Multi-Paradigm Collaborative Adversarial Attack Against Multi-Modal Large Language Models

Supplementary Material

In this supplementary material, we provide the algorithm of the proposed MPCAttack, a detailed description of the dataset and evaluation prompt, the experimental results on the MME dataset, the ablation experiments for Multi-Paradigm, the parameter experiments of the loss function, and the visualization of the attack on adversarial examples on closed-source MLLMs.

7. A Detailed Algorithm of MPCAttack

The detailed description of the proposed MPCAttack is shown in Algorithm 1.

Algorithm 1 MPCAttack

Input: Source image x_s, target image x_t, multi-paradigm image encoders f_{c_I}, f_m, f_v, multi-modal understanding text generator f_{mg}, cross-modal alignment text encoder f_{c_T}, loss function $\mathcal{L}$, weighting factor λ, image processing $\mathcal{T}$, perturbation budget ϵ, iterations N, step size α, momentum factor μ.

Output: Adversarial image x_{adv}.

Initialize: $x_{adv}^0 = x_s + \delta^0$, $g^0 = 0$; // Initialize adversarial image x_{adv} with random noise δ^0
for $n = 0$ to $N - 1$ **do**
  $\hat{x}_{adv}^n = \mathcal{T}(x_{adv}^n)$; // Perform random crop
  Get multi-paradigm aggregated features z_{adv}, z_s, z_t from $\hat{x}_{adv}^n, x_s, x_t$ by Eq. (2), (3), (4), (5);
  Compute loss $\mathcal{L}$ by Eq. (6);
  $g^{n+1} = \mu \cdot g^n + \frac{\nabla_{\mathcal{L}}}{\|\nabla_{\mathcal{L}}\|}$;
  $\delta^{n+1} = \text{Clip}(\delta^n + \alpha \cdot \text{sign}(g^{n+1}), -\epsilon, \epsilon)$;
  $x_{adv}^{n+1} = x_{adv}^n + \delta^{n+1}$
end for
return $x_{adv} = x_{adv}^N$

8. Detailed Datasets and Evaluation Prompt

ImageNet. The ImageNet dataset, which is the most commonly used dataset in the field of adversarial attacks, we adopt 1,000 images from the ImageNet subset used in the NIPS 2017 Adversarial Attacks and Defenses Competition following [25, 29].

Flickr30K. The Flickr30K dataset contains approximately 31,783 images collected from Flickr, and each image is accompanied by five natural language descriptions written by humans. As a commonly used multimodal dataset, it has ground truth captions. Therefore, we randomly select 1000 images along with their corresponding captions. During the generation of adversarial examples, we sequentially select source images and select target images in reverse order to produce the adversarial examples. During the evaluation phase, we directly use the ground truth captions as the source and target descriptions to compute similarity scores with the adversarial descriptions.

MME. The MME dataset provides a comprehensive benchmark for evaluating the multimodal understanding capabilities of large language models, where each sample contains a yes/no type question designed to directly reflect the accuracy of the model's responses. It comprises 1,187 images, each paired with two questions corresponding to "yes" and "no" answers. During the generation of adversarial examples, we sequentially select source images and select target images in reverse order to produce the adversarial examples. During the evaluation phase, for targeted attacks, the adversarial examples are paired with the target image's question and fed into the victim black-box MLLMs, which provide a yes or no answer; the attack is considered successful if the model's response indicates misclassification. For untargeted attacks, the adversarial examples are paired with the source image's question and similarly evaluated against the victim black-box MLLMs. Notably, we only use the question corresponding to the "yes" answer for evaluation, as in practice we found that the "no" question often does not correspond to the associated image, making it incapable of reliably reflecting whether the adversarial sample successfully succeeds and thus unsuitable for a proper assessment.

Evaluation prompt. Following [25], we adopt the same way to evaluate the adversarial performance. Below is the detailed evaluation prompt used to assess semantic similarity between textual inputs. The "{text1}" and "{text2}" are used as placeholders for text inputs. The evaluation prompt template is shown in Figure 7.

9. Detailed Models

Table 5 lists the models used by each paradigm. Here, "main" indicates the primary model used, while "ablation" represents the other models employed in the ablation study in Figure 5. Because the current commonly used image encoders of MLLMs tend to be pre-trained models that are trained using the cross-modal alignment paradigm. Therefore, we mainly adopt the CLIP framework as the main paradigm and integrate three different-sized CLIP models.

