

Visual-Invariance-Augmented Feature Optimal Alignment for Transferable Adversarial Attacks against Closed-Source MLLMs

Xiaojun Jia, Simeng Qin, Yiming Li, Jie Liao, Sensen Gao,
Ke Ma, Yang Liu, *Senior Member, IEEE*, and Xiaochun Cao *Senior Member, IEEE*

Abstract—Multimodal large language models (MLLMs) remain vulnerable to transferable adversarial examples, especially in black-box scenarios where only open-source surrogate models are accessible. Existing targeted transfer attacks mainly align adversarial and target samples using global image-level features, such as the [CLS] embedding of vision-language encoders. However, such coarse alignment insufficiently exploits patch-level visual structures, leading to limited transferability across heterogeneous closed-source MLLMs. To address this issue, we propose a visual-invariance-augmented feature optimal alignment attack with adaptive unbalanced transport, called IAU-FOA, for improving targeted transferability against closed-source MLLMs. IAU-FOA aligns adversarial and target samples from both global and local perspectives: a cosine-based objective narrows their global semantic gap, while patch tokens are clustered into compact local patterns and matched through optimal transport for fine-grained feature alignment. Balanced optimal transport enforces fixed marginal masses even for local clusters without reliable counterparts, potentially introducing misleading alignment gradients. We therefore introduce confidence-adaptive unbalanced transport to relax these constraints for weakly matched clusters, aiming to reduce unreliable local alignment and improve adversarial transferability. Moreover, we study the effect of input transformations on adversarial transferability and propose a visual-invariance augmentation strategy that applies bidirectional pixel-intensity rescaling and per-channel white-balance adjustment to simulate exposure, contrast, illumination, and color-temperature variations, thereby encouraging adversarial perturbations to generalize across different visual encoders. Extensive experiments on open-source and closed-source MLLMs show that IAU-FOA consistently outperforms state-of-the-art transferable attack methods. The code is released at <https://github.com/jiaxiaojunQAQ/IAU-FOA>.

Index Terms—Multimodal large language models, transferable adversarial attacks, feature optimal alignment, adaptive unbalanced optimal transport, visual-invariance augmentation.

I. INTRODUCTION

Xiaojun Jia, Yang Liu, Yiming Li, and Sensen Gao are with Nanyang Technological University, Singapore (e-mail: jiaxiaojunqaq@gmail.com; yangliu@ntu.edu.sg; liyiming.tech@gmail.com; sensen.gao2002@gmail.com).

Simeng Qin is with Northeastern University, Shenyang, Liaoning, China (e-mail: qinsimeng@neuq.edu.cn).

Jie Liao is with Chongqing University, Chongqing, China (e-mail: liaojie@cqu.edu.cn).

Ke Ma is with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: make@ucas.ac.cn).

Xiaochun Cao is with the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China (e-mail: caoxiaochun@mail.sysu.edu.cn).

Multimodal large language models (MLLMs) [6, 55] have become a representative framework for connecting visual perception with language reasoning [15, 21]. In most current systems, an input image is first encoded into visual tokens or embeddings by a vision encoder, and these visual representations are then projected into the language model for multimodal understanding and generation [22, 29, 33, 46, 62]. In particular, many MLLMs adopt pre-trained vision encoders, such as Contrastive Language-Image Pre-training (CLIP) [40], to extract image features and align them with language embeddings. This modular design has enabled strong performance on a wide range of vision-language applications. Beyond advances in open-source MLLMs, commercial closed-source MLLMs such as GPT-5.5, Claude-Opus-4.8, and Gemini-3.5-Flash have also been widely deployed in practical multimodal services. However, since MLLMs heavily rely on vision encoders to perceive visual inputs, they may also inherit the adversarial vulnerabilities of these visual front ends [34, 45, 56]. Recent studies [4, 7, 30, 37] have shown that MLLMs can be vulnerable to adversarial examples [8], which are typically generated by adding small and visually imperceptible perturbations to clean images, such that the perturbed images remain natural to humans but change the model's responses. Such vulnerability raises security and safety concerns for real-world foundation-model services, where manipulated visual inputs may mislead downstream multimodal reasoning. According to the attack objective, adversarial attacks on MLLMs can be broadly divided into untargeted and targeted attacks. Untargeted attacks aim to induce incorrect responses, whereas targeted attacks attempt to drive the model toward attacker-specified semantics or outputs [19, 57].

As illustrated in Fig. 1, targeted adversarial examples can preserve the visual appearance of the source image while steering commercial MLLMs toward target-related descriptions. In closed-source scenarios, the gradients and parameters of target MLLMs are inaccessible, making direct optimization infeasible. A practical alternative is to craft adversarial examples on open-source surrogate models and then transfer them to the target systems. This property, known as adversarial transferability [9], measures whether adversarial examples generated for one model can remain effective on other unseen models. Transferability is important for both untargeted and targeted attacks, but it is particularly difficult to achieve in the targeted black-box setting, where the perturbation must consistently convey the target semantics across heterogeneous MLLMs

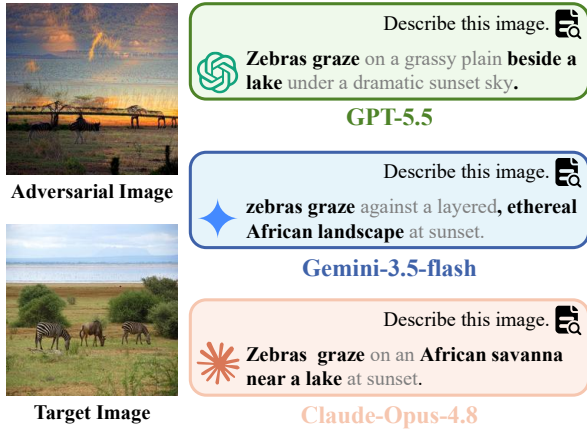


Fig. 1. Examples of targeted transfer attacks against commercial MLLMs. The adversarial images remain visually close to the source images, but induce target-related responses to the prompt “Describe this image”.

rather than merely causing prediction errors [2, 63]. Therefore, an effective targeted attack should not simply exploit surrogate-specific weaknesses, but should learn perturbations that preserve transferable visual semantics across different visual encoders and preprocessing pipelines.

Previous targeted transfer attack methods [23, 24, 45] typically generate adversarial examples by minimizing a contrastive loss between the global features of the adversarial image and the target image, where the global representation is often given by the [CLS] token of an open-source image encoder such as CLIP. By pulling the adversarial image toward the target image in the surrogate feature space, these methods can improve targeted transferability to some extent. However, their adversarial transferability and attack performance against closed-source MLLMs remain limited, mainly due to two limitations. First, most existing methods rely primarily on global image-level features. Although global feature alignment can guide adversarial examples toward the target semantics, it compresses the visual content into a single representation and overlooks rich patch-level structures. As a result, the generated perturbations may capture coarse semantic similarity while failing to sufficiently exploit fine-grained visual patterns that are important for transfer across heterogeneous MLLMs. Even when local features are aligned, exact marginal constraints can force matches between clusters without reliable counterparts. These matches may introduce gradients that favor incidental similarities in a surrogate encoder rather than correspondences shared across models, motivating confidence-adaptive local transport. Second, the choice of input transformations for targeted transfer attacks against MLLMs warrants further investigation. Existing approaches include spatial transformations such as resizing and padding in DI [49], as well as pixel-intensity scaling in SIM [26]. Building on intensity-based augmentation, we investigate the combination of bidirectional intensity scaling and channel-wise gain adjustment to diversify exposure and illumination-color conditions during targeted feature alignment.

To address these issues, we propose **IAU-FOA**, a visual-invariance-augmented feature optimal alignment attack with an adaptive unbalanced transport method. Given a source image

and a target image, IAU-FOA generates an adversarial perturbation by aligning their representations in surrogate vision-language encoders from both global and local perspectives. At the global level, IAU-FOA adopts a coarse-grained feature alignment objective based on cosine similarity, which pulls the image-level representation of the adversarial example toward that of the target sample. Since the global representation is usually encoded by the [CLS] token, this objective provides a stable semantic direction for targeted attacks and encourages the adversarial image to approach the target semantics in the surrogate feature space.

At the local level, IAU-FOA further exploits the patch-token representations produced by Transformer-based vision encoders. While the [CLS] token summarizes global information, the remaining visual tokens contain local patch-level details. Directly using all local tokens, however, can introduce redundancy and noisy correspondences. Therefore, IAU-FOA first clusters patch tokens into compact and discriminative local visual patterns, where each cluster center summarizes the characteristics of a group of related patches. It then formulates the alignment between adversarial and target local patterns as an optimal transport problem, enabling fine-grained feature matching beyond a single global embedding.

To make local alignment more reliable, IAU-FOA further introduces an adaptive unbalanced transport mechanism. Balanced OT requires each cluster to transport its prescribed mass, even when reliable counterparts are absent. The proposed scheme instead adapts marginal regularization to cluster-level matching confidence. High-confidence clusters receive stronger penalties on marginal deviations, while low-confidence clusters receive weaker penalties. This flexibility aims to reduce the interference of unreliable local matches in adversarial optimization and thereby improve adversarial transferability. In addition, when multiple visual encoders are used as surrogate models, IAU-FOA computes an alignment similarity score for each encoder and dynamically adjusts its weight using the ratio between the current and previously recorded scores. For positive scores, larger relative increases receive larger weights, allowing the ensemble to respond to changes in alignment across surrogate encoders.

Finally, IAU-FOA incorporates a visual-invariance augmentation (VIA) strategy during adversarial optimization. Building on existing intensity-scaling augmentation, our strategy combines bidirectional pixel-intensity rescaling with channel-wise gain adjustment to approximate exposure and white-balance variations during targeted alignment. By optimizing adversarial perturbations under these transformed visual conditions, IAU-FOA encourages the perturbations to rely less on surrogate-specific visual statistics and to remain effective across heterogeneous closed-source MLLMs. Our main contributions are summarized as follows:

- We propose **IAU-FOA**, a targeted black-box attack framework for MLLMs that drives adversarial images toward target semantics through a dual-granularity feature matching objective. Specifically, a cosine-based loss provides coarse semantic guidance from global embeddings, while clustered patch tokens are matched through an

optimal-transport formulation to capture discriminative local structures.

- We propose an adaptive unbalanced transport mechanism to improve the reliability of fine-grained matching. It relaxes marginal constraints for weakly matched clusters, aiming to reduce misleading alignment gradients arising from unreliable local correspondences and improve targeted transferability.
- We propose a dynamic weighting strategy for ensemble surrogate optimization. The strategy adjusts each encoder’s contribution according to the relative change in its alignment similarity score, providing a score-based alternative to fixed ensemble weights for targeted transfer.
- We propose a visual-invariance augmentation strategy for transferable adversarial optimization. Through bidirectional intensity scaling and channel-wise white-balance adjustment, the attack is optimized under diverse photometric conditions and becomes less dependent on surrogate-specific visual statistics.
- Extensive experiments on open-source and closed-source MLLMs demonstrate that IAU-FOA consistently outperforms state-of-the-art transferable attack methods.

This paper is an extension of our preliminary conference version, FOA [17], which forms the foundation of the proposed framework. Compared with the original balanced feature optimal alignment framework, we extend this work in four aspects: 1) We introduce visual-invariance-augmented optimization in **Section III-C** to improve robustness against photometric changes and encoder-dependent visual statistics. 2) We develop an adaptive unbalanced transport method for local feature alignment in **Section III-F**, which replaces exact marginal preservation with confidence-dependent penalties on marginal deviations. We add a theoretical analysis of the proposed method in **Section III-H**. 3) We conduct more experiments and analyses, including comparisons with state-of-the-art methods, ablation studies, and evaluation under adversarial defenses. We include additional state-of-the-art adversarial attack methods for comparison in **Section IV-C**. We add ablation studies of the proposed components in **Section IV-E** and evaluate the proposed method under adversarial defenses in **Section IV-D**. 4) We have thoroughly revised the abstract, introduction, method, experiments, and conclusion sections to offer a more detailed overview of our motivation and approach. We have updated all figures and tables to enhance clarity and presentation.

II. RELATED WORK

A. Multimodal Large Language Models

Large language models (LLMs) [60] have achieved impressive success across a wide range of Natural Language Processing (NLP) tasks, demonstrating strong capabilities in language understanding [20], instruction following [13], and complex reasoning [11, 35]. Building upon these advances, multimodal large language models (MLLMs) extend LLMs to vision-language scenarios by integrating visual perception modules with LLM backbones. Specifically, MLLMs align visual representations with the semantic space of LLMs,

enabling the language model to interpret visual inputs and generate grounded textual responses [1, 28, 42]. Representative systems usually connect pre-trained vision encoders and LLMs through lightweight alignment modules, such as the Q-Former in BLIP-2 [22] and the MLP projector in LLaVA [29]. Some open-source models further enhance fine-grained visual perception through versatile visual instruction tuning and high-resolution image processing, as demonstrated by Qwen2-VL [46] and InternVL [3]. Meanwhile, commercial MLLMs, such as GPT-5.5 and Claude-Opus-4.8, have exhibited powerful multimodal understanding, reasoning, and instruction-following capabilities, representing the leading performance of current vision-language systems. However, their closed-source nature, together with substantial differences in model architectures, vision encoders, alignment modules, and training recipes, makes their internal representation spaces difficult to access or approximate, posing challenges for black-box attacks against modern MLLMs.

