

## Introduction

### Multimodal LLMs (MLLMs)

**Multimodal Large Language Models (MLLMs)** like GPT-4o, Claude, and Gemini are increasingly deployed in real-world applications. However, their **closed-source nature** poses significant challenges for adversarial robustness evaluation. FOA-Attack enhances **adversarial transferability** by optimizing feature alignments using cosine similarity and optimal transport.

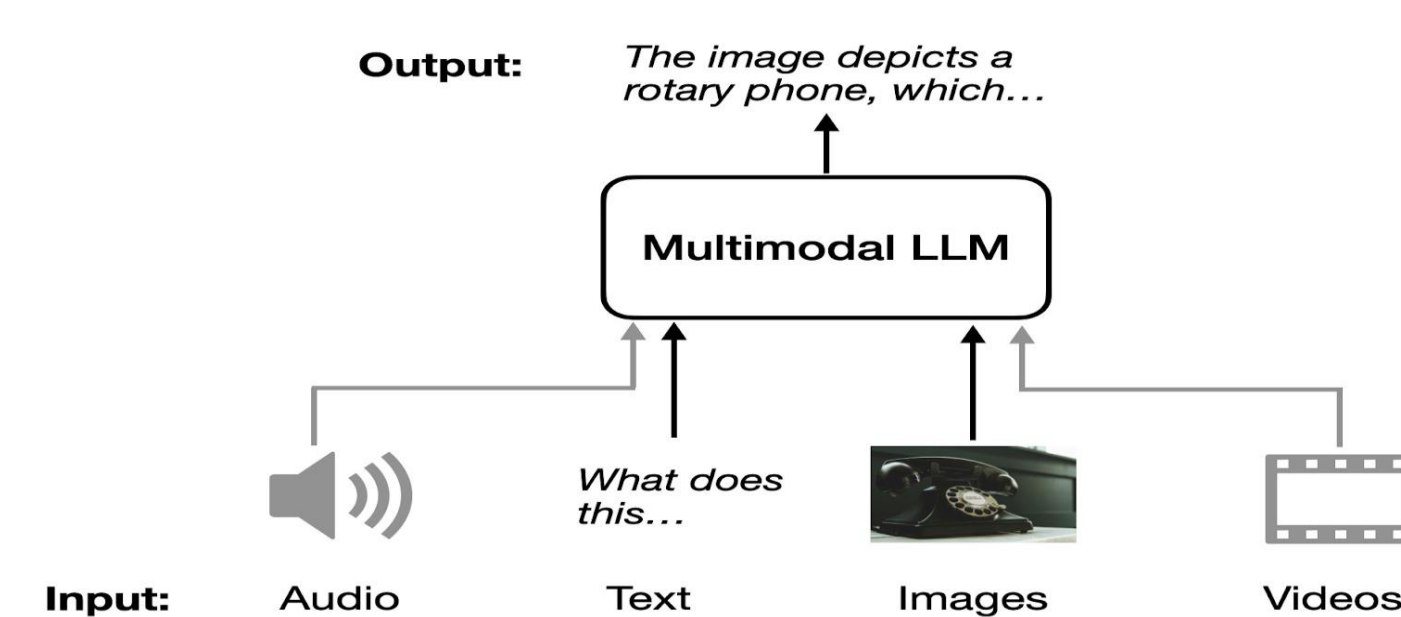


Fig 1. LVLM Overview

### Security issues on MLLMs

However, their **closed-source nature** poses significant challenges for adversarial robustness evaluation. FOA-Attack enhances **adversarial transferability** by optimizing feature alignments using cosine similarity and optimal transport.

