# ENHANCING TRANSFERABLE TARGETED ATTACKS ON MLLMS WITH PROGRESSIVE SEMANTIC CHAIN

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Transfer-based targeted attacks on commercial black-box multimodal large language models (MLLMs) remain challenging because attackers can access only the model's inputs and outputs. Existing methods primarily rely on visual semantic alignment in the embedding space to guide adversarial example generation, but the resulting perturbations often lack sufficiently explicit semantic cues, which limits their transferability. In contrast, textual descriptions can provide more explicit and fine-grained semantic cues for adversarial optimization. Motivated by this insight, we propose Progressive Semantic Chain Attack (PSC-Attack), which introduces a semantic chain to provide explicit semantic guidance for adversarial example generation. Specifically, PSC-Attack decomposes the target semantics into a sequence of concise semantic nodes that capture explicit semantic cues and progressively guide the optimization in the shared vision-language embedding space. Furthermore, we introduce a Minimum-Redundancy Semantic Chain to adaptively select more explicit and fine-grained target semantic nodes while suppressing redundant guidance, thereby enhancing attack transferability. Extensive experiments on closed-source commercial MLLMs demonstrate that PSC-Attack consistently outperforms existing targeted transferable attack methods.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) have demonstrated remarkable capabilities in visual understanding, language generation, and cross-modal reasoning (Yin et al., 2024). Benefiting from advances in vision-language pretraining and alignment, recent open-source models such as BLIP-2, LLaVA, Qwen-VL, and Gemma have achieved strong performance on a wide range of multimodal tasks, including image captioning, visual question answering, and visual reasoning (Li et al., 2023; Liu et al., 2023; Bai et al., 2023; Team et al., 2024; Fu et al., 2025; Li et al., 2024). At the same time, powerful commercial MLLMs, such as GPT-4o, Claude, and Gemini, have further expanded the practical influence of these systems through black-box deployment interfaces. As MLLMs become increasingly capable and widely used, their robustness to adversarial attack has emerged as an important concern (Cui et al., 2024a; Ye & Tang, 2025; Li et al., 2025a; Jiang et al., 2025).

Adversarial attacks on MLLMs exploit the alignment between visual and textual inputs to trigger misleading or harmful outputs (Liu et al., 2025; Cui et al., 2024b). In realistic settings, powerful commercial MLLMs are usually accessed as black boxes, so attackers often rely on transfer-based attacks crafted on open-source surrogate models (Zhao et al., 2023; Dong et al., 2023). Within the black box setting, transferable adversarial attacks aim to generate perturbations that remain effective on unseen closed-source targets, and targeted transferable attacks, which try to make a source image be interpreted as a specific target concept, are particularly challenging (Zhao et al., 2023; Guo et al., 2025; Zhang et al., 2025).

Existing targeted transferable attack methods mainly rely on visual semantic alignment in the embedding space to guide adversarial example generation (Zhao et al., 2023; Zhang et al., 2025). However, as pointed out by M-Attack (Li et al., 2025b), a critical drawback of conventional adversarial strategies is that the generated perturbations often lack clear semantic information, which limits their ability to effectively influence the model's decision-making process and thus hinders attack transferability. This limitation suggests that improving attack transferability requires a form of target guidance that can encode target semantics more explicitly during optimization. In this regard, textual



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

descriptions offer a natural advantage, as they can convey target-relevant semantics more explicitly and in a finer-grained manner, thereby providing a more effective source of semantic guidance for adversarial optimization.

Motivated by this insight, we propose a Progressive Semantic Chain Attack (PSC-Attack), which introduces a semantic chain to provide explicit, progressive target guidance for adversarial example generation. Specifically, PSC-Attack decomposes the target semantics into a sequence of semantic nodes, including object identity, attributes, relations, and scene context, through a predefined semantic chain. Each node captures a distinct semantic cue and progressively guides the optimization process in the shared vision-language embedding space. Experimental comparisons between holistic textual guidance and semantic-chain-based textual guidance show that the latter leads to better targeted attack performance, suggesting that explicit and fine-grained semantic cues provide more effective guidance for adversarial optimization. Nevertheless, such a predefined semantic chain still relies on a fixed and coarse decomposition of target semantics, which may group distinct semantic cues together and fail to explicitly capture fine-grained target semantics. To address this limitation, we introduce a Minimum-Redundancy Semantic Chain (MRSC), which adaptively selects informative and semantically distinct semantic units while suppressing redundant guidance, and progressively composes them into finer-grained target-semantic nodes, thereby enhancing attack transferability.

Our main contributions are summarized as follows:

- We propose a Progressive Semantic Chain Attack (PSC-Attack), which leverages a semantic chain to provide explicit and progressive target-semantic guidance for a transferable targeted adversarial attack.
- We introduce a Minimum-Redundancy Semantic Chain to adaptively select explicit and fine-grained semantic nodes, reducing redundancy in semantic guidance and improving attack transferability.
- Extensive experiments on both closed-source and open-source commercial MLLMs demonstrate the effectiveness of the proposed method over existing targeted transferable attack methods.

## 2 Related work

**Multimodal large language models.** Multimodal Large Language Models (MLLMs) extend LLMs with visual encoders and cross-modal alignment for unified image-text modeling (Zhang et al., 2024; Caffagni et al., 2024; Yin et al., 2024; Singh et al., 2026). Representative systems have evolved from BLIP/BLIP-2, which leverage Q-Former for vision-language integration, to LLaVA with instruction-tuned projections for strong multimodal capability (Zhang et al., 2024; Caffagni et al., 2024; Wu et al., 2026). Their performance is evaluated on benchmarks such as MME and SEED-Bench across perception, cognition, and reasoning (Fu et al., 2025; Li et al., 2024; Wang et al., 2026a). Recent Qwen-VL variants further improve position-aware cross-modal alignment (Zhang et al., 2024; An et al., 2025). Meanwhile, closed-source models (e.g., GPT-4V, Gemini, GPT-4o, Claude-4.5, Gemini-2.0, Gemini-2.5-Pro) push commercial frontiers (Jin et al., 2025; Sun et al., 2026; Li et al., 2026; Kisel et al., 2026), while open-source models (e.g., Qwen2.5-VL, LLaVA-1.5/1.6, Gemma-3) continue advancing multimodal reasoning (Zhou et al., 2026; Wang et al., 2026b; Zhang et al., 2026). This rapid expansion also increases the dependence of downstream applications on reliable cross-modal grounding. Since MLLMs are increasingly deployed in safety-critical and user-facing scenarios, small perturbations or misaligned visual cues can propagate into erroneous textual outputs, making robustness and transferability analysis essential for understanding their practical security boundaries. However, MLLMs remain vulnerable to hallucination, catastrophic forgetting, and adversarial risks in domains such as medical analysis and fake-news detection (Wu et al., 2026; Yin et al., 2024; Ye & Tang, 2025; Yang et al., 2026; Zhao et al., 2026).

**Adversarial attacks on MLLMs.** The security of MLLMs has drawn increasing attention due to cross-modal vulnerabilities (Liu et al., 2024; Jiang et al., 2025; Mei, 2026). Among threat models, transferable targeted attacks on black-box MLLMs are particularly concerning, as they pose practical risks to closed-source systems without requiring parameter access. Existing methods craft targeted perturbations via feature matching or generative semantic manipulation. Feature-based ap-

proaches such as AttackVLM (Zhao et al., 2023) align image embeddings with target texts, while SSA-CWA (Dong et al., 2023) improves transferability through optimized feature matching. Generative methods, including AdvDiffVLM (Guo et al., 2025) and TRAP (Kang et al., 2025), leverage diffusion models to inject stealthy target semantics, and AnyAttack (Zhang et al., 2025) explores a self-supervised targeted framework. More recent work such as M-Attack (Li et al., 2025b) further improves targeted transferability by enriching target-related semantic details and evaluating keyword-level coverage. Nevertheless, such guidance still primarily emphasizes whether target concepts are injected, rather than explicitly modeling their compositional dependencies or progressive semantic structure. This limitation is further amplified when perturbations are optimized mainly toward global representations, which may induce semantic entanglement and limit transferability on advanced closed-source systems such as GPT-4o and Gemini (Ye & Tang, 2025).