B. Adversarial Attacks

Adversarial attacks aim to mislead deep neural networks by adding carefully crafted perturbations to clean inputs [8, 43]. According to the attacker’s access to the victim model, existing adversarial attack methods are commonly categorized into white-box attacks, where model parameters or gradients are available, and black-box attacks, where the attacker has no access to internal model parameters, gradients, or feature representations [36, 58]. Depending on the attack objective, adversarial attacks can also be divided into untargeted attacks and targeted attacks. Untargeted attacks only require the model to produce incorrect outputs, whereas targeted attacks aim to induce specific target semantics or responses. Compared with conventional adversarial attacks on image classification models, attacking MLLMs is substantially more challenging [16, 19]. Classification models usually produce predictions over a fixed label space, allowing attacks to optimize well-defined objectives. In contrast, MLLMs generate open-ended textual responses conditioned on both visual and textual inputs, where the attack goal is often expressed at the semantic or instruction-following level rather than as a discrete label. For closed-source MLLMs, the challenge becomes more pronounced because their parameters, gradients, and intermediate feature representations are unavailable to attackers. A potential solution is to exploit adversarial transferability [31, 39], where adversarial examples generated on accessible surrogate models remain effective on unseen target models. This property enables transfer-based attacks to probe closed-source MLLMs without directly accessing their internal states, making it a practical and promising route for targeted black-box attacks [26, 49].

C. Transfer-Based Targeted Attacks for MLLMs

Recent transfer-based attacks on vision-language models typically optimize adversarial perturbations on accessible surrogate models, such as CLIP, and transfer them to closed-source MLLMs. Specifically, Guo *et al.* [12] propose AdvDiffVLM, which leverages diffusion models to generate natural

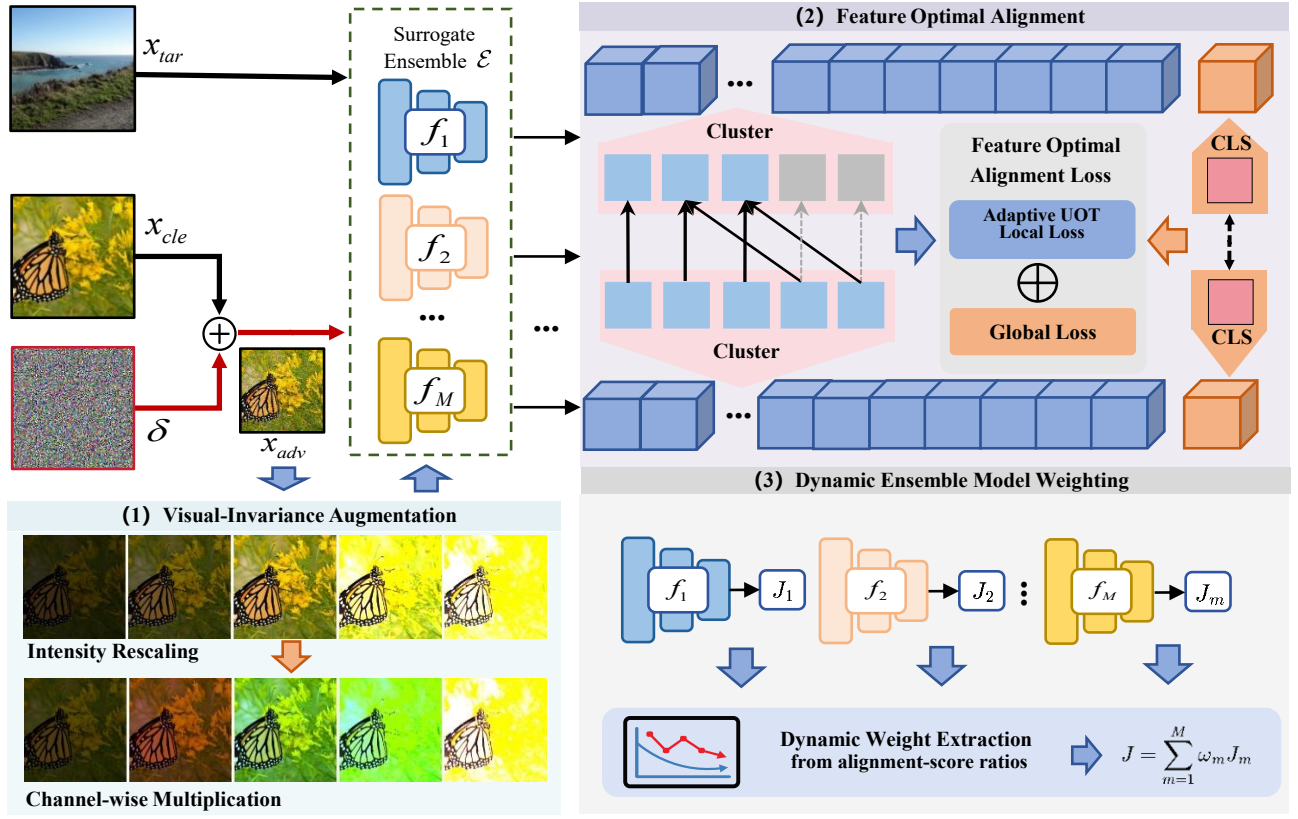


Fig. 2. Overview of IAU-FOA: visual-invariance augmentation, global feature alignment, adaptive unbalanced local transport, and dynamic surrogate weighting.

and transferable targeted adversarial examples for vision-language models. Zhang *et al.* [59] propose AnyAttack, a self-supervised framework for large-scale generation of label-free targeted adversarial examples, improving the scalability of targeted transfer attacks. Xie *et al.* [50] propose Chain of Attack (COA), a transfer-based black-box attack method that iteratively generates adversarial examples through multimodal semantic updates between visual and textual modalities. To better exploit visual details for closed-source MLLM attacks, Li *et al.* [24] propose M-Attack, which constructs a simple yet effective black-box attack baseline by enhancing target-related visual patterns. Zhao *et al.* [61] further propose M-Attack-v2, which emphasizes fine-grained detail targeting to improve attack effectiveness against strong black-box MLLMs. Li *et al.* [23] propose MPCAttack, which collaboratively aggregates multi-paradigm vision-language features to reduce surrogate representation bias and strengthen transfer-based attacks against MLLMs. Despite these advances, existing targeted transfer attacks still show limited effectiveness against advanced closed-source MLLMs, partly because they often rely on surrogate-specific global representations or do not sufficiently account for the reliability of local correspondences. In addition, fine-grained adaptive alignment and visual-invariance optimization remain insufficiently explored.

III. METHOD

Adversarial perturbations learned from a single surrogate may capture non-transferable representation biases, which can be alleviated by optimizing over multiple surrogate image

encoders to emphasize shared feature directions across models [41, 45, 52]. Following this perspective, IAU-FOA adopts a dynamic ensemble framework for targeted transfer attacks against closed-source MLLMs. The proposed framework integrates dual-granularity feature optimal alignment, adaptive unbalanced local transport, dynamic encoder weighting, and visual-invariance augmentation (VIA). By jointly aligning target semantics across multiple surrogate visual encoders and adapting their weights to relative changes in alignment similarity, IAU-FOA encourages adversarial perturbations to capture transferable visual cues shared by foundation models.

A. Problem Formulation

We study targeted transfer-based attacks against closed-source MLLMs using an ensemble of accessible surrogate visual encoders. Let

$$\mathcal{F} = \{f_m\}_{m=1}^M \quad (1)$$

denote M surrogate visual encoders, including vision-language and vision-only models. For an input image $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ denotes the valid image domain, the global and patch-level representations supplied by the m -th feature extractor are denoted as

$$f_m(\mathbf{x}) = [\mathbf{z}_m^g(\mathbf{x}); \mathbf{Z}_m^p(\mathbf{x})] \in \mathbb{R}^{(N_m+1) \times d_m}, \quad (2)$$

where $\mathbf{z}_m^g(\mathbf{x}) \in \mathbb{R}^{d_m}$ is the global representation supplied by the feature extractor (e.g., a [CLS] embedding), and $\mathbf{Z}_m^p(\mathbf{x}) = [\mathbf{z}_{m,1}^p(\mathbf{x}), \dots, \mathbf{z}_{m,N_m}^p(\mathbf{x})]^\top \in \mathbb{R}^{N_m \times d_m}$ denotes the

patch-token features. Given a source image $\mathbf{x}_s$ and a target image $\mathbf{x}_{tar}$, the attacker generates an adversarial example

$$\mathbf{x}_{adv} = \mathbf{x}_s + \delta, \quad \|\delta\|_\infty \leq \epsilon, \quad (3)$$

where δ is the bounded adversarial perturbation. The goal is to make the adversarial image near the target image in the feature spaces of all surrogate encoders:

$$\min_{\|\delta\|_\infty \leq \epsilon} \sum_{m=1}^M \mathcal{D}_m(f_m(\mathbf{x}_{adv}), f_m(\mathbf{x}_{tar})), \quad (4)$$

where $\mathcal{D}_m(\cdot, \cdot)$ denotes the feature-alignment objective on the m -th encoder. Since the victim MLLM is closed-source, its parameters and gradients are unavailable, and the perturbation optimized on $\mathcal{F}$ is directly transferred to unseen MLLMs. IAU-FOA optimizes the adversarial perturbation by reducing the feature-alignment loss between the adversarial image and the target image from both global and local perspectives.

B. Overview of the Proposed Framework

The proposed IAU-FOA consists of four components: visual-invariance augmentation, dual-granularity feature alignment, adaptive unbalanced optimal transport, and dynamic ensemble weighting. The current target view is passed through each surrogate encoder to obtain global and clustered local target features. We use $\mathbf{x}_{tar}$ below to denote the target view supplied to the alignment objective, and suppress spatial preprocessing in the notation. During optimization, multiple photometrically transformed copies of the adversarial image are generated. For each copy, IAU-FOA computes a weighted ensemble objective over surrogate encoders and averages the gradients from all valid copies. The loss minimized at each iteration can be written as

$$\min_{\|\delta\|_\infty \leq \epsilon} \frac{1}{|\mathcal{B}|} \sum_{\mathcal{T} \in \mathcal{B}} \sum_{m=1}^M \omega_m^{(k(\mathcal{T}))} \left(\mathcal{L}_g^m(\mathcal{T}(\mathbf{x}_{adv}), \mathbf{x}_{tar}) + \lambda_l \mathcal{L}_l^m(\mathcal{T}(\mathbf{x}_{adv}), \mathbf{x}_{tar}) \right), \quad (5)$$

where $\mathcal{B}$ is the set of visual-invariance transformations sampled at the current iteration, $k(\mathcal{T})$ indexes the objective evaluation for the transformed copy $\mathcal{T}(\mathbf{x}_{adv})$, $\omega_m^{(k(\mathcal{T}))}$ is the corresponding dynamic weight of the m -th surrogate encoder, and λ_l controls the contribution of local alignment. The weights are updated at each objective evaluation and held fixed during backpropagation. Eq. (5) summarizes the optimization used to update the perturbation. The outer average makes the perturbation effective under different photometric views $\mathcal{T}(\mathbf{x}_{adv})$, which reduces overfitting to the visual statistics of a single input rendering. For each transformed copy, the inner summation aggregates the attack loss over all surrogate encoders. Within each surrogate, $\mathcal{L}_g^m$ measures the global feature alignment loss to the target image, while $\mathcal{L}_l^m$ measures the transported local alignment loss between clustered patch-level structures. Thus, minimizing Eq. (5) drives $\mathbf{x}_{adv}$ toward the target image simultaneously at the semantic level and at the local visual-pattern level. The weight $\omega_m^{(k(\mathcal{T}))}$ is computed from relative changes in alignment similarity scores, as specified in Sec. III-G.

C. Visual-Invariance Augmentation

Real imaging conditions and closed-source MLLM visual preprocessing pipelines often introduce photometric variations, such as changes in exposure, brightness, illumination color, and white balance. Accurately modeling the full physical image formation process is difficult for transferable optimization. Instead, we use two simple, differentiable, and low-cost operations, namely global intensity scaling and channel-wise gain adjustment, to cover these common visual variations. $\mathcal{T}(\mathbf{x}_{adv})$ denotes a transformed view of the adversarial image. To keep the notation compact, we use $\mathcal{T}$ as a generic augmentation operator in the overall objective, and instantiate it here with photometric transformations. Specifically, each transformation combines an intensity scale s and a channel-wise gain vector $\beta = [\beta_R, \beta_G, \beta_B]^\top$. We enumerate the intensity scale s over

$$\mathcal{S} = \{2^k \mid k = -h, \dots, h\}, \quad (6)$$

where h is a nonnegative integer controlling the range of intensity scaling, and k indexes the $2h + 1$ scales. Scaling darkens ($s < 1$), brightens ($s > 1$), or preserves ($s = 1$) image intensity. By default, $h = 2$, giving five bidirectional intensity scales $\mathcal{S} = \{0.25, 0.5, 1, 2, 4\}$. Given $s \in \mathcal{S}$ and $\beta \in \mathbb{R}^3$, the concrete operator used in Eq. (5) is

$$\mathcal{T}(\mathbf{x}; s, \beta) = \beta \odot (s\mathbf{x}), \quad (7)$$

where $\odot$ denotes channel-wise multiplication. For an RGB image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, this operation is defined as

$$[\mathcal{T}(\mathbf{x}; s, \beta)]_{u,v,c} = \beta_c \cdot s \cdot \mathbf{x}_{u,v,c}, \quad c \in \{R, G, B\}. \quad (8)$$

Here, H and W denote the image height and width, u and v index the spatial pixel location, and c indexes the RGB channel. The global scale s is shared across all pixels and channels, while β_c is the gain applied to all pixels in channel c . For each transformed copy, the channel gains are independently sampled as

$$\beta_c \sim \mathcal{U}(1 - s_\beta, 1 + s_\beta), \quad c \in \{R, G, B\}. \quad (9)$$

Here, $\mathcal{U}(a, b)$ denotes the uniform distribution over $[a, b]$, and s_β controls the amplitude of channel-wise gain variations around one. We set $s_\beta = 0.5$ by default, giving a gain range of $[0.5, 1.5]$; $s_\beta = 0$ disables channel-wise gain adjustment.

The global scale s approximates changes in brightness and exposure, while the channel gains β approximate white-balance and illumination-color shifts. The transformation set $\mathcal{B}$ in Eq. (5) is constructed by enumerating $s \in \mathcal{S}$ and sampling a channel-gain vector β for each scale. These transformations simulate exposure, illumination, and color-temperature changes. In each iteration, gradients from all transformed copies are averaged before updating the perturbation, which encourages the adversarial example to remain effective under heterogeneous visual preprocessing pipelines. Each transformed copy is then passed to the surrogate encoders to generate adversarial examples.

