DIVERSIFY AND AGREE: RELIABLE SURROGATE EXPANSION FOR TRANSFERABLE ADVERSARIAL ATTACKS ON MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) remain vulnerable to transfer-based adversarial attacks, where imperceptible perturbations optimized on accessible surrogate models induce unseen MLLMs to produce attacker-specified responses. While enlarging the surrogate ensemble with diverse representation paradigms improves transferability, the set of accessible surrogates is limited in practice, and collecting or training additional models is costly. In this work, we ask whether richer transferable guidance can be extracted from an existing surrogate set without introducing new models. To this end, we propose DA-Attack, a *diversify-and-agree* framework comprising two components. Randomized Surrogate Expansion (RSE) diversifies each surrogate by randomly pruning a small fraction of its parameters at every iteration, turning a single model into a stream of online-sampled variants. Since these variants may yield noisy or conflicting gradients, Intra-Surrogate Neighborhood Consensus (ISNC) measures the directional agreement among gradients from the original surrogate and its variants, and assigns greater weight to directions consistently supported across the neighborhood, producing a reliable consensus update. Extensive experiments on eight open-source and four closed-source MLLMs demonstrate that DA-Attack achieves state-of-the-art attack performance. For example, it raises the average targeted attack success rate across the 12 victim models from 70.4% to 82.1% on NIPS 2017, and from 28.5% to 45.6% on Flickr30K, compared with the strongest baseline. *The anonymized source code is provided in the supplementary materials.*

1 INTRODUCTION

Multimodal large language models (MLLMs) (Singh et al., 2026; Lu et al., 2024a; Zheng et al., 2026) have rapidly evolved into general-purpose systems that connect visual perception with language understanding and reasoning. By integrating powerful visual encoders with large language models, they enable a broad range of applications, including image captioning (Dwibedi et al., 2024), visual question answering (Li et al., 2025a), and complex visual reasoning Chen et al. (2025). However, carefully crafted, imperceptible perturbations added to an image can cause an MLLM to misinterpret its content and produce attacker-specified responses (Zhao et al., 2023; Guo et al., 2026; Wang et al., 2025b). Such manipulation may corrupt the model's subsequent reasoning while leaving the input visually unchanged to human observers. Rigorous adversarial testing under realistic black-box constraints is therefore essential for revealing weaknesses in multimodal representations and assessing the security of MLLMs (Zheng et al., 2025; Wu et al., 2025).

In this context, current transfer-based attacks (Jia et al., 2025b; Wang et al., 2026a; Xu et al., 2025; Lu et al., 2026) provide a practical way to challenge MLLMs. Specifically, an adversarial image is optimized on accessible surrogate models and then directly transferred to an unseen target MLLM. Recent works have improved this process from several complementary perspectives. At the input level, M-Attack (Li et al., 2025b) uses random local views so that the perturbation captures recognizable details of the target image instead of appearing as nearly uniform noise. At the feature level, FOA-Attack (Jia et al., 2025b) jointly aligns global representations and fine-grained patch features, providing richer semantic guidance for transfer. At the gradient level, M-Attack-V2 (Zhao et al., 2026) aggregates multiple crop gradients within each iteration to reduce unstable updates caused

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.

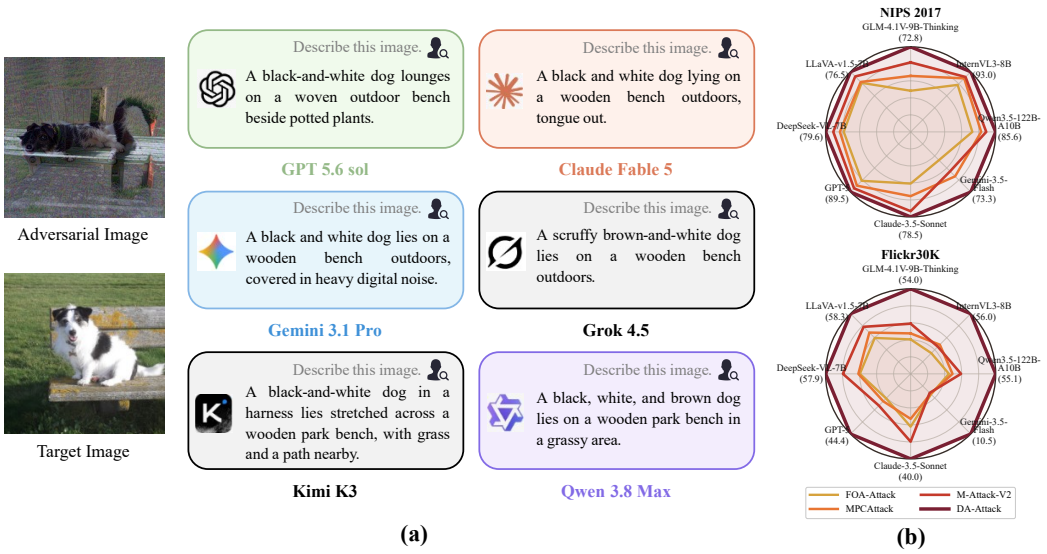


Figure 1: **(a)** Adversarial examples generated by DA-Attack and responses from advanced MLLMs. The MLLMs interpret the adversarial image as semantically aligned with the corresponding target image. **(b)** Attack success rates (%) for a subset of the evaluated MLLMs on NIPS 2017 and Flickr30K.

by random local transformations. At the surrogate level, MPCAttack (Li et al., 2026) broadens the representation space by combining models trained under different learning paradigms. For example, cross-modal alignment models such as CLIP (Radford et al., 2021), multimodal understanding models such as InternVL (Zhu et al., 2025), and visual self-supervised models such as DINOv2 Oquab et al. (2024) capture image–text correspondences, high-level semantic relations, and visual structures, respectively. Their collaborative optimization reduces the bias of any single learning paradigm and improves transfer across heterogeneous MLLMs.