Evaluation Prompt

Rate the semantic similarity between the following two texts on a scale from 0 to 1.
Criteria for similarity measurement:
1. **Main Subject Consistency:** If both descriptions refer to the same key subject or object (e.g., a person, food, an event), they should receive a higher similarity score.
2. **Relevant Description**: If the descriptions are related to the same context or topic, they should also contribute to a higher similarity score.
3. **Ignore Fine-Grained Details:** Do not penalize differences in **phrasing, sentence structure, or minor variations in detail**. Focus on **whether both descriptions fundamentally describe the same thing.**
4. **Partial Matches:** If one description contains extra information but does not contradict the other, they should still have a high similarity score.
5. **Similarity Score Range:**
- **1.0**: Nearly identical in meaning.
- **0.8-0.9**: Same subject, with highly related descriptions.
- **0.7-0.8**: Same subject, core meaning aligned, even if some details differ.
- **0.5-0.7**: Same subject but different perspectives or missing details.
- **0.3-0.5**: Related but not highly similar (same general theme but different descriptions).
- **0.0-0.2**: Completely different subjects or unrelated meanings.

Text 1: {text1}
Text 2: {text2}
Output only a single number between 0 and 1. Do not include any explanation or additional text.

Figure 7. Evaluation prompt template.

Table 5. The models used by each paradigm.

Cross-Modal Alignment Paradigm	clip-vit-base-patch16 (main) clip-vit-base-patch32 (main) CLIP-ViT-G-14-laion2B-s12B-b42K (main) siglip2-base-patch16-224 (ablation) siglip2-base-patch32-256 (ablation) siglip2-giant-opt-patch16-256 (ablation)
Multimodal Understanding Paradigm	InternVL3-1B (main) InternVL3-2B (ablation)
Visual Self-Supervised Paradigm	dinov2-base(main) dinov3-base (ablation)

10. Experimental Results on MME dataset

Tables 6 and 7 present the attack success rates of different methods on both open-source and closed-source MLLMs. Across all settings, MPCAttack consistently delivers the highest ASR, demonstrating strong generalization in both targeted and untargeted attack scenarios. On open-source MLLMs, MPCAttack yields clear performance gains, particularly on InternVL3-8B and glm-4.1v-9b-thinking, where it surpasses FOA-Attack by large margins in both attack types. Its average ASR improves by 5.90% in targeted and 9.29% in untargeted settings, confirming the effectiveness of multi-paradigm collaborative optimization. On closed-source MLLMs, the superiority of MPCAttack is even more pronounced. It achieves the highest ASR on all models, outperforming FOA-Attack by 5.35% (targeted) and 6.22% (untargeted) on average. The improvements on challenging models such as GPT-5 and Claude-4.5 indicate stronger transferability to unseen architectures with distinct training pipelines. The results demonstrate that MPCAttack enhances attack transferability across diverse model families, highlighting the advantage of aggregating multi-paradigm representations for globally optimized adversarial perturbations.

11. Attack performance under defense mechanisms and diverse settings

We evaluate our method under representative defense strategies (e.g., DiffPure) as well as diverse input conditions (e.g., additive noise). As shown in Table 8, MPCAttack largely preserves its attack effectiveness under these challenging settings. Moreover, compared with M-Attack and FOA-Attack, MPCAttack exhibits significantly less performance degradation when facing defense mechanisms and environmental variations, demonstrating its superior robustness.

12. Ablation Study for Multi-Paradigm

The results in Table 9 clearly demonstrate the effectiveness of incorporating the Multi-Paradigm mechanism. For both M-Attack and FOA-Attack, adding the Multi-Paradigm module consistently improves ASR and semantic similarity across all victim models in both targeted and untargeted settings. These gains indicate that integrating complementary paradigm-specific representations can enrich gradient signals and provide more coherent optimization guidance, thereby enhancing the overall attack strength. However, despite these improvements, the Multi-Paradigm–enhanced

variants still fall short of MPCAttack. This gap highlights an important limitation of directly extending existing methods: although they benefit from additional paradigms, their optimization remains largely independent across feature types. Such optimization can lead to redundant gradient directions, restricting their ability to fully leverage complementary information. In contrast, MPCAttack jointly optimizes multi-paradigm features in a unified framework, enabling coordinated gradient alignment and globally consistent perturbation updates. This integrated optimization yields more effective and transferable adversarial perturbations, outperforming all baselines by a clear margin.

