SIMATTACK: A SIMPLE AND EFFECTIVE TRANSFERABLE UNTARGETED ADVERSARIAL ATTACK ON LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated strong multimodal understanding and reasoning capabilities, yet remain vulnerable to transferable adversarial attacks, underscoring the need for rigorous robustness evaluation under black-box settings. Existing transfer-based untargeted attacks often rely on ground-truth textual annotations to obtain transferable optimization signals. However, requiring such annotations for every attacked input imposes a strong practical assumption, since they are generally unavailable in real-world scenarios. To address this limitation, we propose **SimAttack**, a simple and effective transfer-based black-box untargeted attack that derives transferable optimization signals solely from surrogate vision encoder representations without ground-truth textual supervision. By leveraging visual representations from different local regions and encoder layers, SimAttack reduces reliance on a limited subset of surrogate encoder representations and improves adversarial transferability. Extensive experiments demonstrate that our method consistently outperforms existing state-of-the-art attack methods across heterogeneous LVLMs and diverse downstream tasks.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal perception, understanding, and reasoning, enabling their use in diverse tasks and applications such as visual question answering (Li et al., 2023a; Liu et al., 2023), autonomous driving (Sima et al., 2024; Xie et al., 2025b), and embodied agents (Driess et al., 2023; Mu et al., 2023). However, LVLMs remain vulnerable to imperceptible adversarial perturbations applied to visual inputs (Cui et al., 2024; Wang et al., 2024), which can disrupt their visual understanding or induce incorrect responses, raising serious security concerns. As LVLMs are increasingly deployed in safety-critical applications, rigorously evaluating their adversarial robustness is esstential for ensuring their security and reliability. Furthermore, to better reflect real-world threat scenarios, such evaluations should be conducted under black-box settings, where adversaries typically have no access to the target model's internal parameters or gradients (Zhao et al., 2023).

Transfer-based attacks have become a common paradigm for evaluating LVLM robustness under black-box settings, where adversarial examples are optimized solely on accessible surrogate models and then directly transferred to unseen victim LVLMs (Zhang et al., 2025a; Xie et al., 2025a; Li et al., 2025; Jia et al., 2025). Within this paradigm, untargeted attacks aim to disrupt the LVLM's correct understanding of the original visual content, causing its output to fall outside the set of semantically valid responses jointly supported by the visual input and prompt. This objective therefore enables direct assessment of an LVLM's robustness in maintaining reliable multimodal understanding and inference under adversarial perturbations. Without predefined target semantics, untargeted attacks can explore a broader space of perturbation directions, yet the absence of effective guidance may cause the optimization to overfit surrogate-specific features or limited visual attributes, thereby compromising transferability (Mei et al., 2026a; Qian et al., 2026). The open-ended output space of LVLMs, where multiple semantically valid responses can be generated from the same visual input and prompt, further impedes the search for transferable optimization signals (Li et al., 2026b).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To obtain more transferable optimization signals, existing transfer-based untargeted attacks have incorporated ground-truth textual annotations as semantic references for the original image to guide adversarial optimization toward meaningful semantic disruption. MABA (Li et al., 2026b) leverages the ground-truth reference captions of each image to guide adversarial visual representations away from the semantics of the original visual content, while SAF (Qian et al., 2026) dynamic attack uses the ground-truth text label to construct the language-modeling loss for identifying influential adversarial features across models. However, assuming access to ground-truth textual annotations for every attacked input is unrealistic in practical black-box settings, as such informations are typically unavailable for arbitrary inputs encountered at attack time. This motivates a stricter threat model and raises a key question: **how can transfer-based black-box untargeted attacks achieve strong cross-model transferability and maintain attack effectiveness across diverse downstream tasks without access to ground-truth textual annotations?**

To address this question, we propose SimAttack, a transfer-based black-box untargeted attack that generates adversarial examples solely using surrogate vision encoder representations, without relying on ground-truth textual annotations. Recent gray-box studies have shown that disrupting representations extracted by LVLM vision encoder can effectively degrade its performance across diverse downstream tasks, indicating their potential as a task-agnostic basis for attack optimization (Mei et al., 2026b;a). Under the transfer-based black-box setting, the victim LVLM is inaccessible, so the key challenge is to derive transferable optimization signals from surrogate vision encoder representations. To this end, SimAttack incorporates visual representations from different local regions and encoder layers into attack optimization, reducing its reliance on surrogate-specific representations and making the attack less sensitive to variations in victim models and task demands, thereby improving cross-model transferability and task generalization. Specifically, SimAttack combines Correspondence-Preserving Local Region Sampling (CPLRS), which resamples spatially corresponding local regions across iterations, with All-Layer Patch Feature Disruption (ALPFD), which minimizes the average cosine similarity between corresponding clean and adversarial patch features across all encoder layers. Extensive experiments across heterogeneous LVLMs and diverse downstream tasks demonstrate that SimAttack achieves stronger cross-model transferability and task generalization than existing state-of-the-art attacks.

2 Related Work

2.1 Adversarial Attacks on LVLMs

Existing adversarial attacks on LVLMs can be broadly categorized into white-box, gray-box, and black-box settings according to the accessibility of the victim model. White-box attacks assume full access to the victim LVLM, and typically optimizes adversarial examples by directly using gradients from the victim model (Schlarmann & Hein, 2023; Cui et al., 2024; Luo et al., 2024). In the gray-box setting, attackers have limited access to selected internal components or intermediate representations of the victim LVLM. Existing studies typically exploit visual representations extracted by the accessible vision encoder of the victim LVLM to optimize adversarial perturbations. VEAttack (Mei et al., 2026b) performs downstream-agnostic untargeted attacks by directly disrupting patch token representations, while PA-Attack (Mei et al., 2026a) further improves this paradigm through prototype-anchored guidance and two-stage attention refinement to enhance attack effectiveness and task generalization. By contrast, in the black-box setting, attackers have no access to the victim LVLM's internal parameters, gradients, or model components. This setting is particularly relevant to practical LVLM deployments, where model internals are typically inaccessible to external users. Black-box attacks can be broadly categorized into query-based and transfer-based methods. Query-based attacks (Liu et al., 2024a; Fang et al., 2026) iteratively use output feedback from the victim LVLM to generate adversarial examples, whereas transfer-based attacks (Zhao et al., 2023; Zhang et al., 2025a) optimize adversarial examples on accessible surrogate models and transfer them to the victim LVLM. Compared with query-based attacks, transfer-based attacks do not query the victim LVLM during adversarial example generation, thereby avoiding query overhead and making them well-suited to query-restricted settings.

