TIGA: IMPROVING TARGETED TRANSFERABILITY ON MLLMS THROUGH GLOBAL-REGIONAL ALIGNMENT AND NEIGHBORHOOD GRADIENT FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box targeted adversarial transfer attacks on Multimodal Large Language Models (MLLMs) aim to induce outputs of attacker-specified semantics through carefully crafted adversarial examples without accessing any internal information of the victim models. Existing methods typically generate adversarial examples by aligning adversarial and target features on accessible surrogate models, utilizing model ensembles and random cropping to further enhance adversarial transferability. However, feature alignment of previous work may overlook coarse regional information relevant to language generation, while gradients evaluated at individual inputs may provide insufficient optimization guidance under local input variations. To address these limitations, we propose the TIGA framework, which improves targeted transferability from both feature and optimization perspectives. Specifically, Global-Regional Collaborative Alignment combines global features with patch representations within each region of a 2×2 grid, preserving spatial information without strict patch-wise correspondence. Moreover, Neighborhood-Aware Gradient Fusion combines normalized gradients from the current adversarial input and a noise-perturbed neighbor, providing complementary guidance for targeted alignment under local input variations. Experiments on six closed-source and six open-source MLLMs demonstrate that TIGA outperforms existing methods in both attack success rates and output semantic similarities, with particularly pronounced gains on the closed-source models.

1 INTRODUCTION

Multimodal large language models (MLLMs) Li et al. (2024); Bai et al. (2025); Hong et al. (2025a) integrate visual perception with language understanding and generation, supporting a broad range of vision-language tasks. MLLMs typically use a vision encoder to extract image features, which are mapped into the embedding space of the language model through a multimodal projector to jointly process visual and textual information, demonstrating remarkable capabilities in visual understanding and reasoning. These advances are reflected in both open-source models, such as Qwen-VL Bai et al. (2023) and closed-source models, *e.g.*, GPT, and Claude. Despite their impressive generalization abilities, MLLMs remain vulnerable to various security threats Liu et al. (2025), including jailbreak attacks Niu et al. (2024); Gong et al. (2025); Ma et al. (2025); Hu et al. (2026), adversarial attacks Dong et al. (2023); Luo et al. (2024); Wang et al. (2025); Li et al. (2026a), prompt injection Qraitem et al. (2024); Yang et al. (2026a), and data poisoning Lyu et al. (2024; 2025). Among these threats, visual adversarial attacks aim to manipulate model responses by introducing carefully crafted adversarial perturbations to input clean images while leaving the textual prompts unchanged.

Since deep neural networks are found to be vulnerable to adversarial examples Szegedy et al. (2013), numerous research has explored adversarial attacks across various settings Sharif et al. (2016); Eykholt et al. (2018); Dong et al. (2018); Kong et al. (2020); Zhong & Deng (2020); Wang & He (2021); Wang et al. (2021; 2023); Gao et al. (2024); Yang et al. (2025a); Jia et al. (2025a); Ren et al. (2025); Zhu et al. (2025b;a); Guo et al. (2026); Qiu et al. (2026); Wang et al. (2026). According to the attacker's access to the victim model, existing adversarial attacks can be broadly categorized into white-box, gray-box, and black-box attack scenarios depending on the attacker's access to the victim models. **White-box** attacks Croce & Hein (2020); Schlarmann & Hein (2023); Bhagwatkar et al. (2024) assume full access to the victim model, including its architecture, param-

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.

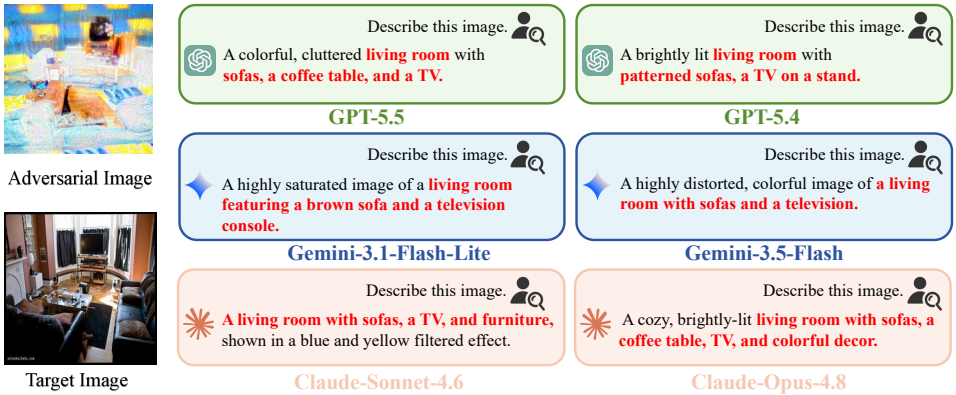

Figure 1: Representative examples of targeted transfer attacks on commercial MLLMs.

eters, and gradient information, enabling high attack success rates. **Gray-box** attacks Mei et al. (2026b;a) assume partial access to the vision encoder of the victim model, whose weights are open-source, achieving strong adversarial transferability across models sharing the same vision encoder. **Black-box attacks** Li et al. (2025); Jia et al. (2025b); Li et al. (2026b) assume no access to any information of the victim models, *e.g.*, architectures, parameters, or gradients, thus posing the most realistic and challenging threat. According to the attack objective, adversarial attacks can be categorized into untargeted attacks and targeted attacks. Compared with **untargeted attacks** that merely disrupt the original model behavior, **targeted attacks** impose a more stringent objective: the adversarial image should elicit responses conveying attacker-specified semantics across different victim MLLMs. Unlike the image classification task, where attack success can be assessed by checking whether predicted labels match original labels or targeted labels, MLLMs generate open-ended textual responses that may express the same semantic content in different ways. For closed-source MLLMs, adversaries have no knowledge of the models and can only interact through API queries while facing unknown commercial safety mechanisms. A natural solution is to leverage adversarial transferability Papernot et al. (2017), where adversarial examples crafted on the surrogate models remain effective against unseen victim models, making it a practical approach for attacking closed-source MLLMs, and representative examples of targeted transfer attacks are shown in Figure 1.