13. Analysis of Loss-Function Parameter Effects

Since the loss function $\mathcal{L}$ contains two parameters, where τ is a temperature coefficient that controls the sharpness of the similarity distribution, and ω is a balancing factor that regulates the trade-off between positive pairs and negative pairs. Table 10 evaluates the influence of the temperature coefficient τ and the balancing factor ω in the loss function. As τ controls the sharpness of the similarity distribution, overly small values lead to unstable optimization and poor ASR, while excessively large values oversmooth the distribution and weaken contrastive guidance. Moderate values (e.g., τ=0.2–0.4) achieve the best balance and yield consistently strong performance. Meanwhile, ω regulates the trade-off between positive and negative pairs: small values overemphasize source separation and hinder targeted alignment, whereas overly large values bias the optimization toward positive pairs and reduce discriminability. Intermediate settings (e.g., ω=2–3) provide the most effective trade-off.

14. Visualization

Figure 8 shows the description of commercial LVLMs for the adversarial images generated by SADCA. It can be seen that it is capable of effectively attacking various commercial MLLMs.

Table 6. Performance of ASR (%) on open-source MLLMs across both targeted and untargeted settings on MME dataset.

Targeted	Victim Black-box Open-Source Models				
	Qwen2.5-VL-7B-Instruct	InternVL3-8B	LLaVa-1.5-7B	glm-4.1v-9b-thinking	Average
M-Attack	9.94	44.23	35.38	34.46	31.00
FOA-Attack	11.20	47.09	36.98	38.50	33.44
MPCAttack	**11.88**	**59.81**	**42.29**	**43.39**	**39.34**
Untargeted	Victim Black-box Open-Source Models				
	Qwen2.5-VL-7B-Instruct	InternVL3-8B	LLaVa-1.5-7B	glm-4.1v-9b-thinking	Average
M-Attack	26.20	30.92	70.18	24.68	38.00
FOA-Attack	26.20	32.69	71.52	24.68	38.77
MPCAttack	**31.84**	**45.32**	**83.49**	**31.59**	**48.06**

Table 7. Performance of ASR (%) on closed-source MLLMs across both targeted and untargeted settings on MME dataset.

Targeted	Victim Black-box Closed-Source Models				
	GPT-4o	GPT-5	Claude-3.5	Gemini-2.0	Average
M-Attack	53.67	27.91	4.38	25.19	27.79
FOA-Attack	55.91	30.13	6.97	31.06	31.02
MPCAttack	**63.42**	**35.95**	**10.03**	**36.09**	**36.37**
Untargeted	Victim Black-box Closed-Source Models				
	GPT-4o	GPT-5	Claude-3.5	Gemini-2.0	Average
M-Attack	38.34	41.10	59.64	31.91	42.75
FOA-Attack	38.50	39.91	60.13	32.17	42.68
MPCAttack	**48.88**	**45.02**	**64.08**	**37.62**	**48.90**

Table 8. The Average ASR (%) and AvgSim on four black-box open-source models under defense mechanisms and diverse settings.

	DiffPure		Noisy	
Targeted	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)
M-Attack	15.05	0.1656	27.75	0.2541
FOA-Attack	17.43	0.1849	30.85	0.2831
MPCAttack	**38.50**	**0.3330**	**50.85**	**0.4136**
Untargeted	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)
M-Attack	48.28	0.4410	59.23	0.3591
FOA-Attack	53.73	0.4036	64.35	0.3272
MPCAttack	**78.13**	**0.2207**	**83.03**	**0.1742**

Table 9. Ablation study for Multi-Paradigm.