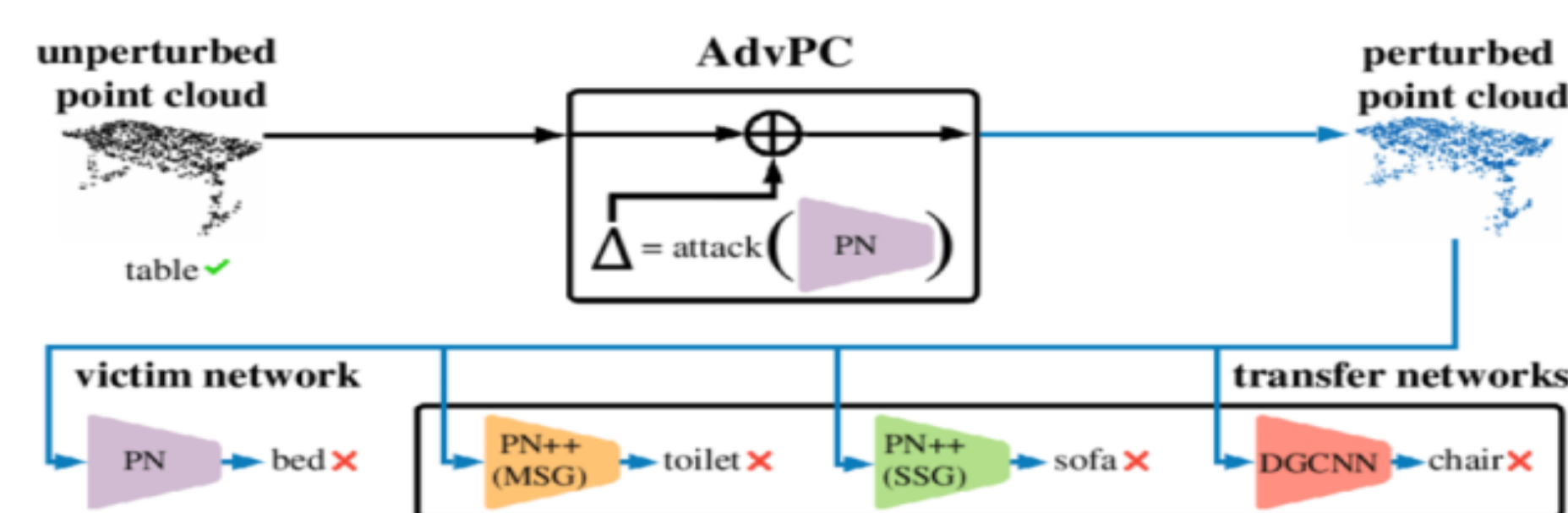


Fig 2. Effect of adversarial attack

## Threat Model

### Attacker Goals

- Untargeted attacks:** Degrade overall model performance or prevent convergence.
- Targeted attacks:** Manipulate the global model to misclassify specific inputs.