**Semantic chain decomposition.** Semantic chain decomposition is closely related to Chain-of-Thought (CoT) reasoning, which decomposes complex tasks into explicit sequential steps to improve generation (Wei et al., 2022). In MLLMs, structured variants such as Compositional CoT and Visual CoT mitigate semantic entanglement by decomposing holistic visual queries into modular semantic operations, enabling more controllable, interpretable, and faithful cross-modal alignment (Mitra et al., 2024). Beyond linear chains, Tree of Thoughts and Graph of Thoughts organize reasoning as branching units or dependency graphs, while Program of Thoughts and Skeleton-of-Thought introduce high-level semantic structures before detailed computation (Yao et al., 2023; Besta et al., 2024; Chen et al., 2023; Ning et al., 2024). These works motivate our use of semantic chains to provide explicit, fine-grained target guidance for adversarial optimization.

# 3 METHOD

Previous work has shown that conventional transferable adversarial attacks often lack explicit target-semantic guidance during perturbation generation, which limits attack transferability. To address this issue, PSC-Attack introduces a progressive semantic-chain-based guidance in the shared vision-language embedding space. Specifically, in Sec. 3.2, we first present the basic PSC-Attack framework, where the adversarial example is progressively optimized toward a sequence of explicit target-semantic nodes while maintaining auxiliary image-level alignment with the target image. Then, in Sec. 3.3, we further improve the chain construction process by replacing the fixed predefined semantic chain with a Minimum-Redundancy Semantic Chain (MRSC), which adaptively selects informative and semantically distinct semantic units and progressively composes them into finer-grained target-guided nodes. In this way, MRSC preserves the progressive optimization framework of PSC-Attack while providing more explicit and less redundant semantic guidance.

## 3.1 PRELIMINARY

Targeted transferable attacks aim to craft a visually imperceptible perturbation on a natural image using white-box surrogate models, such that the resulting adversarial image is interpreted by unseen models as a predefined target content. The perturbation is typically constrained by a small $\ell_\infty$ budget to preserve visual similarity to the original image. To improve transferability across model architectures, prior methods optimize the perturbation over an ensemble of surrogate vision encoders rather than a single model. Following this setting, we use three CLIP image encoders as surrogates, including ViT-B/16, ViT-B/32, and ViT-g/14. The attack objective is defined in the shared embedding space, where the adversarial image is encouraged to approach the target representation.

## 3.2 PROGRESSIVE SEMANTIC CHAIN ATTACK

Given a target image $x_{\mathrm{tar}}$, the proposed PSC-Attack utilizes a semantic chain $\mathcal{C}(x_{\mathrm{tar}})$ defined as:

$$\mathcal{C}(x_{\mathrm{tar}}) = \{c_1, c_2, \ldots, c_k\}, \tag{1}$$

where each node $c_k$ represents a semantic component of the target description. As an initial design, we construct the semantic chain by decomposing a composite target description generated by an image-to-text model into four predefined semantic components, namely object identity, visual attributes, relations, and scene context, denoted as $c_{\mathrm{id}}$, $c_{\mathrm{attr}}$, $c_{\mathrm{rel}}$, and $c_{\mathrm{ctx}}$, respectively.

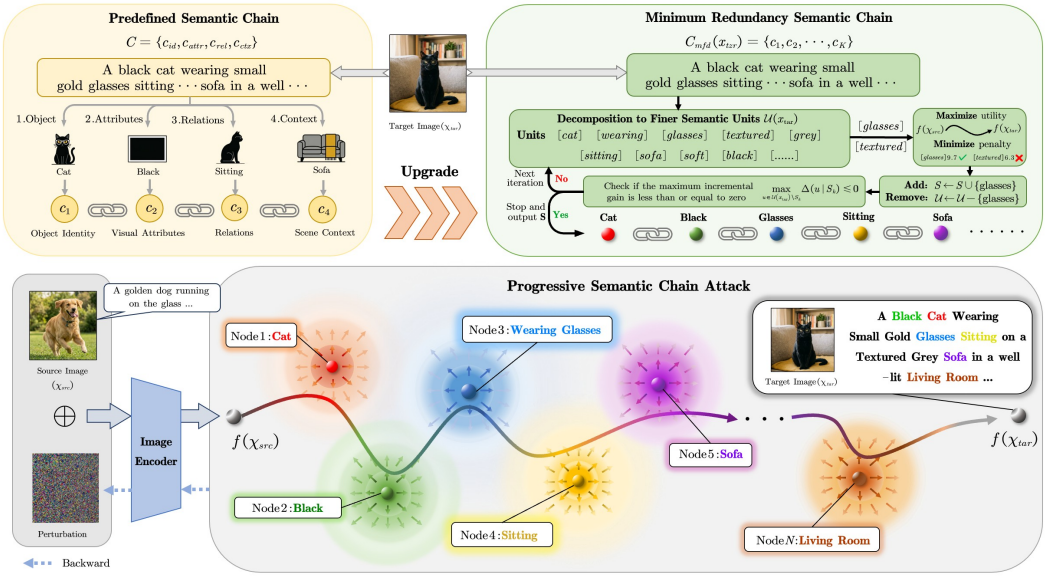


Figure 1: Overview of the proposed Progressive Semantic Chain Attack (PSC-Attack). Given a target image, PSC-Attack first decomposes its caption into coarse semantic components and then constructs a minimum-redundancy semantic chain by selecting explicit, informative, and non-redundant semantic units. During optimization, the adversarial example is progressively guided along the semantic chain in the shared vision-language embedding space, moving from the source concept toward the target concept while preserving image-level alignment with the target image.

Based on the constructed semantic chain, PSC-Attack progressively guides perturbation optimization along the semantic nodes over the $T$ attack iterations, while retaining direct alignment between the adversarial image embedding and the target image embedding. Specifically, at the $t$-th attack iteration, the perturbation is encouraged to approach different semantic nodes with different strengths, such that earlier iterations place more emphasis on earlier nodes in the chain, whereas later iterations increasingly focus on later ones. Let $\phi_i(\cdot)$ denote the image encoder of the $i$-th surrogate vision-language model, where $i = 1, \ldots, N$, and let $\psi(\cdot)$ denote the text encoder. At the $t$-th attack iteration, for the adversarial image $x_{\text{adv}}^{(t)}$ and the $k$-th semantic node $c_k$, their corresponding embeddings are

$$z_{i,t}^{\text{img}} = \phi_i(x_{\text{adv}}^{(t)}), \qquad z_k^{\text{text}} = \psi(c_k). \tag{2}$$

We then normalize these embeddings as

$$\bar{z}_{i,t}^{\text{img}} = \frac{z_{i,t}^{\text{img}}}{\|z_{i,t}^{\text{img}}\|_2}, \qquad \bar{z}_k^{\text{text}} = \frac{z_k^{\text{text}}}{\|z_k^{\text{text}}\|_2}. \tag{3}$$

Based on these embeddings, we define the semantic guidance loss at iteration $t$ as

$$\mathcal{L}_{\text{chain}}^{(t)} = \sum_{i=1}^{N} \sum_{k=1}^{K} w_i \lambda_k^{(t)} \mathcal{D}\left(\bar{z}_{i,t}^{\text{img}}, \bar{z}_k^{\text{text}}\right), \tag{4}$$

where $w_i$ is the weight of the $i$-th surrogate model, $\lambda_k^{(t)}$ is the weight of the $k$-th semantic node at iteration $t$, and

$$\mathcal{D}(a, b) = 1 - a^\top b \tag{5}$$

denotes cosine distance. The time-varying weights $\lambda_k^{(t)}$ progressively shift the optimization focus across semantic nodes over attack iterations:

$$\lambda_k^{(t)} = \frac{\exp\left(-\tau \left| \frac{t}{T} - \frac{k-1}{K-1} \right|\right)}{\sum_{j=1}^{K} \exp\left(-\tau \left| \frac{t}{T} - \frac{j-1}{K-1} \right|\right)}, \tag{6}$$

Table 1: Comparison of target semantic guidance strategies for targeted transferable attacks on GPT-4o.

| Method | ASR↑ | AvgSim↑ |
|---|---|---|
| Baseline | 0.53 | 0.467 |
| Holistic textual guidance | 0.61 | 0.503 |
| Predefined semantic chain | **0.89** | **0.648** |

where $T$ is the total number of attack iterations and $\tau$ controls the sharpness of the transition. In this way, different semantic nodes dominate the optimization at different stages, yielding a progressive semantic guidance process.

