Dual Feature-Relational Alignment for Transferable Targeted Attacks on MLLMs

Dongshen Han[1], **Xiaorong Tian**[1], **Sheng Zheng**[1,*], **Zishan Huang**[1], **Boyu Deng**[1]
Wei Dong[2], **Zhijie Li**[2], **Yang Yang**[1], **Chaoning Zhang**[1]
[1]University of Electronic Science and Technology of China
[2]Xi'an University of Architecture and Technology
zszhx2021@gmail.com

Abstract

Multimodal large language models (MLLMs) remain vulnerable to transferable adversarial examples. However, targeted transferable attacks are particularly challenging, requiring adversarial examples to predict specific target semantics and generalize from surrogate models to unseen black-box MLLMs. Existing methods primarily focus on aligning visual features, while the relational information among these features is overlooked. To bridge this gap, we propose Dual Feature-Relational Alignment Attack (DFRA-Attack), which couples feature alignment with relational alignment to improve targeted transferability. Specifically, semantic-aware alignment employs saliency-guided shared masking to construct consistent masked views of adversarial and target images. Building on these shared masked views, relational-aware alignment further captures relations among visible regions by aligning feature relational matrices and attention interaction patterns. Furthermore, we adopt a two-stage optimization strategy that transitions from masked-view to full-image optimization. Extensive experiments on open-source, closed-source, and reasoning-oriented MLLMs demonstrate that DFRA-Attack consistently outperforms state-of-the-art targeted transfer attacks in terms of both attack success rate and semantic similarity.

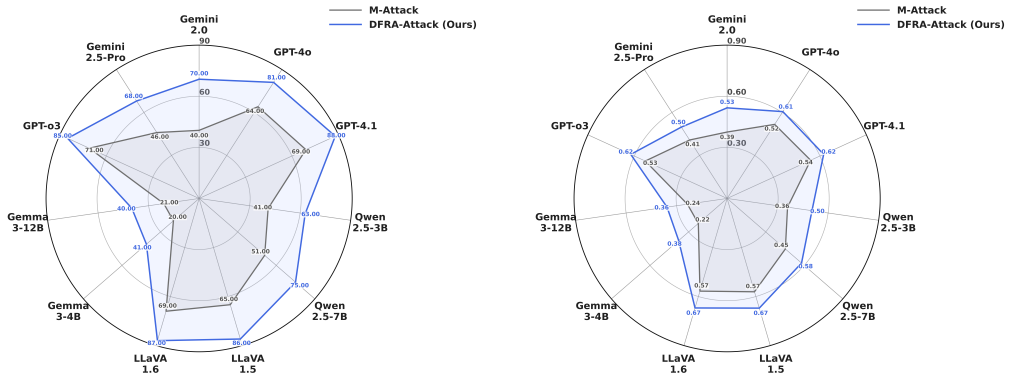
1 Introduction

Multimodal large language models (MLLMs) have achieved strong performance on diverse vision-language tasks, such as image captioning [2, 23], visual question answering [43, 16], and multimodal reasoning [40, 26]. Early vision-language models establish transferable visual-textual representations [35, 22, 23], while recent MLLMs further connect visual encoders with large language models for instruction following and multimodal understanding [2, 29, 6, 38]. Commercial systems such as GPT-4, Claude, and Gemini also demonstrate strong multimodal capabilities in real-world applications [1, 4, 37]. Despite these advances, MLLMs remain vulnerable to adversarial examples, where visually imperceptible perturbations can mislead models into incorrect or attacker-desired responses [11, 33, 7]. Since many advanced MLLMs are closed-source, black-box robustness evaluation through transferable attacks has become an important and realistic setting [36, 39, 41].

Among different adversarial settings, *targeted transferable attacks* are especially challenging. In this setting, perturbations are optimized on accessible surrogate models but are expected to induce a specific target semantic response on unseen black-box MLLMs. This requires adversarial examples to encode target-consistent semantics while generalizing across different model architectures. However, achieving these objectives simultaneously is difficult, as different MLLMs may rely on

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) ASR results. (b) AvgSim results.

Figure 1: Performance comparison between DFRA-Attack and M-Attack across different MLLMs in terms of (a) ASR and (b) AvgSim.

different visual encoders, alignment modules, and multimodal reasoning mechanisms [35, 23, 29]. Previous studies also show that surrogate-model bias and cross-architecture discrepancies can significantly affect black-box attack performance [47, 46].

Existing transferable attacks mainly improve transferability through global feature alignment, local region matching, or transformation-based optimization. Global alignment methods optimize image-level similarity between adversarial and target images, providing a direct target-guided objective [47, 13, 17]. However, such holistic objectives may over-rely on surrogate-specific global representations and leave fine-grained target evidence under-constrained. Recent local matching methods attempt to inject target semantics by aligning regions or patches [19, 25]. While these methods incorporate local visual information, they primarily focus on matching individual visual features. In particular, the relationships among fine-grained region features and their explicit alignment remain insufficiently explored. Consequently, adversarial perturbations may capture individual target cues without adequately preserving the relations among these cues, limiting the representation of target-consistent regional semantics.

To address this gap, we propose **Dual Feature-Relational Alignment Attack (DFRA-Attack)**, a targeted transferable attack framework for MLLMs that jointly considers semantic features and their relations. Specifically, semantic-aware alignment employs saliency-guided shared masking to construct consistent masked views of adversarial and target images, promoting the alignment of fine-grained semantic features. However, feature alignment alone does not explicitly characterize how different visible regions relate to one another. We therefore introduce relational-aware alignment to complement feature-level alignment by matching feature relational matrices and attention interaction patterns among visible regions. In addition, since masked views may not capture all target information available in the complete image, we adopt a two-stage optimization strategy that transitions from masked-view optimization to full-image optimization. Extensive experiments on diverse black-box MLLMs demonstrate that DFRA-Attack achieves significantly stronger targeted transferability and higher semantic fidelity than state-of-the-art baselines. As shown in Figure 1, DFRA-Attack consistently achieves stronger attack performance than M-Attack across diverse MLLMs, demonstrating superior targeted transferability and semantic fidelity. Our contributions are summarized as follows:

- We propose **DFRA-Attack**, a targeted transferable attack framework for MLLMs that couples semantic-aware alignment with relational-aware alignment to improve targeted transferability.
- We introduce semantic-aware alignment, which employs saliency-guided shared masking to promote the alignment of fine-grained semantic features.
- We introduce relational-aware alignment, which captures relations among visible regions by aligning feature relational matrices and attention interaction patterns.

