, and Changick Kim. Breaking temporal consistency: Generating video universal adversarial perturbations using image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4325–4334, 2023.

[Lee *et al.*, 2024] Seunghun Lee, Jiwan Seo, Kiljoon Han, Minwoo Choi, and Sunghoon Im. Context-aware video instance segmentation. *arXiv preprint arXiv:2407.03010*, 2024.

[Li and Wang, 2021] Haoxuan Li and Zheng Wang. Universal sparse adversarial attack on video recognition models. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, 12(3):1–15, 2021.

[Li *et al.*, 2023] Ping Li, Yu Zhang, Li Yuan, Jian Zhao, Xianghua Xu, and Xiaoqin Zhang. Adversarial attacks on video object segmentation with hard region discovery, 2023.

[Li *et al.*, 2024] Fengpeng Li, Kemou Li, Haiwei Wu, Jinyu Tian, and Jiantao Zhou. Dat: Improving adversarial robustness via generative amplitude mix-up in frequency domain. *Advances in Neural Information Processing Systems*, 37:127099–127128, 2024.

[Lu *et al.*, 2017] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *arXiv preprint arXiv:1712.02494*, 2017.

[Mu *et al.*, 2021] Ronghui Mu, Wenjie Ruan, Leandro Soriano Marcolino, and Qiang Ni. Sparse adversarial video attacks with spatial transformations. *arXiv preprint arXiv:2111.05468*, 2021.

[Nguyen Tiet Nguyen *et al.*, 2024] Khoi Nguyen Tiet Nguyen, Wenyu Zhang, Kangkang Lu, Yuhuan Wu, Xingjian Zheng, Hui Li Tan, and Liangli Zhen. A survey and evaluation of adversarial attacks for object detection. *arXiv e-prints*, pages arXiv–2408, 2024.

[Pan *et al.*, 2024] Yi Pan, Jun-Jie Huang, Zihan Chen, Wentao Zhao, and Ziyue Wang. Svastin: Sparse video adversarial attack via spatio-temporal invertible neural networks. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Pony *et al.*, 2021] Roi Pony, Itay Naeh, and Shie Mannor. Over-the-air adversarial flickering attacks against video recognition networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 515–524, 2021.

[Rahaman *et al.*, 2019] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[Schlarmann *et al.*, 2024] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Wang *et al.*, 2021] Zeyuan Wang, Chaofeng Sha, and Su Yang. Reinforcement learning based sparse black-box adversarial attack on video recognition models. *arXiv preprint arXiv:2108.13872*, 2021.

[Wang *et al.*, 2023] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 4057–4066, 2023.

[Wei *et al.*, 2018] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.

[Wu *et al.*, 2024] Junfeng Wu, Yi Parcel Jiang, Qihao Liu, Zehuan Yuan, Xiang Bai, and Song Bai. General object foundation model for images and videos at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3783–3795, 2024.

[Yan *et al.*, 2025] Bin Yan, Martin Sundermeyer, David Joseph Tan, Huchuan Lu, and Federico Tombari. Towards real-time open-vocabulary video instance segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1861–1871. IEEE, 2025.

[Yin *et al.*, 2023] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36:52936–52956, 2023.

[Zhang *et al.*, 2023a] Chaoning Zhang, Fachrina Dewi Puspitasari, Sheng Zheng, Chenghao Li, Yu Qiao, Taegoo Kang, Xinru Shan, Chenshuang Zhang, Caiyan Qin, Francois Rameau, et al. A survey on segment anything model (sam): Vision foundation model meets prompt engineering. *arXiv preprint arXiv:2306.06211*, 2023.

[Zhang *et al.*, 2023b] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023.

[Zhang *et al.*, 2025] Peng-Fei Zhang, Guangdong Bai, and Zi Huang. Maa: Meticulous adversarial attack against vision-language pre-trained models. *arXiv preprint arXiv:2502.08079*, 2025.

[Zheng *et al.*, 2024a] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. Syncvis: Synchronized video instance segmentation. *Advances in Neural Information Processing Systems*, 37:49461–49483, 2024.

[Zheng *et al.*, 2024b] Sheng Zheng, Chaoning Zhang, and Xinhong Hao. Black-box targeted adversarial attack on segment anything (sam). *IEEE Transactions on Multimedia*, 2024.

[Zhou *et al.*, 2023a] Dawei Zhou, Nannan Wang, Heng Yang, Xinbo Gao, and Tongliang Liu. Phase-aware adversarial defense for improving adversarial robustness. In *International Conference on Machine Learning*, pages 42724–42741. PMLR, 2023.

[Zhou *et al.*, 2023b] Tao Zhou, Qi Ye, Wenhan Luo, Kaihao Zhang, Zhiguo Shi, and Jiming Chen. F&f attack: Adversarial attack against multiple object trackers by inducing false negatives and false positives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4573–4583, 2023.

[Zhu *et al.*, 2024] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. Clip-vis: Adapting clip for open-vocabulary video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.