In addition to semantic guidance, we retain direct image-level target alignment to preserve consistency with the target image itself. Let

$$z_{i,\mathrm{tar}}^{\mathrm{img}} = \phi_i(x_{\mathrm{tar}}), \qquad \bar{z}_{i,\mathrm{tar}}^{\mathrm{img}} = \frac{z_{i,\mathrm{tar}}^{\mathrm{img}}}{\|z_{i,\mathrm{tar}}^{\mathrm{img}}\|_2}. \tag{7}$$

The corresponding image-level loss at iteration $t$ is

$$\mathcal{L}_{\mathrm{img}}^{(t)} = \sum_{i=1}^{N} w_i \mathcal{D}\Big(\bar{z}_{i,t}^{\mathrm{img}}, \bar{z}_{i,\mathrm{tar}}^{\mathrm{img}}\Big). \tag{8}$$

The final optimization objective at iteration $t$ combines semantic-chain guidance and image-level alignment:

$$\mathcal{L}_{\mathrm{total}}^{(t)} = \alpha \mathcal{L}_{\mathrm{img}}^{(t)} + (1-\alpha)\mathcal{L}_{\mathrm{chain}}^{(t)}, \tag{9}$$

where $\alpha$ is the trade-off coefficient between image-level target alignment and semantic-chain guidance.

Table 1 illustrates the effect of different strategies of target semantic guidance on target transfer attack performance. Compared with the baseline, incorporating the holistic target textual description improves both ASR and AvgSim, while using the predefined semantic chain yields the best performance, demonstrating the advantage of concise and structured semantic guidance. The results show that progressive guidance with concise semantic nodes is more effective than directly optimizing toward a holistic entangled target description, as it provides more explicit and fine-grained semantics for each attack stage and leads to better target alignment and transferability.

### 3.3 Minimum-Redundancy semantic chain

Although the predefined semantic chain already demonstrates the benefit of semantic decomposition, it still relies on a coarse and fixed partition of target semantics into a small set of predefined categories. Such a design may merge distinct semantic cues into the same component and fail to explicitly capture fine-grained target semantics, thereby limiting the completeness and independence of the resulting guidance chain. To address this limitation, we introduce a Minimum-Redundancy Semantic Chain (MRSC), which replaces the fixed predefined decomposition with a target-adaptive chain construction strategy. Instead of directly predefined coarse semantic nodes, MRSC first decomposes the target description into finer candidate semantic units and then progressively selects informative and semantically distinct units to compose target-guided semantic nodes. In this way, the progressive optimization framework of PSC-Attack remains unchanged, while the semantic chain itself becomes more explicit, fine-grained, and less redundant.

Specifically, we first obtain a textual description of the target image and decompose it into a set of candidate semantic units

$$\mathcal{U}(x_{\mathrm{tar}}) = \{u_1, u_2, \ldots, u_L\}, \tag{10}$$

where each unit $u_\ell$ corresponds to a target-grounded semantic component, such as an object category, an attribute phrase, a relation phrase, a scene descriptor, or a fine-grained modifier.

Starting from the core object unit, we progressively build the semantic chain by selecting additional units from $\mathcal{U}(x_{\text{tar}})$. Let $S_k$ denote the set of semantic units selected up to the $k$-th step, with the initialization

$$S_1 = \{u_{\text{obj}}\}. \tag{11}$$

At each step, the selected units are composed into a target-grounded textual node

$$c_k = \text{Compose}(S_k), \tag{12}$$

where $\text{Compose}(\cdot)$ converts the selected semantic units into a textual description. In this way, the nodes in the chain are not defined by fixed attribute templates, but are constructed progressively from the semantic content of the target itself.

To determine whether a candidate unit $u$ should be added to the current set $S_k$, we define its incremental gain as

$$\Delta(u \mid S_k) = \text{sim}(x_{\text{tar}}, \text{Compose}(S_k \cup \{u\})) - \text{sim}(x_{\text{tar}}, \text{Compose}(S_k)) - \beta\, \text{Red}(u, S_k), \tag{13}$$

where $\text{sim}(\cdot,\cdot)$ measures image-text similarity, $\beta$ is redundancy penalty coefficient and $\text{Red}(u, S_k)$ penalizes semantic redundancy. Accordingly, a candidate unit is preferred if it introduces additional target-relevant semantics while remaining sufficiently different from the units already selected.

We instantiate

$$\text{sim}(x, c) = \cos(\phi_{\text{ref}}(x), \psi(c))\,, \tag{14}$$

and

$$\text{Red}(u, S_k) = \max_{v \in S_k} \cos(\psi(u), \psi(v))\,. \tag{15}$$

The next unit is selected greedily by

$$u^\star = \arg \max_{u \in \mathcal{U}(x_{\text{tar}}) \setminus S_k} \Delta(u \mid S_k), \tag{16}$$

and the selected set is updated as

$$S_{k+1} = S_k \cup \{u^\star\}. \tag{17}$$

The process terminates when no remaining unit yields a positive gain:

$$\max_{u \in \mathcal{U}(x_{\text{tar}}) \setminus S_k} \Delta(u \mid S_k) \le 0. \tag{18}$$

The resulting semantic nodes form the Minimum-Redundancy Semantic Chain

$$\mathcal{C}_{\text{mrsc}}(x_{\text{tar}}) = \{c_1, c_2, \ldots, c_K\}, \tag{19}$$

where $K$ is determined adaptively by the target image. Compared with the predefined semantic chain, this design constructs a more refined and complete guiding chain with more independent and comprehensive fine-grained semantics. As a result, it enables the perturbation to capture a broader range of fine-grained target features while reducing redundant semantic guidance, thereby further improving the transferability of targeted attacks.

# 4 Experiments

## 4.1 Experimental setup

**Dataset and implementation details.** Following prior work (Li et al., 2025b), we evaluate on 100 randomly sampled images from the *NIPS 2017 Adversarial Attacks and Defenses Competition* dataset (Kurakin et al., 2017). To provide stronger statistical evidence, we additionally report results on the 1000 images (see Appendix). Target images are sampled from MSCOCO validation (Lin et al., 2014) under a non-identical source-target constraint. We use ViT-B/16, ViT-B/32, and ViT-g-14-laion2B-s12B-b42K as surrogate models, optimizing each adversarial example for 300 iterations with $\epsilon = 16/255$ under the $\ell_\infty$ norm. Transferability is evaluated on open-source MLLMs (Qwen2.5-VL-3B/7B, LLaVA-1.5/1.6-7B, Gemma-3-4B/12B), closed-source MLLMs (Claude-4.5, Gemini-2.0, GPT-4o, GPT-4.1), and reasoning-oriented MLLMs (Gemini-2.5-Pro, Claude-4.5-thinking). ClipCap (Mokady et al., 2021) is used to generate target descriptions for semantic-chain construction.

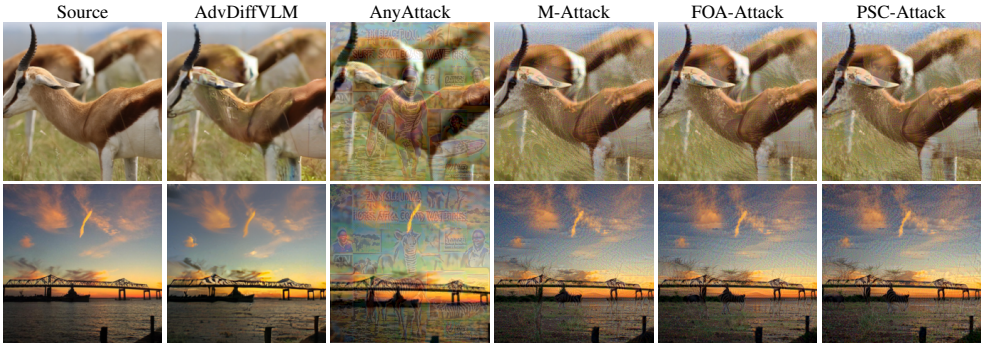


Figure 2: Qualitative comparison of adversarial examples and perturbations.