- We adopt a two-stage optimization strategy that transitions from masked-view to full-image optimization. Extensive experiments demonstrate the effectiveness of DFRA-Attack across open-source, closed-source, and reasoning-oriented MLLMs.

2 Related Work

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) extend large language models with visual perception, enabling image captioning, visual question answering, instruction following, and multimodal reasoning [44, 24, 34, 14]. Early vision-language models such as CLIP [35] and BLIP [22, 23] learn transferable image-text representations through large-scale pre-training, providing important foundations for recent MLLMs. Open-source MLLMs further connect visual encoders with language models through projection layers or adapters [2, 48, 12]. Representative systems include the LLaVA series [29, 30, 3], the Qwen-VL family [6, 42], and Gemma-based multimodal models [38]. In parallel, closed-source MLLMs such as GPT-4o [1], Claude [4], and Gemini [37, 10] have shown strong multimodal reasoning and instruction-following abilities in practical applications. However, their internal architectures and parameters are usually inaccessible, making robustness evaluation more difficult [21, 31]. Recent studies show that MLLMs can still be vulnerable to adversarial inputs [11, 28, 20]. Therefore, transferable attacks provide a practical way to evaluate MLLM robustness under realistic black-box settings.

Adversarial Attacks on MLLMs. Adversarial attacks aim to mislead MLLMs via crafted visual perturbations [32, 8, 33, 7]. Given the restricted access to commercial models, transfer-based attacks, which optimize perturbations on accessible surrogates, have emerged as a practical evaluation paradigm [36, 39, 41, 45]. Existing targeted transfer attacks improve transferability through various strategies: AttackVLM [47] aligns global visual features, SSA-CWA [9] utilizes spectrum-based augmentation, while AdvDiffVLM [17] and AnyAttack [46] leverage generative models for natural adversarial noise. Recently, M-Attack [25] explores local semantic guidance via random cropping. However, such independent or stochastic local views break the spatial comparability between adversarial and target images, and largely ignore inter-region structural dependencies [18, 15, 19]. In contrast, DFRA-Attack jointly aligns local semantic features and their feature-level and attention-level relations through shared masked views to improve targeted transferability.

3 Methodology

We propose **DFRA-Attack**, a targeted transferable attack framework for MLLMs that combines semantic-aware alignment with relational-aware alignment. As shown in Fig. 2, DFRA-Attack consists of three components: semantic-aware alignment, relational-aware alignment, and a two-stage optimization strategy. Specifically, semantic-aware alignment employs saliency-guided shared masking to construct consistent masked views for aligning semantic features. Based on these shared masked views, relational-aware alignment further captures relations among visible regions through feature relational matrices and attention interaction patterns. Finally, we adopt a two-stage optimization strategy that first performs alignment under shared masked views and then continues optimization on the full images.

3.1 Preliminary

In a targeted transferable attack, the attacker is given a source image x_s and a target image x_{tar}. The goal is to generate an adversarial image x_{adv} from x_s such that its visual features approach those of x_{tar} under surrogate visual encoders. We define

$$x_{\mathrm{adv}} = x_s + \delta, \qquad \|\delta\|_\infty \le \epsilon, \tag{1}$$

where δ denotes the adversarial perturbation and ϵ is the ℓ_∞ perturbation budget.

To improve transferability, we employ an ensemble of K surrogate visual encoders, denoted as $\{f^k\}_{k=1}^{K}$. For the k-th surrogate encoder, we denote its image-level feature by $g^k(\cdot)$ and its regional features by $P^k(\cdot) \in \mathbb{R}^{N^k \times d}$, where N^k denotes the number of image regions and d is the feature dimension.

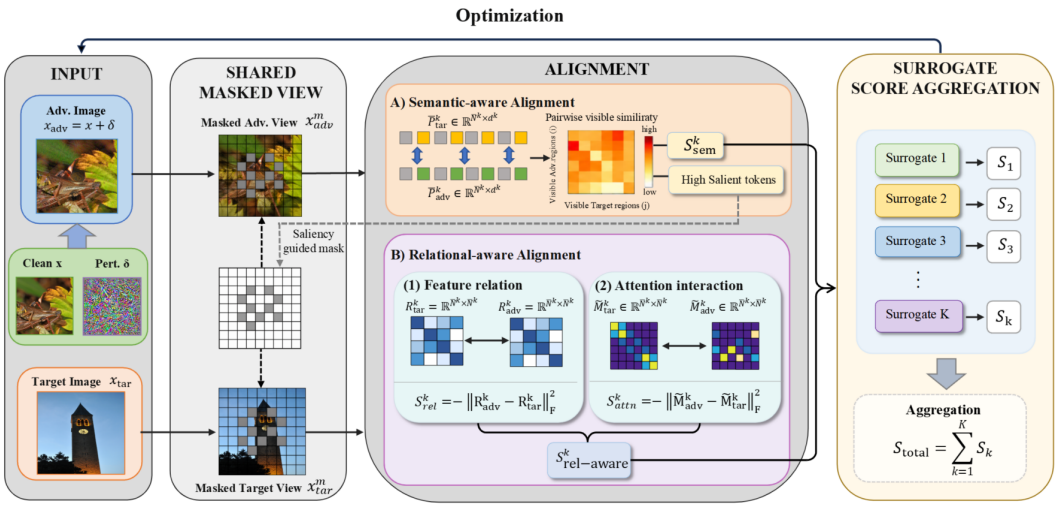


Figure 2: Overview of DFRA-Attack. DFRA-Attack combines semantic-aware alignment and relational-aware alignment to improve targeted transferability on MLLMs. Semantic-aware alignment matches regional features between the adversarial and target images, while relational-aware alignment captures feature relationships and attention interactions among visible regions.