To achieve the semantic adversarial transfer objective, a common strategy is to use a target image as a semantic reference and optimize the adversarial images by aligning their representations on accessible surrogate vision encoders. To enhance transferability, existing methods employ model ensembles to incorporate information from multiple encoders and random cropping to optimize over diverse spatial views Li et al. (2025); Jia et al. (2025b). Local feature matching further extends this approach by incorporating fine-grained target information beyond global representations. Despite these impressive advances, existing targeted transfer attacks still face two key limitations concerning semantic alignment and optimization guidance: **(1) Global-regional mismatch.** Existing targeted transfer attacks mainly encourage the adversarial image to match the target image at the global feature level. However, MLLMs represent visual inputs not only through holistic semantics, but also through region-level tokens that preserve localized visual information. This discrepancy between the attack objective and the visual representation used by MLLMs may cause the adversarial example to capture only coarse target semantics, thereby limiting its transferability across different victim models. **(2) Incomplete gradient guidance under local input variations.** Existing methods commonly employ random cropping to diversify the visual context. However, the gradient obtained from each sampled view reflects the target-alignment direction at a particular configuration, without explicitly accounting for how this direction varies at nearby inputs. As a result, the optimization may overlook complementary guidance available in the local input neighborhood.

To mitigate the two limitations, we propose **T**argeted transfer with **I**nput-neighborhood gradient fusion and **G**lobal-regional **A**lignment (**TIGA**), a targeted transferable attack framework that improves semantic transfer from both representation and optimization perspectives. TIGA first introduces **Global-Regional Collaborative Alignment** to jointly exploit image-level and region-level target semantics, thereby reducing the global-regional mismatch. It further employs **Neighborhood-Aware Gradient Fusion**, which incorporates complementary gradient information from noise-perturbed neighboring inputs, thereby reducing reliance on gradients evaluated at individual input configurations and providing more comprehensive guidance for target alignment. Together, these two compo-

nents enable the adversarial example to capture richer target semantics while preserving them more consistently across heterogeneous victim MLLMs. Our main contributions are as follows:

- We propose **TIGA**, a targeted transfer attack framework for MLLMs that improves adversarial transferability by jointly enhancing target semantic alignment and optimization guidance, enabling effective attacks against unseen victim models.
- We introduce **Global-Regional Collaborative Alignment (GRCA)**, which jointly aligns global and regional representations to provide semantic guidance across multiple spatial granularities, and **Neighborhood-Aware Gradient Fusion (NGF)**, which integrates the gradient at the current adversarial input with one gradient from its local neighborhood to reduce reliance on single-point gradient guidance.
- Extensive experiments on multiple closed-source and open-source MLLMs demonstrate that TIGA consistently improves targeted transfer performance over existing attacks. Comprehensive ablation studies and analyses further validate the effectiveness of GRCA and NGF and provide insights into regional semantic alignment and neighborhood-aware optimization for adversarial transferability.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models(MLLMs) integrate visual and textual inputs to generate coherent textual outputs and have attracted substantial research interest in recent years Liu et al. (2023); Li et al. (2024); Hong et al. (2025a); Yang et al. (2026b). A typical MLLM generally consists of three core components: a vision encoder Radford et al. (2021), a large language model (LLM) backbone, and a multimodal projector. The vision encoder extracts visual information from the input image, and the projector maps it into the embedding space of the LLM. The projected visual features are then combined with textual token embeddings and fed into the LLM backbone for multimodal understanding. MLLMs demonstrate remarkable capabilities across a wide range of vision-language tasks, including image captioning Hu et al. (2022); Chen et al. (2024), visual question answering Dai et al. (2023); Hong et al. (2025b), and reasoning Yang et al. (2025b); Hao et al. (2025).

2.2 Adversarial Attacks for MLLMs

Pioneering transfer-based attacks on MLLMs Zhao et al. (2023); Yin et al. (2023) craft adversarial examples by optimizing against surrogate vision encoders, *e.g*. CLIP Radford et al. (2021), to disrupt image-text alignment and degrade downstream task performance. However, these methods mainly focus on untargeted settings, limiting their transferability and weakening efficiency in targeted scenarios, especially on MLLMs. To achieve effective targeted attacks against MLLMs, a series of works have explored and proposed more sophisticated methods. COA Xie et al. (2025) iteratively optimizes adversarial examples utilizing multimodal embeddings with similarity-based contrastive triple loss. AnyAttack Zhang et al. (2025) pretrains a generator on the LAION-400M dataset Schuhmann et al. (2021) and fine-tunes it on downstream tasks through a self-supervised framework. AdvDiffVLM Guo et al. (2024) adopts diffusion models Ho et al. (2020); Song et al. (2020); Rombach et al. (2022) to generate more transferable adversarial examples. To further enhance target semantic richness of perturbations, M-Attack Li et al. (2025) uses random cropping to better align the semantics of adversarial examples with those of targeted images. FOA-Attack Jia et al. (2025b) incorporates patch-level alignment through optimal transport and dynamic ensemble weighting to improve transferability. M-AttackV2 Zhao et al. (2026) further proposes Multi-Crop Alignment and Auxiliary Target Alignment to strengthen transfer directions. MPC-Attack Li et al. (2026b) combines multiple paradigm models and optimizes perturbations using contrastive learning.