Table 2: Quantitative comparison of targeted attack performance on closed-source and closed-source reasoning MLLMs. The best results are highlighted in bold.

| Method | Surrogate | Closed-source MLLMs | | | | | | | | Closed-source reasoning MLLMs | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gemini-2.0 | | GPT-4o | | GPT-4.1 | | Claude Sonnet 4.5 | | Gemini-2.5-Pro | | Claude-4.5-thinking | |
| | | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim |
| AdvDiffVLM | Ensemble | 0.08 | 0.104 | 0.08 | 0.145 | 0.09 | 0.149 | 0.01 | 0.048 | 0.05 | 0.087 | 0.03 | 0.054 |
| SSA-CWA | Ensemble | 0.14 | 0.181 | 0.11 | 0.177 | 0.12 | 0.211 | 0.01 | 0.076 | 0.08 | 0.112 | 0.02 | 0.043 |
| AnyAttack | Ensemble | 0.29 | 0.223 | 0.21 | 0.321 | 0.28 | 0.298 | 0.02 | 0.066 | 0.09 | 0.209 | 0.02 | 0.032 |
| M-Attack | Ensemble | 0.40 | 0.390 | 0.64 | 0.517 | 0.69 | 0.540 | 0.03 | 0.072 | 0.46 | 0.408 | 0.03 | 0.077 |
| FOA-Attack | Ensemble | 0.60 | 0.483 | 0.72 | 0.589 | 0.69 | 0.590 | 0.03 | 0.088 | 0.48 | 0.444 | 0.04 | 0.082 |
| PSC-Attack | Ensemble | **0.68** | **0.535** | **0.89** | **0.648** | **0.81** | **0.612** | **0.10** | **0.122** | **0.65** | **0.523** | **0.08** | **0.119** |

**Competitive methods.** We compare PSC-Attack with four representative targeted transferable attack baselines for MLLMs: SSA-CWA, AdvDiffVLM, AnyAttack, M-Attack, and FOA-Attack. For fairness, we use official implementations and follow the original attack budgets, optimization settings, and preprocessing pipelines whenever possible. We report both quantitative and qualitative results.

**Evaluation metrics.** Following caption-level semantic matching protocols (Mokady et al., 2021), we compare descriptions generated from adversarial and target images using GPTScore (Fu et al., 2024). An attack is considered successful if the score exceeds 0.5; *Attack Success Rate* (ASR) denotes the success ratio, while *Average Similarity* (AvgSim) is the mean GPTScore over the evaluation set. As a supplementary metric, we report *Keyword Matching Rate* (KMR) following Li et al. (2025b): GPT-4o identifies which manually annotated target keywords are expressed in the generated description, and $\text{KMR}_\alpha$, $\text{KMR}_\beta$, and $\text{KMR}_\gamma$ correspond to coverage thresholds of 0.25, 0.5, and 1.0, respectively.

### 4.2 Quantitative Comparisons

**Closed-source and reasoning MLLMs.** Table 2 reports the targeted transfer results on closed-source and reasoning MLLMs. PSC-Attack consistently achieves the best performance across all evaluated models, while the degree of vulnerability differs substantially across targets. For closed-source MLLMs, Claude-4.5 is the most resistant model, yet PSC-Attack still obtains the highest ASR of 0.10 and AvgSim of 0.122. On Gemini-2.0, GPT-4o, and GPT-4.1, PSC-Attack reaches ASRs of 0.68, 0.89, and 0.81, with corresponding AvgSim values of 0.535, 0.648, and 0.612, respectively. Compared with the strongest baseline, FOA-Attack, these results correspond to ASR gains of 0.08, 0.17, and 0.12, and AvgSim gains of 0.052, 0.059, and 0.022. For reasoning MLLMs, PSC-Attack achieves an ASR of 0.65 and an AvgSim of 0.523 on Gemini-2.5-Pro, improving over FOA-Attack by 0.17 in ASR and 0.079 in AvgSim. On the more robust Claude-4.5-thinking, PSC-Attack remains the best-performing method, with an ASR of 0.08 and an AvgSim of 0.119. These results indicate that PSC-Attack provides stronger targeted transferability and target-semantic alignment across commercial and reasoning MLLMs.

Table 3: Quantitative comparison of targeted attack performance on open-source MLLMs. The best results are highlighted in bold.

| Method | Surrogate | Qwen2.5-VL-3B | | Qwen2.5-VL-7B | | LLaVA-1.5-7B | | LLaVA-1.6-7B | | Gemma-3-4B | | Gemma-3-12B | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim | ASR | AvgSim |
| AdvDiffVLM | Ensemble | 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 | Ensemble | 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 | Ensemble | 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 | Ensemble | 0.41 | 0.359 | 0.51 | 0.453 | 0.65 | 0.571 | 0.69 | 0.567 | 0.20 | 0.224 | 0.21 | 0.236 |
| FOA-Attack | Ensemble | 0.44 | 0.419 | 0.57 | 0.484 | 0.70 | 0.571 | 0.71 | 0.609 | 0.24 | 0.245 | 0.23 | 0.263 |
| PSC-Attack | Ensemble | **0.63** | **0.432** | **0.66** | **0.532** | **0.80** | **0.644** | **0.76** | **0.638** | **0.25** | **0.268** | **0.26** | **0.272** |

Table 4: Ablation study on the trade-off coefficient $\alpha$.

| $\alpha$ | ASR↑ | AvgSim↑ |
|---|---|---|
| 0.1 | 0.70 | 0.505 |
| **0.3** | **0.89** | **0.648** |
| 0.5 | 0.80 | 0.562 |
| 0.7 | 0.75 | 0.532 |
| 0.9 | 0.72 | 0.517 |

Table 5: Ablation study on the redundancy penalty coefficient $\beta$.

| $\beta$ | ASR↑ | AvgSim↑ |
|---|---|---|
| 0 | 0.78 | 0.555 |
| 0.1 | 0.83 | 0.573 |
| **0.3** | **0.89** | **0.648** |
| 0.5 | 0.80 | 0.562 |
| 1.0 | 0.74 | 0.532 |

**Open-source MLLMs.** Table 3 presents the targeted transfer results on open-source MLLMs, where PSC-Attack consistently surpasses prior methods, including AdvDiffVLM (Guo et al., 2025), SSA-CWA (Dong et al., 2023), AnyAttack (Zhang et al., 2025), M-Attack (Li et al., 2025b), and FOA-Attack Jia et al. (2025). Its advantages are most evident on the Qwen and LLaVA families and remain consistent on the more challenging Gemma3 models. Specifically, PSC-Attack obtains ASRs of 0.63 and 0.66 on Qwen-2.5-VL-3B/7B, with AvgSim values of 0.432 and 0.532, respectively; ASRs of 0.80 and 0.76 on LLaVA-1.5/1.6-7B, with AvgSim values of 0.644 and 0.638; and ASRs of 0.25 and 0.26 on Gemma3-4B/12B, with AvgSim values of 0.268 and 0.272. Relative to FOA-Attack, PSC-Attack improves ASR by 0.19, 0.09, 0.10, 0.05, 0.01, and 0.03, and AvgSim by 0.013, 0.048, 0.073, 0.029, 0.023, and 0.009 across the six models. These results further confirm that PSC-Attack enhances both targeted transferability and target-semantic fidelity on open-source MLLMs.