D. Global Feature Alignment

The global branch uses the encoder-specific global feature to provide coarse-grained semantic guidance. For the m -th encoder, the global feature is first normalized as

$$\hat{\mathbf{z}}_m^g(\mathbf{x}) = \frac{\mathbf{z}_m^g(\mathbf{x})}{\|\mathbf{z}_m^g(\mathbf{x})\|_2}. \quad (10)$$

Let $\mathbf{X}_m = \hat{\mathbf{z}}_m^g(\mathbf{x}_{adv})$ and $\mathbf{Y}_m = \hat{\mathbf{z}}_m^g(\mathbf{x}_{tar})$ denote the normalized global features of the adversarial and target images, respectively. The global alignment loss is defined by the cosine distance between these two features:

$$\begin{aligned} \mathcal{L}_g^m(\mathbf{x}_{adv}, \mathbf{x}_{tar}) &= 1 - \cos(\mathbf{X}_m, \mathbf{Y}_m) \\ &= 1 - \frac{\langle \mathbf{X}_m, \mathbf{Y}_m \rangle}{\|\mathbf{X}_m\|_2 \|\mathbf{Y}_m\|_2}, \end{aligned} \quad (11)$$

where $\langle \mathbf{X}_m, \mathbf{Y}_m \rangle$ denotes the inner product between the two feature vectors, and $\|\cdot\|_2$ denotes the ℓ_2 norm, i.e., the Euclidean length of a vector. Minimizing Eq. (11) increases the semantic similarity between the adversarial image and the target image in the global feature representation space.

E. Local Clustering Feature Alignment

Patch tokens retain spatially distributed visual details beyond a single global representation. However, all patch tokens also include substantial redundant image information, and directly matching them may introduce considerable noisy correspondences, degrading the quality of the generated adversarial examples. To reduce redundancy while preserving discriminative local information, we cluster patch-token features into compact local visual patterns before feature alignment.

For the m -th encoder, the fine-grained local features of the adversarial and target images are extracted from the patch-token feature matrices:

$$\begin{aligned} \mathbf{X}_m^{loc} &= \mathbf{Z}_m^p(\mathbf{x}_{adv}) \in \mathbb{R}^{N_m \times d_m}, \\ \mathbf{Y}_m^{loc} &= \mathbf{Z}_m^p(\mathbf{x}_{tar}) \in \mathbb{R}^{N_m \times d_m}, \end{aligned} \quad (12)$$

where each row corresponds to the feature of one image patch. We then apply clustering to the local features of the adversarial and target images, respectively. Here, $\text{KMeans}(\mathbf{A}, K)$ denotes the standard K-means clustering operator applied to the row vectors of $\mathbf{A}$, which partitions them into K clusters by minimizing the within-cluster squared Euclidean distance and returns the cluster centroids:

$$\begin{aligned} \mathbf{X}_m^{clu} &= \text{KMeans}(\mathbf{X}_m^{loc}, K), \\ \mathbf{Y}_m^{clu} &= \text{KMeans}(\mathbf{Y}_m^{loc}, K), \end{aligned} \quad (13)$$

where $\mathbf{X}_m^{clu} = [\mathbf{x}_{m,1}^{clu}, \dots, \mathbf{x}_{m,K}^{clu}]^\top \in \mathbb{R}^{K \times d_m}$ and $\mathbf{Y}_m^{clu} = [\mathbf{y}_{m,1}^{clu}, \dots, \mathbf{y}_{m,K}^{clu}]^\top \in \mathbb{R}^{K \times d_m}$ are the adversarial and target cluster centers. We assign uniform reference mass to each cluster center on both sides. The mass vectors $\mathbf{a}_m^{adv}$ and $\mathbf{a}_m^t$ are defined independently of cluster cardinalities as

$$a_{m,k}^{adv} = \frac{1}{K}, \quad a_{m,k}^t = \frac{1}{K}, \quad k = 1, \dots, K. \quad (14)$$

Both mass vectors are normalized to sum to one. Following the theory of optimal transport [44], we model fine-grained local alignment as a transportation problem between two sets of clustered local patterns. For the m -th

encoder, the adversarial local distribution is represented by $\{(\mathbf{x}_{m,i}^{clu}, a_{m,i}^{adv})\}_{i=1}^K$, while the target local distribution is represented by $\{(\mathbf{y}_{m,j}^{clu}, a_{m,j}^t)\}_{j=1}^K$. The goal is to find a transport plan that moves the mass of adversarial local patterns to target local patterns with minimum matching cost. Let $\mathbf{P}_m(i, j)$ denote the mass transported from the i -th adversarial cluster to the j -th target cluster, and let $c_{ij} = c(\mathbf{x}_{m,i}^{clu}, \mathbf{y}_{m,j}^{clu})$ be their matching cost. The balanced OT form of local alignment is

$$\begin{aligned} \min_{\mathbf{P}_m} \quad & \sum_{i=1}^K \sum_{j=1}^K \mathbf{P}_m(i, j) c_{ij}, \\ \text{s.t.} \quad & \sum_{j=1}^K \mathbf{P}_m(i, j) = a_{m,i}^{adv}, \quad \forall i, \\ & \sum_{i=1}^K \mathbf{P}_m(i, j) = a_{m,j}^t, \quad \forall j, \\ & \mathbf{P}_m(i, j) \geq 0, \quad \forall i, j. \end{aligned} \quad (15)$$

The marginal constraints in Eq. (15) preserve the uniform reference masses on both sides. However, strict balanced constraints require prescribed masses even for clusters without reliable counterparts, potentially introducing misleading local alignment gradients. This motivates the adaptive unbalanced formulation introduced below. Specifically, the ground cost is defined by the cosine distance between local cluster centers. We therefore first construct the pairwise similarity matrix

$$\mathbf{S}_m(i, j) = \cos(\mathbf{x}_{m,i}^{clu}, \mathbf{y}_{m,j}^{clu}). \quad (16)$$

Here, $\mathbf{S}_m(i, j)$ measures the cosine similarity between the i -th adversarial cluster and the j -th target cluster, and the corresponding transport cost is set to $c_{ij} = 1 - \mathbf{S}_m(i, j)$. Minimizing this objective favors low-cost correspondences subject to the prescribed marginals.

F. Adaptive Unbalanced Optimal Transport

The balanced formulation in Eq. (15) provides a basic local alignment objective in which every cluster mass is exactly transported according to fixed marginals. With entropic regularization, the balanced OT problem can be solved using Sinkhorn iterations. However, exact marginal preservation can be too restrictive for targeted adversarial optimization: background clusters, ambiguous regions, or surrogate-specific patch patterns may be forced to match target clusters even when no reliable correspondence exists. Such forced matches may steer perturbations toward incidental similarities specific to the surrogate, limiting their transfer to other encoders. We therefore use an adaptive unbalanced Sinkhorn solver with weaker marginal regularization for low-confidence clusters and stronger regularization for high-confidence clusters. This design aims to reduce unreliable alignment by allowing transported masses to adapt to the available correspondences.

Starting from the cosine-distance cost $c_{ij} = 1 - \mathbf{S}_m(i, j)$ defined above, the entropic transport kernel is

$$\mathbf{K}_m(i, j) = \exp\left(-\frac{1 - \mathbf{S}_m(i, j)}{\varepsilon_{ot}}\right), \quad (17)$$

where $\varepsilon_{ot} > 0$ is the entropy regularization coefficient, also serving as the temperature of the Sinkhorn kernel, and is set to 0.1 in all experiments. For each adversarial-side cluster on the adaptively relaxed side, the confidence is estimated by its best target counterpart:

$$r_i = \max \left\{ 0, \max_j \mathbf{S}_m(i, j) \right\}. \quad (18)$$

We normalize the raw confidence with a stabilized mean:

$$\hat{r}_i = \frac{r_i}{\frac{1}{K} \sum_{k=1}^K r_k + \eta}. \quad (19)$$

Here, $\eta = 10^{-8}$. When the mean raw confidence is well above η , the normalized confidences have a mean close to one, modulating marginal regularization around its base level. If all raw confidences are zero, every $\hat{r}_i$ is zero; with $\gamma = 1$, this gives $\rho_i^a = \phi_i^a = 0$ while the target-side penalty remains unchanged. We use $\rho > 0$ as a hyperparameter controlling the KL penalty on marginal deviations, with $\rho = 0.2$ by default. Accordingly, the base unbalanced relaxation coefficient is

$$\phi = \frac{\rho}{\rho + \varepsilon_{ot}}, \quad (20)$$

where a larger ρ gives a larger ϕ and enforces marginal preservation more strongly. The default settings give $\phi = 2/3$. With mean-normalized adaptive confidence, the adversarial-side relaxation coefficient becomes

$$\begin{aligned} \rho_i^a &= \rho [(1 - \gamma) + \gamma \hat{r}_i], \\ \phi_i^a &= \frac{\rho [(1 - \gamma) + \gamma \hat{r}_i]}{\rho [(1 - \gamma) + \gamma \hat{r}_i] + \varepsilon_{ot}}. \end{aligned} \quad (21)$$

where $\gamma \in [0, 1]$ controls the strength of confidence modulation and is set to 1.0 in all experiments. When $\gamma = 0$, Eq. (21) reduces to $\phi_i^a = \rho / (\rho + \varepsilon_{ot}) = \phi$, exactly matching plain unbalanced OT. With the adopted setting $\gamma = 1$, we have $\rho_i^a = \rho \hat{r}_i$ and $\phi_i^a = \rho \hat{r}_i / (\rho \hat{r}_i + \varepsilon_{ot})$. Thus, high-confidence clusters with $\hat{r}_i > 1$ receive stronger marginal constraints, whereas low-confidence clusters with $\hat{r}_i < 1$ are relaxed further. We denote $\phi^a = [\phi_1^a, \dots, \phi_K^a]^\top$. The target side uses the base coefficient ϕ .

When ρ and all ρ_i^a tend to infinity, the solution approaches entropy-regularized balanced OT. With $\gamma = 1$, however, a zero-confidence cluster remains unpenalized for any ρ .

For fixed similarities and confidence coefficients, we numerically approximate the minimizer of the entropy-regularized problem

$$\begin{aligned} \min_{\mathbf{P} \geq 0} \quad & \sum_{i,j} P_{ij} c_{ij} + \varepsilon_{ot} \sum_{i,j} P_{ij} (\log P_{ij} - 1) \\ & + \sum_i \rho_i^a d((\mathbf{P}\mathbf{1})_i \mid a_{m,i}^{adv}) \\ & + \rho \sum_j d((\mathbf{P}^\top \mathbf{1})_j \mid a_{m,j}^t), \end{aligned}$$

where $d(x \mid y) = x \log(x/y) - x + y$ is the scalar generalized KL divergence for $y > 0$, with $0 \log(0/y) = 0$. The entropy uses the same zero-mass convention. Confidence coefficients are fixed during each transport solve.

We retain the convention that rows correspond to adversarial clusters and columns to target clusters. The implementation stores the transposed matrix, with target clusters first. Under the notation used here, it initializes $\mathbf{u}^{(0)} = \mathbf{v}^{(0)} = \mathbf{1}$ and updates target-side scaling before adversarial-side scaling:

$$\begin{aligned} \mathbf{v}^{(\ell+1)} &= \left(\mathbf{a}_m^t \oslash (\mathbf{K}_m^\top \mathbf{u}^{(\ell)}) \right)^{\circ \phi}, \\ \mathbf{u}^{(\ell+1)} &= \left(\mathbf{a}_m^{adv} \oslash (\mathbf{K}_m \mathbf{v}^{(\ell+1)}) \right)^{\circ \phi^a}, \end{aligned} \quad (22)$$

where $\oslash$ denotes element-wise division and $\circ$ denotes element-wise exponentiation; ϕ^a acts element-wise on the adversarial-side row scaling $\mathbf{u}$, while the scalar exponent ϕ is applied to every entry of the target-side column scaling $\mathbf{v}$. The solver runs for at most 100 iterations and stops when $K^{-1} \|\mathbf{v}^{(\ell+1)} - \mathbf{v}^{(\ell)}\|_1 < 10^{-2}$. We use the final scaling vectors to form the numerical plan $\mathbf{P}_m^*$. The resulting transport plan balances matching costs with confidence-dependent penalties on marginal deviations:

$$\mathbf{P}_m^* = \text{diag}(\mathbf{u}) \mathbf{K}_m \text{diag}(\mathbf{v}). \quad (23)$$

To compute the local alignment term, the implementation normalizes the transport plan by its total mass:

$$\tilde{\mathbf{P}}_m = \frac{\mathbf{P}_m^*}{\sum_{i,j} \mathbf{P}_m^*(i, j) + \eta}, \quad \eta = 10^{-8}.$$

The transport plan and its confidence modulation are computed without gradient tracking. The local alignment loss is the normalized transported cosine distance:

$$\mathcal{L}_l^m = \sum_{i,j} \tilde{\mathbf{P}}_m(i, j) (1 - \mathbf{S}_m(i, j)). \quad (24)$$

The loss in Eq. (24) measures local matching cost using the normalized plan. The entropy and marginal penalties determine the raw plan $\mathbf{P}_m^*$ and are not added to this loss. Gradients of the local loss pass through the similarities while the transport weights are held fixed. Thus, the normalized plan weights the pairwise similarity gradients that guide the perturbation update. Confidence-adaptive relaxation changes these weights to accommodate unreliable correspondences; its effect on transferability is evaluated experimentally.

Hence, the final loss of the proposed IAU-FOA can be formulated as a combination of the global semantic alignment loss and the adaptive local transport loss. For the m -th surrogate image encoder, the loss is defined as:

$$\mathcal{J}_m = \mathcal{L}_g^m + \lambda_l \mathcal{L}_l^m, \quad (25)$$

where λ_l is a hyperparameter that balances the contribution of local transport alignment. We use the same fixed number of clusters for adversarial and target patch features throughout optimization, with $K = 10$ by default. The sensitivity to K is evaluated in Fig. 4.