2.2 Transfer-based Black-box Attacks on LVLMs

Transfer-based black-box attacks on LVLMs can be categorized into targeted and untargeted attacks based on their attack objectives. Targeted attacks aim to steer victim LVLMs toward predefined target semantics. AttackVLM (Zhao et al., 2023) crafts adversarial examples by aligning representations of adversarial images with those of predefined target texts or images. M-Attack (Li et al., 2025) improves transferability through local-to-global and local-to-local feature matching with model ensembling strategy. FOA-Attack (Jia et al., 2025) optimizes adversarial perturbations through jointly aligning global and local features through optimal transport and dynamically balancing multiple surrogate models. MPCAttack (Li et al., 2026a) extends transferable attacks by collaboratively optimizing aggregated representations from multiple learning paradigms and supports both targeted and untargeted settings. Untargeted attacks instead aim to disrupt the original visual semantics without specifying target semantics in advance. MABA (Li et al., 2026b) combines discriminative feature suppression with modality alignment breaking and uses the ground-truth reference caption of each image to guide adversarial optimization. SAF (Qian et al., 2026) uses ground-truth text label to compute a language-modeling loss for feature attribution and identifies shared adversarial features via augmented Integrated Gradients, updating feature importance periodically to disrupt emerging salient features. Although these methods improve the transferability of untargeted attacks through richer textual semantic guidance, they require ground-truth textual annotations for each attacked input, which are unavailable for arbitrary inputs in practical black-box settings.

In contrast, SimAttack considers a stricter black-box setting without access to ground-truth textual annotations and performs attack optimization solely on visual representations extracted from the publicly available surrogate vision encoder.

3 Method

3.1 Preliminaries

LVLMs. We consider a common architecture for the victim LVLM V, consisting of a vision encoder E, a vision-language projector P, and a large language model (LLM) G. Given an image x and a text prompt p, the vision encoder extracts a sequence of visual representations H_v from image x, which are then mapped by the projector into a sequence of visual tokens $Z_v = P(H_v)$ compatible with the LLM input embedding space. The text prompt p is tokenized and embedded into a sequence of text tokens Z_p. The LLM jointly processes the visual and text tokens to perform multimodal inference and generate a textual response y. The overall LVLM can be represented as $y = V(x, p)$.

Threat Model. We consider a black-box setting in which the attacker has no access to the parameters, gradients, or internal representations of V and is not provided with any ground-truth textual annotations associated with the input image-text pair. The attacker is only permitted to perturb the input image x to craft an adversarial image $x_{\text{adv}} = x+\delta$, while keeping the text prompt p unchanged. Here, δ denotes the adversarial perturbation and satisfies the ℓ_∞-norm constraint $\|\delta\|_\infty \leq \epsilon$, where ϵ denotes the perturbation budget. The adversary aims to induce the victim LVLM V to generate a response that is semantically incorrect with respect to the original image x and the text prompt p. Formally, let $\mathcal{Y}(x, p)$ denote the set of semantically valid responses supported by the original image-text pair (x, p). The attack objective is then formulated as:

$$V(x_{\text{adv}}, p) \notin \mathcal{Y}(x, p) \tag{1}$$

Optimization Formulation. To achieve the attack objective defined above, we formulate adversarial example generation as a constrained optimization problem that disrupts the visual representations extracted by the victim vision encoder E. Such representation changes propagate through the vision-language projector P and can subsequently affect the LLM's multimodal inference and response generation. If E is accessible, the corresponding optimization problem could be formulated as:

$$\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\text{adv}} \left(E(x), E(x+\delta)\right), \tag{2}$$

where $\mathcal{L}_{\text{adv}}$ denotes the loss defined over the representations of the original and adversarial images.

Since the internal representations and gradients of the victim LVLM are inaccessible, the above objective cannot be directly optimized on the E. We therefore adopt a transfer-based strategy and

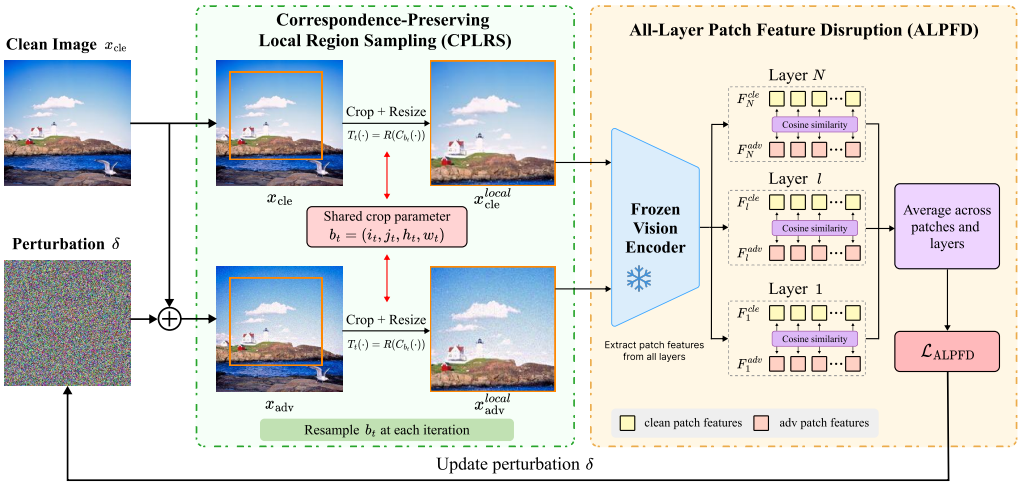


Figure 1: **Overall framework of our proposed SimAttack**. At each iteration, SimAttack first employs CPLRS to construct spatially corresponding local views of the clean and adversarial images. ALPFD then extract patch features from all layers of the frozen vision encoder and computes cosine similarities between corresponding clean and adversarial patches. These similarities are averaged across patches and layers to obtain the loss $\mathcal{L}_{\text{ALPFD}}$, which is used to update the perturbation.

employ a publicly available pretrained vision encoder E_s as the surrogate model. Accordingly, we optimize the adversarial perturbation on E_s using the following surrogate objective, and the resulting adversarial image is directly transferred to the victim LVLM:

$$\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}_{\text{adv}}\left(E_s(x), E_s(x+\delta)\right), \tag{3}$$

3.2 Overview

Under the above formulation, the key is to enhance the transferability of adversarial examples optimized on the surrogate model to unseen victim models. To this end, SimAttack introduces two components. CPLRS resamples spatially corresponding local regions across iterations to use diverse local regions while preserving clean-adversarial spatial correspondence, allowing the optimization to focus on perturbation-induced feature changes and thereby improving attack transferability. All-Layer Patch Feature Disruption (ALPFD) uniformly exploits patch features from all encoder layers without predefined layer preferences, reducing dependence on layer-specific surrogate representations and improving transferability across unknown LVLMs and tasks. As illustrated in Figure 1, SimAttack combines CPLRS and ALPFD to construct the surrogate attack objective on a single frozen pretrained vision encoder E_s.