### 4.3 Qualitative comparisons

Figure 2 presents representative adversarial examples and their corresponding perturbation maps for different attack methods. AdvDiffVLM introduces more visible color distortions over the image, while AnyAttack yields more concentrated perturbations around object contours and other salient regions. In contrast, M-Attack and PSC-Attack preserve the visual appearance of the clean images more effectively under the same perturbation budget. Their perturbation maps show lower overall intensity and weaker visible patterns, making the perturbations less perceptible. Compared with M-Attack, PSC-Attack further produces smoother perturbations that remain visually closer to the original images while still achieving effective targeted transfer.

### 4.4 Ablation study

**Effect of trade-off coefficient $\alpha$.** We analyze the effect of $\alpha$ in PSC-Attack, where $\alpha$ balances image-level alignment and semantic-chain guidance in the overall objective (Eq. 9). Table 4 reports the results for different values of $\alpha$ with all other settings fixed. The best performance is achieved at $\alpha = 0.3$, yielding an ASR of 0.89 and an AvgSim of 0.648. As $\alpha$ increases beyond 0.3, both ASR and AvgSim gradually decline, indicating that reducing the influence of semantic-chain guidance weakens the attack by limiting fine-grained target-semantic alignment during optimization. Conversely, when $\alpha$ is too small (e.g., $\alpha = 0.1$), the contribution of image-level alignment becomes insufficient, which also leads to a slight performance drop. These results suggest that neither image-level alignment nor semantic-chain guidance alone is sufficient, and that a proper balance between them is important for achieving strong transfer attack performance.

### 4.5 Additional analysis

**Visualization of semantic guidance.** We visualize the attention maps corresponding to several representative semantic nodes from different stages of the semantic chain during the optimization process of PSC-Attack. As shown in Fig. 3, different semantic-node descriptions attend to different target-relevant regions in the adversarial image. Specifically, compared with the entangled attention map produced by the holistic textual description, "Two white horses, adorned with red feather plumes, wear bridles and harnesses and stand side by side on a city street.", the decomposed semantic nodes, such as "red feather plumes" and "two white horses with bridles", more explicitly highlight the approximate regions of their corresponding objects. This observation suggests that decomposing the target description into multiple semantic units provides more explicit and fine-grained semantic guidance for perturbation optimization than directly using a single entangled target description. In addition, the visual responses of different semantic nodes appear complementary, suggesting that these decomposed semantic units focus on different aspects of the target semantics rather than repeatedly emphasizing the same information.



Figure 3: Visualization of semantic guidance in PSC-Attack. From left to right: the target image, the source image, the attention map for the semantic node "red feather plumes", the attention map for the semantic node "two white horses with bridles", and the attention map for the holistic target description. Compared with the holistic description, decomposed semantic nodes attend to more explicit and localized target-relevant regions.

**Effect of perturbation budget.** The main experiments adopt an $\ell_\infty$ perturbation budget of $\epsilon = 16/255$. To justify this choice, we further evaluate PSC-Attack under different perturbation budgets. Following Li et al. (2025b), we vary $\epsilon$ over $\{4/255, 8/255, 16/255\}$ while keeping all other settings fixed. As shown in Table 6, increasing $\epsilon$ consistently improves both ASR and AvgSim. Specifically, ASR increases from 0.20 at $\epsilon = 4/255$ to 0.70 at $8/255$ and further to 0.89 at $16/255$, while AvgSim improves from 0.133 to 0.350 and 0.648, respectively. These results indicate that a small perturbation budget restricts the optimization space for target-semantic transfer, whereas a larger budget allows the adversarial example to better approach the target semantics. Based on these results, we adopt $\epsilon = 16/255$ as the default perturbation budget in the main experiments.

Table 6: Effect of the perturbation budget $\epsilon$.

| $\epsilon$ | $4/255$ | $8/255$ | $16/255$ |
|---|---|---|---|
| ASR | 0.20 | 0.70 | **0.89** |
| AvgSim | 0.133 | 0.350 | **0.648** |

## 5 Conclusion

In this paper, we proposed PSC-Attack for targeted transferable attacks against black-box MLLMs. We found that progressive semantic-chain guidance provides more explicit and fine-grained target semantics than holistic target guidance, leading to stronger targeted transferability. Building on this observation, we further introduced a Minimum-Redundancy Semantic Chain to adaptively select informative and semantically distinct semantic units while reducing redundant guidance. Extensive experiments across open-source, closed-source, and reasoning-oriented MLLMs demonstrate that PSC-Attack consistently improves both attack success rate and semantic similarity over existing methods. These results show the effectiveness of explicit, progressive, and low-redundancy semantic guidance for enhancing targeted transferability in black-box MLLMs.

## AI USE STATEMENT

We used generative AI tools solely to assist with language editing, grammar checking, and consistency in wording. We did not use generative AI tools for research ideation, method development, experimental design, or scientific analysis. The authors reviewed all AI-assisted suggestions and take full responsibility for the accuracy, integrity, and final content of this work.

## REFERENCES

Jisu An, Junseok Lee, Jeoungeun Lee, and Yongseok Son. Towards LLM-centric multimodal fusion: A survey on integration strategies and techniques, 2025.

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

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 17682–17690, 2024.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13590–13618, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2023.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 958–979. IEEE, 2024a.

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

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?, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, Mexico City, Mexico, jun 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.365.

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. doi: 10.1109/TIFS.2024.3518072.

Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source MLLMs via feature optimal alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 57993–58028. Curran Associates, Inc., 2025. doi: 10.52202/085713-1944.

Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. Survey of adversarial robustness in multimodal large language models, 2025.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and Lizhuang Ma. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025. doi: 10.1007/s44267-025-00099-6.

Hangoo Kang, Jehyeok Yeon, and Gagandeep Singh. TRAP: Targeted redirecting of agentic preferences. In *Advances in Neural Information Processing Systems*, volume 38, pp. 169565–169590. Curran Associates, Inc., 2025. doi: 10.52202/085713-5645.

Nikita Kisel, Illia Volkov, Klara Janouskova, and Jiri Matas. Multimodal large language models as image classifiers, 2026.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. NIPS 2017: Defense against adversarial attack. Kaggle Competition, 2017.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024. doi: 10.1109/CVPR52733.2024.01263.

Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, Jianguo Cao, James M. Rehg, Heng Ji, Ismini Lourentzou, and Xu Cao. Toward cognitive supersensing in multimodal large language model, 2026.

Jing Li, Jingyuan Li, Guo Yang, Lie Yang, Haozhuang Chi, and Lichao Yang. Applications of large language models and multimodal large models in autonomous driving: A comprehensive review. *Drones*, 9(4):238, 2025a. doi: 10.3390/drones9040238.

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

Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of GPT-4.5/4o/o1. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45018–45051. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1500.

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*, pp. 740–755. Springer International Publishing, 2014.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19525–19545, 2025. doi: 10.1109/TNNLS.2025.3592935.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and texts, 2024.

Liangyu Mei. Research on multimodal large language models for visual question answering: Advances and challenges. *Applied and Computational Engineering*, 223:47–53, 2026.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431, 2024.

Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Xuefei Ning, Zinan Lin, Zixuan Zhou, Zifu Wang, Huazhong Yang, and Yu Wang. Skeleton-of-Thought: Prompting LLMs for Efficient Parallel Generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Gurpreet Singh, Lamia Qamar, Nicholas Valentino Volta, Amruta Velamuri, and Aya Khanyile. Vision–language foundation models and multimodal large language models: A comprehensive survey of architectures, benchmarks, and open challenges. *Preprints*, February 2026. doi: 10.20944/preprints202602.0467.v2.

Haoxiang Sun, Tao Wang, Li Yuan, Jian Zhao, and Jiancheng Lv. From structure to synergy: A survey of vision-language perception paradigm evolution in multimodal large language models. *Information Fusion*, pp. 104285, 2026.

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.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. M4U: Evaluating multilingual understanding and reasoning for large multimodal models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 382–392, 2026a.

Xinshun Wang, Peiming Li, Ziyi Wang, Zhongbin Fang, Zhichao Deng, Songtao Wu, Jason Li, and Mengyuan Liu. Superman: Unifying skeleton and vision for human motion perception and generation. *arXiv preprint arXiv:2602.02401*, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022.