G. Dynamic Ensemble Model Weighting

To improve adversarial transferability, adversarial examples are generated using the ensemble loss on several surrogate encoders. It is computed as follows:

$$\mathcal{J} = \sum_{m=1}^M \omega_m \mathcal{J}_m. \quad (26)$$

where $\mathcal{J}_m$ is the distance loss in Eq. (25). Dynamic weights are computed from alignment similarity scores, which are distinct from these distance losses. For the k -th evaluation of the ensemble objective, define

$$A_m^{(k)} = \cos(\hat{\mathbf{z}}_m^g(\mathbf{x}^c), \hat{\mathbf{z}}_m^g(\mathbf{x}_{tar})) + \lambda_l \sum_{i,j} \tilde{\mathbf{P}}_m(i,j) \mathbf{S}_m(i,j), \quad (27)$$

where $\mathbf{x}^c$ is the current augmented input and the local similarities and transport plan are evaluated for this input. Larger $A_m^{(k)}$ indicates stronger feature alignment. The dynamic weights are calculated from the score ratios defined below. We compute the similarity-score ratio between consecutive evaluations as

$$q_m^{(k)} = \frac{A_m^{(k)}}{A_m^{(k-1)} + \eta}, \quad (28)$$

where $\eta = 10^{-8}$ stabilizes the denominator. For the first evaluation, we set $A_m^{(0)} = A_m^{(1)}$. Here, k indexes successive objective evaluations, including evaluations of different augmented views within the same attack iteration. The ratio reflects score changes across evaluations, which can also arise from differences between augmented views. The weights are

$$\omega_m^{(k)} = M \cdot \frac{\exp(q_m^{(k)}/\tau_w)}{\sum_{n=1}^M \exp(q_n^{(k)}/\tau_w)}, \quad (29)$$

where $\tau_w = 1.0$ and the factor M ensures that the weights sum to M , matching the scale of equal weighting. For positive scores with $A_m^{(k-1)}$ well above η , a larger ratio indicates a larger current score relative to the preceding evaluation and receives a larger weight. This rule adapts the ensemble to observed score changes. The complete optimization procedure is summarized in Algorithm S1 in the Appendix.

H. Transferability Analysis

Prior studies relate adversarial transferability to gradient similarity and model smoothness [53], knowledge transferability [25], and ensemble transferability error [54]. We compare balanced OT with uniform local matching at a fixed feasible perturbation δ , holding the encoder, augmentation view, target features, and cluster assignments fixed.

Write $S_{ij} = \mathbf{S}_m(i,j)$ and $c_{ij} = 1 - S_{ij}$. Assume that the pairwise similarities are differentiable at δ . Let $\mathcal{J}_v$ be a differentiable victim alignment objective, with larger values indicating stronger target alignment, and set $\mathbf{g}_v = \nabla \mathcal{J}_v(\delta)$ and $\mathbf{g}_{ij} = \nabla_\delta S_{ij}$. For a nonnegative unit-mass plan $\mathbf{Q}$ held fixed during differentiation, define

$$\mathbf{G}(\mathbf{Q}) := \sum_{i,j} Q_{ij} \mathbf{g}_{ij}. \quad (30)$$

Here, $\mathbf{G}(\mathbf{Q})$ is the negative gradient of the local alignment loss with the plan held fixed. The uniform reference plan is $Q_{0,ij} = 1/K^2$.

Assumption 1 (Surrogate cost separation). *The $K \geq 2$ clusters admit a semantic correspondence permutation π , specified*

independently of the transport solution. The surrogate costs distinguish these correspondences with a gap $\Delta > 0$:

$$c_{ij} \geq c_{i\pi(i)} + \Delta, \quad j \neq \pi(i). \quad (31)$$

Assumption 2 (Cross-model gradient separation). *There exist $A > 0$, $B \geq 0$, and deviations $|e_{ij}| \leq \xi$ with $\xi \geq 0$ such that*

$$\langle \mathbf{g}_v, \mathbf{g}_{ij} \rangle = \begin{cases} A + e_{ij}, & j = \pi(i), \\ -B + e_{ij}, & j \neq \pi(i). \end{cases} \quad (32)$$

Define the noncorresponding mass

$$q(\mathbf{Q}) = \sum_i \sum_{j \neq \pi(i)} Q_{ij}, \quad q_0 := q(\mathbf{Q}_0) = 1 - \frac{1}{K}. \quad (33)$$

Theorem 1. *Let $H(\mathbf{Q}) = -\sum_{i,j} Q_{ij} \log Q_{ij}$, with $0 \log 0 = 0$. For $\varepsilon_{ot} \geq 0$, let $\mathbf{Q}_{ot}$ minimize*

$$\sum_{i,j} Q_{ij} c_{ij} - \varepsilon_{ot} H(\mathbf{Q}) \quad (34)$$

over nonnegative balanced plans satisfying $\sum_j Q_{ij} = 1/K$ for every i and $\sum_i Q_{ij} = 1/K$ for every j .

Under Assumption 1,

$$q(\mathbf{Q}_{ot}) \leq \bar{q} := \min \left\{ 1, \frac{\varepsilon_{ot} \log K}{\Delta} \right\}. \quad (35)$$

If Assumption 2 also holds and

$$\frac{\varepsilon_{ot} \log K}{\Delta} + \frac{2\xi}{A+B} < 1 - \frac{1}{K}, \quad (36)$$

then

$$\langle \mathbf{g}_v, \mathbf{G}(\mathbf{Q}_{ot}) - \mathbf{G}(\mathbf{Q}_0) \rangle \geq \kappa, \quad (37)$$

where

$$\kappa := (A+B)(q_0 - \bar{q}) - 2\xi > 0. \quad (38)$$

Under these conditions, let

$$\mathbf{d}_r = \mathbf{h} + \lambda_l \mathbf{G}(\mathbf{Q}_r), \quad r \in \{ot, 0\}, \quad (39)$$

where $\mathbf{h}$ is the common gradient contribution from the other branches and $\lambda_l > 0$ is the common local weight. Suppose $\|\mathbf{d}_r\|_2 \leq D$ for some $D > 0$, and $\mathcal{J}_v$ has an L -Lipschitz gradient along both update segments, with $L \geq 0$. Choose $\alpha > 0$ so that both updates remain feasible and, when $L > 0$, $\alpha < \lambda_l \kappa / (LD^2)$. Then

$$\mathcal{J}_v(\delta + \alpha \mathbf{d}_{ot}) - \mathcal{J}_v(\delta + \alpha \mathbf{d}_0) \geq \tau, \quad (40)$$

where

$$\tau := \alpha \lambda_l \kappa - L \alpha^2 D^2 > 0. \quad (41)$$

Thus, under the stated assumptions and step-size condition, OT produces stronger target alignment on the victim model than uniform local matching after one feasible ordinary gradient step from the same perturbation, with an objective gap of at least $\tau > 0$. This conditional advantage suggests that OT-based local alignment can improve the effectiveness of adversarial examples for targeted transfer, consistent with the empirical gains of the complete attack reported in Section IV. The proof and discussion of projected sign updates are provided in the Appendix (Section S1).

IV. EXPERIMENTS

A. Experimental Setup

Datasets. Following prior studies [4, 17, 24], we adopt 1,000 clean source images resized to $224 \times 224 \times 3$ from the NIPS 2017 Adversarial Attacks and Defenses Competition dataset¹. For the targeted setting, 1,000 images are randomly sampled from the MSCOCO validation set [27] as target images.

Implementation Settings. Following [23, 45], we generate adversarial examples using five surrogate models: three CLIP variants, ViT-B/16 [40], ViT-B/32 [40], and ViT-G/14 (LAION-2B) [40], together with InternVL3-1B [64] and DI-NOv2 [38]. For IAU-FOA, unless otherwise specified, the perturbation budget is $\epsilon = 16/255$ under the ℓ_∞ norm, with a step size of $1/255$ and 300 attack iterations. The number of local clusters is fixed at $K = 10$ throughout optimization. Local cluster reference masses are uniform, and the local-loss weight is $\lambda_l = 2.0$. Adaptive UOT uses $\rho = 0.2$, $\epsilon_{ot} = 0.1$, and $\gamma = 1$, with mean-normalized confidence and total-mass normalization of the transport plan. Dynamic weighting uses $\tau_w = 1.0$ and a numerical stabilizer of 10^{-8} . VIA uses intensity scales $\{0.25, 0.5, 1, 2, 4\}$ ($h = 2$) and a channel-gain amplitude of $s_\beta = 0.5$. Our main comparison covers twelve MLLMs, comprising six open-source models (Qwen2.5-VL-7B/72B, InternVL3-14B/78B, MiniCPM-o-2.6, LLaVA-1.6-7B) and six closed-source models (Claude-Sonnet-4.6/Opus-4.8, GPT-5.4/5.5, Gemini-3.1-Flash-Lite/3.5-Flash). An additional evaluation on five recent commercial MLLMs brings the total to seventeen targets. For all evaluations, the prompt is fixed as “Describe this image, no longer than 25 words”. Experiments are conducted on an Ubuntu system equipped with an NVIDIA H100 Tensor Core GPU with 80 GB of GPU memory.

Competitive Methods. We compare the proposed IAU-FOA with seven representative targeted transfer attacks for MLLMs, including AdvDiffVLM [12], AnyAttack [59], COA [50], M-Attack [24], FOA-Attack [17], M-Attack-v2 [61], and MPCAttack [23]. For baselines other than AdvDiffVLM, we use the same surrogate models, perturbation budget, and attack iterations as IAU-FOA; the remaining method-specific settings follow their original papers. AdvDiffVLM follows the settings of its original paper [12] and is not constrained by the ℓ_∞ budget of $16/255$. All methods use the same evaluation protocol described below.

Evaluation metrics. Following [17, 24], we evaluate targeted transferability with an LLM-as-a-judge protocol. For each adversarial–target pair, the target MLLM first produces captions for both images, and GPTScore is then used to quantify their semantic similarity by using GPT-4o. We regard an attack as successful when the similarity score is at least 0.5, indicating that the adversarial image is judged to share the target subject. In addition to the attack success rate (ASR), we report the average similarity score (AvgSim) to reflect overall target alignment; higher values of both metrics indicate better targeted transferability.

B. Hyper-parameter Selection

All results in Figs. 3 and 4 are evaluated on 100 source–target image pairs per target model. We investigate the sensitivity of IAU-FOA to three hyperparameters: the marginal-relaxation penalty weight ρ in adaptive unbalanced OT, the local-loss weighting factor λ_l in the dual-granularity alignment objective, and the number of local clusters K . We vary one hyperparameter at a time while keeping all other settings fixed. For ρ and λ_l , we report the arithmetic means of ASR and AvgSim over two models in each family: Gemini-3.5-Flash and Gemini-3.1-Flash-Lite for Gemini, GPT-5.5 and GPT-5.4 for GPT, and Claude-Sonnet-4.6 and Claude-Opus-4.8 for Claude. For K , we report ASR on all six commercial targets in the main evaluation.

As shown in Fig. 3(a), we vary ρ over $\{0.1, 0.2, 0.3, 0.4\}$. The setting $\rho = 0.2$ yields the highest reported ASR and AvgSim for Gemini, GPT, and Claude, achieving family-mean ASRs of 85.0%, 81.0%, and 90.5% for Gemini, GPT, and Claude, respectively, with corresponding mean AvgSim values of 65.8%, 63.2%, and 68.7%. These family-mean ASRs agree with the corresponding two-model averages of the full-method results in Table V. A smaller value may over-relax the marginal constraints, whereas a larger value moves the transport plan closer to balanced OT and may retain unreliable local correspondences. Fig. 3(b) evaluates $\lambda_l \in \{1.0, 2.0, 3.0, 4.0\}$. Setting $\lambda_l = 2.0$ also yields the highest reported values of both metrics for Gemini, GPT, and Claude, with the same ASR and AvgSim values reported above. These results favor an intermediate local-loss weight: too little emphasis on local alignment may limit its contribution, whereas too much may reduce the benefit of global semantic alignment. Based on these results, we use $\rho = 0.2$ and $\lambda_l = 2.0$ in all subsequent experiments.

Figure 4 further evaluates $K \in \{6, 8, 10, 12, 14, 16\}$, which controls the granularity of the clustered patch features used for local transport alignment. The highest mean ASR is 0.855 at $K = 10$, compared with 0.822 and 0.793 at $K = 6$ and $K = 8$, respectively. Increasing K beyond 10 yields mean ASRs between 0.828 and 0.838, indicating that finer clustering does not consistently improve aggregate transferability. The preferred setting nevertheless varies by target: $K = 10$ achieves the highest ASR on GPT-5.5, GPT-5.4, and Cl-S-4.6 and ties for the best result on G-3.5-F, whereas G-3.1-FL and Cl-O-4.8 peak at $K = 14$ with ASRs of 0.910 and 0.940, respectively. Based on the best mean performance across the six targets, we use a fixed $K = 10$ throughout optimization unless otherwise specified.