At each iteration, CPLRS samples a new set of crop parameters and applies the same crop-and-resize transformation to the clean image x_{cle} and the current adversarial image $x_{\text{adv}} = x_{\text{cle}} + \delta$, yielding a pair of spatially corresponding local views. The paired local views are then fed into the frozen surrogate vision encoder, from which ALPFD extracts patch features across all encoder layers. Then ALPFD computes the cosine similarity between corresponding clean and adversarial patch features at each layer and averages these similarities over all patches and encoder layers. The resulting averaged similarity is used to instantiate the surrogate adversarial loss defined above. We then optimize this loss using I-FGSM algorithm (Kurakin et al., 2018) and project the updated perturbation onto the feasible ℓ_∞ set after each iteration. After optimization, the resulting adversarial image is directly transferred to unseen victim LVLMs for robustness evaluation. A detailed description of the our method is provided in the Appendix A.

3.3 Correspondence-Preserving Local Region Sampling

A straightforward implementation of the above formulation is to disrupt the visual representations extracted from the entire clean and adversarial images. However, optimizing over the full image can lead to early optimization saturation and limited transferability, whereas randomly sampling

local regions across iterations provides diverse and fine-grained visual information that improves transferability (Li et al., 2025). Since the adversarial image retains the same underlying visual content of the clean image, their local views should preserve spatial correspondence so that the feature discrepancy primarily reflect changes induced by the adversarial perturbation. This motivates us to sample spatially corresponding local regions from the clean and adversarial images within each iteration, while resampling the region across iterations to retain local view diversity. Based on this principle, we propose **Correspondence-Preserving Local Region Sampling (CPLRS)**.

At the t-th iteration, let $x_{\text{adv}}^t = x_{\text{cle}} + \delta^t$ denote the current adversarial image. Given predefined crop scale and aspect ratio ranges, CPLRS randomly samples a set of crop parameters $b_t = (i_t, j_t, h_t, w_t)$, where i_t and j_t denote the top and left coordinates of the cropped region, and h_t and w_t denote its height and width, respectively. Based on the sampled parameters b_t, the crop-and-resize transformation is then defined as:

$$T_t(\cdot) = R(C_{b_t}(\cdot)), \tag{4}$$

where $C_{b_t}(\cdot)$ crops the region determined by b_t, and $R(\cdot)$ resizes the cropped region to the original resolution of the clean image x_{cle}.

To preserve spatial correspondence, CPLRS applies the same transformation T_t to both the clean and adversarial images, yielding a pair of local views:

$$x_{\text{cle}}^{t,\text{local}} = T_t(x_{\text{cle}}), \qquad x_{\text{adv}}^{t,\text{local}} = T_t(x_{\text{adv}}^t). \tag{5}$$

Thus, the clean and adversarial local views correspond to the same image region within each iteration, while a new set of crop parameters b_t is resampled at every iteration to expose the attack optimization to diverse local regions.

3.4 All-Layer Patch Feature Disruption

A pretrained vision encoder yields hierarchical visual representations across its layers, where different depths encode distinct visual information and exert varying contributions to downstream tasks (Lin et al., 2025; Chen et al., 2025). Meanwhile, adversarial transferability is affected by both the choice of encoder layers used for optimization and their relative contributions to the attack objective (Huang et al., 2019; Liu et al., 2025). Since selecting only a subset of layers or assigning unequal weights imposes a predefined layer preference whose effectiveness may vary across models and tasks, we avoid introducing such preferences when the victim LVLM and downstream task are unknown. Building on Block-wise Similarity Attack (BSA) (Yin et al., 2023), we therefore formulate **All-Layer Patch Feature Disruption (ALPFD)**, which computes the cosine similarities between corresponding clean and adversarial patch features at every encoder layer and uniformly averages them over all patch features and layers, without imposing a predefined layer preference.

Given the paired local views $x_{\text{cle}}^{t,\text{local}}$ and $x_{\text{adv}}^{t,\text{local}}$ generated by CPLRS at the t-th iteration, ALPFD feeds them into the frozen surrogate vision encoder E_s to extract patch features from all encoder layers. For a vision encoder with L layers, each layer outputs N patch features for the input image. Let $f_{l,i}^{t,\text{cle}}$ and $f_{l,i}^{t,\text{adv}}$ denote the i-th patch features output by the l-th layer for $x_{\text{cle}}^{t,\text{local}}$ and $x_{\text{adv}}^{t,\text{local}}$, respectively, where $l \in \{1, \ldots, L\}$ and $i \in \{1, \ldots, N\}$.