Hao Wu, Junlong Tong, Xudong Wang, Yang Tan, Changyu Zeng, Anastasia Antsiferova, and Xiaoyu Shen. From Data to Model: A Survey of the Compression Lifecycle in MLLMs. TechRxiv preprint, 2026.

Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. Model merging in LLMs, MLLMs, and beyond: Methods, theories, applications, and opportunities. *ACM Computing Surveys*, 58(8):1–41, 2026.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey. *arXiv preprint arXiv:2504.21051*, 2025.

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.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, 2024. doi: 10.18653/v1/2024.findings-acl.738.

Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19900–19909, 2025.

Jingrui Zhang, Feng Liang, Yong Zhang, Wei Wang, Runhao Zeng, and Xiping Hu. Sparse shortcuts: Facilitating efficient fusion in multimodal large language models. *arXiv preprint arXiv:2602.00505*, 2026.

Wei Zhao, Zhe Li, Yige Li, and Jun Sun. Q-MLLM: Vector Quantization for Robust Multimodal Large Language Model Security. In *Network and Distributed System Security (NDSS) Symposium*, 2026. doi: 10.14722/ndss.2026.230407.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *arXiv preprint arXiv:2305.16934*, 2023.

Feizhong Zhou, Xingyue Liu, Zhiying Yang, Zhipeng Li, and Hanguang Xiao. Enhancing medical MLLMs with dual vision encoders and MoE-based modality projector. *Knowledge-Based Systems*, pp. 115275, 2026.

# A Appendix

## A.1 Additional Implementation Details

We provide additional details on the dataset construction, target sampling, attack configuration, and evaluation models used in the main experiments. Following prior work (Li et al., 2025b), the primary evaluation is conducted on 100 randomly selected images from the *NIPS 2017 Adversarial Attacks and Defenses Competition* dataset (Kurakin et al., 2017). To further examine statistical robustness and compare with state-of-the-art methods, we additionally conduct a larger-scale evaluation on 1000 images, reported in Appendix A.3. For targeted attacks, target images are randomly sampled from the MSCOCO validation set (Lin et al., 2014), with the constraint that no target image coincides with its paired clean source image.

We use ViT-B/16, ViT-B/32, and ViT-g-14-laion2B-s12B-b42K as surrogate models, all initialized from officially released open-source checkpoints. The adversarial perturbation is constrained by an $\ell_\infty$ budget of $\epsilon = 16/255$, and each attack is optimized for 300 iterations. Transferability is evaluated across three categories of target models: open-source MLLMs, closed-source MLLMs, and reasoning MLLMs. The open-source models include Qwen2.5-VL-3B/7B, LLaVA-1.5/1.6-7B, and Gemma-3-4B/12B. The closed-source models include Claude 4.5 sonnet, Gemini-2.0, GPT-4o, and GPT-4.1. The reasoning MLLMs include Gemini-2.5-Pro and Claude-4.5-thinking. For semantic-chain construction, we follow prior work and use ClipCap (Mokady et al., 2021) as the image-to-text model to generate target descriptions. All experiments were conducted on a local GPU server equipped with NVIDIA RTX 4090 GPUs.

## A.2 Additional qualitative comparisons

Figure 4 compares adversarial examples and perturbation maps across targeted transferable attacks. While all adversarial examples remain visually close to the source images, the perturbation maps show distinct patterns. AdvDiffVLM yields diffuse reconstruction-like residuals, whereas AnyAttack introduces stronger target-specific structures that are more visually conspicuous. M-Attack produces finer high-frequency noise with limited semantic structure. PSC-Attack retains subtle perturbations while preserving more coherent target-oriented contours, indicating that semantic-chain guidance improves target transfer without noticeably increasing perceptual distortion.

## A.3 Large-Scale evaluation on 1000 images

To further examine the statistical stability of SF-Attack, we conduct a large-scale evaluation on 1000 source-target image pairs under the same protocol as the main experiments. As shown in Table 7, PSC-Attack consistently outperforms M-Attack across all evaluated closed-source MLLMs in terms of both ASR and AvgSim. These results demonstrate that PSC-Attack maintains strong targeted transferability under a larger sample size and further validate the robustness and effectiveness of the proposed semantic guidance mechanism.

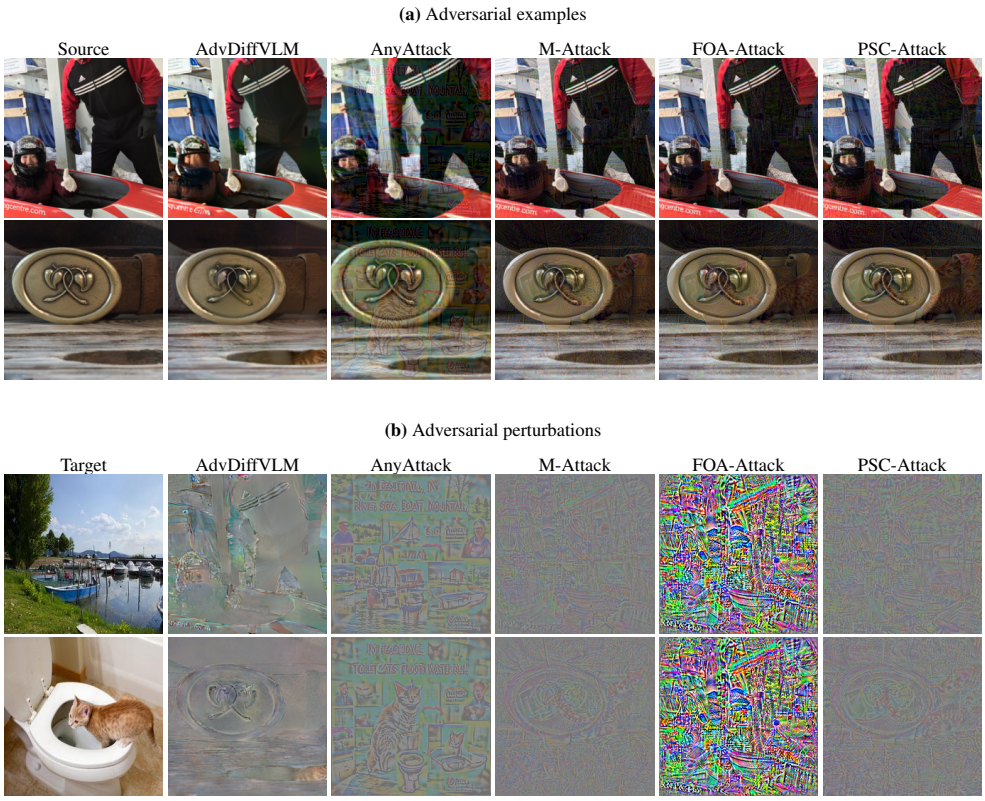


Figure 4: Qualitative comparison of adversarial examples and perturbations on additional samples. PSC-Attack preserves visual similarity to the source images while producing perturbation patterns more consistent with target-oriented semantic transfer.

Table 7: 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 | 0.495 | 0.428 | 0.640 | 0.523 | 0.678 | 0.546 | 0.035 | 0.069 |
| PSC-Attack | **0.722** | **0.514** | **0.826** | **0.588** | **0.875** | **0.648** | **0.131** | **0.152** |

### A.4 Analysis of keyword matching rate.

Table 8 shows that PSC-Attack consistently achieves the best keyword matching performance on both GPT-4o and Gemini-2.0. On GPT-4o, it obtains the highest $KMR_\alpha$, $KMR_\beta$, and $KMR_\gamma$ scores, reaching 0.86, 0.59, and 0.15, outperforming the strongest baseline M-Attack by 0.04, 0.05, and 0.02. On Gemini-2.0, PSC-Attack also ranks first on $KMR_\alpha$ and $KMR_\beta$, with 0.84 and 0.56, and achieves the best $KMR_\gamma$ score of 0.13. The gains are more evident over weaker baselines, showing that PSC-Attack better preserves target-related keywords and semantics. The smaller margin under $KMR_\gamma$ suggests that strict keyword matching remains challenging.