C. Comparative Experimental Results

Comparisons with different attack methods. All results in Tables I and II are evaluated on the same 1,000 source–target image pairs per target model. We compare the proposed IAU-FOA with seven targeted transfer attack baselines, including COA, AnyAttack, AdvDiffVLM, M-Attack, FOA, M-Attack-v2, and MPCAttack, across both open-source and closed-source MLLMs. As shown in Table I, IAU-FOA achieves the highest ASR and AvgSim on all six open-source targets

¹<https://nips.cc/Conferences/2017/CompetitionTrack>

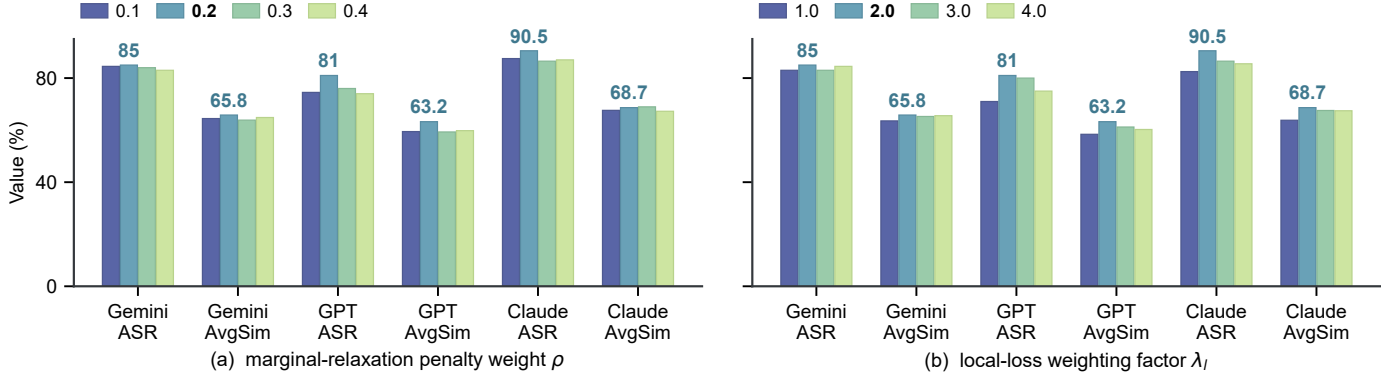


Fig. 3. Sensitivity to (a) the marginal-relaxation penalty ρ and (b) the local-loss weight λ_l , evaluated on 100 source–target image pairs per model. ASR and AvgSim are averaged over two models per family (Gemini, GPT, and Claude). Bold legend entries mark the selected settings, $\rho = 0.2$ and $\lambda_l = 2.0$; numerical labels show metric values.

TABLE I
TARGETED TRANSFER TO SIX OPEN-SOURCE MLLMs ON THE SAME 1,000 SOURCE–TARGET IMAGE PAIRS PER MODEL. ASR AND AVGSIM MEASURE ATTACK SUCCESS AND MEAN SIMILARITY, RESPECTIVELY; BOLD MARKS COLUMN MAXIMA.

Method	Target MLLMs												Average	
	Qwen2.5-VL-7B		InternVL3-14B		MiniCPM-o-2.6 (8B)		LLaVA-1.6-7B		Qwen2.5-VL-72B		InternVL3-78B		Overall	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
COA [50]	0.001	0.018	0.002	0.018	0.002	0.020	0.004	0.016	0.001	0.020	0.000	0.015	0.002	0.018
AnyAttack [59]	0.011	0.039	0.011	0.032	0.014	0.036	0.023	0.036	0.005	0.042	0.008	0.028	0.012	0.036
AdvDiffVLM [12]	0.309	0.313	0.315	0.318	0.293	0.292	0.309	0.296	0.314	0.327	0.194	0.239	0.289	0.298
M-Attack [24]	0.569	0.472	0.661	0.523	0.632	0.520	0.680	0.562	0.571	0.476	0.440	0.390	0.592	0.491
FOA [17]	0.575	0.488	0.703	0.561	0.692	0.551	0.725	0.598	0.617	0.506	0.490	0.422	0.634	0.521
M-Attack-v2 [61]	0.705	0.564	0.809	0.631	0.814	0.634	0.812	0.666	0.693	0.563	0.648	0.535	0.747	0.599
MPCAttack [23]	0.697	0.559	0.853	0.659	0.836	0.649	0.839	0.686	0.714	0.577	0.772	0.606	0.785	0.623
IAU-FOA (Ours)	0.821	0.645	0.910	0.710	0.866	0.692	0.885	0.723	0.823	0.653	0.871	0.685	0.863	0.685

TABLE II
TARGETED TRANSFER TO SIX COMMERCIAL MLLMs ON THE SAME 1,000 SOURCE–TARGET IMAGE PAIRS PER MODEL. ASR AND AVGSIM MEASURE ATTACK SUCCESS AND MEAN SIMILARITY, RESPECTIVELY; BOLD MARKS COLUMN MAXIMA.

Method	Target MLLMs												Average	
	Gemini-3.5-Flash		Gemini-3.1-Flash-Lite		GPT-5.5		GPT-5.4		Claude-Sonnet-4.6		Claude-Opus-4.8		Overall	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
COA [50]	0.001	0.010	0.000	0.013	0.002	0.019	0.001	0.018	0.000	0.013	0.002	0.018	0.001	0.015
AnyAttack [59]	0.004	0.019	0.004	0.028	0.009	0.032	0.005	0.027	0.008	0.026	0.006	0.029	0.006	0.027
AdvDiffVLM [12]	0.055	0.129	0.068	0.127	0.116	0.171	0.112	0.172	0.277	0.297	0.324	0.313	0.159	0.202
M-Attack [24]	0.309	0.321	0.312	0.333	0.285	0.292	0.219	0.255	0.425	0.384	0.535	0.452	0.348	0.339
FOA [17]	0.340	0.350	0.342	0.353	0.350	0.332	0.296	0.296	0.474	0.417	0.581	0.484	0.397	0.372
M-Attack-v2 [61]	0.525	0.455	0.506	0.450	0.454	0.401	0.383	0.357	0.630	0.511	0.723	0.568	0.537	0.457
MPCAttack [23]	0.673	0.547	0.675	0.542	0.453	0.408	0.390	0.365	0.621	0.514	0.724	0.581	0.589	0.493
IAU-FOA (Ours)	0.776	0.616	0.789	0.622	0.696	0.561	0.672	0.553	0.744	0.605	0.822	0.641	0.750	0.600

across the Qwen2.5-VL, InternVL3, MiniCPM-o, and LLaVA model families. Its reported average ASR and AvgSim are 86.3% and 68.5%, respectively, compared with 78.5% and 62.3% for MPCAttack, the strongest baseline on average. Compared with the best-performing baseline for each target and metric, IAU-FOA improves ASR by 3.0–11.6 percentage points and AvgSim by 3.7–8.1 percentage points. The largest gains in both metrics occur on Qwen2.5-VL-7B, where IAU-FOA achieves an ASR of 82.1% and an AvgSim of 64.5%, compared with 70.5% and 56.4% for M-Attack-v2.

As reported in Table II, IAU-FOA also ranks first in both metrics on all six commercial targets from the Gemini, GPT, and Claude families. Its reported average ASR and AvgSim

are 75.0% and 60.0%, respectively, compared with 58.9% and 49.3% for MPCAttack. Relative to the strongest competing method for each target and metric, the improvements range from 9.8 to 28.2 percentage points in ASR and from 6.0 to 18.8 percentage points in AvgSim. The largest gains occur on GPT-5.4, where IAU-FOA improves ASR from 39.0% to 67.2% and AvgSim from 36.5% to 55.3%. These results demonstrate that the proposed method outperforms previous methods in both attack success rate and average target-semantic alignment.

Evaluation on recent commercial MLLMs. In August 2026, we further evaluate IAU-FOA on five recent commercial MLLMs: OpenAI’s GPT-5.6-Luna, GPT-5.6-Sol, and GPT-

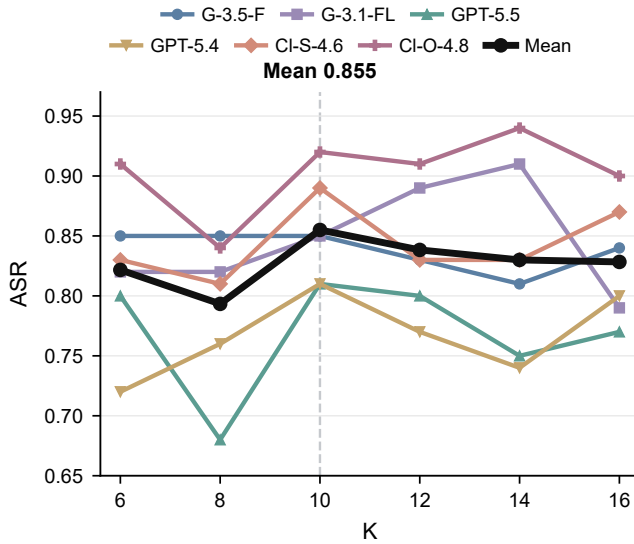


Fig. 4. Sensitivity to the number of local clusters K , evaluated on 100 source–target image pairs per model. Colored curves show ASR on six commercial MLLMs; the thick black curve shows their mean. The dashed line marks the default $K = 10$, which maximizes mean ASR among the tested values.

5.6-Terra, Anthropic’s Claude-Opus-5, and Google’s Gemini-3.7-Flash. These targets span three providers and multiple capability and efficiency tiers. We use the same 1,000 source–target image pairs per target model as the main comparison and compare against M-Attack-v2 and MPCAttack, its two strongest baselines. As shown in Table III, IAU-FOA achieves the highest ASR and AvgSim on all five targets, averaging 66.7% and 54.2%, respectively. Compared with MPCAttack, the strongest baseline on average, these results represent gains of 21.5 percentage points in ASR and 13.4 points in AvgSim. Across individual targets, improvements over the stronger baseline range from 7.3 to 30.7 percentage points in ASR and from 4.0 to 19.4 points in AvgSim. The largest ASR gain occurs on GPT-5.6-Luna, increasing from 25.0% to 55.7%. On GPT-5.6-Sol and GPT-5.6-Terra, ASR improves by 29.1 and 29.7 percentage points over the respective strongest baselines, while the gains on Claude-Opus-5 and Gemini-3.7-Flash are 7.3 and 9.9 points. These results demonstrate improved transferability across all five targets, with the largest gains on the GPT-5.6 family.

D. Evaluation under Adversarial Defenses

We conduct comparative experiments under six input-transformation and purification defenses: JPEG compression [5], total-variation minimization (TVM) [10], bit-depth reduction (BDR) [51], median filtering [51], ComDefend [18], and DISCO [14]. JPEG compression uses a quality factor of 75, BDR reduces the color depth to 4 bits, and median filtering uses a 3×3 kernel. All settings, including the undefended setting, are evaluated on 100 source–target image pairs per target model. Table IV reports the mean ASR across the six commercial MLLMs for each defense; Def. Avg. averages the six defended settings, excluding the undefended setting. IAU-FOA achieves the highest ASR in every defended setting, with an average of 0.688, compared with 0.495 for MPCAttack and

TABLE III
TARGETED TRANSFER TO FIVE RECENT COMMERCIAL MLLMS (AUGUST 2026). ASR AND AVGSIM ARE EVALUATED ON 1,000 SOURCE–TARGET IMAGE PAIRS PER MODEL; BOLD MARKS ROW MAXIMA.

Target MLLM	Metric	M-Attack-v2 [61]	MPCAttack [23]	IAU-FOA (Ours)
GPT-5.6-Luna	ASR	0.245	0.250	0.557
	AvgSim	0.275	0.277	0.471
GPT-5.6-Sol	ASR	0.323	0.315	0.614
	AvgSim	0.326	0.321	0.509
GPT-5.6-Terra	ASR	0.296	0.299	0.596
	AvgSim	0.306	0.307	0.492
Claude-Opus-5	ASR	0.669	0.683	0.756
	AvgSim	0.541	0.552	0.592
Gemini-3.7-Flash	ASR	0.553	0.713	0.812
	AvgSim	0.480	0.582	0.647
Average	ASR	0.417	0.452	0.667
	AvgSim	0.385	0.408	0.542

TABLE IV
ROBUSTNESS AGAINST INPUT DEFENSES: MEAN ASR ACROSS SIX COMMERCIAL MLLMS. ALL SETTINGS USE 100 SOURCE–TARGET IMAGE PAIRS PER MODEL. DEF. AVG. AVERAGES THE SIX DEFENDED SETTINGS; BOLD MARKS ROW MAXIMA.

Defense	M-Attack-v2 [61]	MPCAttack [23]	IAU-FOA(Ours)
No defense	0.638	0.690	0.855
JPEG ₇₅ [5]	0.557	0.640	0.803
BDR _{4-bit} [51]	0.508	0.605	0.778
Median _{3×3} [51]	0.343	0.382	0.605
TVM [10]	0.495	0.567	0.737
ComDefend [18]	0.370	0.393	0.618
DISCO [14]	0.297	0.382	0.587
Def. Avg.	0.428	0.495	0.688

0.428 for M-Attack-v2. These results correspond to average improvements of 19.3 and 26.0 percentage points, respectively. The gains over MPCAttack range from 16.3 to 22.5 percentage points across individual defenses. Under DISCO, IAU-FOA achieves an ASR of 0.587, exceeding MPCAttack (0.382) and M-Attack-v2 (0.297). These results show that the advantage of the proposed method persists after both conventional pre-processing and learned purification.

E. Ablation Study

To evaluate the contribution of each component in IAU-FOA, we compare the full method with four ablated variants: w/o OT removes the local alignment branch and retains only global alignment; w/o Adaptive UOT replaces adaptive unbalanced transport with balanced OT; w/o VIA disables visual-invariance augmentation; and w/o Dynamic ensemble fixes all surrogate weights to one. All variants use the same 100 source–target image pairs per target model against six commercial MLLMs, with all remaining attack and evaluation settings unchanged. Table V reports per-model ASR and the mean ASR and AvgSim across the six targets.

The full method achieves the highest mean ASR and AvgSim of 0.855 and 0.659, respectively. Removing local OT

TABLE V

COMPONENT ABLATION ON SIX COMMERCIAL MLLMs USING THE SAME 100 SOURCE-TARGET IMAGE PAIRS PER MODEL. G/CL DENOTE GEMINI/CLAUDE; F/FL/S/O DENOTE FLASH/FLASH-LITE/SONNET/OPUS. W/O OT REMOVES LOCAL ALIGNMENT; W/O ADAPTIVE UOT USES BOT; W/O DYNAMIC ENSEMBLE USES EQUAL WEIGHTING.