ALPFD computes the cosine similarity between each pair of corresponding clean and adversarial patch features and averages these similarities over all patch features and encoder layers. The resulting adversarial loss is defined as:

$$\mathcal{L}_{\text{ALPFD}} = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N}\sum_{i=1}^{N}\cos\left(f_{l,i}^{t,\text{cle}}, f_{l,i}^{t,\text{adv}}\right). \tag{6}$$

Accordingly, the surrogate adversarial loss defined in the optimization formulation is instantiated as $\mathcal{L}_{\text{adv}} = -\mathcal{L}_{\text{ALPFD}}$. Since $x_{\text{adv}}^{t,\text{local}}$ is obtained from the full adversarial image $x_{\text{adv}}^t = x + \delta^t$ through the CPLRS transformation, the gradient of $\mathcal{L}_{\text{adv}}$ is backpropagated to the corresponding spatial region of x_{adv}^t, while pixels outside this region receive no gradient at iteration t. The adversarial perturbation is then updated using I-FGSM, by maximizing the loss $\mathcal{L}_{\text{adv}}$ subject to the ℓ_∞ constraint $\|\delta\|_\infty \leq \epsilon$. As CPLRS resamples different regions across iterations, these localized perturbation updates accumulate across different spatial regions of the full adversarial image. The optimization stops after a predefined number of attack iterations, with the resulting adversarial image directly transferred to unseen victim LVLMs to assess their adversarial robustness.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Tasks and datasets. Following prior work (Mei et al., 2026b), we evaluate task generalization across three representative tasks: image captioning, visual question answering (VQA), and hallucination evaluation. For image captioning, we use the COCO 2014 (Lin et al., 2014) and Flickr30k (Young et al., 2014) datasets. For VQA, we use the VQAv2 (Goyal et al., 2017) and TextVQA (Singh et al., 2019) datasets. For hallucination evaluation, following (Mei et al., 2026b), we use images from the COCO 2014 dataset paired with the questions and annotations from POPE (Li et al., 2023b) dataset. We randomly sample 1,000 images for each dataset. Specifically, COCO 2014 and Flickr30k are sampled from the Karpathy test split (Karpathy & Fei-Fei, 2015). For VQAv2, TextVQA, and POPE, we first sample images and retain all valid question-annotation pairs associated with the sampled images. To ensure a fair comparison, all methods are evaluated on exactly the same sampled images and their corresponding question-annotation pairs.

Evaluation metrics. We evaluate attack effectiveness under an untargeted criterion by measuring the degradation of standard task-specific metrics on adversarial images relative to their clean counterparts. For image captioning, we report CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), CLIPScore and RefCLIPScore (Hessel et al., 2021) to evaluate reference-based caption quality and image-text semantic consistency. For visual question answering, we report the VQA Accuracy (Goyal et al., 2017; Singh et al., 2019). For hallucination evaluation, we report the F1-score on POPE (Li et al., 2023b). To facilitate comparison across tasks with different metric scales, we further report the Score Reduction Rate (SRR) (Mei et al., 2026a), defined as $\mathrm{SRR} = 100(1 - S_{adv}/S_{clean})$, where S_{clean} and S_{adv} denote the corresponding task scores on clean and adversarial images, respectively. A higher SRR indicates stronger attack effectiveness.

Victim LVLMs. As adopted in previous studies (Mei et al., 2026a; Li et al., 2026b), we evaluate our method and several baselines on five publicly available LVLMs: Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), MiniCPM-o-2.6 (Yao et al., 2025), Phi-4-Multimodal (Abouelenin et al., 2025), and LLaVA-v1.6-Vicuna-7B (Liu et al., 2024b). These victim LVLMs span diverse model families, architectures, and training recipes, enabling a comprehensive evaluation of adversarial transferability across heterogeneous models.

Baselines. We compare SimAttack with three representative and strong LVLM adversarial attack methods: PA-Attack (Mei et al., 2026a), MABA (Li et al., 2026b), and MPCAttack (Li et al., 2026a). PA-Attack is a state-of-the-art gray-box untargeted attack that relies only on the victim LVLM's vision encoder. Given the methodological similarity between PA-Attack and SimAttack, both of which rely solely on the vision encoder for adversarial optimization, we adapt PA-Attack to the transfer-based black-box setting for a fair comparison. MABA is a state-of-the-art transfer-based black-box untargeted attack that requires ground-truth textual annotations to guide adversarial optimization. MPCAttack is a state-of-the-art transfer-based attack method that reports superior performance over prior transferable attacks under both targeted and untargeted settings.

Implementation details. Following MABA (Li et al., 2026b), we set the perturbation budget to $\epsilon = 16/255$ under the ℓ_∞ norm and use 30 optimization steps with a step size of $\alpha = 3/255$ for all methods. SimAttack uses a frozen CLIP-L/14 as the surrogate vision encoder, with the crop scale range set to $[0.5, 0.9]$ and the aspect ratio range set to the default $[3/4, 4/3]$. The ablation study on crop scale range is provided in Appendix. For PA-Attack, we use CLIP-L/14 as the surrogate vision encoder in place of the victim LVLM's vision encoder. MABA and MPCAttack retain their original surrogate configurations. We employ VLMEvalKit (Duan et al., 2024) to load and query all victim LVLMs through a unified inference interface, facilitating consistent and reproducible evaluation across models. For each downstream task, the same task-specific prompt is used across all victim LVLMs to ensure consistent evaluation settings, with the detailed prompt settings provided in Appendix D. All experiments are conducted on a single NVIDIA GeForce RTX 4090 GPU.

Table 1: **Comparison of SimAttack with different attack methods across five datasets and five victim LVLMs**. Dataset columns report task-specific metric scores(%; lower is better): the averaged captioning metrics for COCO and Flickr30K, VQA Accuracy for VQAv2 and TextVQA, and F1 for POPE. Average SRR denotes the mean SRR across the five datasets (%; higher is better). Clean rows are shaded gray. Bold indicates the best result in each column.