3 Methodology

3.1 Prelimineries

Let x_{clean} denote a benign source image and x_{target} a chosen target image. A targeted transfer-based attack seeks a perturbation δ bounded by $\|\delta\|_\infty \leq \epsilon$, where ϵ is the ℓ_∞ budget, such that the adversarial example $x_{adv} = x_{clean} + \delta$ induces a victim MLLM to generate an output whose semantics align with x_{target} rather than x_{clean}. Under the strict black-box attack setting, the victim

MLLM is inaccessible. We therefore optimize δ using the surrogate vision encoder $f(\cdot)$ and formulate the attack as an optimization problem in Eq.1, which encourages the feature representation of x_{adv} towards the target image semantic:

$$\delta^* = \underset{\|\delta\|_\infty \leq \epsilon}{\arg\min}\, L\big(f(x_{adv}),\, f(x_{target})\big), \tag{1}$$

where L measures the discrepancy between the adversarial and target image features.

3.2 MOTIVATION

> ***Observation 1:*** *Global feature alignment alone may overlook regional visual information that contributes to semantic understanding in MLLMs.*

Adversarial examples optimized on surrogate models are expected to transfer to unseen victim models, analogous to generalization in the normal model training process Lin et al. (2020); Wang & He (2021). In the strict black-box setting, all the information of the victim model, *e.g.*, parameters and language-generation objective, is inaccessible, so perturbations are optimized through a surrogate feature alignment objective, and the features aligned by this objective should therefore reflect the visual information used by the victim model for text generation.

Representative MLLMs condition language generation on sequences of patch-derived visual tokens rather than a single CLS-based global descriptor. For example, LLaVA-1.5 Liu et al. (2023) projects patch features through an MLP connector, while Qwen2.5-VL Bai et al. (2025) and InternVL3 Zhu et al. (2025c) aggregate neighboring patch features before passing them to the language model backbones. Although these architectures process visual tokens differently, they all provide the language backbone with visual information from multiple image regions. Therefore, language generation in these MLLMs may depend not only on global image semantics but also on visual information from different image regions.

Existing transfer-based adversarial attacks mainly rely on global feature alignment, typically using CLS-based representations. Recent works such as FOA-Attack Jia et al. (2025b) exploit local patch features through clustering and optimal transport, but their local alignment focuses on fine-grained alignment, rather than the spatial organization of visual tokens. Moreover, global feature alignment alone does not ensure that the regional visual features of adversarial examples used for language generation are also aligned with those of the target images. Matching a global descriptor imposes no explicit requirement on the correspondence between individual image regions. Consequently, an adversarial image may achieve high global similarity while its regional features still retain information associated with the source image. Such information may continue to influence the generated response, limiting the attack efficacy. **This motivates retaining CLS-based global guidance while incorporating regional feature alignment, so that both overall semantics and spatially organized visual information can contribute to the attack objective.**

> ***Observation 2:*** *Gradients evaluated at the current adversarial input may provide incomplete guidance for target alignment under local input variations.*

Beyond the choice of alignment features, targeted transfer also depends on the gradient information used to guide adversarial optimization. Random cropping provides spatially diverse views of the adversarial image, allowing the optimization to incorporate target-alignment information from different image regions. However, the gradient computed for each sampled view describes the local target-alignment direction at a specific input configuration, without explicitly accounting for how this direction varies at nearby inputs. Consequently, the optimization may insufficiently account for target-alignment directions in the local input neighborhood.

To incorporate this complementary information, we consider target-alignment gradients from both the current adversarial input and its local neighborhood. The current-input gradient provides direct guidance toward the target representation, while gradients from noise-perturbed neighbors provide additional guidance under local input variations. **This motivates neighborhood-aware gradient fusion, which combines these gradients to reduce reliance on gradients evaluated at individual input configurations while retaining direct target-alignment guidance.**

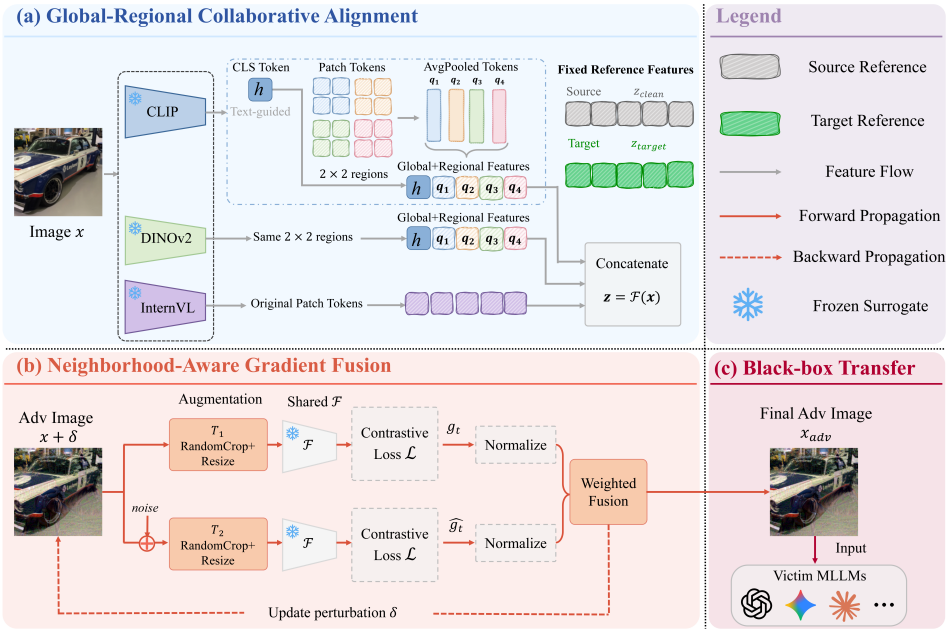

Figure 2: Overview of the TIGA Framework.