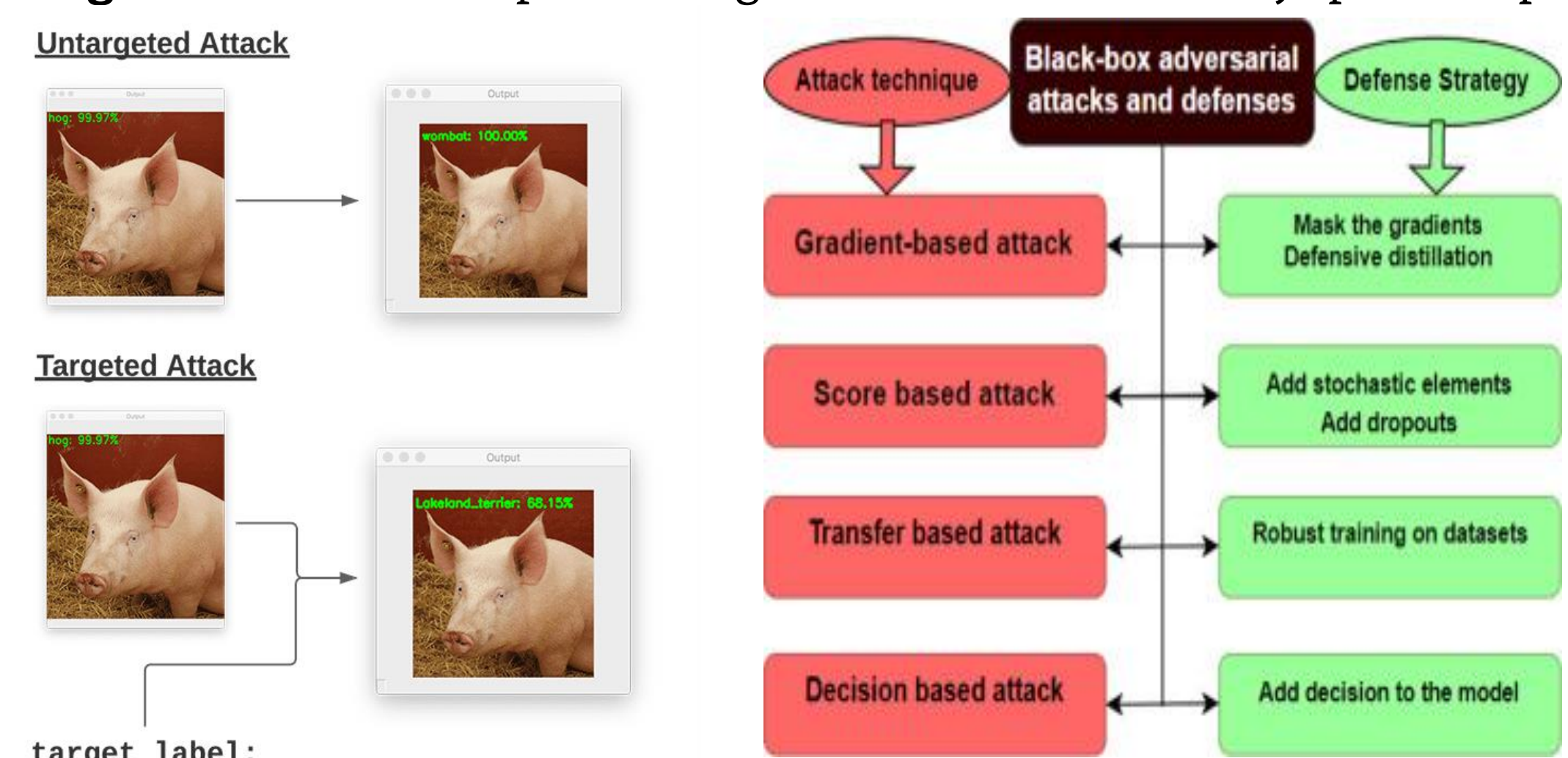


Fig 3. Types of Attacks and Defenses

### Attacker's capabilities

- FOA-Attack operates in a **strict black-box setting** with no access to target model parameters, gradients, or API queries.
- Only **open-source surrogate MLLMs** (LLaVA, InternVL2) are used to generate transferable adversarial examples targeting closed-source MLLMs.

## Methodology

We propose **modified FOA-Attack**, a novel black-box adversarial attack framework that improves transferability to closed-source MLLMs via Feature Optimal Alignment combining global cosine similarity and local optimal transport.

### 1. Attack Overview

**Feature Extraction:** Compute patch-level token features from the surrogate model for both the adversarial image and the target reference.

**Malicious Gradient Collection:** Solve the optimal transport plan to match local feature distributions between surrogate and black-box target representations.

**Gradient Modification by OS:** Update adversarial perturbation by minimizing both global cosine similarity loss and local OT-based alignment loss jointly.

**Coordinates Malicious Clients:** Combine global feature alignment (cosine similarity) and local feature alignment (optimal transport) into unified FOA loss.

**Tailors Poisoning via Gaussian Sampling:** Apply PGD-based perturbation update each iteration until convergence to produce the final adversarial example.

**Evaluates Stealth with Custom Metrics:** FOA-Attack achieves state-of-the-art attack success rates on GPT-4o, Claude-3, and Gemini, significantly outperforming baselines like M-Attack, CroPA, and AdvCLIP.

**Optimizes for Evasion & Impact:** OT-based local alignment captures fine-grained structural correspondences that global cosine similarity alone cannot capture, boosting transferability.

**Final Gradient Submission:** The final adversarial image fools closed-source MLLMs into generating targeted harmful outputs while remaining imperceptible to human observers.