Although broader representation paradigms and larger surrogate ensembles (Huang et al., 2025) can improve transferability, the set of accessible surrogates remains limited in practice. Moreover, every additional model requires extra storage and repeated forward and backward passes, which substantially increases memory use and computation costs. We therefore ask a more practical question. *Can we extract richer transferable guidance from the existing surrogate set and thereby improve attacks on MLLMs without additional models?*

In this work, we propose a simple yet effective approach to improve transferability across MLLMs. Our key motivation is to diversify each surrogate through random parameter pruning and stabilize the attack using the directions shared across its variants. Based on this idea, we introduce DA-Attack, a *diversify-and-agree* framework that generates more offensive adversarial examples for MLLMs. Specifically, DA-Attack consists of two components, Randomized Surrogate Expansion (RSE) and Intra-Surrogate Neighborhood Consensus (ISNC). Motivated by recent advances in model self-ensembling (Huang et al., 2023; Su et al., 2026), RSE randomly masks a small fraction of parameters at each iteration, turning each surrogate into a diverse set of model variants without collecting or training additional models. However, diversity alone is insufficient because different variants may produce noisy or conflicting gradients. ISNC therefore compares the directions produced by the original surrogate and its randomized variants. Directions that receive consistent support across variants are assigned greater weights, allowing their shared signal to guide a more reliable attack update. In short, RSE diversifies the surrogates, while ISNC finds where their attack directions agree. Together, they make richer use of existing surrogates and greatly improve transferability across MLLMs. As shown in Figure 1, DA-Attack successfully shifts the semantics perceived by advanced MLLMs toward the target images and achieves superior attack performance across various MLLMs. We summarize the main contributions as follows:

- We analyze the practical bottleneck in transferable attacks against MLLMs. Even multi-paradigm attacks rely on a limited set of accessible surrogates, whereas collecting or train-

ing additional models is costly. We therefore study how to extract richer and more reliable attack directions from the same surrogate set.

- We propose DA-Attack, a simple diversify-and-agree framework for transferable adversarial attacks. RSE expands each surrogate into diverse parameter variants through random pruning, while ISNC stabilizes optimization by placing more trust in the directions shared across these variants.
- Extensive experiments demonstrate that DA-Attack achieves state-of-the-art attack performance across open-source and closed-source MLLMs, raising the average targeted attack success rate over 12 victim models from 70.4% to 82.1% on NIPS 2017 and from 28.5% to 45.6% on Flickr30K, compared with the strongest baseline. We further validate the effectiveness of DA-Attack on recent advanced MLLMs such as GPT 6 Astra, Claude Fable 5.1, and Kimi K3.

2 Related Work

Multimodal Large Language Models. MLLMs combine visual encoders with LLMs to interpret images and generate open-ended responses. With large-scale image–text pretraining (Zhai et al., 2023) and multimodal instruction tuning (Liu et al., 2023), they have progressed from image captioning and visual question answering to multi-image, video, 3D understanding, and complex visual reasoning tasks. Open-source models such as BLIP-2 (Li et al., 2023), Qwen-VL (Bai et al., 2025), DeepSeek (Lu et al., 2024a) and InternVL (Zhu et al., 2025) make multimodal research broadly accessible, while closed-source models such as GPT, Claude, and Gemini exhibit stronger multimodal understanding and reasoning capabilities. With the rapid development and growing adoption of MLLMs, their security has received increasing attention (Jia et al., 2025a; Liu et al., 2025).

Adversarial Attacks on MLLMs. Deep neural networks are vulnerable to adversarial examples (Goodfellow et al., 2014; Nie et al., 2026), and such attacks can also pose threats in physical-world scenarios (Wang et al., 2026b; 2025a; Yong et al., 2026). In real-world settings, attackers typically have no prior knowledge of the target MLLM's architecture, parameters, or training process. Transfer-based attacks, which optimize adversarial images on publicly available large-scale visual foundation models such as CLIP (Radford et al., 2021), therefore provide a practical way to challenge black-box MLLMs. Early work such as AnyAttack (Zhang et al., 2025) trains a general-purpose adversarial perturbation generator. Although it performs well on retrieval tasks with CLIP encoders, it transfers poorly to MLLMs. Recent methods improve attack performance through local crop matching, more stable gradient updates, and more diverse surrogate models. Specifically, M-Attack (Li et al., 2025b) uses local views to embed recognizable target details into perturbations, while FOA-Attack (Jia et al., 2025b) further aligns global representations and patch-level features. M-Attack-V2 (Zhao et al., 2026) stabilizes local matching by aggregating gradients from multiple crops, whereas PRAF-Attack (Wang et al., 2026a) gradually introduces target details from coarse to fine and exploits intermediate-layer features. MPCAttack (Li et al., 2026) combines surrogates trained through cross-modal alignment, multimodal understanding, and visual self-supervision to improve transfer across different MLLMs. In this work, we further improves transferability by generating diverse parameter variants from the existing surrogate models.

3 Method

DA-Attack extracts multiple attack directions from a fixed surrogate set and places greater weight on directions shared across parameter variants. RSE generates the variants online, and ISNC weights their gradients by directional agreement before combining surrogates. The resulting consensus drives a targeted transfer attack against an unseen MLLM. The overall framework of DA-Attack is shown in Figure 2.

3.1 Preliminaries

DA-Attack operates in a practical black-box setting. Let $x_s \in [0,1]^{H\times W\times 3}$ denote a source image, x_t a target image, and ϵ the perturbation budget. We seek a perturbation δ that forms $x_{\mathrm{adv}} = x_s + \delta$,

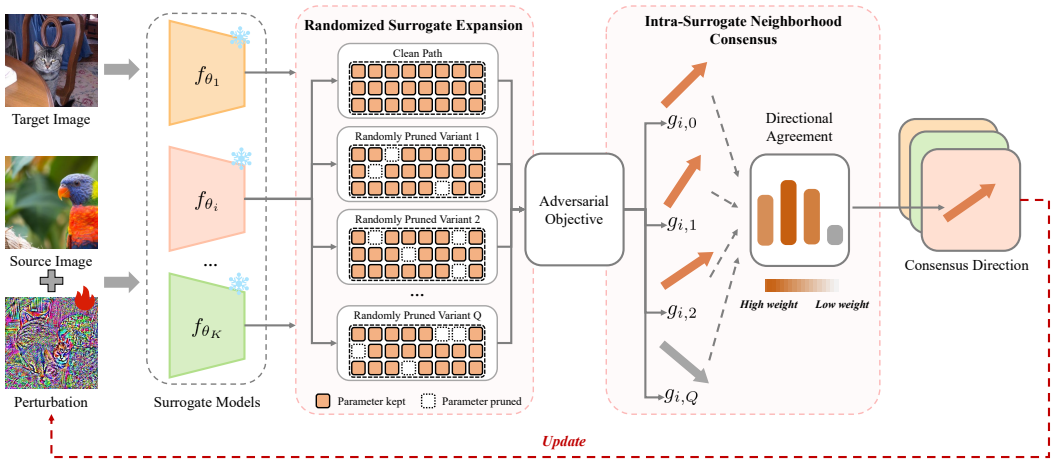


Figure 2: Overview of the proposed DA-Attack. DA-Attack diversifies each surrogate with RSE and uses ISNC to reinforce attack directions shared across its variants, improving transferability to unseen MLLMs.

with $\|\delta\|_\infty \leq \epsilon$, and induces the victim MLLM to interpret x_{adv} as semantically similar to x_t. The attacker receives no information from the victim MLLM during optimization.