Following common target-guided attacks, we use the cosine similarity between the image-level features of the adversarial and target images as a basic alignment objective:

$$\mathcal{S}^k_{\text{glb}} = \left\langle g^k(x_{\text{adv}}), g^k(x_{\text{tar}}) \right\rangle, \tag{2}$$

where $\langle \mathbf{u}, \mathbf{v} \rangle = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}$ denotes the cosine similarity between two vectors. While this objective provides direct target guidance, it does not explicitly align regional semantic features or constrain the relationships among visible regions. To address these limitations, DFRA-Attack introduces semantic-aware alignment and relational-aware alignment, as described below.

3.2 Semantic-aware Alignment

We utlize the region-level feature representations $P^k(\cdot)$ defined in Sec. 3.1 to align fine-grained semantic features under a shared masked view. Image-level features provide a global representation of the image but do not explicitly align the fine-grained semantic features of individual regions. To address this limitation, we employ saliency-guided shared masking to construct consistent masked views of the adversarial and target images. By applying the same mask to both images, we establish a common spatial reference for comparing their visible regions and aligning their fine-grained semantic features.

To construct the shared mask, we estimate the saliency of the target image using multiple surrogate encoders. Their saliency maps are mapped to a common spatial grid and combined to obtain a cross-surrogate saliency map. The shared mask ratio specifies the proportion of regions to be masked, with regions selected according to their saliency scores, prioritizing those with the highest scores. The same mask is applied to both the target and adversarial images, ensuring spatial consistency between their visible regions and enabling regional feature alignment under the same visual context.

Let $m(\cdot)$ denote the shared masking operation at the current iteration. We apply the same mask to both images:

$$x^m_{\text{tar}} = m(x_{\text{tar}}), \qquad x^m_{\text{adv}} = m(x_{\text{adv}}). \tag{3}$$

The shared mask ensures that the target and adversarial images are compared under the same visible spatial regions.

For the k-th surrogate encoder, we extract region-level features from the masked images:

$$P^k_{\text{tar}} = P^k(x^m_{\text{tar}}), \qquad P^k_{\text{adv}} = P^k(x^m_{\text{adv}}). \tag{4}$$

After excluding masked regions, we denote the retained feature representations by $\bar{P}^k_{\mathrm{tar}}, \bar{P}^k_{\mathrm{adv}} \in \mathbb{R}^{\bar{N}^k \times d^k}$, where $\bar{N}^k$ denotes the number of visible regions for the k-th surrogate encoder.

We align fine-grained semantic features under the shared masked view. Specifically, we compute pairwise cosine similarities between visible target and adversarial region features and use them to establish soft correspondences. The semantic-aware alignment score is defined as

$$\mathcal{S}^k_{\mathrm{sem}} = \sum_i \sum_j w^k_{ij} \left\langle \bar{P}^k_{\mathrm{tar},i}, \bar{P}^k_{\mathrm{adv},j} \right\rangle, \tag{5}$$

where $\bar{P}^k_{\mathrm{tar},i}$ and $\bar{P}^k_{\mathrm{adv},j}$ denote the feature representations of the i-th visible target region and the j-th visible adversarial region, respectively. The cosine similarity $\langle \cdot, \cdot \rangle$ is defined in Eq. (2). The soft correspondence weight w^k_{ij} is obtained by applying a softmax to pairwise cosine similarities. Maximizing this score encourages the adversarial image to align its fine-grained semantic features with those of the target image under consistent visible regions.

3.3 Relational-aware Alignment

While semantic-aware alignment promotes the matching of individual visible-region features, it does not explicitly constrain the relationships among these regions. We therefore introduce relational-aware alignment to further align the feature relationships and attention interactions among visible regions.

Given the visible region-level feature representations $\bar{P}^k_{\mathrm{tar}}$ and $\bar{P}^k_{\mathrm{adv}}$, we construct feature relational matrices:

$$R^k_{\mathrm{tar}} = \bar{P}^k_{\mathrm{tar}} (\bar{P}^k_{\mathrm{tar}})^\top, \qquad R^k_{\mathrm{adv}} = \bar{P}^k_{\mathrm{adv}} (\bar{P}^k_{\mathrm{adv}})^\top. \tag{6}$$

Each entry of a relational matrix represents the inner product between a pair of visible-region feature vectors, capturing their pairwise feature affinities. We define the feature relational alignment score as

$$\mathcal{S}^k_{\mathrm{rel}} = - \left\| R^k_{\mathrm{adv}} - R^k_{\mathrm{tar}} \right\|^2_F. \tag{7}$$

Maximizing this score encourages the adversarial image to preserve the pairwise feature relationships of the target image.

In addition to feature-level relationships, we introduce an auxiliary attention consistency term to capture interactions among visible regions. For transformer-based surrogate encoders, we extract the last-layer self-attention maps and retain the entries corresponding to interactions among visible regions. Let the resulting attention maps for the target and adversarial images be denoted by $\tilde{M}^k_{\mathrm{tar}}$ and $\tilde{M}^k_{\mathrm{adv}}$, respectively. The attention interaction alignment score is defined as

$$\mathcal{S}^k_{\mathrm{attn}} = - \left\| \tilde{M}^k_{\mathrm{adv}} - \tilde{M}^k_{\mathrm{tar}} \right\|^2_F. \tag{8}$$

We combine the feature-level and attention-level terms to obtain the overall relational-aware alignment score:

$$\mathcal{S}^k_{\mathrm{rel-aware}} = \lambda_{\mathrm{rel}} \mathcal{S}^k_{\mathrm{rel}} + \lambda_{\mathrm{attn}} \mathcal{S}^k_{\mathrm{attn}}, \tag{9}$$

where λ_{rel} and λ_{attn} control the contributions of feature relational alignment and attention interaction alignment, respectively. This objective complements semantic-aware alignment by encouraging the adversarial image to match both the feature relationships and attention interaction patterns of the target image.