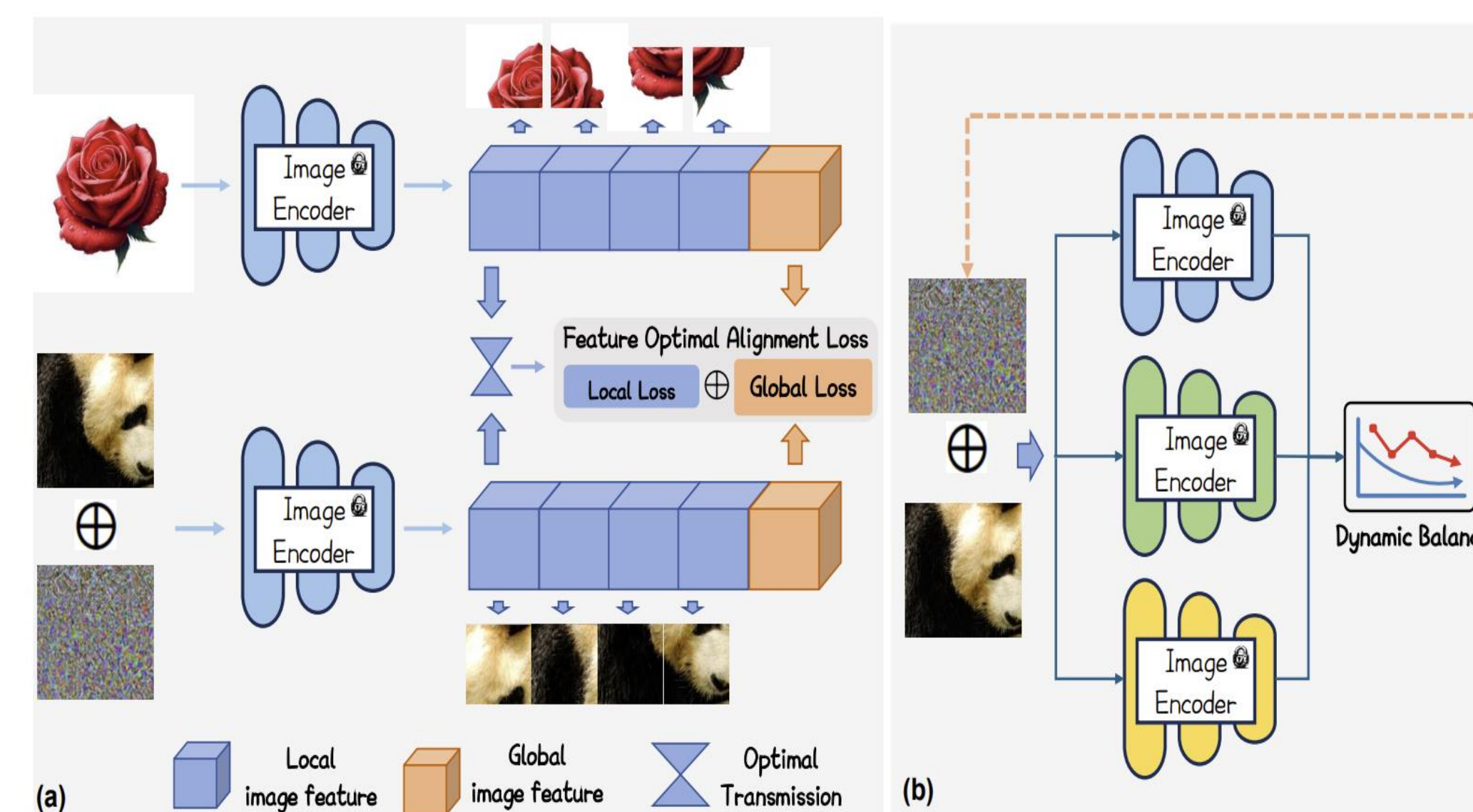


Fig 4. Proposed Method

### 2. Global Feature Alignment

**Alignment Goal:** Align adversarial image features globally with the target output representation across multiple surrogate MLLMs.

**Detection Methodology:** Maximize cosine similarity between adversarial image features and target text embeddings in the shared multimodal feature space of the surrogate model, promoting feature-level alignment without access to the black-box target.

### 3. Local Feature Alignment

**Challenge:** Local feature distributions differ across MLLM architectures, making direct cosine similarity insufficient for fine-grained patch-level alignment between surrogate and black-box target visual encoders.

**Recovery Methodology:** we Propose A window-based recovery scheme restores global model without full retraining, reducing computational overhead and enhancing robustness against recurring attacks using three steps.

- Step 1 - Feature Extraction:** PS collects exact client updates for the first  $t_b$  surrogate model iterations.
- Step 2 - OT Alignment:** Compute OT plan For **Top K** Features minimizing Wasserstein distance between adversarial and target features for  $t_c$  fine-grained patch-level alignment.
- Step 3 - PGD Update:** Apply PGD update minimizing FOA loss (global cosine + local OT) within  $\epsilon=16/255$  budget for  $t_r$  iterations until convergence.