Model	Method	COCO	Flickr30K	VQAv2	TextVQA	POPE	Average SRR
LLaVA-v1.6-7B	Clean	67.81	62.08	76.30	48.57	84.45	0.00
	PA-Attack (Mei et al., 2026a)	59.48	55.81	66.57	36.51	76.61	13.85
	MABA (Li et al., 2026b)	55.52	51.39	63.02	28.53	68.72	22.52
	MPCAttack (Li et al., 2026a)	36.28	37.59	47.77	22.55	55.44	42.25
	SimAttack	**29.99**	**29.91**	**38.74**	**16.48**	**35.17**	**56.25**
MiniCPM-o 2.6	Clean	58.85	60.47	77.39	61.48	85.18	0.00
	PA-Attack (Mei et al., 2026a)	52.33	54.26	67.13	47.28	78.97	13.00
	MABA (Li et al., 2026b)	52.45	52.91	65.23	38.37	73.30	18.12
	MPCAttack (Li et al., 2026a)	31.95	34.30	45.67	29.20	55.63	43.43
	SimAttack	**27.66**	**28.58**	**42.98**	**26.18**	**46.00**	**50.72**
Qwen2.5-VL-7B	Clean	59.89	58.93	78.19	66.88	82.07	0.00
	PA-Attack (Mei et al., 2026a)	54.26	54.03	68.94	50.04	74.20	12.86
	MABA (Li et al., 2026b)	53.89	52.10	66.21	43.53	69.71	17.38
	MPCAttack (Li et al., 2026a)	38.47	39.22	52.63	34.56	58.88	35.70
	SimAttack	**29.49**	**29.93**	**49.53**	**33.20**	**49.51**	**45.33**
Phi-4-Multimodal	Clean	67.86	62.41	72.34	51.09	82.64	0.00
	PA-Attack (Mei et al., 2026a)	60.51	55.45	63.77	37.81	75.81	13.62
	MABA (Li et al., 2026b)	59.09	54.24	61.65	31.47	70.44	18.79
	MPCAttack (Li et al., 2026a)	36.88	37.36	45.03	23.77	56.47	41.74
	SimAttack	**30.60**	**31.15**	**41.77**	**23.16**	**46.11**	**49.23**
InternVL3-8B	Clean	79.97	69.86	74.89	59.37	89.25	0.00
	PA-Attack (Mei et al., 2026a)	69.34	60.87	65.64	45.23	81.48	14.21
	MABA (Li et al., 2026b)	65.12	56.80	63.52	36.94	77.12	20.76
	MPCAttack (Li et al., 2026a)	37.66	37.29	43.85	28.10	56.61	46.05
	SimAttack	**29.07**	**29.54**	**38.82**	**23.21**	**38.63**	**57.43**

4.2 Main Results

Table 1 reports the attack performance of all methods across different datasets and victim LVLMs. Overall, SimAttack achieves the highest Average SRR for every victim LVLM and consistently outperforms all baselines across all 25 dataset-LVLM combinations. Compared with the strongest baseline, MPCAttack, SimAttack improves the average SRR across all victim LVLMs by 23.8%. We attribute the superior performance of SimAttack to its joint exploitation of visual representations across different local regions and encoder layers during attack optimization. By avoiding concentration on a narrow subset of surrogate representations, the resulting adversarial examples remain effective despite differences in victim models and task demands, leading to stronger cross-model transferability and task generalization, and ultimately higher attack effectiveness.

Across victim LVLMs, SimAttack achieves its highest attack effectiveness on InternVL3-8B and the lowest on Qwen2.5-VL-7B, while maintaining strong performance on the remaining models. This variation may stem from differences in model families, architectures, and training recipes, which can affect the transferability of adversarial examples across victim LVLMs. Detailed architectural configurations of these victim models are provided in Appendix D. When averaged across victim LVLMs, SimAttack achieves the strongest attack effectiveness on image captioning, followed by VQA and hallucination evaluation. This difference may arise because image captioning relies more directly on comprehensive visual semantics, whereas VQA and hallucination evaluation additionally involve prompt-conditioned reasoning and task-specific decision processes, making them comparatively less susceptible to adversarial attacks. The visualization of adversarial examples and perturbations of these methods are provided in Appendix C.

4.3 Ablation Studies

To assess the contribution of individual design choices in SimAttack, we conduct ablation studies across three downstream tasks and two victim LVLMs. Specifically, we evaluate image captioning on COCO 2014, visual question answering on VQAv2, and hallucination evaluation on POPE. For

each task, we randomly select 200 images from the corresponding dataset and evaluate all ablation variants on Qwen2.5-VL-7B and Phi-4-Multimodal. Unless otherwise specified, all other experimental settings follow the main experiments.

Ablation of SimAttack Components. We investigate the contribution of the two key components of SimAttack by removing CPLRS and ALPFD individually. Specifically, w/o CPLRS uses the entire clean and current adversarial images at each iteration and retains ALPFD, while w/o ALPFD keeps CPLRS unchanged and uses only the final layer patch features for optimization. As shown in Table 2, removing either component consistently degrades attack performance across downstream tasks and victim LVLMs. The average SRR decreases from 49.14% for the full SimAttack to 23.03% without CPLRS and 27.84% without ALPFD. The full SimAttack achieves the best performance in every task-LVLM combination, demonstrating that both CPLRS and ALPFD contribute substantially to attack effectiveness, that their joint use yields the strongest overall performance.

Table 2: **Ablation of SimAttack Components**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task-LVLM pairs. Bold marks the best result within each column.

Variant	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
w/o CPLRS	33.02	19.51	16.24	30.82	20.98	17.59	23.03
w/o ALPFD	34.68	21.91	23.12	39.02	26.31	22.02	27.84
SimAttack	**63.52**	**36.18**	**43.26**	**65.28**	**40.72**	**45.90**	**49.14**

Ablation of CPLRS. To isolate the effect of CPLRS, we disable ALPFD and use the same attack objective for all variants, defined as the average cosine similarity between clean and adversarial patch features from the final layer of the surrogate vision encoder. As shown in Table 3, we compare three variants: No Crop, which directly uses the entire clean and current adversarial images at each iteration; Independent Crop, which independently crops and resizes the clean and current adversarial images to obtain local views; and CPLRS, which applies the same crop-and-resize operation to both images to obtain spatially corresponding local views. Compared with No Crop, Independent Crop increases the AvgSRR from 13.26% to 19.15%, indicating the benefit of local region sampling. More importantly, CPLRS further improves the AvgSRR to 27.84% and consistently outperforms Independent Crop across all downstream tasks and victim LVLMs. These results show that preserving spatial correspondence between clean and adversarial local views provides a substantial additional gain beyond local region sampling alone.

Table 3: **Ablation of CPLRS**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task-LVLM pairs. Bold marks the best result within each column.

Variant	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
No Crop	17.98	11.57	10.36	17.63	14.01	7.98	13.26
Independent Crop	27.02	14.53	13.58	28.57	17.18	14.04	19.15
CPLRS	**34.68**	**21.91**	**23.12**	**39.02**	**26.31**	**22.02**	**27.84**

Ablation of ALPFD. We investigate ALPFD from two aspects: layer selection and layer weighting. All variants adopt the 24-layer CLIP-L/14 under the No Crop setting to isolate the effect of CPLRS. For layer selection, we consider six variants that differ in which encoder layers supply patch features for optimization: Final Layer selects only the final layer of the encoder; Shallow Layers, Middle Layers, and Deep Layers use three contiguous layer subsets corresponding to layers 1-8, 9-16, and 17-24, respectively; Uniform Sampling selects eight layers spanning the shallow, middle, and deep regions of the encoder; and All Layers uses all encoder layers of the encoder. As shown in Table 4, Shallow Layers, Middle Layers, and Deep Layers all achieve higher AvgSRR than the Final Layer, but none of them consistently outperforms the others across victim LVLMs

Table 4: **Ablation of Layer Selection in ALPFD**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task-LVLM pairs. Bold marks the best result within each column.