Following prior works (Li et al., 2025b; Jia et al., 2025b; Zhao et al., 2026), we optimize δ on K accessible surrogate encoders $\mathcal{F} = \{f_{\theta_i}\}_{i=1}^{K}$, where θ_i denotes the parameters of surrogate i. Each surrogate extracts adversarial, source, and target features:

$$z_i^a = f_{\theta_i}(x_{\text{adv}}), \qquad z_i^s = f_{\theta_i}(x_s), \qquad z_i^t = f_{\theta_i}(x_t). \tag{1}$$

The attack pulls z_i^a toward the target feature z_i^t and separates it from the source feature z_i^s. Let $s_i^t = \cos(z_i^a, z_i^t)$ and $s_i^s = \cos(z_i^a, z_i^s)$. Following MPCAttack (Li et al., 2026), we optimize the binary contrastive objective

$$\begin{aligned} J_i(\delta;\theta_i) &= -\log \frac{\exp(\omega s_i^t/\tau)}{\exp(\omega s_i^t/\tau) + \exp(s_i^s/\tau)}, \\ \min_{\|\delta\|_\infty \leq \epsilon} \quad & \frac{1}{K}\sum_{i=1}^{K} J_i(\delta;\theta_i), \end{aligned} \tag{2}$$

where τ is the temperature and ω controls target attraction. Optimizing the perturbation across multiple surrogate models improves its transferability to victim MLLMs.

3.2 Randomized Surrogate Expansion

Larger surrogate ensembles often improve transferability, but collecting and running additional models is computational expensive. RSE instead turns each surrogate into a set of model variants through lightweight parameter masking. Since these variants are sampled online throughout optimization, the attack continually encounters new parameter configurations, substantially enriching surrogate diversity.

Specifically, for surrogate f_{θ_i}, RSE independently samples Q binary masks at iteration t:

$$M_{i,q}^{(t)} \sim \text{Bernoulli}(1-p), \qquad \widetilde{\theta}_{i,q}^{(t)} = M_{i,q}^{(t)} \odot \theta_i, \quad q = 1, \ldots, Q. \tag{3}$$

Here, p is the pruning probability, and $\odot$ denotes element-wise multiplication. Each entry of $M_{i,q}^{(t)}$ equals one with probability $1-p$ and zero with probability p, which determines whether RSE retains or prunes the corresponding parameter. In practice, RSE masks the weights and biases of attention and multilayer perceptron (MLP) linear layers and the affine parameters of LayerNorm. It leaves patch embeddings, class and positional embeddings, and the final visual projection unchanged, following RaPA (Su et al., 2026). All mask entries outside this scope remain one.

We preserve the original surrogate as the clean path, with $q = 0$ and $\widetilde{\theta}_{i,0}^{(t)} = \theta_i$, to anchor the neighborhood to the unmodified model. Each iteration therefore provides one clean path and Q randomized paths per surrogate.

Each path computes all three features under one parameter variant:

$$z_{i,q}^{a} = f_{\widetilde{\theta}_{i,q}^{(t)}}(x_{\mathrm{adv}}), \qquad z_{i,q}^{s} = f_{\widetilde{\theta}_{i,q}^{(t)}}(x_s), \qquad z_{i,q}^{t} = f_{\widetilde{\theta}_{i,q}^{(t)}}(x_t). \tag{4}$$

Sharing $\widetilde{\theta}_{i,q}^{(t)}$ across the triplet keeps the three features in the same feature space. Let $J_{i,q}$ denote the objective in Eq. 2 evaluated on path q. Its gradient with respect to the perturbation is

$$g_{i,q}^{(t)} = \nabla_{\delta} J_{i,q}, \qquad q = 0, \ldots, Q. \tag{5}$$

The clean path and its Q randomized paths form a parameter neighborhood. Next, we detail how ISNC measures agreement within this neighborhood to recover a more consistent attack direction.

3.3 Intra-Surrogate Neighborhood Consensus

RSE broadens the direction set, but gradients from random paths can conflict. Uniform averaging can cancel useful components and weaken the resulting update. Therefore, ISNC is devised to assign greater weight to directions that recur across parameter variants.

For surrogate i, ISNC compares the $Q + 1$ gradients in Eq. 5. We omit the iteration superscript below for clarity. The directional agreement between paths q and r is

$$A_i(q, r) = \frac{\langle g_{i,q}, g_{i,r} \rangle}{\|g_{i,q}\|_2 \, \|g_{i,r}\|_2}, \qquad q, r \in \{0, \ldots, Q\}. \tag{6}$$

ISNC scores each path by its mean agreement with the other paths and normalizes the scores into weights:

$$a_{i,q} = \frac{1}{Q} \sum_{r \neq q} A_i(q, r), \qquad \alpha_{i,q} = \frac{\max(a_{i,q}, 0) + \rho}{\sum_{\ell=0}^{Q} \big[\max(a_{i,\ell}, 0) + \rho\big]}. \tag{7}$$

The positive part removes negative agreement from the confidence score, while the smoothing constant $\rho > 0$ keeps every path active.

To prevent differences in gradient scale from dominating the aggregation, we normalize each gradient by its mean absolute value before applying the consensus weights:

$$\widehat{g}_{i,q} = \frac{g_{i,q}}{\mathrm{mean}(|g_{i,q}|)}, \qquad \bar{g}_i = \sum_{q=0}^{Q} \alpha_{i,q} \widehat{g}_{i,q}. \tag{8}$$

The weighted sum $\bar{g}_i$ gives one consensus direction for surrogate i. Averaging these directions across the K surrogates yields the final attack gradient:

$$g^{(t)} = \frac{1}{K} \sum_{i=1}^{K} \bar{g}_i. \tag{9}$$

3.4 Attack and Evaluation Pipeline

The final gradient drives a standard momentum-based iterative attack under the ℓ_∞ constraint. After optimization, we transfer x_{adv} directly to an unseen victim MLLM.

We follow the evaluation protocol of MPCAttack. The victim MLLM describes the adversarial and target images under the same prompt, and a separate language model (e.g., GPT-4o mini) scores the semantic similarity between the two descriptions. We count the attack as successful when this score exceeds a predefined threshold. If a dataset provides a ground-truth description for the target image, we use it in place of a generated target description.