Variant	Attack success rate per target MLLM						Overall	
	G-3.5-F	G-3.1-FL	GPT-5.5	GPT-5.4	CL-S-4.6	CL-O-4.8	ASR	AvgSim
IAU-FOA (Ours)	0.850	0.850	0.810	0.810	0.890	0.920	0.855	0.659
w/o OT	0.680	0.730	0.650	0.550	0.760	0.780	0.692	0.543
w/o Adaptive UOT	0.840	0.800	0.780	0.750	0.850	0.850	0.812	0.635
w/o VIA	0.770	0.750	0.610	0.500	0.630	0.790	0.675	0.551
w/o Dynamic ensemble	0.850	0.800	0.770	0.730	0.830	0.820	0.800	0.646

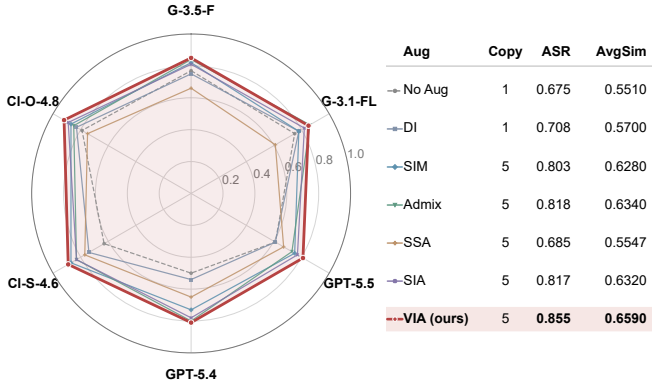


Fig. 5. Input augmentation strategies on six commercial MLLMs, evaluated on 100 source-target image pairs per model. The radar chart shows ASR on Gemini-3.5-Flash (G-3.5-F), Gemini-3.1-Flash-Lite (G-3.1-FL), GPT-5.5, GPT-5.4, Claude-Sonnet-4.6 (CL-S-4.6), and Claude-Opus-4.8 (CL-O-4.8). The table reports input copies (Copy) and mean ASR and AvgSim across targets.

alignment reduces these metrics by 16.3 and 11.6 percentage points, while replacing Adaptive UOT with balanced OT reduces them by 4.3 and 2.4 points, supporting both local matching and confidence-adaptive transport. Removing VIA causes the largest ASR decrease of 18.0 points, together with a 10.8-point decrease in AvgSim. Equal ensemble weighting reduces mean ASR and AvgSim by 5.5 and 1.3 points, respectively; ASR decreases on five targets and remains unchanged on G-3.5-F. Overall, all four ablations reduce average performance, with VIA contributing most to ASR and local OT alignment contributing most to AvgSim under the evaluated settings.

F. Analysis of the Proposed Method

We conduct all comparative experiments in this subsection on 100 source-target image pairs per model.

Visual-Invariance Augmentation. As shown in Fig. 5, we compare the proposed visual-invariance augmentation (VIA) with No Aug and five input-diversification strategies (DI [49], SIM [26], Admix [47], SSA [32], and SIA [48]) on six commercial MLLMs, while keeping all other attack settings fixed. VIA achieves the highest ASR on all six targets and the best aggregate performance, with an average ASR of 0.855 and an AvgSim of 0.6590. Compared with No Aug (0.675 ASR and 0.5510 AvgSim), it improves the two metrics by 18.0 and

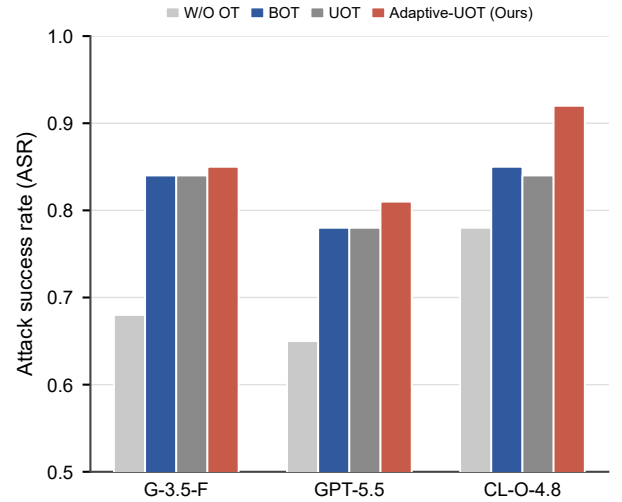


Fig. 6. OT-based local alignment on three commercial MLLMs, evaluated on 100 source-target image pairs per model. W/O OT removes the local alignment branch; BOT and UOT use balanced and plain unbalanced OT, respectively. Adaptive-UOT is our confidence-aware formulation and achieves the highest ASR on all three targets.

10.80 percentage points, respectively. Admix is the strongest competing augmentation on average, with an ASR of 0.818 and an AvgSim of 0.6340; VIA exceeds these scores by 3.7 and 2.50 percentage points. SIM, Admix, SSA, and SIA all use five transformed copies, matching VIA, yet obtain lower average ASR and AvgSim. This comparison indicates that the gains depend on the augmentation strategy, beyond the number of augmented views. VIA varies intensity and channel-wise gains while preserving spatial content, encouraging the perturbation to remain effective under exposure, illumination-color, and white-balance variations. The improvements across the Gemini, GPT, and Claude families support the effectiveness of these photometric transformations for targeted transferability. **Adaptive Unbalanced Optimal Transport.** As shown in Fig. 6, we compare Adaptive-UOT with W/O OT, BOT, and UOT while keeping all other attack settings fixed. W/O OT removes the local alignment branch, and BOT imposes balanced transport constraints; these correspond to w/o OT and w/o Adaptive UOT in Table V. UOT relaxes the marginal constraints without confidence-aware adaptation. Adaptive-UOT achieves the highest ASR on all three targets, reaching 0.85, 0.81, and 0.92 on G-3.5-F, GPT-5.5, and CL-O-4.8, respectively. Its mean ASR is 0.860, compared with 0.703,

TABLE VI
ASR COMPARISON OF ENSEMBLE WEIGHTING STRATEGIES. RESULTS USE 100 SOURCE-TARGET IMAGE PAIRS PER MODEL; ONLY THE WEIGHTING RULE VARIES. G/CL DENOTE GEMINI/CLAUDE; F/FL/S/O DENOTE FLASH/FLASH-LITE/SONNET/OPUS. BOLD MARKS COLUMN MAXIMA.

Weighting	G-3.5-F	G-3.1-FL	GPT-5.5	GPT-5.4	CL-S-4.6	CL-O-4.8	ASR
Equal	0.850	0.800	0.770	0.730	0.830	0.820	0.800
Random	0.810	0.800	0.760	0.770	0.890	0.870	0.817
Score-Prop.	0.850	0.830	0.800	0.740	0.860	0.910	0.832
Dynamic (Ours)	0.850	0.850	0.810	0.810	0.890	0.920	0.855

0.823, and 0.820 for W/O OT, BOT, and UOT, corresponding to gains of 15.7, 3.7, and 4.0 percentage points. All transport formulations outperform W/O OT on every target, supporting OT-based local alignment. Plain UOT matches BOT on G-3.5-F and GPT-5.5 but performs slightly worse on CL-O-4.8 (0.84 vs. 0.85). Adaptive-UOT’s largest gains occur on CL-O-4.8: 7.0 percentage points over BOT and 8.0 over UOT. These results support adapting marginal regularization to matching confidence; uniform relaxation alone yields no improvement over BOT here. These experimental results demonstrate the effectiveness of the proposed Adaptive Unbalanced Optimal Transport in improving targeted transferability.

Dynamic Ensemble Model Weighting. As shown in Table VI, we compare dynamic ensemble weighting with three alternatives while keeping all other attack settings fixed. Using the score notation in Sec. III-G, Equal Weighting sets every weight to one, and Random Weighting samples symmetric Dirichlet weights scaled to sum to M . Score-Proportional Weighting assigns $\omega_m^{(k)} \propto \max(A_m^{(k)}, 0)$ with the same normalization, reverting to equal weights when the sum of nonnegative scores is at most 10^{-8} . Thus, this control weights encoders by their current nonnegative alignment scores. Dynamic weighting instead uses the current-to-history score ratio in Eqs. (28) and (29). Dynamic weighting achieves the highest mean ASR of 0.855 across the six target models, compared with 0.800, 0.817, and 0.832 for Equal, Random, and Score-Proportional Weighting, respectively. These results correspond to gains of 5.5, 3.8, and 2.3 percentage points. It ranks first on all six targets, with outright best results on G-3.1-FL, GPT-5.5, GPT-5.4, and CL-O-4.8, and ties on G-3.5-F and CL-S-4.6. The largest gain over Score-Proportional Weighting occurs on GPT-5.4, where ASR increases from 0.740 to 0.810. These experimental results demonstrate the effectiveness of the proposed Dynamic Ensemble Model Weighting in improving targeted transferability. We further analyze visual fidelity and conduct experiments to examine the effect of the number of attack iterations, with results provided in the Appendix (Figs. S1 and S2). We also examine the association between source-target CLIP similarity and attack success, with results reported in the Appendix (Table S1).

V. CONCLUSION

In this paper, we investigate targeted adversarial transferability across heterogeneous MLLMs and propose IAU-FOA to improve black-box attack performance through global and

local feature alignment. Specifically, global alignment provides semantic guidance, while optimal transport matches clustered patch features to exploit fine-grained visual structures. To address unreliable local correspondences under balanced transport constraints, we propose adaptive unbalanced optimal transport, which adjusts marginal regularization according to matching confidence. We also provide a conditional theoretical analysis of how local transport alignment can improve targeted transfer. Furthermore, we develop a dynamic ensemble weighting strategy that adjusts each surrogate encoder’s contribution according to relative changes in its alignment similarity score. Finally, we propose a visual-invariance augmentation strategy that combines bidirectional intensity scaling and channel-wise white-balance adjustment to improve transferability under varying photometric conditions. Extensive experiments on open-source and closed-source MLLMs demonstrate that IAU-FOA outperforms the compared transferable attack methods. Ablation studies and component analyses further demonstrate the effectiveness of the proposed designs.

REFERENCES

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2022.
- [3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.
- [4] 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.
- [5] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.
- [6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2025.
- [7] 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. *arXiv preprint arXiv:2403.12445*, 2024.

- [8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [9] Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqian Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*, 2023.
- [10] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.
- [11] Meng-Hao Guo, Jiajun Xu, Yi Zhang, Jiaxi Song, Haoyang Peng, Yi-Xuan Deng, Xinzhi Dong, Kiyohiro Nakayama, Zhengyang Geng, Chen Wang, et al. R-bench: Graduate-level multi-disciplinary benchmarks for llm & mllm complex reasoning evaluation. *arXiv preprint arXiv:2505.02018*, 2025.
- [12] 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*, 20:1333–1348, 2025.
- [13] Keno Harada, Yudai Yamazaki, Masachika Taniguchi, Edison Marrese-Taylor, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. When instructions multiply: Measuring and estimating llm capabilities of multiple instructions following. *arXiv preprint arXiv:2509.21051*, 2025.
- [14] Chih-Hui Ho and Nuno Vasconcelos. Disco: Adversarial defense with local implicit functions. *Advances in neural information processing systems*, 35:23818–23837, 2022.
- [15] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.
- [16] Bhavuk Jain, Serkan Ö Arik, and Hardeo K Thakur. Adversarial attacks on multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2603.27918*, 2026.
- [17] 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*, 38:57993–58028, 2025.
- [18] Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6085. IEEE, 2019.
- [19] Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. Survey of adversarial robustness in multimodal large language models. *arXiv preprint arXiv:2503.13962*, 2025.
- [20] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290, 2023.
- [21] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.
- [22] 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 *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [23] Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiaojun 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*, pages 30065–30075, 2026.
- [24] 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/o1. *Advances in Neural Information Processing Systems*, 38:45018–45051, 2025.
- [25] Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *International Conference on Machine Learning*, pages 6577–6587. PMLR, 2021.
- [26] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.
- [27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.
- [28] 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*, pages 26296–26306, 2024.
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [30] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and texts. *arXiv preprint arXiv:2402.00357*, 2024.
- [31] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

- [32] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European conference on computer vision*, pages 549–566. Springer, 2022.
- [33] 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*, 2024.
- [34] Hui Lu, Yi Yu, Yiming Yang, Chenyu Yi, Xueyi Ke, Qixing Zhang, Bingquan Shen, Alex Kot, and Xudong Jiang. Make anything match your target: Universal adversarial perturbations against closed-source mllms via multi-crop routed meta optimization. *arXiv preprint arXiv:2601.23179*, 2026.
- [35] Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhui Chen. General-reasoner: Advancing llm reasoning across all domains. *Advances in Neural Information Processing Systems*, 38:56596–56618, 2025.
- [36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [37] Hefei Mei, Zirui Wang, Shen You, Minjing Dong, and Chang Xu. Veattack: Downstream-agnostic vision encoder attack against large vision language models. *arXiv preprint arXiv:2505.17440*, 2025.
- [38] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
- [39] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.
- [40] 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*, pages 8748–8763. PmLR, 2021.
- [41] Temurbek Rahmatullaev, Polina Druzhinina, Nikita Kurdiukov, Matvey Mikhalechuk, Andrey Kuznetsov, and Anton Razzhigaev. Universal adversarial attack on aligned multimodal llms. *arXiv preprint arXiv:2502.07987*, 2025.
- [42] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.
- [43] 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.
- [44] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.
- [45] 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*, 2026.
- [46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [47] 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)*, pages 16138–16147. IEEE, 2021.
- [48] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4584–4596. IEEE, 2023.
- [49] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.
- [50] Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14679–14689, 2025.
- [51] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.
- [52] Bo Yang, Hengwei Zhang, Jindong Wang, Yuchen Ren, Chenhao Lin, Chao Shen, and Zhengyu Zhao. Use as many surrogates as you want: Selective ensemble attack to unleash transferability without sacrificing resource efficiency. *arXiv preprint arXiv:2505.12644*, 2025.
- [53] Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655, 2021.
- [54] Wei Yao, Zeliang Zhang, Huayi Tang, and Yong Liu. Understanding model ensemble in transferable adversarial attack. *arXiv preprint arXiv:2410.06851*, 2024.
- [55] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.
- [56] Leitao Yuan, Qinghua Mao, Daizong Liu, Kun Wang, Wenjie Wang, Yan Teng, Jing Shao, and Dongrui Liu.