Variant	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
Final Layer	17.98	11.57	10.36	17.63	14.01	7.98	13.26
Shallow Layers	26.98	19.45	17.36	15.96	14.26	11.63	17.61
Middle Layers	25.98	19.02	15.90	23.31	16.07	13.35	18.94
Deep Layers	21.07	12.61	12.71	22.54	15.41	10.27	15.77
Uniform Sampling	30.50	**20.35**	**18.43**	26.34	18.97	15.17	19.96
All Layers	**33.02**	19.51	16.24	**30.82**	**20.98**	**17.59**	**23.03**

Table 5: **Ablation of Layer Weighting in ALPFD**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task-LVLM pairs. Bold marks the best result within each column.

Variant	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
Final Layer Similarity	25.71	16.18	13.90	24.11	17.46	12.42	18.30
Random	30.84	**20.84**	15.64	29.82	19.21	15.64	22.00
Equal	**33.02**	19.51	**16.24**	**30.82**	**20.98**	**17.59**	**23.03**

and downstream tasks. Uniform Sampling further increases the AvgSRR to 19.96%, while All Layers achieves the strongest overall AvgSRR of 23.03% and the best performance in four of the six settings. These results indicate that no particular layer or layer subset consistently provides stronger attack performance across models and tasks. We therefore use all encoder layers in ALPFD.

For layer weighting, we fix the layer selection to All Layers and compare three variants. Equal assigns the same weight to every layer; Random assigns randomly generated weights; and Final Layer Similarity, following the weighting strategy in HQA-VLAttack (Liu et al., 2025), assigns weights according to the cosine similarity between each layer's CLS feature and that of the final layer. As shown in Table 5, Equal achieves the highest AvgSRR of 23.03% and the best performance in five of the six settings, while Random performs better only on Qwen2.5-VL for VQAv2 and Final Layer Similarity remains consistently weaker. These results further indicate that assigning unequal importance to particular layers does not consistently improve attack effectiveness across models and tasks. We therefore apply equal weighting to all encoder layers in ALPFD.

5 Conclusion

In this work, we proposed SimAttack, a simple and effective transfer-based black-box untargeted attack that derives transferable optimization signals solely from a pretrained surrogate vision encoder and requires no ground-truth textual annotations. SimAttack leverages visual representations from different local regions and encoder layers to improve transferability. Specifically, CPLRS preserves spatial correspondence between clean and adversarial local regions, while ALPFD uniformly disrupts corresponding patch features across all encoder layers. Extensive experiments demonstrate that SimAttack consistently outperforms existing attack method across heterogeneous LVLMs and diverse downstream tasks.

Limitations. Although ALPFD uniformly aggregates features across all encoder layers to avoid imposing a predefined layer preference, it remains unclear which encoder layer features can consistently support transfer across different victim LVLMs and downstream tasks. This limitation motivates further study of how to identify and exploit transferable features at different layers of the surrogate vision encoder, so that the contribution of each layer can be adaptively adjusted without access to the victim LVLM and downstream task.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure as follows: assisting with translation, and cleaning and reformatting datasets. We have not used generative AI tools for any other tasks with required disclosure. Additionally, we used generative AI tools for tasks with recommended disclosure, including creating or editing software code, summarising or analysing existing literature, editing a research paper to improve readability, formatting references, and sourcing/searching for information. We have carefully reviewed all AI-assisted work. Specifically, AI-assisted dataset cleaning and reformatting were manually checked against the original data to ensure data integrity and formatting consistency. Specifically, AI-assisted translation was manually reviewed to ensure translation accuracy and correct use of terminology. All AI-created or edited code was reviewed and tested by the authors to verify its correctness and reproducibility. AI-summarised literature and sourced information were cross-checked against the original sources. AI-polished text and formatted references were reviewed to ensure compliance with submission requirements. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI tools.

ETHICS STATEMENT

This work investigates adversarial attacks against LVLMs to evaluate their robustness under black-box settings. All experiments are conducted on publicly available benchmark datasets and open-source models. This research does not involve human subjects, crowdsourcing data collection, or the acquisition of personally identifiable information. We acknowledge that transferable adversarial attacks may have potential security implications if misused against real-world deployed multimodal systems. We advocate for the responsible use of such attack methodologies to uncover model vulnerabilities and to facilitate the development of more robust and secure multimodal AI systems.

REPRODUCIBILITY STATEMENT

Section 3 elaborates on the full design of SimAttack, including its attack pipeline and core component design. Section 4 provides a complete specification of the experimental setup, including the downstream tasks, datasets, victim LVLMs, evaluation metrics, attack configurations, and optimization hyperparameters. Visualization of Adversarial Examples and Perturbations, additional ablation studies and experimental details are presented in the Appendix. To support reproducibility, our supplementary materials include an anonymous reference implementation of SimAttack, along with the scripts and configuration files required to reproduce the reported experiments and evaluation results.

REFERENCES

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, volume 9909 of *Lecture Notes in Computer Science*, pp. 382–398. Springer, 2016.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6677–6695, Suzhou, China, 2025. Association for Computational Linguistics.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201. Association for Computing Machinery, 2024.

Xiang Fang, Wanlong Fang, and Changshuo Wang. Unveiling the fragility of vision-language models: Multi-modal adversarial synergy via texture-constrained perturbations and cross-modal optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3867–3875, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4733–4742, 2019.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 64643–64678, 2025.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In Roman V. Yampolskiy (ed.), *Artificial Intelligence Safety and Security*, chapter 8, pp. 99–112. Chapman and Hall/CRC, 2018.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, 2023b. Association for Computational Linguistics.

Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30065–30075, 2026a.

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of GPT-4.5/4o/o1. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50286–50319, 2025.

Zhichao Li, Hongshan Yang, Zhibo Wang, Huiyu Xu, Junhong Lai, Yaopeng Wang, Kui Ren, and Chun Chen. On evaluating the robustness of large vision-language models via untargeted modality alignment breaking adversarial attack. In *35th USENIX Security Symposium (USENIX Security 26)*, pp. 4801–4820, Baltimore, MD, 2026b. USENIX Association.