Table 1: Performance of ASR (%) and AvgSim on open-source MLLMs on NIPS dataset. The best results are shown in bold, and the second-best results are underlined.

Methods	Qwen2.5-VL-7B		InternVL3-8B		LLaVA-v1.5-7B		GLM-4.1V-9B-Thinking	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	0.90	0.0293	1.00	0.0258	1.10	0.0272	1.30	0.0318
M-Attack	14.80	0.1753	68.10	0.5331	57.50	0.4830	31.10	0.2979
FOA-Attack	18.30	0.1955	72.70	0.5606	62.40	0.5074	35.20	0.3226
MPCAttack	21.50	0.2259	82.80	0.6119	63.50	0.5181	47.90	0.3972
M-Attack-V2	29.80	0.2695	84.90	0.6315	70.50	0.5611	59.50	0.4680
DA-Attack	**64.90**	**0.5142**	**93.00**	**0.6996**	**76.50**	**0.6021**	**72.80**	**0.5614**
Methods	DeepSeek-VL-7B		MiniCPM-V-4.5		Ovis2-8B		Qwen3.5-122B-A10B	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	0.80	0.0261	1.00	0.0310	1.10	0.0308	0.70	0.0227
M-Attack	56.90	0.4617	62.80	0.4943	69.30	0.5286	57.70	0.4749
FOA-Attack	62.60	0.4905	65.20	0.5229	72.80	0.5539	62.10	0.4971
MPCAttack	66.20	0.5147	74.50	0.5636	81.00	0.5992	71.00	0.5402
M-Attack-V2	72.40	0.5608	80.30	0.6018	83.00	0.6129	76.10	0.5804
DA-Attack	**79.60**	**0.6026**	**89.40**	**0.6561**	**91.50**	**0.6710**	**85.60**	**0.6366**

Table 2: Performance of ASR (%) and AvgSim on closed-source MLLMs on NIPS dataset.

Methods	GPT-4o		GPT-5		Claude-3.5-Sonnet		Gemini-3.5-Flash	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	0.40	0.0244	0.60	0.0238	0.30	0.0174	0.30	0.0164
M-Attack	65.00	0.5197	68.00	0.5304	44.70	0.3867	29.90	0.2935
FOA-Attack	70.30	0.5462	72.80	0.5587	47.70	0.4052	34.50	0.3260
MPCAttack	78.80	0.5970	80.00	0.5905	59.30	0.4755	54.22	0.4577
M-Attack-V2	82.60	0.6318	84.10	0.6242	73.20	0.5601	48.10	0.4108
DA-Attack	**90.20**	**0.6756**	**89.50**	**0.6679**	**78.50**	**0.5947**	**73.35**	**0.5570**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Datasets. Following (Jia et al., 2025b; Li et al., 2026), we randomly sample 1,000 images from the NIPS 2017 Adversarial Attacks and Defenses Competition dataset (Kurakin et al., 2018) as source images and 1,000 images from ImageNet (Deng et al., 2009) as target images. In addition, we evaluate the attack on Flickr30K (Young et al., 2014), which provides ground-truth captions for each image. We randomly sample 1,000 Flickr30K images and pair them at random to form source–target pairs. This setting is more challenging because evaluation relies on the ground-truth captions of the target images. All ablation experiments are conducted on 100 images randomly selected from the NIPS dataset.

Victim models. We perform attacks on eight open-source MLLMs (Qwen2.5-VL-7B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), LLaVA-v1.5-7B (Liu et al., 2024), GLM-4.1V-9B-Thinking (Hong et al., 2025), DeepSeek-VL-7B (Lu et al., 2024a), MiniCPM-V-4.5 (Yu et al., 2025), Ovis2-8B (Lu et al., 2024b), and Qwen3.5-122B-A10B) and four closed-source models (GPT-4o (Hurst et al., 2024), GPT-5, Claude-3.5-Sonnet, and Gemini-3.5-Flash (Team et al., 2023)). All 12 MLLMs share the same text prompt: "*Describe this image in one concise sentence, no longer than 20 words*".

Evaluation metrics. Following previous works (Zhao et al., 2026; Li et al., 2026), we adopt the LLM-as-a-judge framework, in which the victim MLLM generates captions for both the adversarial and target images, and GPT-4o-mini assigns a GPTScore based on their semantic similarity. An attack is considered successful when GPTScore exceeds 0.5, indicating that the victim MLLM interprets the adversarial image as semantically similar to the target. We report the attack success rate (ASR) and AvgSim across 1000 attempts to evaluate the attack performance.

Table 3: Performance of ASR (%) and AvgSim on open-source MLLMs on Flickr30K dataset.

Methods	Qwen2.5-VL-7B		InternVL3-8B		LLaVA-v1.5-7B		GLM-4.1V-9B-Thinking	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	1.82	0.0618	2.24	0.0718	2.36	0.0739	1.80	0.0763
M-Attack	4.36	0.0950	18.80	0.2177	34.74	0.3371	19.30	0.2408
FOA-Attack	4.88	0.0978	19.26	0.2212	34.82	0.3389	21.88	0.2565
MPCAttack	4.58	0.0943	27.18	0.2736	40.04	0.3636	25.68	0.2835
M-Attack-V2	4.96	0.0999	25.32	0.2634	45.62	0.3992	31.84	0.3177
DA-Attack	**15.18**	**0.1876**	**56.00**	**0.4566**	**58.34**	**0.4679**	**54.00**	**0.4461**
Methods	DeepSeek-VL-7B		MiniCPM-V-4.5		Ovis2-8B		Qwen3.5-122B-A10B	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	1.98	0.0738	2.80	0.0809	2.70	0.0727	1.70	0.0590
M-Attack	32.06	0.3263	21.98	0.2480	27.86	0.2890	24.54	0.2586
FOA-Attack	34.54	0.3436	21.94	0.2505	29.08	0.2947	24.80	0.2642
MPCAttack	35.58	0.3529	24.96	0.2681	35.12	0.3295	27.10	0.2772
M-Attack-V2	46.06	0.4132	33.00	0.3153	40.70	0.3615	32.62	0.3127
DA-Attack	**57.86**	**0.4772**	**53.38**	**0.4422**	**61.06**	**0.4789**	**55.06**	**0.4519**

Table 4: Performance of ASR (%) and AvgSim on closed-source MLLMs on Flickr30K dataset.