3.3 METHOD

We build TIGA upon MPC-Attack Li et al. (2026b), retaining its multi-paradigm ensemble and contrastive objective while introducing Global-Regional Collaborative Alignment and Neighborhood-Aware Gradient Fusion to improve transferability, and the overall framework is shown in Figure 2.

Global-Regional Collaborative Alignment (GRCA). Observation 1 suggests that global feature alignment alone may overlook regional visual information. We therefore introduce a lightweight regional representation to preserve such information during optimization. However, on the one hand, directly aligning individual patch tokens is unsuitable because the source and target images generally lack patch-level spatial correspondence. On the other hand, aggregating all patch tokens into a single representation discards their spatial organization. Inspired by the grid-based spatial pooling strategy in PSPNet Zhao et al. (2017), we partition the patch feature map into a coarse 2×2 grid and aggregate the tokens within each region independently. This design retains coarse spatial organization while avoiding explicit patch-level correspondence between the source and target images. We adopt a coarse 2×2 partition to account for the distinct resolutions of various ensembled vision encoders. For example, CLIP ViT-B/32 produces only a 7×7 patch feature map, where a finer partition would leave only a few patch tokens in each region.

Formally, let $h(x)$ denote the global representation, and $P = [p_1, \ldots, p_N] \in \mathbb{R}^{N\times D}$ denote the patch features extracted by a vision encoder, where $N = HW$ is the number of patches arranged in an $H \times W$ spatial grid, and D is the feature dimension. We partition the spatial feature map into four sub-regions and aggregate the patch features within each region by average pooling. Let Q_r denote the patch set of the r-th region, where $r \in \{1, 2, 3, 4\}$ indexes the four regions from left to right and top to bottom, and $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization; the corresponding regional feature is computed in Eq. 2:

$$q_r = \mathrm{Norm}\left(\frac{1}{|Q_r|}\sum_{p_j\in Q_r} p_j\right). \tag{2}$$

Following the spatial order defined above, we concatenate the four pooled features and further combine them with the global feature as Eq.3:

$$\psi(x) = \mathrm{Concat}(q_1, q_2, q_3, q_4), \qquad f(x) = \mathrm{Concat}(h(x), \psi(x)). \tag{3}$$

The resulting representation $f(x)$ preserves the global semantic guidance while retaining coarse spatial information from different image regions. Consequently, the alignment objective is no longer

restricted to image-level semantics, but also captures target-oriented information at the regional level without imposing explicit patch-to-patch correspondence. For CLIP and DINOv2, we augment their global representations with the proposed regional features, while the InternVL branch retains its original representation since it is a native VLM architecture. Let $f_{clip}(x)$, $f_{internvl}(x)$, and $f_{dino}(x)$ denote the representations obtained from the CLIP, InternVL, and DINOv2 branches, respectively. The three representations are concatenated to form the final representation, as defined in Eq.4:

$$\mathcal{F}(x) = \text{Concat}\left(f_{clip}(x),\ f_{internvl}(x),\ f_{dino}(x)\right). \tag{4}$$

To optimize the aggregated representation, we employ the contrastive matching objective. Specifically, we denote the source, target, and adversarial representations as $z_{clean} = \mathcal{F}(x_{clean})$, $z_{target} = \mathcal{F}(x_{target})$, and $z_{adv} = \mathcal{F}(x_{adv})$, respectively. The objective is formulated as Eq.5:

$$\mathcal{L} = -\log \frac{\exp\left(\omega \cdot \text{sim}(z_{adv}, z_{target})/\tau\right)}{\exp\left(\text{sim}(z_{adv}, z_{target})/\tau\right) + \exp\left(\text{sim}(z_{adv}, z_{clean})/\tau\right)}, \tag{5}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature coefficient, and ω is a balancing factor controlling the relative contributions of target attraction and source repulsion. The adversarial example is thereby optimized to approach the target representation while moving away from the source representation, with both global and regional information jointly involved in the optimization.

Neighborhood-Aware Gradient Fusion (NGF). Following Observation 2, we aim to reduce the dependence of each update on the gradient evaluated at a single adversarial point. Specifically, at each iteration, we construct an asymmetric gradient pair by evaluating one gradient at the current adversarial example and another at a perturbed counterpart. Both branches use the same RandomCropResize configuration. The unperturbed gradient preserves direct optimization toward the target, while the perturbed gradient provides complementary information about how the optimization behaves under a small input variation. Let $T(\cdot)$ denote the RandomCropResize transformation, which is independently sampled for the two branches from the same configuration. We construct the perturbed counterpart as $\hat{x}^t_{adv} = x^t_{adv} + \eta_t$, where $\eta_t \sim \mathcal{U}(-\xi \cdot \epsilon, \xi \cdot \epsilon)$ and ξ controls the noise strength. Accordingly, the gradients of the two branches are computed as in Eq.6:

$$g_t = \nabla_{\delta_t} \mathcal{L}\Big(\mathcal{F}(T(x^t_{adv}))\Big), \qquad \hat{g}_t = \nabla_{\delta_t} \mathcal{L}\Big(\mathcal{F}(T(\hat{x}^t_{adv}))\Big). \tag{6}$$

Since the two branch gradients may have different magnitudes, we first normalize them to place their contributions on a comparable scale and prevent gradient magnitude differences from dominating the subsequent fusion, as illustrated in Eq.7:

$$\bar{g}_t = \frac{g_t}{\text{Mean}(|g_t|) + \epsilon_0}, \qquad \bar{\hat{g}}_t = \frac{\hat{g}_t}{\text{Mean}(|\hat{g}_t|) + \epsilon_0}, \tag{7}$$

where $\epsilon_0 = 10^{-12}$ is a small positive constant introduced to avoid division by zero. We then perform a weighted fusion of the two normalized gradients according to Eq.8:

$$g_t^{\text{update}} = \lambda \bar{g}_t + (1-\lambda)\bar{\hat{g}}_t, \tag{8}$$

where $\lambda \in [0, 1]$ controls the relative contribution of the unperturbed and perturbed gradients. The overall detailed TIGA optimization procedure is summarized in the Appendix A.2.