3.4 Two-stage Optimization Strategy

Semantic-aware alignment and relational-aware alignment are performed under shared masked views, enabling the alignment of fine-grained semantic features and their relations. We further adopt a two-stage optimization strategy that transitions from masked-view optimization to full-image optimization.

For the k-th surrogate encoder, the overall alignment objective is defined as

$$\mathcal{S}^k = \mathcal{S}^k_{\mathrm{glb}} + \lambda_{\mathrm{sem}} \mathcal{S}^k_{\mathrm{sem}} + \mathcal{S}^k_{\mathrm{rel-aware}}, \tag{10}$$

where λ_{sem} controls the contribution of semantic-aware alignment, and $\mathcal{S}^k_{\text{rel}-\text{aware}}$ denotes the relational-aware alignment score.

Let T denote the transition iteration. During the masked-view optimization stage, a shared mask is applied to both adversarial and target images:

$$x^t_{\text{tar}} = m_{\rho_t}(x_{\text{tar}}), \qquad x^t_{\text{adv}} = m_{\rho_t}(x_{\text{adv}}), \quad t < T, \tag{11}$$

where ρ_t denotes the mask ratio at iteration t. After iteration T, the shared mask is removed, and the optimization continues on the full images. The semantic-aware and relational-aware alignment objectives are retained to guide adversarial optimization using complete visual information. Finally, the alignment scores from multiple surrogate encoders are aggregated to guide the overall adversarial optimization.

4 Experiments

4.1 Experimental Setup

Datasets and baselines. Following the transfer attack protocol in [47], we construct 100 semantically distinct source-target image pairs from the NIPS 2017 Adversarial Attacks and Defenses Competition dataset [5] and the MSCOCO validation set [27]. We compare DFRA-Attack with AdvDiffVLM [17], SSA-CWA [13], AnyAttack [46], M-Attack [25]. We use M-Attack [25] as the reference baseline and follow its experimental settings for all methods. All images are resized to 224×224. The perturbation budget is set to $\epsilon = 16/255$ under the ℓ_∞ norm. All adversarial image generation and open-source MLLM evaluations are conducted on NVIDIA RTX 4090 GPUs, while closed-source MLLMs are evaluated through official APIs.

Evaluation metrics. We evaluate attack performance using Attack Success Rate (ASR), Average Semantic Similarity (AvgSim), and Keyword Matching Rate (KMR). ASR and AvgSim follow the LLM-as-a-judge protocol in [25], where GPT-4o scores the semantic similarity between the generated caption of the adversarial image and the target image. An attack is considered successful when the similarity score exceeds 0.5. KMR measures whether generated descriptions cover manually annotated target keywords, with threshold variants KMR_α, KMR_β, and KMR_γ corresponding to matching ratios of at least 0.25, 0.50, and 1.00. By combining continuous semantic scores with discrete keyword constraints, this multi-level evaluation protocol ensures a rigorous and comprehensive assessment of targeted transferability.

Table 1: Quantitative comparison of targeted attack performance on open-source MLLMs.

Method	**Qwen2.5-VL-3B**		**Qwen2.5-VL-7B**		**LLaVA-1.5-7B**		**LLaVA-1.6-7B**		**Gemma3-4B**		**Gemma3-12B**	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
AdvDiffVLM [17]	0.04	0.105	0.11	0.160	0.13	0.160	0.12	0.134	0.05	0.095	0.04	0.119
SSA-CWA [13]	0.07	0.172	0.13	0.155	0.17	0.212	0.16	0.152	0.09	0.111	0.07	0.133
AnyAttack [46]	0.16	0.211	0.27	0.278	0.33	0.312	0.42	0.312	0.11	0.162	0.13	0.178
M-Attack [25]	0.41	0.359	0.51	0.453	0.65	0.571	0.69	0.567	0.20	0.224	0.21	0.236
DFRA-Attack (Ours)	0.63	0.504	0.75	0.579	0.86	0.671	0.87	0.670	0.41	0.377	0.40	0.357

4.2 Quantitative Comparison

Open-source MLLMs. As shown in Table 1, DFRA-Attack consistently outperforms all baseline methods across all evaluated open-source MLLMs, including the Qwen, LLaVA, and Gemma series. Compared with M-Attack, DFRA-Attack improves ASR by 0.22 and 0.24 on Qwen2.5-VL-3B and Qwen2.5-VL-7B, and by 0.21 and 0.18 on LLaVA-1.5-7B and LLaVA-1.6-7B, respectively. On the Gemma series, DFRA-Attack also achieves clear improvements, increasing ASR by 0.21 on Gemma3-4B and 0.19 on Gemma3-12B. In terms of AvgSim, DFRA-Attack consistently surpasses M-Attack across all evaluated models, with improvements of 0.145 and 0.126 on the Qwen models, 0.100 and 0.103 on the LLaVA models, and 0.153 and 0.121 on the Gemma models. These results demonstrate that DFRA-Attack effectively improves both attack success rate and semantic similarity across diverse open-source MLLM architectures.

Closed-source MLLMs. Table 2 reports results on closed-source MLLMs, including Claude Sonnet 4.5, Gemini-2.0-Flash, GPT-4o, and GPT-4.1. DFRA-Attack achieves ASR scores of 0.13, 0.70, 0.81, and 0.88 on these four models, respectively. Compared with the strongest baseline M-Attack, DFRA-Attack improves the ASR by 0.10, 0.30, 0.17, and 0.19, respectively. In terms of AvgSim, DFRA-Attack obtains scores of 0.184, 0.532, 0.606, and 0.624, outperforming M-Attack by absolute margins of 0.112, 0.142, 0.089, and 0.084. These results demonstrate that DFRA-Attack achieves strong targeted transferability and semantic similarity across diverse closed-source MLLMs under black-box settings.

Table 2: Quantitative comparison of targeted attack performance on closed-source MLLMs.