Methods	GPT-4o		GPT-5		Claude-3.5-Sonnet		Gemini-3.5-Flash	
	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)	ASR (↑)	AvgSim (↑)
AnyAttack	1.90	0.0670	1.72	0.0573	1.64	0.0605	1.40	0.0560
M-Attack	16.82	0.2008	16.56	0.2072	22.04	0.2417	3.36	0.0825
FOA-Attack	18.20	0.2143	18.14	0.2200	24.82	0.2567	3.56	0.0861
MPCAttack	19.52	0.2226	20.26	0.2247	21.14	0.2258	3.48	0.0801
M-Attack-V2	23.46	0.2472	23.46	0.2525	31.86	0.3078	3.32	0.0854
DA-Attack	**40.98**	**0.3683**	**44.44**	**0.3941**	**39.96**	**0.3617**	**10.49**	**0.1562**

Implementation details. The surrogate ensemble consists of three CLIP models (Radford et al., 2021; Cherti et al., 2023) (ViT-B/16, ViT-B/32, and LAION ViT-G/14), InternVL3-1B, and DINOv2 (Oquab et al., 2024). RSE is applied only to the three CLIP encoders. At each iteration, we keep one clean path and sample $Q = 3$ randomized paths using a default pruning probability of $p = 0.03$. The perturbation budget ϵ under the ℓ_∞-norm constraint, step size, and number of attack iterations are set to $16/255$, $1/255$, and 300, respectively.

4.2 Comparison Results

We compare DA-Attack with five recent transfer-based adversarial attacks: AnyAttack (Zhang et al., 2025), M-Attack (Li et al., 2025b), FOA-Attack (Jia et al., 2025b), MPCAttack (Li et al., 2026), and M-Attack-V2 (Zhao et al., 2026). On NIPS 2017, the results across eight open-source and four closed-source MLLMs are summarized in Tables 1 and 2. DA-Attack achieves the highest ASR and AvgSim on every open-source model. The strongest prior baseline, M-Attack-V2, stabilizes local matching by averaging gradients over multiple source crops, thereby reducing the high variance induced by different crop gradients. In contrast, DA-Attack adopts a simple yet remarkably effective diversify-and-agree strategy. It expands surrogate diversity and updates the perturbation along directions supported by multiple surrogate variants, reducing reliance on any particular surrogate and improving transfer to unseen black-box MLLMs. On GPT-4o, GPT-5, Claude-3.5-Sonnet, and Gemini-3.5-Flash, DA-Attack achieves ASRs of 90.2%, 89.5%, 78.5%, and 73.35%, outperforming the strongest baseline on each model by 7.6%, 5.4%, 5.3%, and 19.1%, respectively. These results show that DA-Attack poses a broad transfer-based threat to both open- and closed-source MLLMs.

We further report attack results on Flickr30K in Tables 3 and 4. Unlike the NIPS setting, Flickr30K uses ground-truth target captions as references for comparing the description generated from the adversarial image. The attack must therefore match target semantics expressed by an external reference rather than another caption produced by the victim MLLM, making the evaluation more challenging. Under this stricter setting, the gains of DA-Attack become more pronounced. For example, on open-source models, it improves ASR over the strongest baseline by 28.8% on InternVL3-8B, 22.4% on Qwen3.5-122B-A10B, and 22.2% on GLM-4.1V-9B-Thinking. On closed-source models, the cor-



Figure 3: Visualization of adversarial images and perturbations.

Table 5: Ablation results of RSE and ISNC. The ASR (%) and AvgSim are reported. $*$ indicates that the features extracted by different surrogate models are concatenated before computing the attack loss, following MPCAttack. RSE-all applies RSE to all five surrogates. CSC denotes cross-surrogate consensus, which weights gradients according to their directional agreement across different surrogate models.

Variants	Qwen2.5-VL-7B		InternVL3-8B		LLaVA-v1.5-7B		GLM-4.1V-9B		Average	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
Baseline	22.00	0.2350	91.00	0.6490	66.00	0.5600	55.00	0.4540	58.50	0.4745
+RSE-all*	64.00	0.5040	83.00	0.5950	63.00	0.5180	71.00	0.5460	70.25	0.5408
+RSE*	57.00	0.4790	91.00	0.6570	70.00	0.5550	77.00	0.5680	73.75	0.5648
+RSE	65.00	0.5100	97.00	0.7120	77.00	0.5850	77.00	0.5660	79.00	0.5933
+CSC	43.00	0.3490	96.00	0.7120	**84.00**	**0.6430**	60.00	0.4660	70.75	0.5425
+RSE+ISNC+CSC	69.00	**0.5420**	96.00	0.7270	76.00	0.6210	76.00	0.5710	79.25	0.6153
+RSE+ISNC	**71.00**	0.5260	**98.00**	**0.7330**	81.00	0.6280	**84.00**	**0.6060**	**83.50**	**0.6233**

responding gains reach 21.0% on GPT-5, 17.5% on GPT-4o, and 8.1% on Claude-3.5-Sonnet. These improvements highlight the importance of surrogate diversity and consistent gradient directions in transferable adversarial attacks.

4.3 Ablation Study

Effects of RSE and ISNC. DA-Attack follows the default setting of MPCAttack, containing three CLIP encoders with the self-supervised visual model DINOv2 and the multimodal understanding model InternVL3-1B. Table 5 reports the ablation results.

Applying RSE to all surrogate models (+RSE-all*) improves the average ASR from 58.50% to 70.25%. Restricting RSE to the three CLIP encoders (+RSE*) further raises it to 73.75%. Notably, +RSE-all* performs worse than the baseline on InternVL3-8B and LLaVA-v1.5-7B, which suggests that InternVL and DINOv2 are more useful as stable semantic anchors, while the CLIP encoders provide sufficient parameter diversity. Since MPCAttack concatenates features extracted by different surrogates before computing the attack loss, semantic discrepancies across different models may interfere with attack objective, especially when some features are produced by randomly pruned models. We therefore make a simple change: each surrogate computes its own attack objective, while RSE remains applied to the three CLIP encoders. This variant (+RSE) largely boost the average ASR to 79.00%.