Frequency-domain regularized adversarial alignment for transferable attacks against closed-source mllms. *arXiv preprint arXiv:2605.21541*, 2026.

- [57] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.
- [58] Chiyu Zhang, Lu Zhou, Xiaogang Xu, Jiafei Wu, and Zhe Liu. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Computing Surveys*, 58(2):1–42, 2025.
- [59] 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 Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025.
- [60] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.
- [61] 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.
- [62] Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 18378–18394, 2024.
- [63] Hegui Zhu, Xiaoyan Sui, Yuchen Ren, Yanmeng Jia, and Libo Zhang. Boosting transferability of targeted adversarial examples with non-robust feature alignment. *Expert Systems with Applications*, 227:120248, 2023.
- [64] 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.

Visual-Invariance-Augmented Feature Optimal Alignment for Transferable Adversarial Attacks against Closed-Source MLLMs

Appendix

Xiaojun Jia, Simeng Qin, Yiming Li, Jie Liao, Sensen Gao,
Ke Ma, Yang Liu, and Xiaochun Cao

S1. PROOFS AND DETAILS OF THE TRANSFERABILITY ANALYSIS

This section proves Theorem 1 and discusses projected sign updates. Throughout the local comparison, the encoder, augmentation view, target features, cluster assignments, and transport weights are held fixed. In particular, the frozen-plan local loss satisfies

$$-\nabla_{\delta} \mathcal{L}_l(\delta; \mathbf{Q}) = -\nabla_{\delta} \sum_{i,j} Q_{ij}(1 - S_{ij}) = \sum_{i,j} Q_{ij} \mathbf{g}_{ij} = \mathbf{G}(\mathbf{Q}). \quad (\text{S1})$$

Proof of Theorem 1

Proof. Let $C(\mathbf{Q}) = \sum_{i,j} Q_{ij} c_{ij}$, $a_i = b_j = 1/K$, and $Q_{ij}^{\dagger} = K^{-1} \mathbf{1}\{j = \pi(i)\}$. The feasible plan $\mathbf{Q}^{\dagger}$ and Assumption 1 give

$$C(\mathbf{Q}_{ot}) - C(\mathbf{Q}^{\dagger}) = \sum_{i,j} Q_{ot,ij} (c_{ij} - c_{i\pi(i)}) \geq \Delta q(\mathbf{Q}_{ot}). \quad (\text{S2})$$

Optimality of $\mathbf{Q}_{ot}$ gives

$$C(\mathbf{Q}_{ot}) - C(\mathbf{Q}^{\dagger}) \leq \varepsilon_{ot} [H(\mathbf{Q}_{ot}) - H(\mathbf{Q}^{\dagger})] \leq \varepsilon_{ot} \log K, \quad (\text{S3})$$

where $H(\mathbf{Q}^{\dagger}) = \log K$ and $H(\mathbf{Q}_{ot}) \leq 2 \log K$. Hence Eq. (35) holds.

Let $E(\mathbf{Q}) = \sum_{i,j} Q_{ij} e_{ij}$, so $|E(\mathbf{Q})| \leq \xi$ for unit-mass plans. Assumption 2 yields

$$\langle \mathbf{g}_v, \mathbf{G}(\mathbf{Q}) \rangle = A[1 - q(\mathbf{Q})] - Bq(\mathbf{Q}) + E(\mathbf{Q}) = A - (A + B)q(\mathbf{Q}) + E(\mathbf{Q}). \quad (\text{S4})$$

Subtracting the two plan contributions gives

$$\begin{aligned} \langle \mathbf{g}_v, \mathbf{G}(\mathbf{Q}_{ot}) - \mathbf{G}(\mathbf{Q}_0) \rangle &= (A + B)[q_0 - q(\mathbf{Q}_{ot})] + E(\mathbf{Q}_{ot}) - E(\mathbf{Q}_0) \\ &\geq (A + B)(q_0 - \bar{q}) - 2\xi = \kappa. \end{aligned} \quad (\text{S5})$$

The condition in Eq. (36) gives $\bar{q} < q_0 - 2\xi/(A + B)$, hence $\kappa > 0$.

Since $\mathbf{h}$ cancels, $\langle \mathbf{g}_v, \mathbf{d}_{ot} - \mathbf{d}_0 \rangle \geq \lambda_l \kappa$. For $r \in \{ot, 0\}$, L -smoothness gives

$$|\mathcal{J}_v(\delta + \alpha \mathbf{d}_r) - \mathcal{J}_v(\delta) - \alpha \langle \mathbf{g}_v, \mathbf{d}_r \rangle| \leq \frac{L\alpha^2}{2} \|\mathbf{d}_r\|_2^2. \quad (\text{S6})$$

Subtracting the uniform-step upper bound from the OT-step lower bound yields

$$\begin{aligned} \mathcal{J}_v(\delta + \alpha \mathbf{d}_{ot}) - \mathcal{J}_v(\delta + \alpha \mathbf{d}_0) &\geq \alpha \lambda_l \kappa - \frac{L\alpha^2}{2} (\|\mathbf{d}_{ot}\|_2^2 + \|\mathbf{d}_0\|_2^2) \\ &\geq \alpha \lambda_l \kappa - L\alpha^2 D^2 = \tau. \end{aligned} \quad (\text{S7})$$

The step-size condition ensures $\tau > 0$ for $L > 0$; for $L = 0$, $\tau = \alpha \lambda_l \kappa > 0$. $\square$

Finite Solver Accuracy

For a balanced numerical plan with exact prescribed marginals, suppose its regularized objective is at most that of $\mathbf{Q}^{\dagger}$ plus $\zeta \geq 0$. The same argument gives

$$q(\mathbf{Q}) \leq \min \left\{ 1, \frac{\varepsilon_{ot} \log K + \zeta}{\Delta} \right\}. \quad (\text{S8})$$

For this inexact plan, the corresponding sufficient condition becomes $(\varepsilon_{ot} \log K + \zeta)/\Delta + 2\xi/(A + B) < q_0$. Thus, the optimization error must also fit within the available comparison margin. The scaling-vector stopping threshold in Section III-F is not itself a bound on ζ .

Relation to the Implemented Update

For a raw plan $\mathbf{P}$ with total mass $s = \sum_{i,j} P_{ij} > 0$, let $\mathbf{Q} = \mathbf{P}/s$. The implementation uses $\tilde{\mathbf{P}} = \mathbf{P}/(s + \eta) = \theta\mathbf{Q}$, where $\theta = s/(s + \eta) > 0$. Thus, with the local-loss weight λ_l , its local ascent contribution is $\lambda_l\theta\mathbf{G}(\mathbf{Q})$, corresponding to an effective local weight $\lambda_l\theta$ for a fixed plan. If two compared plans have different factors θ_1 and θ_0 , this difference must be retained. For $\|\mathbf{g}_{ij}\|_2 \leq G_{\max}$, and a unit-mass directional advantage bounded below by κ , one has

$$\langle \mathbf{g}_v, \theta_1\mathbf{G}(\mathbf{Q}_1) - \theta_0\mathbf{G}(\mathbf{Q}_0) \rangle \geq \theta_1\kappa - |\theta_1 - \theta_0| \|\mathbf{g}_v\|_2 G_{\max}. \quad (\text{S9})$$

This follows by adding and subtracting $\theta_1\mathbf{G}(\mathbf{Q}_0)$. For directions $\mathbf{d}_r = \mathbf{h} + \lambda_l\theta_r\mathbf{G}(\mathbf{Q}_r)$, the finite-step comparison uses λ_l times this corrected first-order bound and bounds on the resulting directions. Holding the remaining branch contributions and ensemble weights fixed isolates the effect of the local matching rule. The same decomposition can be summed over encoders and views with fixed nonnegative weights; it does not compare complete trajectories with changing weights or cluster assignments.

Algorithm S1 uses projected sign updates. Let $\mathcal{D}$ denote the feasible perturbation box, including the image-value constraints, and define the actual displacement

$$\mathbf{s}_r = \Pi_{\mathcal{D}}(\delta + \alpha \text{sign}(\mathbf{d}_r)) - \delta. \quad (\text{S10})$$

If $\mathcal{J}_v$ is L -smooth along these displacement segments, Taylor's inequality gives

$$\mathcal{J}_v(\delta + \mathbf{s}_1) - \mathcal{J}_v(\delta + \mathbf{s}_0) \geq \langle \mathbf{g}_v, \mathbf{s}_1 - \mathbf{s}_0 \rangle - \frac{L}{2} (\|\mathbf{s}_1\|_2^2 + \|\mathbf{s}_0\|_2^2). \quad (\text{S11})$$

To connect this displacement comparison to the preceding direction bound, let ν be a lower bound on $\langle \mathbf{g}_v, \mathbf{d}_1 - \mathbf{d}_0 \rangle$. For the ordinary directions of Theorem 1, $\nu = \lambda_l\kappa$; with the implemented normalization, use $\nu = \lambda_l[\theta_1\kappa - |\theta_1 - \theta_0| \|\mathbf{g}_v\|_2 G_{\max}]$ from Eq. (S9) instead. Define $\mathbf{b}_r := \mathbf{s}_r - \alpha\mathbf{d}_r$, the change in displacement caused by taking signs and projecting. Then

$$\begin{aligned} \langle \mathbf{g}_v, \mathbf{s}_1 - \mathbf{s}_0 \rangle &= \alpha \langle \mathbf{g}_v, \mathbf{d}_1 - \mathbf{d}_0 \rangle + \langle \mathbf{g}_v, \mathbf{b}_1 - \mathbf{b}_0 \rangle \\ &\geq \alpha\nu - \|\mathbf{g}_v\|_2 \|\mathbf{b}_1 - \mathbf{b}_0\|_2. \end{aligned} \quad (\text{S12})$$

Combining Eqs. (S11) and (S12) therefore gives

$$\begin{aligned} \mathcal{J}_v(\delta + \mathbf{s}_1) - \mathcal{J}_v(\delta + \mathbf{s}_0) &\geq \alpha\nu - \|\mathbf{g}_v\|_2 \|\mathbf{b}_1 - \mathbf{b}_0\|_2 - \frac{L}{2} (\|\mathbf{s}_1\|_2^2 + \|\mathbf{s}_0\|_2^2). \end{aligned} \quad (\text{S13})$$

Thus, a positive raw-direction margin transfers to a strict advantage for the implemented update whenever the displacement-change term and the curvature term together are smaller than $\alpha\nu$. This is an additional sufficient condition, not a consequence of the transport mass bound. In particular, making α small does not automatically remove the effect of taking signs, since that displacement change can also be of order α . The equations identify the extra quantities needed to pass from ordinary gradients to projected sign steps, without assuming that the ordering is preserved. Theorem 1 explains a local mechanism under explicit assumptions; it does not prove unconditional improvement of the full attack or discrete ASR. Whether these assumptions hold for a particular closed-source victim is not established by the proof. The reported experiments evaluate the complete method.

A Constructive Example of Possible Improvement

The sufficient conditions are nonempty, and improvement is possible even with sign updates. Consider a scalar perturbation $t \in [-0.01, 0.01]$, two surrogate clusters, and target features $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^3$. Define unit surrogate features

$$\begin{aligned} \mathbf{z}_1(t) &= (a(t), b(t), \sqrt{1 - a(t)^2 - b(t)^2}), \\ \mathbf{z}_2(t) &= (b(t), a(t), \sqrt{1 - a(t)^2 - b(t)^2}), \quad a(t) = 0.9 + t, \quad b(t) = 0.1 - 2t. \end{aligned} \quad (\text{S14})$$

These features are smooth throughout the feasible interval. Their cosine similarities to the targets are $S_{11} = S_{22} = a(t)$ and $S_{12} = S_{21} = b(t)$. Let the victim feature be $\mathbf{z}_v(t) = (t, \sqrt{1 - t^2})$ with target $(1, 0)$, giving the distinct victim alignment objective $\mathcal{J}_v(t) = t$. At $t = 0$, the identity correspondence has cost gap $\Delta = 0.8$; its pairwise derivatives are 1, whereas the off-diagonal derivatives are -2 . Thus, Eq. (32) holds with $A = 1$, $B = 2$, and $\xi = 0$.

For balanced entropic OT with $\varepsilon_{ot} = 0.1$, symmetry and the uniform marginals give the exact off-diagonal mass

$$q(\mathbf{Q}_{ot}) = \frac{1}{1 + \exp(8)}, \quad G(\mathbf{Q}_{ot}) = 1 - 3q(\mathbf{Q}_{ot}) > 0, \quad G(\mathbf{Q}_0) = -\frac{1}{2}. \quad (\text{S15})$$

With no other branch, the same positive local weight, and step size $\alpha = 0.01$, the projected sign steps from $t = 0$ yield $t_{ot} = 0.01$ and $t_0 = -0.01$. Hence $\mathcal{J}_v(t_{ot}) > \mathcal{J}_v(0) > \mathcal{J}_v(t_0)$. This explicit cosine-feature construction establishes that OT can improve a victim objective relative to uniform matching, including under a sign update. It is an existence example, not a model of the evaluated MLLMs or evidence that their gradients satisfy the assumptions.