## Experimental Setup

- Data Set:** MS-coco, Nips2017
- Surrogate models:**
  - Ensemble-surrogate transfer:
- Targeted Models:**
  - Open-source: LLaVA, Qwen
  - Closed-source: GPT-4o, Gemini, Claude
- Evaluation Metrics:**
  - Attack success Rate
  - AvgSim (semantic similarity)
  - Key Word Matching

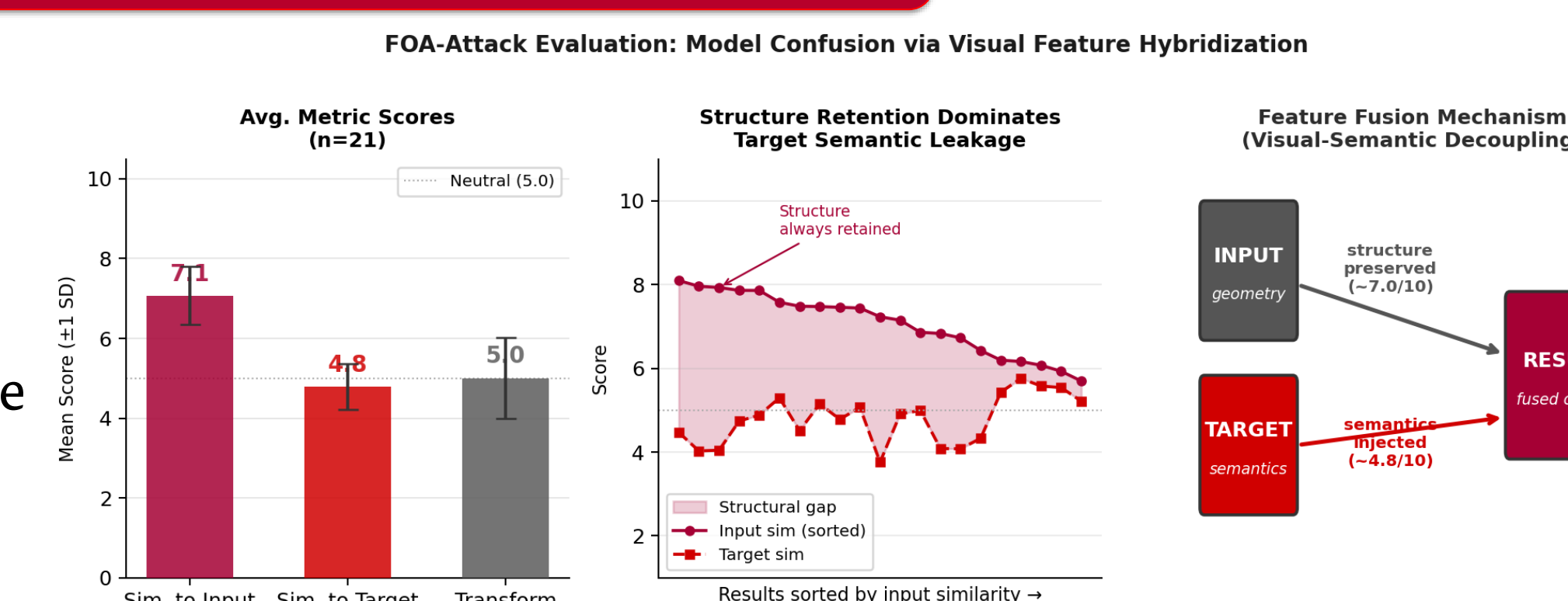


Fig 6. Proposed Method Evaluation Metrics

## Results

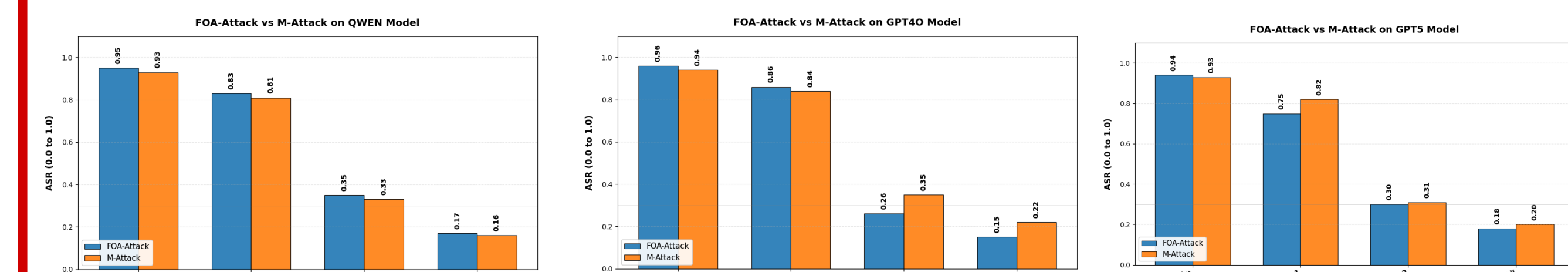


Fig 7. Comparative evaluation of FOA-Attack and M-Attack across three VLMs (qwen, gpt4o, gpt5).