Before introducing ISNC, we first apply gradient consensus across surrogate models, denoted as CSC. CSC follows the same agreement-based weighting procedure as ISNC, but achieves consensus across different surrogate models rather than parameter variants of the same surrogate. Applying CSC alone improves the average ASR and AvgSim from 58.50% and 0.4745 to 70.75% and 0.5425, respectively. Building on RSE, ISNC achieves the strongest attack performance, with an average ASR of 83.50% and an AvgSim of 0.6233. Further applying CSC after RSE+ISNC instead reduces

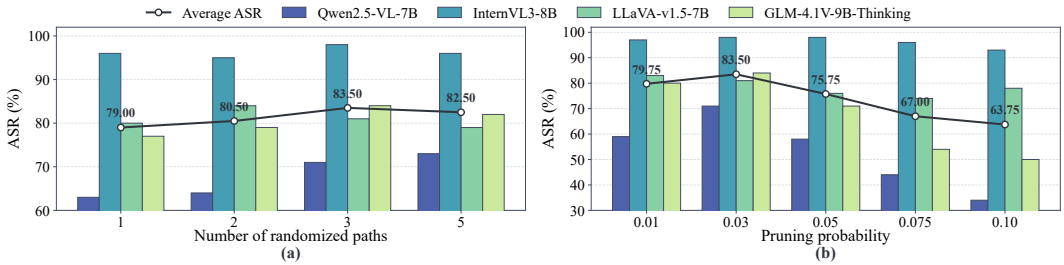


Figure 4: (a) Impact of the number of randomized paths Q; (b) Impact of the pruning probability p.

Table 6: Performance of DA-Attack on recent advanced MLLMs on NIPS dataset.

GPT 6 Astra		GPT 5.6 sol		Claude Fable 5.1		Claude Opus 5		Kimi K3	
ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
60.00	0.483	65.00	0.509	57.00	0.454	29.00	0.265	83.00	0.617

these results to 79.25% and 0.6153. This suggests that preserving diverse gradient directions from different representation paradigms is beneficial for adversarial transferability.

Effects of Q and p. Figure 4 studies the number of randomized paths Q and the pruning probability p. Increasing Q from 1 to 3 raises the average ASR across the four MLLMs from 79.00% to 83.50%, while further increasing Q brings only marginal gains. For the pruning probability, $p = 0.03$ performs best, whereas larger values weaken the attack performance. An overly large p may disrupt feature extraction and discard useful visual information, thereby reducing transferability. We therefore set $Q = 3$ and $p = 0.03$ by default.

4.4 Evaluation on Recent Advanced MLLMs

To examine the transferability of DA-Attack to recent advanced MLLMs, we conduct an additional evaluation on 100 randomly sampled source–target pairs under the NIPS setting. The results are reported in Table 6.

Despite the continued improvement in the reasoning and understanding capabilities of MLLMs, adversarial attacks can still effectively manipulate their outputs. For example, DA-Attack achieves ASRs of 60.00% and 57.00% on the recent GPT 6 Astra and Claude Fable 5.1, respectively. Moreover, improved model capabilities do not necessarily imply greater adversarial robustness: Claude Opus 5, though less capable overall than Claude Fable 5.1, has a lower ASR of 29.00%. These findings underscore the urgent need for effective adversarial defenses for MLLMs.

5 Conclusion

In this paper, we identify a practical bottleneck in transferable adversarial attacks against MLLMs: the set of accessible surrogate models is limited, while collecting or training additional models is costly. To address this, we propose DA-Attack, a diversify-and-agree framework that extracts richer transferable guidance from existing surrogates. RSE diversifies each surrogate into online-sampled parameter variants via random pruning, and ISNC reinforces the attack directions consistently shared across these variants, yielding reliable consensus updates. Extensive experiments on both open-source and closed-source MLLMs demonstrate the superior attack performance of DA-Attack.

Our findings reveal that stronger reasoning capabilities do not necessarily confer adversarial robustness, underscoring the urgent need for principled multimodal defenses. Nevertheless, two limitations remain. First, the current pruning strategy is uniform and heuristic, and exploring structure-aware or adaptive surrogate expansion is a promising direction. Second, prevailing evaluations of attacks on MLLMs rely on ASR under the LLM-as-a-judge protocol, which mainly captures perception-level semantic errors. How adversarial perturbations further affect the reasoning and understanding of MLLMs calls for more comprehensive evaluation metrics in future work.

AI USE STATEMENT

We used generative AI tools to assist with the implementation of Randomized Surrogate Expansion. We also used LLMs to prepare an initial textual draft of the Method section based on the authors' existing method and technical specifications, with the aim of improving clarity and identifying potential errors in its description. Additionally, we used LLMs to polish the English writing in the Introduction and Experiments sections, primarily to correct grammatical errors and improve coherence and readability.

We did not use generative AI tools for literature retrieval, research ideation, experiment design, synthetic dataset generation, or proving mathematical claims.

We reviewed all AI-assisted work. The AI-assisted RSE code was verified and tested for correctness by two authors. The authors reviewed and revised all AI-assisted text, checking the method description against the implementation and ensuring that the language revisions preserved the intended technical meaning and accurately reflected the experimental settings and results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work investigates adversarial vulnerabilities in MLLMs to support security assessment and future defense research. The code has been anonymized and included in the supplementary materials to support reproducibility.

REFERENCES

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.

Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and Junxian He. Bring reason to vision: Understanding perception and reasoning through model merging. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9803–9817. PMLR, 2025.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, June 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Debidatta Dwibedi, Vidhi Jain, Jonathan Tompson, Andrew Zisserman, and Yusuf Aytar. FlexCap: Describe anything in images in controllable detail. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

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

Qi Guo, Xiaojun Jia, Shanmin Pang, Simeng Qin, Lin Wang, Ju Jia, Yang Liu, and Qing Guo. PhysPatch: A physically realizable and transferable adversarial patch attack for multimodal large language models-based autonomous driving systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4412–4420, 2026.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Hanxun Huang, Sarah Monazam Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on CLIP. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25204–25234. PMLR, 2025.

Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-SEA: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20514–20523, June 2023.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xiaojun Jia, Sensen Gao, Qing Guo, Simeng Qin, Ke Ma, Yihao Huang, Yang Liu, Ivor W. Tsang, and Xiaochun Cao. Semantic-aligned adversarial evolution triangle for high-transferability vision-language attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10): 8489–8505, 2025a.

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. *arXiv preprint arXiv:2505.21494*, 2025b.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In Sergio Escalera and Markus Weimer (eds.), *The NIPS '17 Competition: Building Intelligent Systems*, The Springer Series on Challenges in Machine Learning, pp. 195–231. Springer, Cham, 2018.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-Interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

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, 2023.

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, June 2026.

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, 2025b.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision-language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19525–19545, 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, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, June 2024.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. DeepSeek-VL: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Hui Lu, Yi Yu, Yiming Yang, Chenyu Yi, Xueyi Ke, Qixing Zhang, Bingquan Shen, Alex Kot, and Xudong Jiang. Universal adversarial attacks against closed-source MLLMs via target-view routed meta optimization. *arXiv preprint arXiv:2601.23179*, 2026.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024b.