Method	**Claude Sonnet 4.5**		**Gemini-2.0-Flash**		**GPT-4o**		**GPT-4.1**	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
AdvDiffVLM [17]	0.01	0.048	0.08	0.104	0.08	0.145	0.09	0.149
SSA-CWA [13]	0.01	0.076	0.14	0.181	0.11	0.177	0.12	0.211
AnyAttack [46]	0.02	0.066	0.29	0.223	0.21	0.321	0.28	0.298
M-Attack [25]	0.03	0.072	0.40	0.390	0.64	0.517	0.69	0.540
DFRA-Attack (Ours)	0.13	0.184	0.70	0.532	0.81	0.606	0.88	0.624

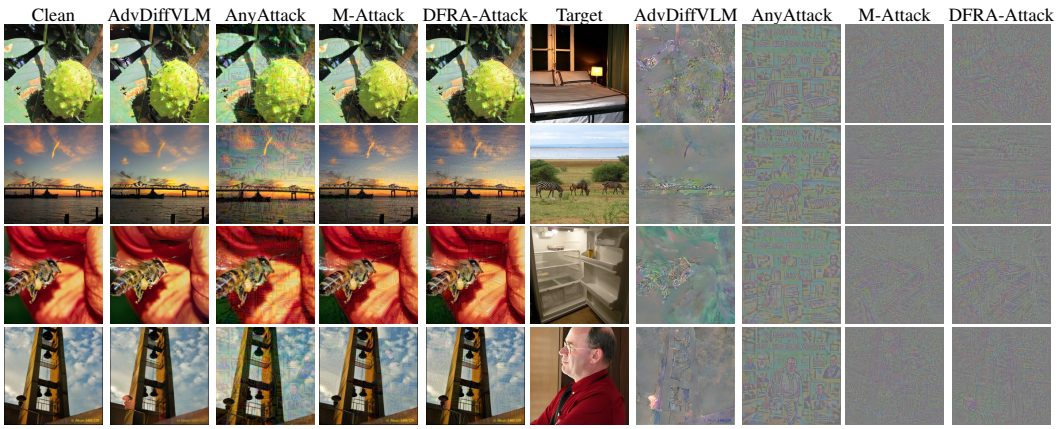
Reasoning MLLMs. To assess performance on reasoning-oriented MLLMs, we evaluate DFRA-Attack on GPT-o3, Gemini-2.5-Pro, and Claude-4.5-thinking (Table 3). Our method achieves 0.85 ASR and 0.623 AvgSim on GPT-o3, alongside 0.68 ASR and 0.498 AvgSim on Gemini-2.5-Pro, indicating that DFRA-Attack effectively transfers target semantics to advanced reasoning models. While Claude-4.5-thinking proves more challenging for all evaluated methods, yielding generally lower success rates, DFRA-Attack still obtains the highest alignment (0.08 ASR and 0.141 AvgSim). Overall, the consistent performance across these models confirms the effectiveness of our approach under complex reasoning settings.

Table 3: Quantitative comparison of targeted attack performance on reasoning MLLMs.

Method	GPT-o3		Gemini-2.5-Pro		Claude-4.5-thinking	
	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
AdvDiffVLM [17]	0.09	0.157	0.05	0.087	0.03	0.054
SSA-CWA [13]	0.11	0.177	0.08	0.112	0.02	0.043
AnyAttack [46]	0.08	0.205	0.09	0.209	0.02	0.032
M-Attack [25]	0.71	0.529	0.46	0.408	0.03	0.077
DFRA-Attack (Ours)	**0.85**	**0.623**	**0.68**	**0.498**	**0.08**	**0.141**

4.3 Qualitative Comparisons

Visual comparisons of the generated adversarial examples and their corresponding noise patterns are provided in Fig. 3. While baselines such as AdvDiffVLM and AnyAttack often introduce globally scattered noise, and M-Attack suffers from localized but blocky artifacts due to its unconstrained random cropping strategy, DFRA-Attack maintains exceptionally high visual fidelity. Moreover, the residual maps demonstrate that our method injects highly regularized and spatially coherent semantic patterns rather than unguided random noise. Specifically, the perturbations generated by DFRA-Attack implicitly outline the semantic contours and structural boundaries of the target objects. This distinct visual behavior confirms the effectiveness of our locality-aware design: the shared masking restricts semantic injection to well-aligned visible regions, successfully avoiding widespread background distortion, while the relational alignment explicitly preserves fine-grained inter-region dependencies. Consequently, DFRA-Attack tightly constrains the optimization space, achieving an optimal balance between visual imperceptibility and strong targeted transferability.



(a) Adversarial examples **(b)** Adversarial perturbations

Figure 3: Qualitative comparison of adversarial examples and perturbations generated by DFRA-Attack and representative targeted transfer attack methods. The left part shows the clean image and adversarial examples, while the right part shows the target image and corresponding perturbation visualizations.

4.4 Ablation Study

We conduct ablation studies to validate the effectiveness of the key components in DFRA-Attack. Unless otherwise specified, all experiments are evaluated on Gemini-2.0-Flash using 100 image pairs under the ℓ_∞ perturbation budget $\epsilon = 16/255$.

Table 4: Ablation study on the alignment components of DFRA-Attack.

Variant	ASR	AvgSim
Baseline	0.40	0.390
w Semantic-aware	0.58	0.459
w Relation-aware	0.59	0.445
DFRA-Attack (Ours)	**0.70**	**0.532**

Table 5: Ablation study on the saliency-guided mask ratio.

Mask Ratio	ASR	AvgSim
0.05	0.65	0.519
0.10	**0.70**	**0.532**
0.15	0.65	0.517
0.20	0.63	0.510

Effect of Alignment Components. To evaluate the contributions of the two alignment components, we compare the baseline with variants incorporating semantic-aware alignment or relational-aware alignment individually, as well as the complete DFRA-Attack. As shown in Table 4, adding semantic-aware alignment improves ASR from 0.40 to 0.58 and AvgSim from 0.390 to 0.459, while adding relational-aware alignment achieves an ASR of 0.59 and an AvgSim of 0.445. These results demonstrate that both components improve targeted transferability. The improvements also suggest that saliency-guided shared masking provides a useful spatial reference for aligning visible features and their relations. Combining both alignment components further improves performance, achieving an ASR of 0.70 and an AvgSim of 0.532. Compared with the semantic-aware variant, DFRA-Attack improves ASR by 0.12 and AvgSim by 0.073; compared with the relational-aware variant, the improvements are 0.11 and 0.087, respectively. These results demonstrate the complementary benefits of the two alignment components.