### A.5 Limitations and Future Work

Although the proposed method achieves strong performance in transferring target-oriented adversarial examples, it introduces additional computational overhead compared with purely visual feature alignment. In particular, the construction of semantic chains, the encoding of multiple textual guidance prompts, and the greedy selection procedure of the Minimum-Redundancy Semantic Chain

Table 8: Keyword Matching Rate comparison across multiple MLLMs. The best results are highlighted in bold.

| Method | GPT-4o | | | Gemini-2.0 | | |
|---|---|---|---|---|---|---|
| | $KMR_\alpha$ | $KMR_\beta$ | $KMR_\gamma$ | $KMR_\alpha$ | $KMR_\beta$ | $KMR_\gamma$ |
| AttackVLM | 0.09 | 0.04 | 0.00 | 0.07 | 0.02 | 0.00 |
| AdvDiffVLM | 0.02 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 |
| SSA-CWA | 0.11 | 0.06 | 0.00 | 0.05 | 0.02 | 0.00 |
| AnyAttack | 0.44 | 0.20 | 0.04 | 0.46 | 0.21 | 0.05 |
| M-Attack | 0.82 | 0.54 | 0.13 | 0.75 | 0.53 | 0.11 |
| PSC-Attack | **0.86** | **0.59** | **0.15** | **0.84** | **0.56** | **0.13** |

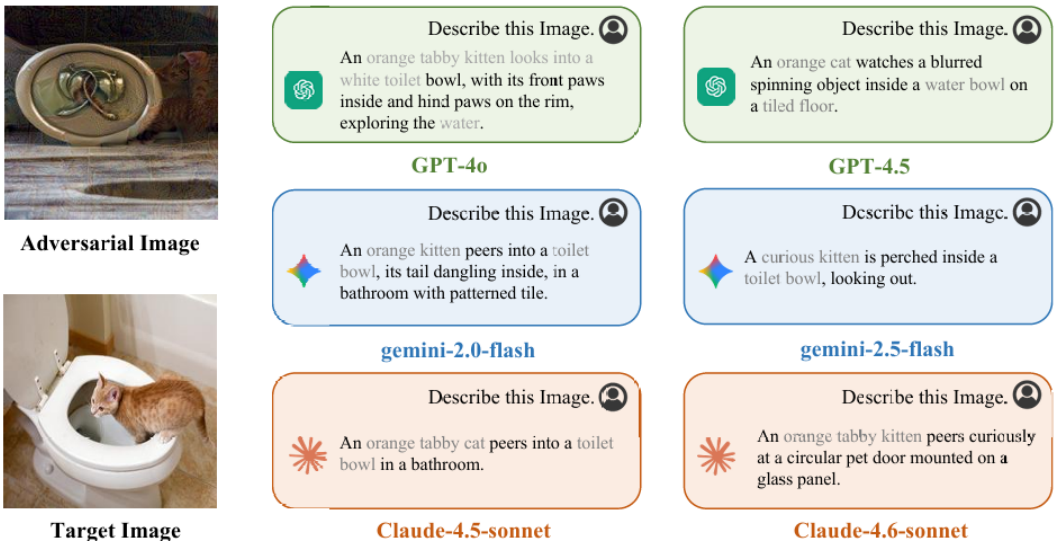


Figure 5: Examples of targeted adversarial transfer generated by PSC-Attack on commercial closed-source MLLMs.

increase the complexity of adversarial example generation. This overhead becomes more noticeable when using long optimization schedules or large surrogate ensembles. Improving the efficiency of text-guided optimization will be an important direction for future work.

### A.6 Commercial MLLM response

We provide qualitative case studies of commercial MLLM responses to targeted adversarial images.

Figure 5 presents a compact multi-model overview, where the same adversarial image induces target-aligned descriptions across different commercial MLLMs. In this example, the target image contains an orange cat near a toilet, and the generated responses consistently capture key target concepts such as “orange cat/kitten”, “toilet”, “bowl”, and “bathroom”. This overview illustrates that PSC-Attack can transfer target-oriented semantics across multiple commercial systems despite differences in model families and response styles.

We further provide a separate case study with the original model-interface screenshots. Figure 6 shows the clean and target images used in this case, where the target semantics correspond to people riding an elephant in a lush outdoor environment. Figures 7, 8, 9, 10, 11, 12, and 13 report the responses from individual commercial MLLMs. These examples show that the models generally shift toward the target semantics, although their descriptions vary in granularity, scene specificity, and exact object counts. Together, these cases complement the aggregate ASR and AvgSim results by illustrating model-specific semantic transfer behavior at the response level. All experiments are



**(a)** Clean image **(b)** Target image

Figure 6: Clean and target images used in the qualitative appendix case.

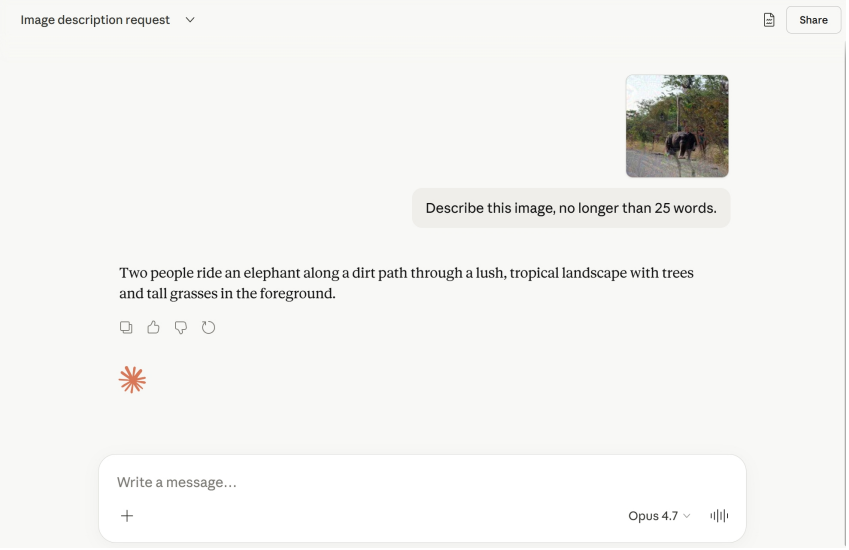


Figure 7: **Opus-4.7 response.** Opus-4.7 accurately describes two people riding an elephant along a dirt path within a tropical landscape. This response strongly matches the target image and demonstrates clear target-oriented semantic transfer.

conducted on NVIDIA RTX 4090 GPUs. The code and implementation details are available at: `https://github.com/dongshenhan/PSC-Attack`.

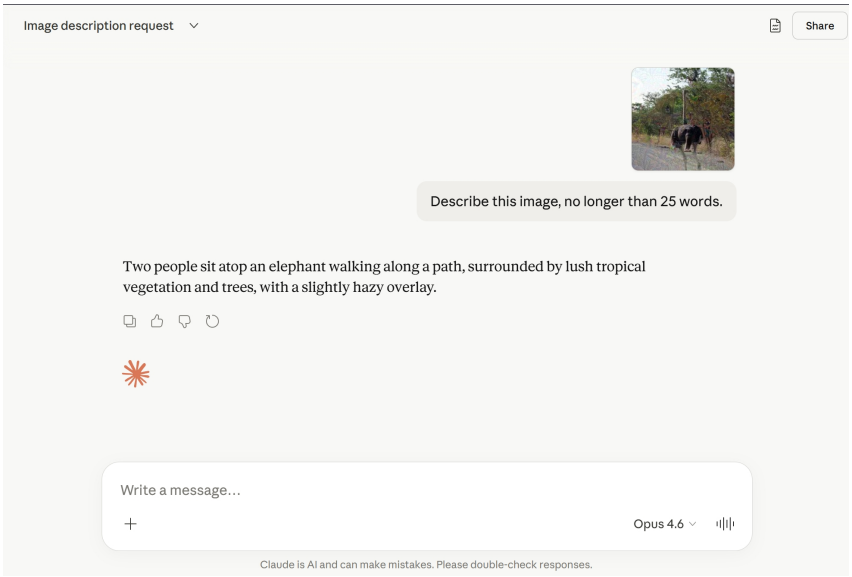


Figure 8: **Opus-4.6 response.** Opus-4.6 identifies two people sitting on an elephant and describes the surrounding tropical vegetation. The response closely matches the target semantics, indicating successful transfer toward the intended elephant-riding scene.