Sen Nie, Jie Zhang, Jianxin Yan, Shiguang Shan, and Xilin Chen. V-attack: Targeting disentangled value features for controllable adversarial attacks on lvlms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42257–42267, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. OpenAI GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2026.

Tongrui Su, Qingbin Li, Shengyu Zhu, Wei Chen, and Xueqi Cheng. RaPA: Enhancing transferable targeted attacks via random parameter pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6538–6548, June 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Haobo Wang, Xiaorong Ma, Weiqi Luo, Xiaojun Jia, and Jiwu Huang. Adversarial attacks against MLLMs via progressive resolution processing and adaptive feature alignment. *arXiv preprint arXiv:2605.09902*, 2026a.

Jian Wang, Fan Li, Song Lv, Lijun He, and Chao Shen. Physically realizable adversarial creating attack against vision-based BEV space 3d object detection. *IEEE Transactions on Image Processing*, 34:538–551, 2025a.

Jian Wang, Lijun He, Yixing Yong, Haixia Bi, and Fan Li. Invisible triggers, visible threats! road-style adversarial creation attack for visual 3d detection in autonomous driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):9903–9911, 2026b.

Lu Wang, Tianyuan Zhang, Yang Qu, Siyuan Liang, Yuwei Chen, Aishan Liu, Xianglong Liu, and Dacheng Tao. Black-box adversarial attack on vision language models for autonomous driving. *arXiv preprint arXiv:2501.13563*, 2025b.

Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Dissecting adversarial robustness of multimodal LM agents. In *International Conference on Learning Representations*, 2025.

Binyan Xu, Xilin Dai, Di Tang, and Kehuan Zhang. One surrogate to fool them all: Universal, transferable, and targeted adversarial attacks with CLIP. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3087–3101. ACM, 2025.

Yixing Yong, Jian Wang, Ming Lei, Lijun He, and Fan Li. Unleashing the representational power of fourier shapes for attacking infrared object detection. *arXiv preprint arXiv:2605.17822*, 2026.

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.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

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, June 2025.

Xiaohan Zhao, Zhaoyi Li, Yaxin Luo, Jiacheng Cui, and Zhiqiang Shen. Pushing the frontier of black-box LVLM attacks via fine-grained detail targeting. *arXiv preprint arXiv:2602.17645*, 2026.

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

Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, and Cong Wang. CALM: Curiosity-driven auditing for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27757–27764, 2025.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

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 Appendix

A.1 GPTScore Evaluation Prompt

We use the following prompt for GPT-4o-mini to evaluate the semantic similarity between two descriptions. {text1} and {text2} are replaced with the descriptions being compared.

GPTScore Evaluation Prompt

Rate the semantic similarity between the following two texts on a scale from 0 to 1.

Criteria for similarity measurement:

1. **Main Subject Consistency:** If both descriptions refer to the same key subject or object (e.g., a person, food, an event), they should receive a higher similarity score.
2. **Relevant Description:** If the descriptions are related to the same context or topic, they should also contribute to a higher similarity score.
3. **Ignore Fine-Grained Details:** Do not penalize differences in **phrasing, sentence structure, or minor variations in detail.** Focus on **whether both descriptions fundamentally describe the same thing.**
4. **Partial Matches:** If one description contains extra information but does not contradict the other, they should still have a high similarity score.
5. **Similarity Score Range:**
   - **1.0:** Nearly identical in meaning.
   - **0.8–0.9:** Same subject, with highly related descriptions.
   - **0.7–0.8:** Same subject, core meaning aligned, even if some details differ.

```
          • 0.5–0.7: Same subject but different perspectives or
            missing details.
          • 0.3–0.5: Related but not highly similar (same general
            theme but different descriptions).
          • 0.0–0.2: Completely different subjects or unrelated
            meanings.
Text 1: {text1}
Text 2: {text2}
Output only a single number between 0 and 1. Do not include any
explanation or additional text.
```

A.2 DA-ATTACK ALGORITHM

Algorithm 1 summarizes DA-Attack, which computes losses and gradients independently for each parameter variant, forms consensus within each surrogate, and averages the resulting surrogate directions to update the perturbation.

Algorithm 1 Optimization process of DA-Attack.

Input: Source image x_s, target image x_t, surrogate encoders $\{f_{\theta_i}\}_{i=1}^{K}$.
Parameters: Iterations N, perturbation budget ϵ, step size η, momentum coefficient μ, randomized paths Q, pruning probability p, crop distribution $\mathcal{D}$, temperature τ, target weight ω, consensus smoothing $\rho > 0$, numerical constant $\epsilon_g > 0$.
Output: Adversarial image x_{adv}.