Junyan Lin, Haoran Chen, Yue Fan, Yingqi Fan, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multi-layer visual feature fusion in multimodal LLMs: Methods, analysis, and best practices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4156–4166, 2025.

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

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's Box: Towards building universal attackers against real-world large vision-language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 52127–52158, 2024a.

Han Liu, Jiaqi Li, Zhi Xu, Xiaotong Zhang, Xiaoming Xu, Fenglong Ma, Yuanman Li, and Hong Yu. HQA-VLAttack: Towards high quality adversarial attack on vision-language pre-trained models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 170967–170986, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-1.6: Improved reasoning, OCR, and world knowledge. LLaVA project release, 2024b.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An Image Is Worth 1000 Lies: Adversarial transferability across prompts on vision-language models. In *International Conference on Learning Representations*, 2024.

Hefei Mei, Zirui Wang, Chang Xu, Jianyuan Guo, and Minjing Dong. PA-Attack: Guiding gray-box attacks on LVLM vision encoders with prototypes and attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15679–15688, 2026a.

Hefei Mei, Zirui Wang, Shen You, Minjing Dong, and Chang Xu. VEAttack: Downstream-agnostic vision encoder attack against large vision language models. In *International Conference on Learning Representations*, 2026b.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. EmbodiedGPT: Vision-language pre-training via embodied chain of thought. In *Advances in Neural Information Processing Systems*, volume 36, pp. 25081–25094, 2023.

Yaguan Qian, Xucheng Zhu, Qiqi Bao, Fei Yu, Shouling Ji, Zhaoquan Gu, Wei Wang, Bin Wang, and Zhen Lei. Exploiting shared adversarial features for dynamic attacks in large vision-language models. *IEEE Transactions on Information Forensics and Security*, 21:592–607, 2026.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3677–3685, 2023.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beßwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *Computer Vision – ECCV 2024*, volume 15110 of *Lecture Notes in Computer Science*, pp. 256–274. Springer, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1072–1081. Association for Computing Machinery, 2024.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689, 2025a.

Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, et al. S4-Driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1622–1632, 2025b.

Yuan Yao, Tianyu Yu, Chongyi Wang, Junbo Cui, Bokai Xu, Hongji Zhu, Tianchi Cai, Fuwei Huang, Tianran Wang, Wenshuo Ma, et al. MiniCPM-o 2.6: A GPT-4o-level MLLM for vision, speech, and multimodal live streaming on your phone. Technical report and model documentation, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52936–52956, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. AnyAttack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19900–19909, 2025a.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. LMMs-Eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, Albuquerque, New Mexico, 2025b. Association for Computational Linguistics.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54111–54138, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A A Detailed Description of SimAttack

The detailed description of the proposed SimAttack is shown in Algorithm 1.

Algorithm 1 SimAttack

Input: Clean image x_{cle}; frozen surrogate vision encoder E_s; crop scale range $\mathcal{S}$; aspect ratio range $\mathcal{A}$; perturbation budget ϵ; iterations K; step size α.

Output: Adversarial image x_{adv}.

Initialize $\delta^0 = 0$, $x_{\mathrm{adv}}^0 = x_{\mathrm{cle}} + \delta^0$; *// Initialize the perturbation and adversarial image*

for $t = 0$ to $K - 1$ **do**

  Sample crop parameters $b_t = (i_t, j_t, h_t, w_t)$ given $\mathcal{S}$ and $\mathcal{A}$;

  Obtain paired local views $x_{\mathrm{cle}}^{t,\mathrm{local}}$ and $x_{\mathrm{adv}}^{t,\mathrm{local}}$ by Eq. (5) using the same b_t; *// CPLRS*

  Extract patch features from all layers of E_s for $x_{\mathrm{cle}}^{t,\mathrm{local}}$ and $x_{\mathrm{adv}}^{t,\mathrm{local}}$;

  Compute loss by Eq. (6) and let $\mathcal{L}_{\mathrm{adv}}^t = -\mathcal{L}_{\mathrm{ALPFD}}^t$; *// ALPFD*

  $g^t = \nabla_{\delta^t} \mathcal{L}_{\mathrm{adv}}^t$;

  $\delta^{t+1} = \mathrm{Clip}\big(\delta^t + \alpha\,\mathrm{sign}(g^t), -\epsilon, \epsilon\big)$;

  $x_{\mathrm{adv}}^{t+1} = x_{\mathrm{cle}} + \delta^{t+1}$;

end for

return x_{adv}.

B Additional Ablation Studies

Unless stated otherwise, all settings are identical to those used in the ablation studies in Section 4.

Ablation of Crop Scale Range. We investigate the effect of the crop scale range in CPLRS while keeping all other settings unchanged. Here, the crop scale denotes the cropped area as a proportion of the original image. As shown in Table 6, attack performance improves significantly when the sampled local cropped area are restricted to medium-to-large scales. The default $[0.5, 0.9]$ achieves the highest AvgSRR of 49.14%, while $[0.6, 0.9]$ yields a comparable attack performance. The close performance of these two ranges also suggests that CPLRS is relatively stable to moderate changes in the crop scale range. We therefore adopt $[0.5, 0.9]$ as the default setting.

Table 6: **Ablation of Crop Scale Range**. Results are reported as SRR (%; higher is better); AvgSRR is the mean SRR across all task-LVLM pairs. Bold marks the best result within each column.

Crop Scale Range	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
$[0.1, 0.3]$	46.35	23.22	20.49	26.64	14.97	10.81	23.75
$[0.3, 0.6]$	58.20	30.29	33.00	51.79	29.40	27.48	38.36
$[0.6, 0.9]$	63.27	35.06	**44.15**	64.58	40.51	**46.16**	48.96
$[0.1, 0.9]$	62.94	34.46	39.22	59.39	36.58	37.64	45.04
$[0.5, 0.9]$	**63.52**	**36.18**	43.26	**65.28**	**40.72**	45.90	**49.14**

Ablation of Surrogate Models. To investigate how surrogate model selection affects attack performance, we compare CLIP-L/14, DINOv2-Base, InternVL3-1B, and CLIP-G/14 under the same attack settings. As shown in Table 7, the relative performance varies across downstream tasks and victim LVLMs, with no single surrogate model consistently performing best in all settings. CLIP-L/14 and CLIP-G/14 achieve highly comparable AvgSRR scores, yet CLIP-L/14 exhibits a more balanced performance profile across all task-LVLM pairs despite employing substantially fewer encoder layers (24 vs. 40). We therefore adopt CLIP-L/14 as the default surrogate model.