4 Experiments

4.1 Experimental Setup

Datasets. Following previous work Li et al. (2025); Jia et al. (2025b); Li et al. (2026b), we utilize 1,000 clean source images from the NIPS 2017 Adversarial Attacks and Defences Competition dataset[1]. The 1,000 target images are randomly sampled from the MSCOCO validation set Lin et al. (2014). For defense experiments and ablation studies, we utilize 100 source-target pairs following Li et al. (2025), which keeps the API cost of black-box evaluation tractable. To mitigate evaluation variance arising from the relatively limited sample size and the inherent variability of MLLM-generated responses, **we repeat each experiment 3 times** and report the average results.

[1] https://nips.cc/Conferences/2017/CompetitionTrack

Table 1: Targeted attack success rates and average similarities on closed-source MLLMs.

Method	GPT-5.5		GPT-5.4		Gemini-3.5-Flash		Gemini-3.1-Flash-Lite		Claude-Sonnet-4.6		Claude-Opus-4.8		Average	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
COA	0.002	0.019	0.001	0.018	0.001	0.010	0.000	0.013	0.000	0.013	0.002	0.018	0.001	0.015
AnyAttack	0.009	0.032	0.005	0.027	0.004	0.019	0.004	0.028	0.008	0.026	0.006	0.029	0.006	0.027
AdvDiffVLM	0.116	0.171	0.112	0.172	0.055	0.129	0.068	0.127	0.277	0.297	0.324	0.313	0.159	0.202
M-Attack	0.285	0.292	0.219	0.255	0.309	0.321	0.312	0.333	0.425	0.384	0.535	0.452	0.347	0.339
FOA-Attack	0.350	0.332	0.296	0.296	0.340	0.350	0.342	0.353	0.474	0.417	0.581	0.484	0.397	0.372
M-AttackV2	0.454	0.401	0.383	0.357	0.525	0.455	0.506	0.450	0.630	0.511	0.723	0.568	0.537	0.457
MPC-Attack	0.453	0.408	0.390	0.365	0.673	0.547	0.675	0.542	0.621	0.514	0.724	0.581	0.589	0.493
Ours	**0.684**	**0.561**	**0.666**	**0.550**	**0.793**	**0.630**	**0.779**	**0.620**	**0.822**	**0.652**	**0.861**	**0.682**	**0.768**	**0.616**

Table 2: Targeted attack success rates and average similarities on open-source MLLMs.

Method	LLaVA-1.6-7B		MiniCPM-o-2.6		Qwen2.5-VL-7B		Qwen2.5-VL-72B		InternVL3-14B		InternVL3-78B		Average	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
COA	0.004	0.016	0.002	0.020	0.001	0.018	0.001	0.020	0.002	0.018	0.000	0.015	0.002	0.018
AnyAttack	0.023	0.036	0.014	0.036	0.011	0.039	0.005	0.042	0.011	0.032	0.008	0.028	0.012	0.036
AdvDiffVLM	0.309	0.296	0.293	0.292	0.309	0.313	0.314	0.327	0.315	0.318	0.194	0.239	0.289	0.298
M-Attack	0.680	0.562	0.632	0.520	0.569	0.472	0.571	0.476	0.661	0.523	0.440	0.390	0.592	0.491
FOA-Attack	0.725	0.598	0.692	0.551	0.575	0.488	0.617	0.506	0.703	0.561	0.490	0.422	0.634	0.521
M-AttackV2	0.812	0.666	0.814	0.634	0.705	0.564	0.693	0.563	0.809	0.631	0.648	0.535	0.747	0.599
MPC-Attack	0.839	0.686	0.836	0.649	0.697	0.559	0.714	0.577	0.853	0.659	0.772	0.606	0.785	0.623
Ours	**0.887**	**0.732**	**0.904**	**0.710**	**0.858**	**0.674**	**0.870**	**0.680**	**0.885**	**0.696**	**0.897**	**0.700**	**0.884**	**0.699**

Parameter Settings. Following Li et al. (2026b), we generate adversarial examples using three CLIP variants Radford et al. (2021), InternVL3-1B Zhu et al. (2025c), and DINOv2 Oquab et al. (2023) as surrogate models, and we set the temperature $\tau = 0.2$ and the balancing factor $\omega = 2.0$. The perturbation budget is $\epsilon = 16/255$ and constrained within an l_∞ ball; the number of iterations is $\mathcal{N} = 300$, the step size is $\alpha = 1/255$, and the momentum decay is $\mu = 0.9$ Dong et al. (2018). For the TIGA-specific hyperparameters, we set the noise strength multiplier to $\xi = 1.5$, the fusion weight to $\lambda = 0.5$, and the regional granularity to 2×2 across all experiments.

Models. We evaluate the transferability performance on twelve MLLMs, including six closed-source models, namely GPT-5.5, GPT-5.4, Gemini-3.1-Flash-Lite, Gemini-3.5-Flash, Claude-Sonnet-4.6, and Claude-Opus-4.8, and six open-source models, namely LLaVA-1.6-7B, MiniCPM-o-2.6, Qwen2.5-VL-7B, Qwen2.5-VL-72B, InternVL3-14B, and InternVL3-78B. For ablation studies, we measure the effectiveness on GPT-5.5, Gemini-3.1-Flash-Lite, and Claude-Opus-4.8. All the prompt text is set to *"Describe this image in one concise sentence, no longer than 20 words."*

Baselines. We compare our method with seven representative state-of-the-art targeted adversarial attack baselines, namely AdvDiffVLM Guo et al. (2024), COA Xie et al. (2025), AnyAttack Zhang et al. (2025), M-Attack Li et al. (2025), FOA-Attack Jia et al. (2025b), M-AttackV2 Zhao et al. (2026), MPC-Attack Li et al. (2026b). We utilize the same surrogate models, perturbation budget, step size, and optimization steps across all methods for a fair comparison. For method-specific hyperparameters, we use the defaults reported in their respective papers. All experiments are conducted on a server equipped with an NVIDIA H100 (80GB) GPU.