Initialize perturbation and momentum: $\delta^{(0)} \leftarrow 0$, $m^{(0)} \leftarrow 0$.
for $t = 0, \ldots, N-1$ **do**
  Form $x_{\mathrm{adv}}^{(t)} \leftarrow x_s + \delta^{(t)}$ and sample a shared crop $T^{(t)} \sim \mathcal{D}$.
  for $i = 1, \ldots, K$ **do**
    Retain the clean path: $\widetilde{\theta}_{i,0}^{(t)} \leftarrow \theta_i$.
    for $q = 1, \ldots, Q$ **do**
      Sample $M_{i,q}^{(t)} \sim \mathrm{Bernoulli}(1-p)$ on prunable parameters; set other entries to 1.
      Construct a randomized variant: $\widetilde{\theta}_{i,q}^{(t)} \leftarrow M_{i,q}^{(t)} \odot \theta_i$.
    end for
    for $q = 0, \ldots, Q$ **do**
      Extract features: $z_{i,q}^{a} \leftarrow f_{\widetilde{\theta}_{i,q}^{(t)}}(T^{(t)}(x_{\mathrm{adv}}^{(t)}))$, $z_{i,q}^{s} \leftarrow f_{\widetilde{\theta}_{i,q}^{(t)}}(x_s)$, $z_{i,q}^{t} \leftarrow f_{\widetilde{\theta}_{i,q}^{(t)}}(x_t)$.
      Compute similarities: $s_{i,q}^{t} \leftarrow \cos(z_{i,q}^{a}, z_{i,q}^{t})$, $s_{i,q}^{s} \leftarrow \cos(z_{i,q}^{a}, z_{i,q}^{s})$.
      Compute the path loss and gradient: $J_{i,q} \leftarrow -\log \frac{\exp(\omega s_{i,q}^{t}/\tau)}{\exp(\omega s_{i,q}^{t}/\tau)+\exp(s_{i,q}^{s}/\tau)}$, $g_{i,q}^{(t)} \leftarrow \nabla_{\delta^{(t)}} J_{i,q}$.
    end for
    Compute gradient agreement: $A_i(q,r) \leftarrow \frac{\langle g_{i,q}^{(t)}, g_{i,r}^{(t)} \rangle}{\|g_{i,q}^{(t)}\|_2 \|g_{i,r}^{(t)}\|_2 + \epsilon_g}$ $(q \neq r)$, $a_{i,q} \leftarrow \frac{1}{Q} \sum_{\substack{r=0 \\ r \neq q}}^{Q} A_i(q,r)$.
    Assign consensus weights: $\alpha_{i,q} \leftarrow \frac{\max(a_{i,q},0)+\rho}{\sum_{\ell=0}^{Q}[\max(a_{i,\ell},0)+\rho]}$ for all q.
    Normalize path gradients: $\widehat{g}_{i,q} \leftarrow \frac{g_{i,q}^{(t)}}{\mathrm{mean}(|g_{i,q}^{(t)}|)+\epsilon_g}$ for all q.
    Aggregate variants within surrogate i: $\bar{g}_i \leftarrow \sum_{q=0}^{Q} \alpha_{i,q} \widehat{g}_{i,q}$.
  end for
  Average across surrogates: $g^{(t)} \leftarrow \frac{1}{K} \sum_{i=1}^{K} \bar{g}_i$.
  Update momentum: $m^{(t+1)} \leftarrow \mu m^{(t)} + g^{(t)}$.
  Update the perturbation: $\delta^{(t+1)} \leftarrow \mathrm{clip}(\delta^{(t)} - \eta\, \mathrm{sign}(m^{(t+1)}), -\epsilon, \epsilon)$.
  Enforce the valid image range: $\delta^{(t+1)} \leftarrow \mathrm{clip}(x_s + \delta^{(t+1)}, 0, 1) - x_s$.
end for
return $x_{\mathrm{adv}} = x_s + \delta^{(N)}$.

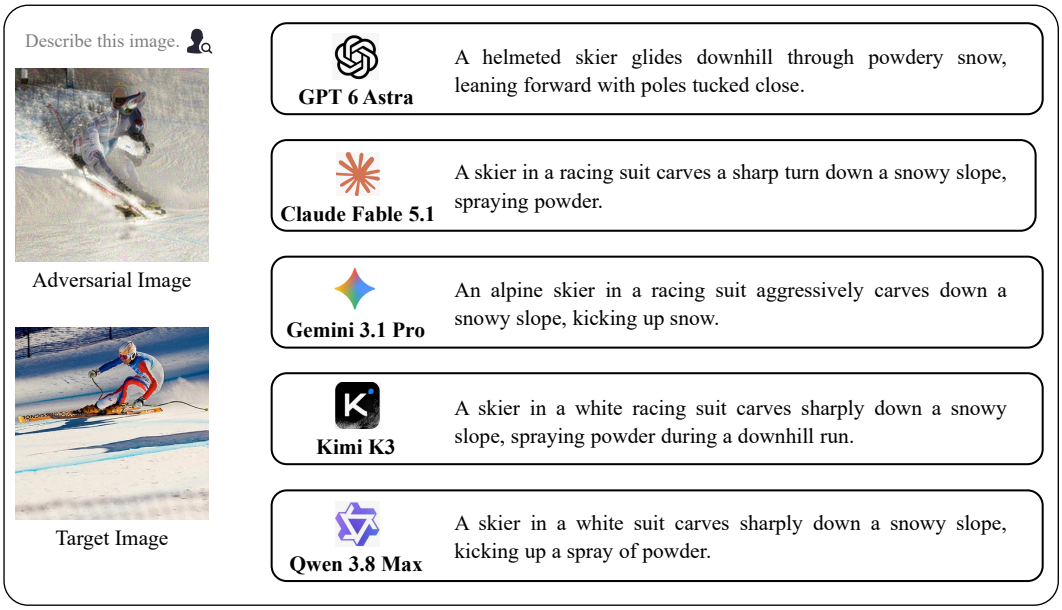


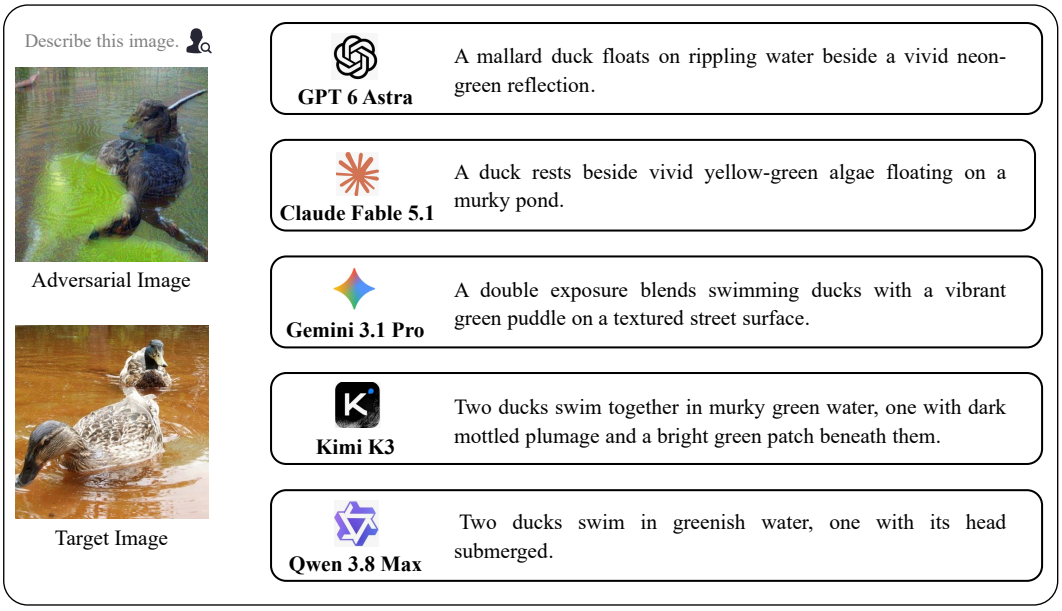
(a)

Figure 5: Additional attack results of DA-Attack. Each example provides an adversarial image, its corresponding target image, and the descriptions of the adversarial image generated by the evaluated MLLMs.

A.3 Additional Attack Results

Figure 5 shows additional examples of adversarial images, their corresponding target images, and the descriptions of the adversarial images generated by the evaluated MLLMs.





(b)

Figure 5: Additional attack results of DA-Attack (continued).