Targeted	Victim Black-box Open-Source Models							
	Qwen2.5-VL-7B		InternVL3-8B		LLaVa-1.5-7B		GLM-4.1V-9B-Thinking	
Methods	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)
M-Attack	17.3	0.1919	68.1	0.5361	59.9	0.4954	31.0	0.3118
M-Attack+Multi-Paradigm	24.2	0.2508	79.7	0.6117	69.6	0.5476	33.0	0.3088
FOA-Attack	20.0	0.2222	72.3	0.5700	63.8	0.5162	38.3	0.3522
FOA-Attack+Multi-Paradigm	**32.7**	0.3061	84.3	0.6454	72.9	0.5707	42.4	0.3760
MPCAttack	32.5	**0.3191**	**88.7**	**0.6678**	**73.9**	**0.5905**	**58.2**	**0.4756**
Untargeted	Victim Black-box Open-Source Models							
	Qwen2.5-VL-7B		InternVL3-8B		LLaVa-1.5-7B		GLM-4.1V-9B-Thinking	
Methods	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)
M-Attack	54.3	0.3982	90.8	0.1417	94.9	0.0819	61.2	0.3696
M-Attack+Multi-Paradigm	67.9	0.3179	96.3	0.0888	97.9	0.0527	59.2	0.3730
FOA-Attack	61.8	0.3589	93.0	0.1204	96.3	0.0659	68.1	0.3227
FOA-Attack+Multi-Paradigm	72.3	0.2821	97.0	0.0790	98.2	0.0448	69.8	0.2989
MPCAttack	**84.9**	**0.1919**	**99.1**	**0.0341**	**99.3**	**0.0186**	**85.1**	**0.1825**

Table 10. Parameter experiments of τ and ω.

Targeted	Victim Black-box Open-Source Models							
	Qwen2.5-VL-7B-Instruct		InternVL3-8B		LLaVa-1.5-7B		GLM-4.1V-9B-Thinking	
Methods	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)	ASR(↑)	AvgSim(↑)
τ=0.05, ω=2	6.6	0.1306	38.1	0.3895	37.1	0.3541	21.3	0.2511
τ=0.1, ω=2	31.2	0.3081	89.7	0.6712	75.1	0.5959	60.1	0.4842
τ=0.2, ω=2	31.7	0.3051	91.3	0.6929	76.5	0.6016	62.9	0.4910
τ=0.3, ω=2	32.3	0.3140	90.9	0.6856	77.6	0.6079	59.8	0.4934
τ=0.4, ω=2	32.1	0.3054	90.6	0.6860	77.3	0.6072	61.5	0.4917
τ=0.5, ω=2	33.2	0.3137	90.5	0.6847	76.7	0.6032	61.2	0.4893
τ=0.2, ω=0.5	11.9	0.1620	36.9	0.3748	31.7	0.3256	22.7	0.2546
τ=0.2, ω=1	26.3	0.2726	82.5	0.6173	68.3	0.5357	53.8	0.4430
τ=0.2, ω=3	34.1	0.3100	90.7	0.6855	76.4	0.6058	59.0	0.4858
τ=0.2, ω=4	30.1	0.2946	89.8	0.6870	76.6	0.6057	59.2	0.4771
τ=0.2, ω=5	14.8	0.1878	77.8	0.5854	62.5	0.5094	41.2	0.3648
Untargeted	Victim Black-box Open-Source Models							
	Qwen2.5-VL-7B-Instruct		InternVL3-8B		LLaVa-1.5-7B		GLM-4.1V-9B-Thinking	
Methods	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)	ASR(↑)	AvgSim(↓)
τ=0.05, ω=2	75.8	0.2525	98.7	0.0440	99.4	0.0282	78.0	0.2464
τ=0.1, ω=2	86.6	0.1775	99.3	0.0322	99.5	0.0195	86.3	0.1665
τ=0.2, ω=2	83.6	0.1836	99.1	0.0332	99.3	0.0209	87.7	0.1620
τ=0.3, ω=2	85.5	0.1847	99.6	0.0306	99.5	0.0188	86.7	0.1621
τ=0.4, ω=2	84.6	0.1886	99.4	0.0304	99.6	0.0186	86.5	0.1688
τ=0.5, ω=2	86.4	0.1812	99.6	0.0316	99.4	0.0198	88.6	0.1610
τ=0.2, ω=0.5	90.3	0.1447	99.8	0.0148	99.6	0.0111	87.8	0.1632
τ=0.2, ω=1	86.4	0.1623	99.8	0.0221	99.8	0.0147	90.1	0.1486
τ=0.2, ω=3	83.7	0.2034	99.1	0.0386	99.5	0.0207	85.7	0.1799
τ=0.2, ω=4	80.8	0.2207	99.2	0.0407	99.1	0.0243	83.4	0.1981
τ=0.2, ω=5	69.3	0.2974	98.9	0.0549	99.3	0.0295	74.9	0.2625