Defenses. To evaluate attack performance under defenses, we consider four methods: JPEG compression (JPEG) Guo et al. (2018), bit-depth reduction (BitRed) Guo et al. (2018), median filtering (Median) Xu et al. (2017), and random resizing and padding (R&P) Xie et al. (2018).

Evaluation Metrics. Following prior work Li et al. (2025); Jia et al. (2025b); Li et al. (2026b), we utilize an LLM-as-a-judge method to evaluate the targeted adversarial transferability. Specifically, for each adversarial-target pair, the victim MLLM generates captions for both images, and GPT-4o quantifies their semantic similarity using GPTScore. A targeted attack is considered successful if the GPTScore reaches the threshold of 0.5. We evaluate targeted transferability using both the attack success rate (ASR) and the average similarity score (AvgSim).

4.2 Comparison results

Attack results against closed-source models. Table 1 reports the targeted attack performance on six closed-source MLLMs. TIGA consistently outperforms all competing methods across all models in terms of both ASR and AvgSim. In particular, TIGA achieves an average ASR of 76.8% and an average AvgSim of 0.616, surpassing the strongest MPC-Attack by 17.9% and 0.123, respectively. The improvement is consistent across different model families. For example, compared with MPC-

Attack, TIGA improves ASR by 23.1% and 27.6% on GPT-5.5 and GPT-5.4, respectively, while achieving gains of 20.1% and 13.7% on Claude-Sonnet-4.6 and Claude-Opus-4.8. Similar improvements are observed on the Gemini models. These results demonstrate that TIGA transfers effectively across heterogeneous closed-source MLLMs without requiring access to their internal information.

Attack results against open-source models. Table 2 presents the targeted attack performance on six open-source MLLMs. TIGA again achieves the best overall performance, reaching an average ASR of 88.4% and an average AvgSim of 0.699, which exceed those of the strongest baseline, MPC-Attack, by 9.9% and 0.076, respectively. The advantage remains consistent across different model architectures and scales. In particular, TIGA improves the ASR over the best-performing MPC-Attack by 16.1% on Qwen2.5-VL-7B, 15.6% on Qwen2.5-VL-72B, and 12.5% on InternVL3-78B. Corresponding improvements are also observed in AvgSim, indicating that the generated adversarial examples not only trigger the target semantics more frequently but also produce responses that are more closely aligned with the desired target. These results further demonstrate the strong transferability of TIGA across diverse open-source MLLMs.

Table 3: Attack results with and without defenses on six closed-source MLLMs. Defense Avg. is computed over the four defenses only.

Method	No defense		JPEG		BitRed		Median		R&P		Defense Avg.	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
MPC-Attack	0.703	0.565	0.665	0.541	0.610	0.506	0.393	0.371	0.658	0.528	0.582	0.487
Ours	**0.852**	**0.663**	**0.838**	**0.659**	**0.810**	**0.643**	**0.715**	**0.568**	**0.835**	**0.655**	**0.799**	**0.631**

Attack results against defenses. Table 3 reports the performance with and without four defenses on six closed-source MLLMs. TIGA consistently outperforms MPC-Attack across all four defenses in terms of both ASR and AvgSim. Compared with the no-defense setting, the average ASR decreases from 70.3% to 58.2% for MPC-Attack, while TIGA exhibits a smaller decrease from 85.2% to 79.9%. Under the defenses, TIGA achieves an average ASR of 79.9% and an AvgSim of 0.631, exceeding MPC-Attack by 21.7% and 0.144, respectively. These results demonstrate that TIGA maintains stronger targeted transferability against the evaluated methods.

4.3 Ablation Study

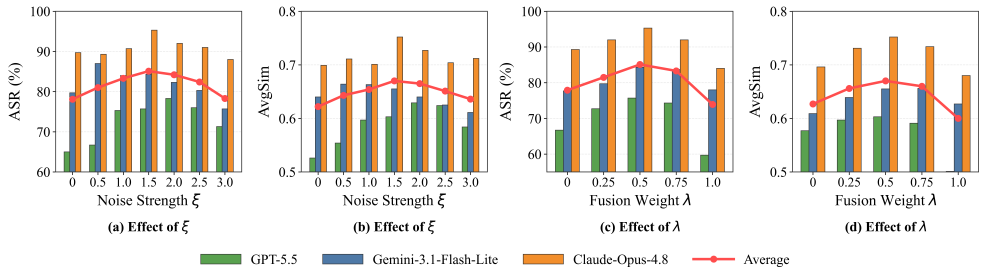

Figure 3: Ablation studies of noise strength ξ and fusion weight λ.

Ablation on noise strength ξ. We investigate the effect of the noise strength ξ, which controls the range of uniformly sampled noise. As shown in Figure 3(a)-(b), with GRCA enabled, both ASR and AvgSim first improve and then decline as ξ increases. Even at $\xi = 0$, the two-branch optimization improves the average ASR from 68.8% to 78.1% and AvgSim from 0.549 to 0.622 relative to the MPC-Attack baseline. Introducing moderate neighborhood noise further improves ASR and AvgSim to 85.1% and 0.670 at $\xi = 1.5$. Further increasing ξ degrades performance, with the average ASR and AvgSim dropping to 78.3% and 0.636 at $\xi = 3.0$. This suggests that moderate noise provides useful neighborhood information, whereas excessive noise introduces additional randomness and weakens the benefit of gradient fusion. We therefore set $\xi = 1.5$ as the default setting.