Effect of the shared mask ratio. We investigate the effect of the mask ratio by varying it from 0.05 to 0.20. As shown in Table 5, DFRA-Attack achieves the best performance at a mask ratio of 0.10, reaching 70.0% ASR and 0.532 AvgSim. In our saliency-guided shared masking strategy, a small mask ratio removes only a limited portion of highly salient regions, which may be insufficient to reduce the influence of dominant responses. In contrast, a larger mask ratio removes more visual information and may weaken regional feature alignment under the masked views. These results reveal a trade-off between suppressing dominant visual cues and retaining sufficient visual information for feature alignment. An initial mask ratio of 0.10 provides a favorable trade-off between the two and is therefore adopted in our experiments.

To further examine the effect of shared masking on regional feature alignment, we generate adversarial images with and without shared masking and compute the cosine similarity between spatially corresponding visual tokens of the adversarial and target images. As shown in Fig. 4, the middle and right images visualize the token-level feature similarity without and with shared masking, respectively. Brighter regions indicate higher feature similarity to the target image. Compared with the variant without shared masking, our saliency-guided shared masking strategy increases the mean token similarity from 0.290 to 0.524. This observation suggests that masking a small portion of highly salient regions while maintaining a shared spatial reference facilitates feature alignment between the adversarial and target images.



Figure 4: Visualization of token-level cosine similarity between the adversarial and target images. From left to right: source image, similarity heatmap without shared masking, and similarity heatmap with shared masking. Brighter regions indicate higher feature similarity.

4.5 Additional Analysis

Table 6: Comparison of KMR across different methods on multiple MLLMs.

Method	**GPT-4o**			**Gemini-2.0**		
	KMR_α	KMR_β	KMR_γ	KMR_α	KMR_β	KMR_γ
AdvDiffVLM [17]	0.02	0.00	0.00	0.02	0.00	0.00
SSA-CWA [13]	0.05	0.02	0.00	0.11	0.06	0.00
AnyAttack [46]	0.46	0.21	0.05	0.44	0.20	0.04
M-Attack [25]	0.75	0.53	0.11	0.82	0.54	0.13
DFRA-Attack (Ours)	0.91	0.72	0.19	0.92	0.77	0.32

Keyword Matching Rate. We report the Keyword Matching Rate (KMR) to assess how well adversarial descriptions cover the target keywords. Following [25], we evaluate KMR_α, KMR_β, and KMR_γ at matching-ratio thresholds of 0.25, 0.50, and 1.00, respectively. As shown in Table 6, DFRA-Attack achieves KMR_α scores of 0.91 on GPT-4o and 0.92 on Gemini-2.0. DFRA-Attack also performs well under the stricter thresholds, reaching a KMR_γ of 0.32 on Gemini-2.0. These results are consistent with our use of saliency-guided masking and relational alignment to guide the generation of target-consistent adversarial descriptions.

Table 7: Effect of combining DFRA with FOA-Attack on GPT-4o and Gemini-2.0-Flash.

Method	**GPT-4o**		**Gemini-2.0-Flash**	
	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
FOA-Attack [19]	0.74	0.589	0.63	0.499
FOA+DFRA	**0.88**	**0.658**	**0.71**	**0.531**

Combination with an Existing Attack. To further evaluate the effectiveness and generalizability of DFRA, we investigate whether its proposed alignment strategies can further enhance the performance of an existing targeted transfer attack. We select FOA-Attack [19], a highly competitive and representative state-of-the-art method for targeted transfer attacks against black-box MLLMs, as our baseline. Specifically, we incorporate DFRA's semantic-aware alignment, relational-aware alignment, and saliency-guided shared masking into the FOA-Attack optimization framework, while retaining its original dynamic surrogate reweighting mechanism. We denote the resulting variant as FOA+DFRA. As shown in Table 7, FOA+DFRA improves both ASR and AvgSim on GPT-4o and Gemini-2.0-Flash. Specifically, on GPT-4o, ASR increases from 0.74 to 0.88, while AvgSim improves from 0.589 to 0.658. On Gemini-2.0-Flash, ASR increases from 0.63 to 0.71, with AvgSim

rising from 0.499 to 0.531. These results demonstrate that augmenting FOA-Attack with DFRA provides additional target guidance through fine-grained semantic feature alignment and inter-region relationship alignment, further improving both attack success rate and semantic similarity.

5 Conclusion

In this paper, we propose DFRA-Attack, a targeted transferable attack framework for MLLMs that combines semantic-aware alignment with relational-aware alignment. Specifically, saliency-guided shared masking provides consistent visible regions for aligning fine-grained semantic features, while relational-aware alignment further captures relationships among visible regions through feature relational matrices and attention interaction patterns. Extensive experiments on open-source, closed-source, and reasoning-oriented MLLMs demonstrate that DFRA-Attack consistently outperforms existing targeted transfer attacks in terms of attack success rate and semantic similarity. Ablation studies validate the contributions of shared masking, feature relational alignment, and attention interaction alignment, while mask ratio experiments reveal a trade-off between suppressing dominant visual cues and retaining sufficient visual information for alignment. Furthermore, experiments integrating DFRA into FOA-Attack demonstrate its potential to improve the performance of existing targeted transfer attacks.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (General Program) under Grant No. 62576267, and in part by the Major Science and Technology Innovation Project of Xianyang under Grant Nos. L2025-ZDKJ-ZDGG-ZYRGZN-002 and L2024-ZDKJ-ZDGG-GY-0010.

References

[1] J Achiam. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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.

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[4] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4, 2024.

[5] A. K B. Hamner and I. Goodfellow. Nips 2017: Defense against adversarial attack. `https://kaggle.com/competitions/nips-2017-defense-against-adversarial-attack`, 2017. Kaggle Competition.