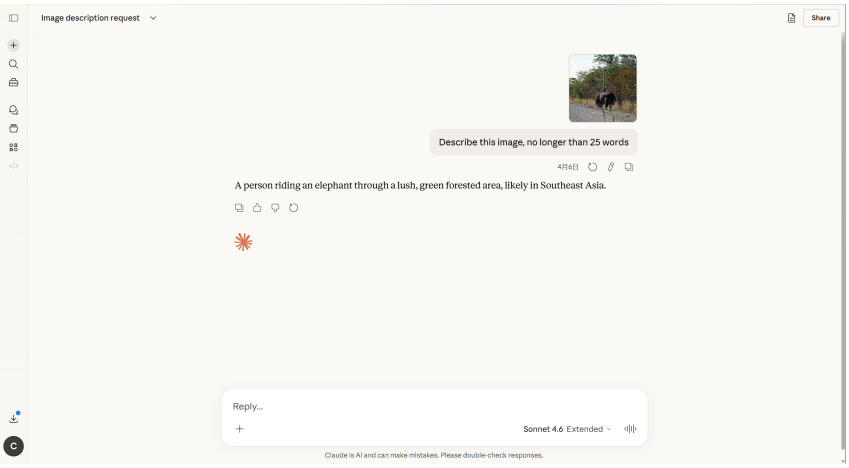


Figure 9: **Sonnet-4.6 response.** Sonnet-4.6 identifies the core target concept of an elephant-riding scene in a lush forested environment. Although it describes a single rider rather than two people, the response remains semantically aligned with the target image.

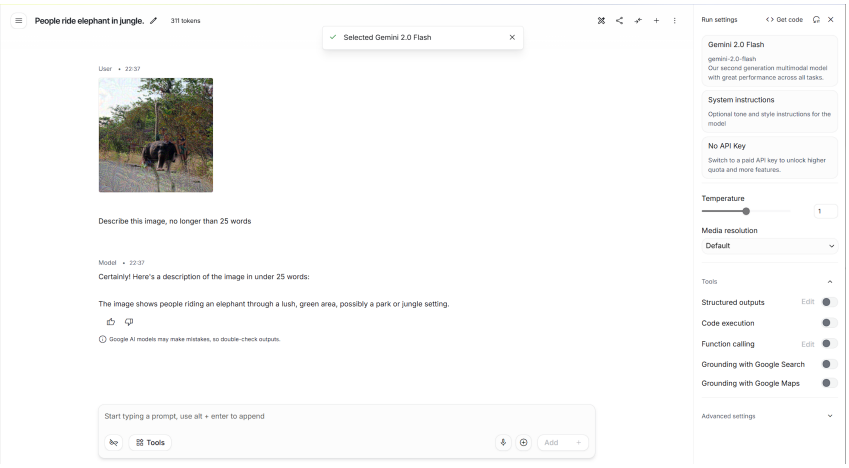


Figure 10: **Gemini-2.0 Flash response.** Gemini-2.0 Flash recognizes people riding an elephant in a lush green area, possibly a park or jungle. The response preserves the main target semantics despite using a relatively general scene description.

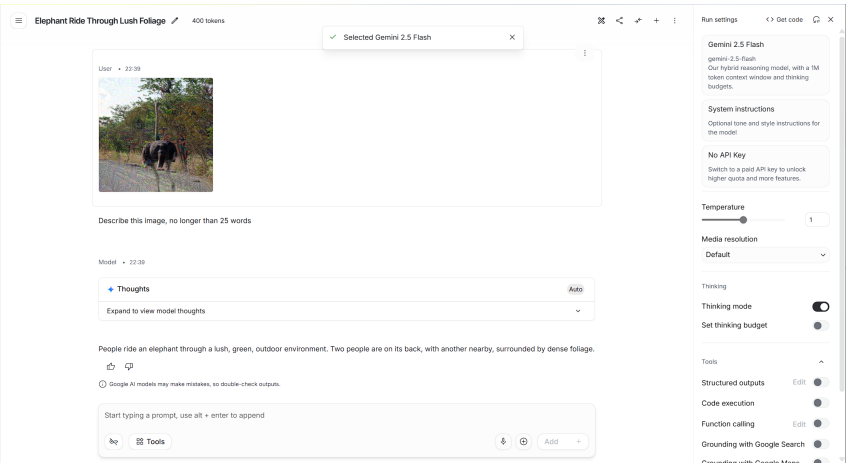


Figure 11: **Gemini-2.5 Flash response.** Gemini-2.5 Flash describes people riding an elephant in a dense outdoor environment and explicitly mentions two people on its back. The response is highly consistent with the target image, with only minor additional contextual inference.

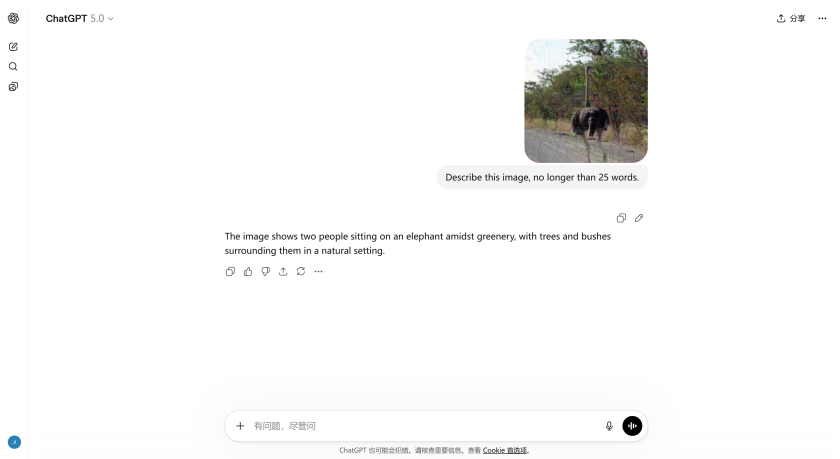


Figure 12: **GPT-5.0 response.** GPT-5.0 identifies two people sitting on an elephant surrounded by greenery, trees, and bushes. The description closely reflects the target image, showing successful semantic transfer to the intended elephant-riding concept.

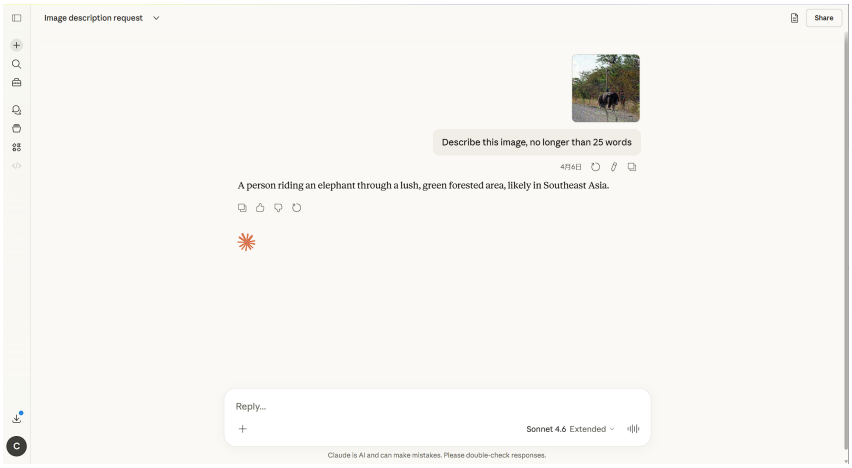


Figure 13: **Sonnet-4.6 response.** Sonnet-4.6 captures the key target concept of a person riding an elephant in a lush forested area. Although it underestimates the number of riders, the generated description remains semantically aligned with the target image.