Ablation on fusion weight λ. We investigate the effect of the fusion weight λ, which balances the unperturbed and noise-perturbed branches. As shown in Fig. 3(c)-(d), both ASR and AvgSim first increase and then decrease as λ varies from 0 to 1. The best overall performance is achieved at $\lambda = 0.5$, with an average ASR of 85.1% and AvgSim of 0.670. In contrast, using only the perturbed branch ($\lambda = 0$) or only the unperturbed branch ($\lambda = 1$) yields lower performance, with average ASRs of 77.9% and 73.9%, respectively. These results indicate that the two branches provide complementary gradient information and are most effective under balanced fusion. We therefore set $\lambda = 0.5$ as the default setting.

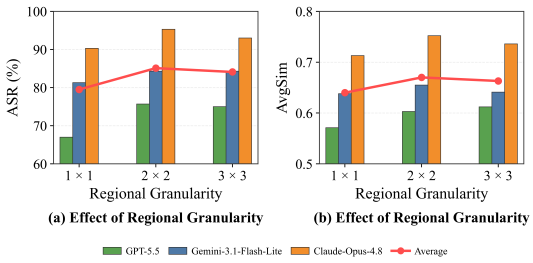

Figure 4: Ablation studies of regional granularity.

Ablation on Regional Granularity. As reported in Figure 4, we investigate the effect of regional granularity by comparing 1×1, 2×2, and 3×3 partitions while keeping all other settings unchanged. The 1×1 variant averages all patch tokens into a single feature, whereas the other variants pool tokens within each region independently, and all variants retain the global feature branch. Compared with 1×1 pooling, the 2×2 partition improves both ASR and AvgSim across all three models, increasing the average ASR from 79.5% to 85.1% and AvgSim from 0.640 to 0.670, which suggests that preserving coarse spatial organization provides more effective alignment guidance than pooling all patch tokens into a single representation. Increasing the granularity to 3×3 yields comparable but slightly lower average performance, with an ASR of 84.1% and AvgSim of 0.663, indicating that finer partitioning does not necessarily improve transferability. We therefore adopt 2×2 as the default partition, as it achieves the best average performance among the evaluated configurations.

Table 4: Ablation study of different components in TIGA. Time denotes the average adversarial example generation time in seconds per image (s/img).

GRCA	NGF	GPT-5.5		Gemini-3.1-Flash-Lite		Claude-Opus-4.8		Average		Time
		ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	
✗	✗	0.523	0.446	0.710	0.570	0.830	0.630	0.688	0.549	**33.58**
✓	✗	0.597	0.501	0.780	0.627	0.840	0.680	0.739	0.603	33.91
✗	✓	0.742	0.597	0.825	0.650	0.931	0.739	0.833	0.662	67.73
✓	✓	**0.757**	**0.603**	**0.843**	**0.655**	**0.953**	**0.752**	**0.851**	**0.670**	69.57

Ablation on individual components. We evaluate the contributions of Global-Regional Collaborative Alignment (GRCA) and Neighborhood-Aware Gradient Fusion (NGF). As shown in Table 4, both components independently improve targeted transferability. GRCA increases the average ASR from 68.8% to 73.9% and AvgSim from 0.549 to 0.603, while NGF achieves larger gains, reaching 83.3% and 0.662, respectively. Combining both components yields the highest average ASR of 85.1% and AvgSim of 0.670. Regarding computational overhead, GRCA introduces negligible additional cost, increasing the generation time from 33.58 s/img to 33.91 s/img. NGF increases the time to 67.73 s/img due to an additional gradient computation. The full TIGA requires 69.57 s/img, approximately twice the baseline cost, highlighting the trade-off between computational overhead and improved transferability.

Ablation on Matched Computational Budgets. To assess the impact of additional computation, we increase MPC-Attack's iterations from 300 to 600, matching TIGA's backward propagation (BP) count. As shown in Table 5, doubling the iterations of MPC-Attack increases its average ASR from 68.8% to 75.4%. Under the matched BP budget, TIGA achieves a higher ASR of 85.1% and AvgSim of 0.670, outperforming MPC-Attack by 9.7% and 0.069, respectively, demonstrating its transferability advantage under matched BP budgets.

Table 5: Attack performance under matched BP count.

Method	BP	Average	
		ASR	AvgSim
MPC-Attack	300	0.688	0.549
	600	0.754	0.601
Ours	600	**0.851**	**0.670**

5 Conclusion

We propose **TIGA**, a targeted transferable attack framework for MLLMs that improves cross-model semantic transfer from both representation and optimization perspectives. Specifically, **Global-Regional Collaborative Alignment** jointly exploits image-level and region-level target semantics to provide more comprehensive adversarial guidance, while **Neighborhood-Aware Gradient Fusion** incorporates complementary gradients from neighboring inputs to reduce reliance on gradients evaluated at individual input configurations. Extensive experiments on both closed-source and open-source MLLMs demonstrate that TIGA consistently outperforms existing SOTA targeted transfer attacks in terms of attack success rate and semantic similarity. These results further reveal the vulnerability of current MLLMs to transferable visual perturbations and underscore the importance of improving their robustness against cross-model adversarial attacks.

AI USE STATEMENT

We used generative AI tools for language polishing and manuscript preparation, including wording revisions and LaTeX formatting, as well as research discussions. Apart from the LLM-based evaluation procedures explicitly reported in this paper, all research ideas, experimental designs, implementations, and analyses were reviewed and validated by the authors. The reported experiments include LLM-as-a-judge evaluation, as described in Section 4. The authors take full responsibility for the final text, results, and scientific claims.