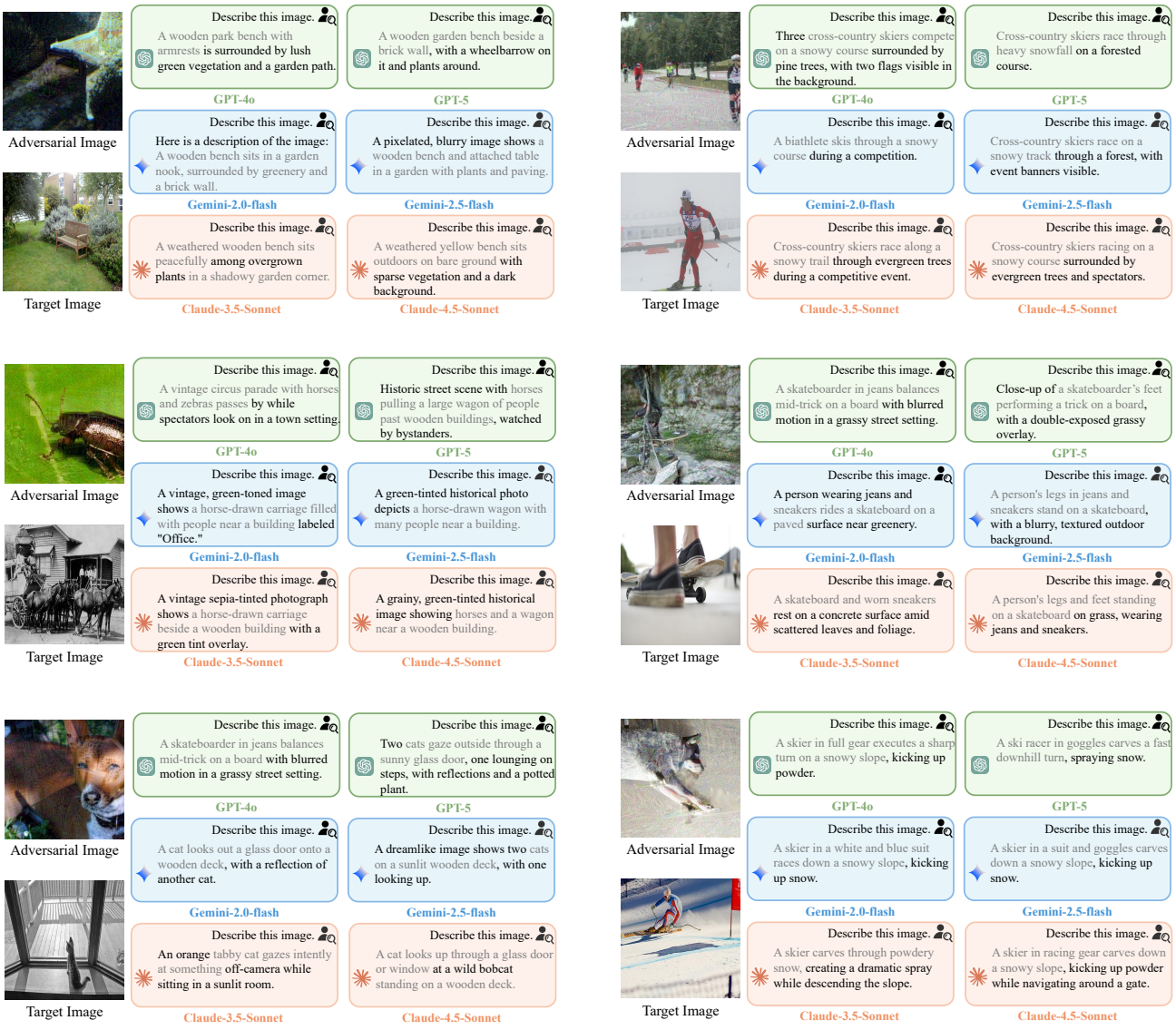

Figure 8. Visualization of adversarial images in attacking commercial MLLMs.