[6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[7] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

[8] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[9] Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, volume 2024, pages 19199–19225, 2024.

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[13] 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.

[14] 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, 2026.

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

[16] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. *arXiv preprint arXiv:2212.10846*, 2022.

[17] 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, 2024.

[18] Dongchen Han, Xiaojun Jia, Yang Bai, Jindong Gu, Yang Liu, and Xiaochun Cao. Ot-attack: Enhancing adversarial transferability of vision-language models via optimal transport optimization. *arXiv preprint arXiv:2312.04403*, 2023.

[19] 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, 2026.

[20] Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. Survey of adversarial robustness in multimodal large language models. *arXiv preprint arXiv:2503.13962*, 2025.

[21] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308. IEEE, 2024.

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[23] 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.

[24] Ming Li, Keyu Chen, Ziqian Bi, Ming Liu, Xinyuan Song, Zekun Jiang, Tianyang Wang, Benji Peng, Qian Niu, Junyu Liu, et al. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991*, 2024.

[25] Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. *Advances in Neural Information Processing Systems*, 38:45018–45051, 2026.

[26] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1915–1929, 2024.

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

[28] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE transactions on neural networks and learning systems*, 2025.

[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] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296. IEEE, 2024.

[31] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024.

[32] 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.

[33] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213*, 1(3):4, 2023.

[34] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. Multilingual large language model: A survey of resources, taxonomy and frontiers. *arXiv preprint arXiv:2404.04925*, 2024.

[35] 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.

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

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

[38] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[39] Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6920–6928, 2024.

[40] Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *arXiv preprint arXiv:2402.14899*, 2024.

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

[42] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[43] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 3081–3089, 2022.

[44] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the association for computational linguistics: ACL 2024*, pages 12401–12430, 2024.

[45] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5005–5013, 2022.

[46] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19900–19909. IEEE, 2025.

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

[48] Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394, 2024.

A Qualitative Examples of Closed-Source MLLM Outputs

To provide a more intuitive qualitative comparison across different closed-source MLLMs, we visualize their outputs on targeted adversarial queries crafted by DFRA-Attack. As shown in Figure A.1, our method demonstrates remarkable cross-model consistency: despite the vast differences in model architectures and reasoning capabilities, the generated responses across the GPT, Gemini, and Claude families all accurately converge on the predefined target concept (e.g., "tennis"). This alignment suggests that the perturbations crafted by DFRA-Attack successfully capture fundamental visual-semantic features that are universally recognized by state-of-the-art MLLMs. To facilitate reproducibility and support future research, our source code and additional experimental assets are publicly available at `https://github.com/xiaorong-ui/DFRA-Attack`.

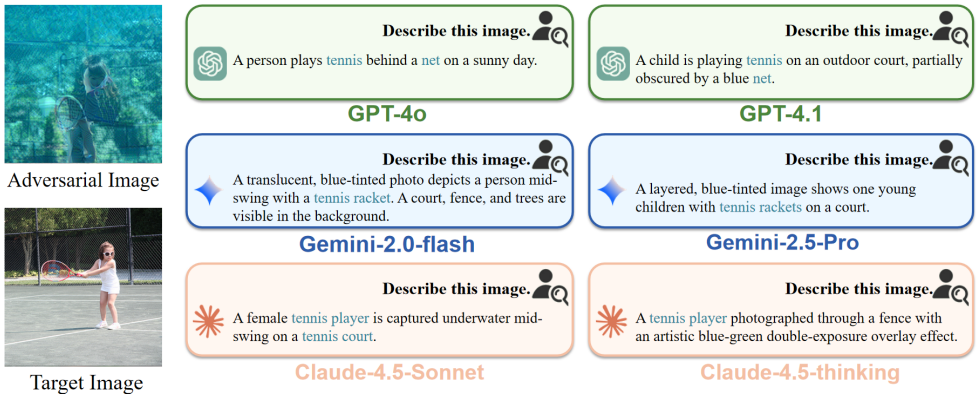


Figure A.1: Examples of closed-source MLLMs outputs on targeted adversarial queries crafted by DFRA-Attack.

B Additional Ablation Studies

B.1 Effect of Hyperparameters

To further evaluate the sensitivity of DFRA-Attack to different alignment objectives, we conduct ablation studies on the weighting hyperparameters λ_{attn}, λ_{rel}, and λ_{sem}. The results on Gemini-2.0-Flash are summarized in Tables B.1, B.2, and B.3. DFRA-Attack achieves the best performance when $\lambda_{\text{attn}} = 0.15$, $\lambda_{\text{rel}} = 0.5$, and $\lambda_{\text{sem}} = 1.2$, reaching 70.0% ASR and 0.532 AvgSim. Smaller weights provide weaker semantic or structural guidance, while overly large weights may overemphasize a single objective and hurt the balance among different alignment signals. These results indicate that DFRA-Attack benefits from a moderate combination of semantic, relational, and attention-based objectives.

Table B.1: Effect of λ_{attn}.

λ_{attn}	ASR	AvgSim
0.05	0.60	0.489
0.10	0.62	0.492
0.15	**0.70**	**0.532**
0.20	0.64	0.495

Table B.2: Effect of λ_{rel}.

λ_{rel}	ASR	AvgSim
0.4	0.61	0.498
0.5	**0.70**	**0.532**
0.6	0.68	0.526
0.7	0.60	0.500

Table B.3: Effect of λ_{sem}.

λ_{sem}	ASR	AvgSim
0.4	0.59	0.488
0.8	0.69	0.518
1.2	**0.70**	**0.532**
1.6	0.60	0.495

B.2 Ablation of Relational-aware Alignment

We ablate the two components of relational-aware alignment to evaluate their individual contributions. As shown in Table B.4, retaining only attention interaction alignment yields an ASR of 0.66 and an AvgSim of 0.523, whereas retaining only feature relational alignment achieves an ASR of

0.68 and an AvgSim of 0.521. Compared with either individual component, combining both components improves both metrics, achieving an ASR of 0.70 and an AvgSim of 0.532. These results indicate that feature relational alignment contributes more to ASR, while attention interaction alignment contributes more to AvgSim. The best performance is obtained by jointly aligning feature relationships and attention interactions.