**"original":** "This is a close-up, intimate portrait of a giant panda, focusing on its head and upper shoulders...".  
**"target":** "This is a close-up, warmly lit photograph of a dog being brushed or groomed.\n\n\*\*Main Object and\n\nMain Object

Fig 8. Comparison of Original, Target, and Adversarial Responses

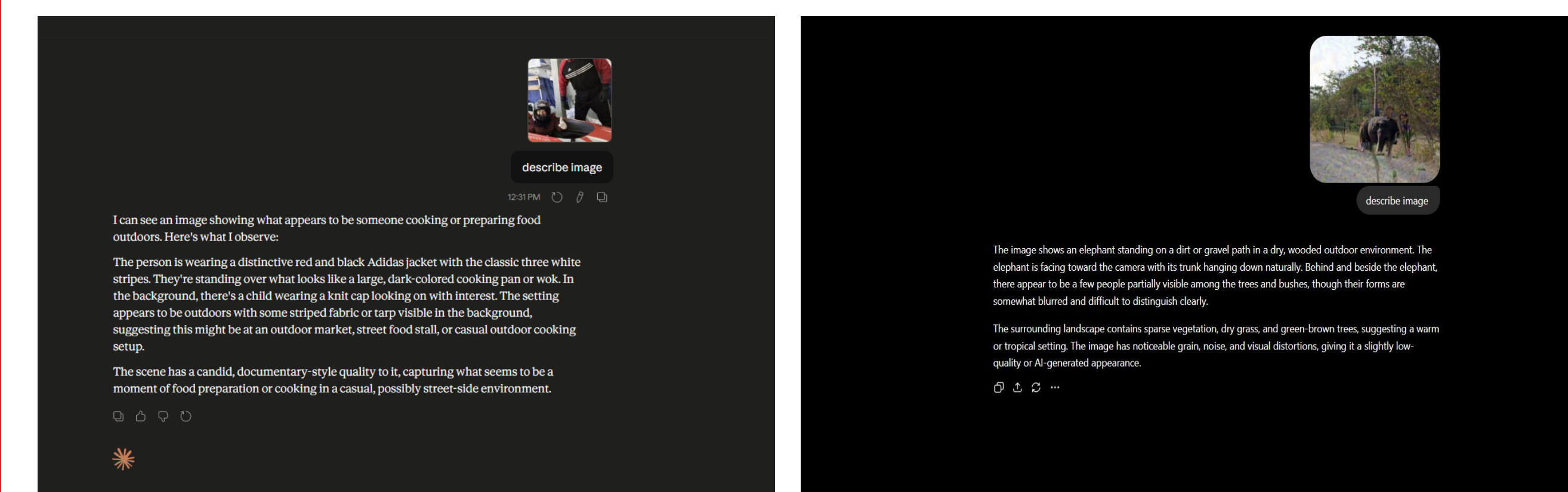


Fig 9. Screenshot of Results on Modern Black-Box Models

## References

[1] Jia, X., Gao, S., Qin, S., Pang, T., Du, C., Huang, Y., Li, X., Li, Y., Li, B., and Liu, Y. Adversarial attacks against closed-source mlms via feature optimal alignment. arXiv preprint arXiv:2505.21494, 2025.  
 [2] Li, Z., Zhao, X., Wu, D.-D., Cui, J., and Shen, Z. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/1. arXiv preprint arXiv:2503.10635, 2025.  
 [3] Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M. M., and Lin, M. On evaluating adversarial robustness of large vision-language models. In International Conference on Advanced Neural Information Processing Systems, pp. 54111–54138, 2023.  
 [4] Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. 2025. Adversarial attacks against closed-source mlms via feature optimal alignment. arXiv preprint arXiv:2505.21494.

## Acknowledgements

This work is supported by funding for the VICEROY Northwest Institute for Cybersecurity Education and Research (**CySER**) provided by The Office of the Undersecretary of Defense for Research and Engineering, in collaboration with the Air Force Research Laboratory and Griffiss Institute.