ETHICS STATEMENT

TIGA investigates targeted transferable adversarial attacks on multimodal large language models (MLLMs) to identify security vulnerabilities and support research on multimodal robustness. Such attacks have dual-use risks: adversarial images could be misused to manipulate model outputs or induce misleading descriptions. Our experiments are conducted in controlled benchmark settings using public image datasets and evaluated MLLMs. We optimize adversarial examples on accessible surrogate models and evaluate their transferability without accessing the internal parameters or gradients of victim models. Attack success measures targeted semantic alignment under the specified image-captioning protocol, rather than demonstrated harm in real-world deployments. Our goal is to advance the understanding of adversarial transferability and inform the development of more robust and reliable MLLMs.

REPRODUCIBILITY STATEMENT

We provide the source code of TIGA in the supplementary materials. Section 3.3 describes the attack formulation, feature alignment objective, and gradient fusion procedure, with detailed pseudocode provided in Appendix A.2. Section 4 specifies the datasets, surrogate and victim models, attack configurations, baselines, and evaluation metrics. The appendix further provides the evaluation prompts and implementation details in Appendix A.1 and Appendix A.6. Reproducing the reported results requires matching the model versions, attack parameters, and evaluation settings, particularly for closed-source MLLMs whose outputs may vary over time.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Rishika Bhagwatkar, Shravan Nayak, Reza Bayat, Alexis Roger, Daniel Z Kaplan, Pouya Bashivan, and Irina Rish. Towards adversarially robust vision-language models: Insights from design choices and prompt formatting techniques. *arXiv preprint arXiv:2407.11121*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models

with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 1625–1634. IEEE, 2018.

Sensen Gao, Xiaojun Jia, Xuhong Ren, Ivor Tsang, and Qing Guo. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *European Conference on Computer Vision*, pp. 442–460. Springer, 2024.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018.

Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 2024.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4412–4420, 2026.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can MLLMs reason in multimodality? EMMA: An enhanced MultiModal ReAsoning benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 22044–22090. PMLR, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

yuyang Hong, Jiaqi Gu, Yang Qi, Lubin Fan, Yue Wu, Ying Wang, Kun Ding, SHIMING XIANG, and Jieping Ye. Knowledge-based visual question answer with multimodal processing, retrieval and filtering. In *Advances in Neural Information Processing Systems*, 2025b.

Xiangdong Hu, Yangyang Jiang, Qin Hu, and Xiaojun Jia. Gambit: A gamified jailbreak framework for multimodal large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pretraining for image captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17959–17968. IEEE, 2022.

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. *Advances in Neural Information Processing Systems*, 2025b.

Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14242–14251. IEEE, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Jingru Li, Wei Ren, and Tianqing Zhu. Seeing no evil: Blinding large vision-language models to safety instructions via adversarial attention hijacking. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18304–18323, 2026a.

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 (CVPR)*, 2026b.

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. *Advances in Neural Information Processing Systems*, 2025.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

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*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 23701–23722, 2024.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024.

Weimin Lyu, Michael Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. Backdooring vision-language models with out-of-distribution data. In *International Conference on Learning Representations*, volume 2025, pp. 87511–87529, 2025.

Teng Ma, Xiaojun Jia, Ranjie Duan, Xinfeng Li, Yihao Huang, Xiaoshuang Jia, Zhixuan Chu, and Wenqi Ren. Heuristic-induced multimodal risk distribution jailbreak attack for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2686–2696. IEEE, 2025.

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 (CVPR)*, pp. 15679–15688, June 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*, volume 2026, pp. 18135–18161, 2026b.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model. *arXiv preprint arXiv:2402.02309*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Chunlin Qiu, Ang Li, Yiheng Duan, Shenyi Zhang, Yuanjie Zhang, Lingchen Zhao, and Qian Wang. Boosting adversarial transferability with low-cost optimization via maximin expected flatness. *IEEE Transactions on Information Forensics and Security*, 2026.

Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving adversarial transferability on vision transformers via forward propagation refinement. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25071–25080. IEEE, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3679–3687. IEEE, 2023.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1924–1933. IEEE, 2021.

Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 16138–16147. IEEE, 2021.

Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4584–4596. IEEE, 2023.

Xiaosen Wang, Shaokang Wang, Zhijin Ge, Yuyang Luo, and Shudong Zhang. Attention! your vision language model could be maliciously manipulated. *Advances in Neural Information Processing Systems*, 38:128627–128658, 2025.

Xiaosen Wang, Zhijin Ge, Bohan Liu, Zheng Fang, Fengfan Zhou, Ruixuan Zhang, Shaokang Wang, and Yuyang Luo. Devling into Adversarial Transferability on Image Classification: Review, Benchmark, and Evaluation. *arXiv preprint arXiv:2602.23117*, 2026.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Hao Yang, Zhuo Ma, Yang Liu, Yilong Yang, Guancheng Wang, and JianFeng Ma. A cross-modal prompt injection attack against large vision-language models with image-only perturbation. *arXiv preprint arXiv:2605.16090*, 2026a.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, et al. Mage-vl: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026b.

Xiao Yang, Longlong Xu, Tianyu Pang, Yinpeng Dong, Yikai Wang, Hang Su, and Jun Zhu. Face3dadv: Exploiting robust adversarial 3d patches on physical face recognition. *International Journal of Computer Vision*, 133(1):353–371, 2025a.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385. IEEE, 2025b.

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

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

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.

Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2020.

Hegui Zhu, Wenqi Cui, Yue Yan, and Ning Han. Reinforcing adversarial transferability via negative class guided example generation. *IEEE Transactions on Information Forensics and Security*, 21: 532–546, 2025a.

Hegui Zhu, Yanmeng Jia, Yue Yan, and Ze Yang. Improving transferability of adversarial examples via statistical attribution-based attacks. *Neural Networks*, 187:107341, 2025b.

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*, 2025c.