Table B.4: Ablation study on the components of relational-aware alignment.

Variant	**ASR**	**AvgSim**
Attention Interaction Only	0.66	0.523
Feature Relation Only	0.68	0.521
DFRA-Attack (Ours)	**0.70**	**0.532**

B.3 Effect of Perturbation Budget

To study the impact of perturbation magnitude, we evaluate DFRA-Attack under different ℓ_∞ perturbation budgets, including $\epsilon = 2$, $\epsilon = 4$, $\epsilon = 8$, and $\epsilon = 16$. As shown in Table B.5, both ASR and AvgSim increase as the perturbation budget becomes larger. When ϵ is small, the perturbation space is highly constrained, making it difficult to inject sufficient target semantics. As ϵ increases from 2 to 16, ASR improves from 2.0% to 70.0%, and AvgSim increases from 0.020 to 0.532. These results show that a larger perturbation budget provides stronger target semantic manipulation, while the performance under smaller budgets reflects the trade-off between imperceptibility and attack effectiveness.

Table B.5: Effect of perturbation budget ϵ.

ϵ	**ASR**	**AvgSim**
2	0.02	0.020
4	0.06	0.098
8	0.27	0.303
16	**0.70**	**0.532**

C Robustness Analysis Across Larger Sample Sizes

Table C.6 evaluates DFRA-Attack using a larger set of 1,000 images. Our method consistently outperforms M-Attack across all metrics, confirming its robust transferability at scale. Notably, DFRA-Attack maintains high success rates on GPT-series models, achieving 0.85 ASR on GPT-4o and 0.90 ASR on GPT-4.1. The performance gains on Gemini-2.0-Flash and Claude Sonnet 4.5 further demonstrate its strong generalization capabilities across diverse closed-source architectures.

Table C.6: Full 1000-image results on closed-source MLLMs.

Method	**Gemini-2.0-Flash**		**GPT-4o**		**GPT-4.1**		**Claude Sonnet 4.5**	
	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim	ASR	AvgSim
M-Attack	0.50	0.428	0.64	0.523	0.68	0.546	0.04	0.069
DFRA-Attack (Ours)	0.72	0.543	0.85	0.625	0.90	0.636	0.16	0.192

D Limitations and Impact Statement

Limitations. Although DFRA-Attack achieves strong targeted transferability across diverse MLLMs, several limitations remain. First, our evaluation is conducted on a finite set of source-target image pairs and selected open-source, closed-source, and reasoning-oriented MLLMs. The results may vary when commercial MLLMs are updated or equipped with different safety filters.

Second, DFRA-Attack relies on surrogate visual encoders, so its performance may be affected by the choice and diversity of surrogate models. Third, our experiments mainly focus on digital ℓ_∞-bounded targeted attacks, while physical-world settings and other perceptual constraints are left for future work. Finally, the LLM-as-a-judge protocol enables scalable semantic evaluation, but may still introduce evaluator bias.

Impact Statement. This work is intended to support robustness evaluation of MLLMs under realistic black-box settings. By revealing that targeted adversarial examples can transfer to closed-source and reasoning-oriented MLLMs, our study can help model developers better understand potential failure modes and design stronger defenses. However, transferable targeted attacks are dual-use and could be misused to manipulate model outputs or bypass safety mechanisms. Therefore, DFRA-Attack should only be used for authorized security testing and research. We encourage future work to study detection, mitigation, and adversarial defense strategies together with attack evaluation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are supported by the methodology in Section 3 and the quantitative, qualitative, ablation, and additional analyses in Sections 4.24.5 and Appendix A.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of DFRA-Attack in Appendix D, including the finite evaluation set, dependence on surrogate encoders, the digital ℓ_∞ threat model, and possible bias from LLM-as-a-judge evaluation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper does not include formal theoretical results, theorems, or proofs. Therefore, this question is not applicable.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the proposed method, objective functions, datasets, baselines, evaluation metrics, perturbation budget, and key ablation settings in Sections 3, 4.1–4.5, and Appendix B. These details provide the information needed to reproduce the reported experimental results under the described evaluation protocol.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The source code for DFRA-Attack, including implementation details and evaluation scripts, is provided in the supplemental material and is also available at `https://github.com/xiaorong-ui/DFRA-Attack`.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We discuss this in detail for all our experiments in Sections 4.14.5 and Appendix A.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: We do not report error bars, confidence intervals, or statistical significance tests. Following prior targeted transfer attack protocols, we evaluate all methods on the same fixed set of 100 source-target image pairs and report aggregate metrics, including ASR, AvgSim, and KMR. We leave multi-seed evaluation and confidence interval analysis to future work.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report that adversarial image generation and open-source MLLM evaluations are conducted on NVIDIA RTX 4090 GPUs, while closed-source MLLMs are evaluated through official APIs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is conducted for authorized robustness evaluation of MLLMs and follows the NeurIPS Code of Ethics. We discuss the dual-use risks of transferable targeted attacks and responsible use in Appendix D.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Appendix D. The positive impact is that DFRA-Attack can help evaluate and improve the robustness of MLLMs under realistic black-box settings. The negative impact is that transferable targeted attacks may be misused to manipulate MLLM outputs or bypass safety mechanisms, so we emphasize authorized robustness testing and future defense research.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk pretrained model, image generator, or scraped dataset. We discuss the dual-use risks of transferable targeted attacks and responsible use in Appendix D.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the datasets, models, APIs, and baseline methods used in our experiments in Section 4.1, and state that these assets are used only for research evaluation under their corresponding licenses and terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The paper will release code for implementing and evaluating the proposed method. The code package will include basic documentation, environment setup instructions, configuration files, and scripts for reproducing the main experiments. We do not introduce new datasets or model weights. All datasets used in the experiments are existing public datasets and are cited in the paper.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research requiring IRB or equivalent review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Our study did not use an LLM to design the experimental design, integrate new experiments or do anything non-standard.