S2. VISUALIZATION OF ADVERSARIAL EXAMPLES

Figure S1 compares the adversarial images and perturbations generated by COA, AdvDiffVLM, AnyAttack, M-Attack-v2, MPCAttack, and IAU-FOA on three source–target pairs. The left panel shows the clean sources and adversarial images, while the right panel displays the targets and corresponding perturbations. IAU-FOA preserves the main structures and colors of the source images and achieves the highest PSNR and SSIM in all three displayed examples. From top to bottom, its PSNR values are 29.8, 30.3, and 32.5 dB, with corresponding SSIM values of 0.808, 0.770, and 0.729. The average PSNR and SSIM are 30.9 dB and 0.769, respectively, compared with 30.1 dB and 0.702 for COA, the strongest baseline on both average metrics. These comparisons illustrate the visual fidelity of IAU-FOA on the displayed examples; aggregate image quality over the full evaluation set is not assessed here.

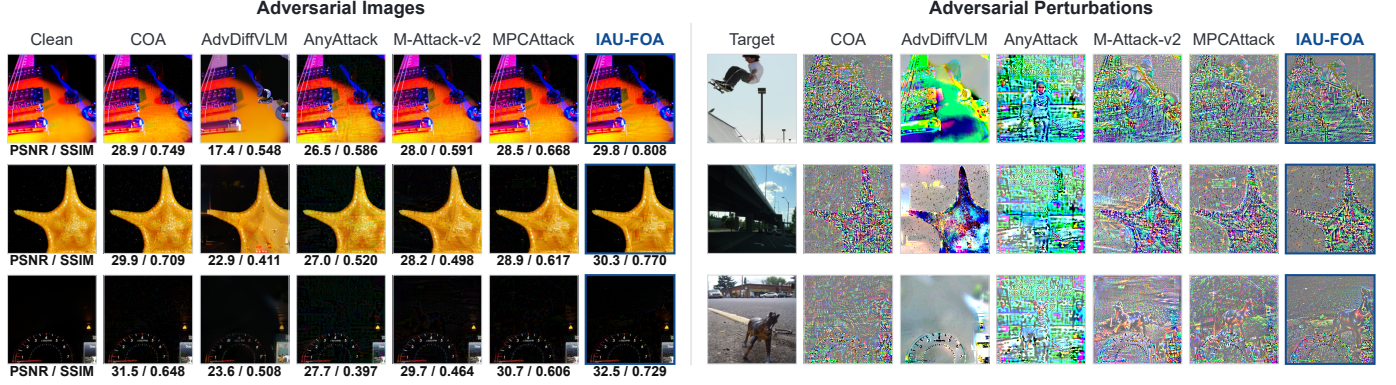


Fig. S1. **Adversarial examples and perturbations for three source–target pairs.** Each row shows one source–target pair. Left: clean source images and adversarial images generated by COA, AdvDiffVLM, AnyAttack, M-Attack-v2, MPCAttack, and IAU-FOA. Right: target images and the corresponding perturbation visualizations. PSNR (dB) and SSIM below each adversarial image are measured against its clean source; higher values indicate better visual fidelity.

S3. EFFECT OF THE NUMBER OF ATTACK ITERATIONS

All results in Fig. S2 are evaluated on 100 source–target image pairs per target model. Figure S2 evaluates the effect of the optimization budget by varying the number of attack iterations over $\{100, 200, 300, 400, 500\}$ while keeping all other settings fixed. Increasing the budget from 100 to 300 iterations improves ASR on all six commercial MLLMs and raises the mean ASR from 69.5% to 85.5%, a gain of 16.0 percentage points. Among the tested budgets, 300 iterations achieve the highest mean ASR. Increasing the budget to 400 iterations lowers ASR on every target and reduces the mean to 82.5%. At 500 iterations, the mean ASR is 84.5%, still 1.0 percentage point below the result at 300 iterations. Relative to 300 iterations, 500 iterations improve ASR on G-3.5-F and GPT-5.4, leave it unchanged on CI-S-4.6, and reduce it on the other three targets. Thus, additional iterations beyond 300 do not consistently improve targeted transferability. Considering the highest observed mean ASR and the lower computational cost relative to larger budgets, we retain 300 iterations as the default setting unless otherwise specified.

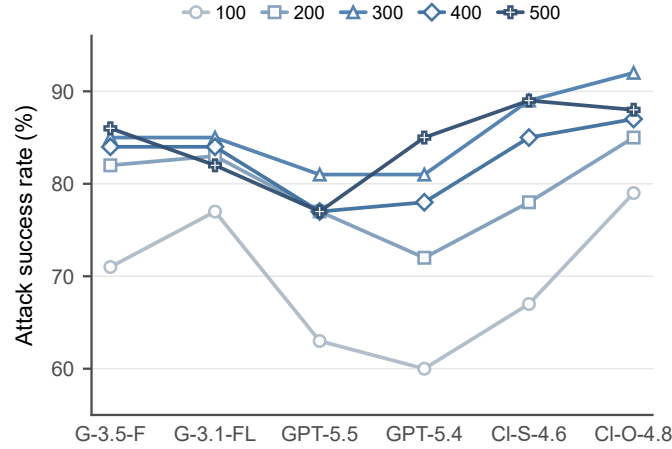


Fig. S2. **Effect of attack iterations on targeted transferability.** ASR on six commercial MLLMs for 100–500 iterations, evaluated on 100 source–target image pairs per model. The highest mean ASR (85.5%) occurs at 300 iterations.

S4. OPTIMIZATION ALGORITHM

Algorithm S1 summarizes the alignment and perturbation updates for supplied source transformations and target views. Spatial view preparation is represented by the input view schedule; visual-invariance augmentation, adaptive transport, and dynamic weighting are shown explicitly. We use $\eta = 10^{-8}$ and $\tau_w = 1.0$; transport plans, confidence modulation, and ensemble weights are held fixed during backpropagation.

Algorithm S1 IAU-FOA Optimization

Input: source image $\mathbf{x}_s$, target image $\mathbf{x}_{tar}$, surrogate encoder ensemble $\mathcal{F} = \{f_m\}_{m=1}^M$, perturbation budget ϵ , iterations T , step size α , local-loss weight λ_l , marginal KL weight ρ , fixed number of local clusters K (default: 10), prepared view schedule $\{\mathcal{V}^{(t)}\}_{t=0}^{T-1}$.

Output: adversarial image $\mathbf{x}_{adv}$.

- 1: Initialize $\delta^{(0)} = \mathbf{0}$, $\mathbf{x}_{adv}^{(0)} = \mathbf{x}_s + \delta^{(0)}$, and evaluation index $k = 1$.
- 2: Set $\mathbf{a}_m^{adv} = \mathbf{a}_m^t = \mathbf{1}/K$, $\varepsilon_{ot} = 0.1$, and $\phi = \rho/(\rho + \varepsilon_{ot})$.
- 3: **for** $t = 0$ to $T - 1$ **do**
- 4: Obtain prepared view pairs $(\mathcal{T}, \mathbf{x}_{tar}^c) \in \mathcal{V}^{(t)}$, with VIA parameters (s, β) in $\mathcal{T}$.
- 5: Set accumulated gradient $\mathbf{g} = \mathbf{0}$ and valid-copy count $n = 0$.
- 6: **for** each $(\mathcal{T}, \mathbf{x}_{tar}^c) \in \mathcal{V}^{(t)}$ **do**
- 7: $\mathbf{x}^c = \mathcal{T}(\mathbf{x}_{adv}^{(t)})$; obtain global target features and $\mathbf{Y}_m^{clu} = \text{KMeans}(\mathbf{Z}_m^p(\mathbf{x}_{tar}^c), K)$ without gradient tracking.
- 8: **for** $m = 1$ to M **do**
- 9: $\mathcal{L}_g^m = 1 - \cos(\hat{\mathbf{z}}_m^g(\mathbf{x}^c), \hat{\mathbf{z}}_m^g(\mathbf{x}_{tar}^c))$.
- 10: $\mathbf{X}_m^{clu} = \text{KMeans}(\mathbf{Z}_m^p(\mathbf{x}^c), K)$.
- 11: $\mathbf{S}_m(i, j) = \cos(\mathbf{x}_{m,i}^{clu}, \mathbf{y}_{m,j}^{clu})$.
- 12: $\mathbf{K}_m(i, j) = \exp(-(1 - \mathbf{S}_m(i, j))/\varepsilon_{ot})$.
- 13: $r_i = \max(\max_j \mathbf{S}_m(i, j), 0)$, $\hat{r}_i = r_i / (\frac{1}{K} \sum_j r_j + \eta)$, $\phi_i^a = \rho \hat{r}_i / (\rho \hat{r}_i + \varepsilon_{ot})$ with $\gamma = 1$.
- 14: Initialize $\mathbf{u} = \mathbf{v} = \mathbf{1}$; update $\mathbf{v}$ then $\mathbf{u}$ by Eq. (22), for at most 100 rounds or until the mean change in $\mathbf{v}$ is below 10^{-2} .
- 15: $\mathbf{P}_m^* = \text{diag}(\mathbf{u}) \mathbf{K}_m \text{diag}(\mathbf{v})$.
- 16: $\tilde{\mathbf{P}}_m = \text{stopgrad}(\mathbf{P}_m^* / (\sum_{i,j} \mathbf{P}_m^*(i, j) + \eta))$.
- 17: $\mathcal{L}_l^m = \sum_{i,j} \tilde{\mathbf{P}}_m(i, j) (1 - \mathbf{S}_m(i, j))$.
- 18: $\mathcal{J}_m^c = \mathcal{L}_g^m + \lambda_l \mathcal{L}_l^m$; $A_m^{(k)} = 1 - \mathcal{L}_g^m + \lambda_l \sum_{i,j} \tilde{\mathbf{P}}_m(i, j) \mathbf{S}_m(i, j)$.
- 19: **end for**
- 20: If $k = 1$, set $A_m^{(0)} = \text{stopgrad}(A_m^{(1)})$ for all m .
- 21: $q_m^{(k)} = \text{stopgrad}(A_m^{(k)}) / (A_m^{(k-1)} + \eta)$, $m = 1, \dots, M$.
- 22: $\omega_m^{(k)} = M \cdot \exp(q_m^{(k)} / \tau_w) / \sum_{j=1}^M \exp(q_j^{(k)} / \tau_w)$, $m = 1, \dots, M$.
- 23: $\mathcal{J}^c = \sum_{m=1}^M \omega_m^{(k)} \mathcal{J}_m^c$; store $A_m^{(k)} \leftarrow \text{stopgrad}(A_m^{(k)})$ for all m ; $k \leftarrow k + 1$.
- 24: $\mathbf{g} \leftarrow \mathbf{g} + \nabla_{\delta} \mathcal{J}^c$; $n \leftarrow n + 1$.
- 25: **end for**
- 26: $\mathbf{g} \leftarrow \mathbf{g} / \max(n, 1)$.
- 27: $\delta^{(t+1)} = \Pi_{[-\epsilon, \epsilon]}(\delta^{(t)} - \alpha \cdot \text{sign}(\mathbf{g}))$.
- 28: $\mathbf{x}_{adv}^{(t+1)} = \mathbf{x}_s + \delta^{(t+1)}$.
- 29: **end for**
- 30: **return** $\mathbf{x}_{adv}^{(T)}$.

S5. SOURCE–TARGET SIMILARITY BY ATTACK OUTCOME

To investigate whether the semantic proximity of a source–target image pair is associated with targeted attack success, we partition the IAU-FOA adversarial examples for each target MLLM into successful and failed subsets and compute the mean CLIP similarity between the corresponding clean source and target images. As reported in Table S1, successful attacks exhibit higher source–target similarity than failed attacks on all 12 target models. The table also reports the corresponding ASR. Within each model category, the Avg column aggregates successful and failed group means separately, weighted by their respective sample counts, and averages ASR across models. The differences in similarity are small and are interpreted as descriptive associations between source–target similarity and attack outcome. Because the analysis groups samples by attack outcome after evaluation, it does not establish that increasing source–target similarity causes higher ASR.

TABLE S1

SOURCE–TARGET CLIP SIMILARITY BY ATTACK OUTCOME. MEAN SIMILARITY FOR SUCCESSFUL AND FAILED IAU-FOA ATTACKS, WITH THE CORRESPONDING ASR. AVG REPORTS SAMPLE-COUNT-WEIGHTED MEANS FOR THE SUCCESSFUL AND FAILED GROUPS, AND THE ARITHMETIC MEAN OF ASR ACROSS MODELS. VALUES ARE REPORTED TO THREE DECIMAL PLACES. Q, I, M, L, G, AND CL DENOTE QWEN2.5-VL, INTERNVL3, MINICPM-O, LLAVA-1.6, GEMINI, AND CLAUDE; F, FL, S, AND O DENOTE FLASH, FLASH-LITE, SONNET, AND OPUS, RESPECTIVELY.

Open-source	Q-7B	Q-72B	I-14B	I-78B	M-2.6	L-7B	Avg
Successful	0.476	0.475	0.474	0.474	0.474	0.474	0.474
Failed	0.462	0.465	0.467	0.469	0.469	0.469	0.466
ASR	0.821	0.823	0.910	0.871	0.866	0.885	0.863
Closed-source	G-3.5-F	G-3.1-FL	GPT-5.5	GPT-5.4	CL-S-4.6	CL-O-4.8	Avg
Successful	0.474	0.475	0.478	0.475	0.476	0.474	0.475
Failed	0.469	0.465	0.463	0.469	0.464	0.470	0.466
ASR	0.776	0.789	0.696	0.672	0.744	0.822	0.750

S6. EVALUATION PROMPT

For reproducibility, Fig. S3 presents the complete prompt used by GPT-4o to compute GPTScore. For each adversarial–target pair, the two placeholders are replaced with the descriptions generated by the victim MLLM for the adversarial image and the corresponding target image. The judge is instructed to focus on subject-level semantic consistency, tolerate minor differences in wording and detail, and return only a scalar score in $[0, 1]$. We use this scalar directly when computing AvgSim and classify an attack as successful when the score is at least 0.5. The same prompt and decision rule are applied to all attack methods and victim models.

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: {input_text1}
Text 2: {input_text2}

Output only a single number between 0 and 1. Do not include any explanation or additional text.

Fig. S3. GPT-4o evaluation prompt for GPTScore. Placeholders contain the victim MLLM’s descriptions of the adversarial and target images.