As a complementary analysis, we examine whether model ensembling can further improve transferability over a single surrogate model. We ensemble CLIP-B/16, CLIP-B/32, and CLIP-G/14 following the surrogate configuration adopted in prior work (Li et al., 2025). As shown in Table 8, the ensemble improves performance across all evaluated task-LVLM pairs and increases the AvgSRR from 49.14% to 55.79%. This result shows that surrogate model ensembling can provide additional transferability gains beyond the default single surrogate model configuration.

Table 7: **Ablation of Surrogate Model Selection**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task–LVLM pairs. Bold marks the best result within each column.

Surrogate Model	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
CLIP-L/14	**63.52**	**36.18**	43.26	**65.28**	40.72	45.90	**49.14**
DINOv2-Base	53.01	23.34	24.19	40.91	25.39	20.67	31.25
InternVL3-1B	62.50	36.05	**49.77**	57.66	36.09	41.52	47.27
CLIP-G/14	61.59	32.88	42.20	64.74	**41.04**	**51.89**	49.06

Table 8: **Ablation of Surrogate Model Ensembling**. SRR (%; higher is better) is reported, with AvgSRR averaged over all task–LVLM pairs. Bold marks the best result within each column.

Variant	Qwen2.5-VL-7B			Phi-4-Multimodal			AvgSRR
	COCO	VQAv2	POPE	COCO	VQAv2	POPE	
Single Model	63.52	36.18	43.26	65.28	40.72	45.90	49.14
Model Ensemble	**70.52**	**39.98**	**52.54**	**68.93**	**45.99**	**56.80**	**55.79**

C Visualization of Adversarial Examples and Perturbations

To qualitatively compare perturbation characteristics under different attacks, we visualizes the adversarial examples and perturbations generated by PA-Attack (Mei et al., 2026a), MABA (Li et al., 2026b), MPCAttack (Li et al., 2026a), and SimAttack. As illustrated in Figure 2, PA-Attack and MABA produce more spatially dispersed perturbations, whereas MPCAttack and SimAttack exhibit more localized and image-dependent perturbation patterns. SimAttack exhibits relatively localized perturbation patterns around object contours and textured regions, while preserving the recognizable semantic content of the clean images. This qualitative observation aligns well with the stronger attack performance reported in Section 4.

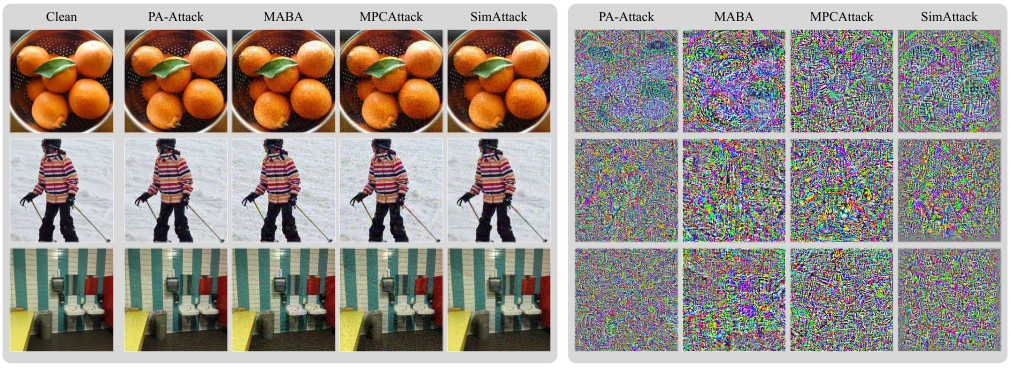


Figure 2: **Visualization of adversarial examples and perturbations.** The left panel presents clean images and adversarial examples generated by PA-Attack, MABA, MPCAttack, and SimAttack, with the right panel showing their corresponding perturbations.

D Additional Experimental Details

Victim LVLM architectures. Table 9 summarizes the architectural configurations of the five victim LVLMs in terms of their vision encoder, vision-language projector, and LLM. Qwen2.5-VL-7B, InternVL3-8B, and MiniCPM-o-2.6 share Qwen2.5-7B as the LLM, while employing different vision cncoders and vision-language projectors (Bai et al., 2025; Zhu et al., 2025; Yao et al., 2025). A complementary configuration is represented by MiniCPM-o-2.6 and Phi-4-Multimodal, both of which adopt SigLIP-400M as the vision encoder but use different vision-language projectors and LLMs (Yao et al., 2025; Abouelenin et al., 2025). LLaVA-v1.6-Vicuna-7B adopts a CLIP ViT-L/14 vision encoder, an MLP-based vision-language projector, and Vicuna-7B-v1.5 as the LLM (Liu et al., 2024b). Together, these architectural variations enable a more comprehensive evaluation of attack transferability across different LVLM architectures.

Table 9: **Architectures of the five victim LVLMs used in our evaluation**.

Victim LVLM	Vision encoder	Vision-language Projector	LLM
Qwen2.5-VL-7B	Qwen2.5-VL ViT	MLP merger	Qwen2.5-7B
InternVL3-8B	InternViT-300M	MLP projector	Qwen2.5-7B
MiniCPM-o-2.6	SigLIP-400M	Perceiver resampler	Qwen2.5-7B
Phi-4-Multimodal	SigLIP-400M	MLP projector	Phi-4-Mini
LLaVA-v1.6-Vicuna-7B	CLIP ViT-L/14	MLP projector	Vicuna-7B-v1.5

Task prompts. We use the same task-specific prompts for all victim LVLMs. For image captioning task, we use the prompt "Provide a one-sentence caption for the provided image.", following the standard setup in LMMs-Eval (Zhang et al., 2025b). For VQA task, we append "Answer the question using a single word or phrase." to the original question, following commonly adopted short-answer VQA evaluation protocols (Duan et al., 2024). For hallucination evaluation task, we similarly append "Answer the question with yes or no." to each question, following the binary question-answering setting of POPE (Li et al., 2023b).