

## 384 Appendix

### 385 A Complementary Details of M-Attack-V2

386 Alg. 2 and Alg. 3 provide detailed update rule of line 13 in Alg. 1. Fig. 8 provides a comparison  
 387 between the entire procedure of M-Attack and M-Attack-V2 under the local-matching framework.  
 388 Notably, M-Attack utilizes a radical crop on the target image, risking unrelated or broken semantics  
 389 for the source image to align. Our ATA anchors more points inside the semantic manifold (blue), and  
 390 provides a mild transformation to provide a coherence sampling from the target semantic manifold.

---

#### Algorithm 2 M-Attack-V2 (Adam variant)

---

**Require:** clean image  $\mathbf{X}_{\text{clean}}$ ; primary target  $\mathbf{X}_{\text{tar}}$ ; **auxiliary set**  $\mathcal{A} = \{\mathbf{X}_{\text{aux}}^{(p)}\}_{p=1}^P$ ; **patch ensemble**<sup>+</sup>  $\Phi^+ = \{\phi_j\}_{j=1}^m$ ; iterations  $n$ , step size  $\alpha$ , perturbation budget  $\epsilon$ ; Adam  $\beta_1, \beta_2, \eta$ ; number of crops  $K$ , auxiliary weight  $\lambda$ ;

- 1:  $\mathbf{X}_{\text{adv}} \leftarrow \mathbf{X}_{\text{clean}}, m \leftarrow 0, v \leftarrow 0$
- 2: **for**  $i = 1$  **to**  $n$  **do**
- 3:   Draw  $K$  transforms  $\{\mathcal{T}_k\}_{k=1}^K \sim \mathcal{D}$
- 4:    $g \leftarrow 0$  ▷ accumulate over crops
- 5:   **for**  $k = 1$  **to**  $K$  **do** ▷ — crop loop —
- 6:     Draw  $\{\tilde{\mathcal{T}}_p\}_{p=0}^P \sim \tilde{\mathcal{D}}$
- 7:     **for**  $j = 1$  **to**  $m$  **do**
- 8:        $y_0 = f(\tilde{\mathcal{T}}_0(\mathbf{X}_{\text{tar}}))$
- 9:        $y_p = f(\tilde{\mathcal{T}}_p(\mathbf{X}_{\text{aux}}^{(p)})), p = 1, \dots, P$
- 10:        $\hat{\mathcal{L}}_k = \mathcal{L}(f_{\phi_j}(\mathcal{T}_k(\mathbf{X}_{\text{adv}})), y_0) + \frac{\lambda}{P} \sum_{p=1}^P \mathcal{L}(f_{\phi_j}(\mathcal{T}_k(\mathbf{X}_{\text{adv}})), y_p)$
- 11:        $g \leftarrow g + \frac{1}{Km} \nabla_{\mathbf{X}_{\text{adv}}} \hat{\mathcal{L}}_k$
- 12:     **end for**
- 13:   **end for** ▷ — Adam update —
- 14:    $m \leftarrow \beta_1 m + (1 - \beta_1)g$
- 15:    $v \leftarrow \beta_2 v + (1 - \beta_2)g^{\odot 2}$
- 16:    $\hat{m} \leftarrow m / (1 - \beta_1^i); \hat{v} \leftarrow v / (1 - \beta_2^i)$
- 17:    $\mathbf{X}_{\text{adv}} \leftarrow \text{clip}_{\mathbf{X}_{\text{clean}}, \epsilon}(\mathbf{X}_{\text{adv}} + \alpha \hat{m} / (\sqrt{\hat{v}} + \eta))$
- 18: **end for**
- 19: **return**  $\mathbf{X}_{\text{adv}}$

---



---

#### Algorithm 3 M-Attack-V2 (MI-FGSM variant)

---

**Require:** clean image  $\mathbf{X}_{\text{clean}}$ ; primary target  $\mathbf{X}_{\text{tar}}$ ; **auxiliary set**  $\mathcal{A} = \{\mathbf{X}_{\text{aux}}^{(p)}\}_{p=1}^P$ ; **patch ensemble**<sup>+</sup>  $\Phi^+ = \{\phi_j\}_{j=1}^m$ ; iterations  $n$ , step size  $\alpha$ , perturbation budget  $\epsilon$ ; momentum decay  $\gamma$ ; number of crops  $K$ , auxiliary weight  $\lambda$ ;

- 1:  $\mathbf{X}_{\text{adv}} \leftarrow \mathbf{X}_{\text{clean}}, \mu \leftarrow 0$
- 2: **for**  $i = 1$  **to**  $n$  **do**
- 3:   Draw  $K$  transforms  $\{\mathcal{T}_k\}_{k=1}^K \sim \mathcal{D}$
- 4:    $g \leftarrow 0$
- 5:   **for**  $k = 1$  **to**  $K$  **do**
- 6:     Draw  $\{\tilde{\mathcal{T}}_p\}_{p=0}^P \sim \tilde{\mathcal{D}}$
- 7:     **for**  $j = 1$  **to**  $m$  **do**
- 8:        $y_0 = f(\tilde{\mathcal{T}}_0(\mathbf{X}_{\text{tar}}))$
- 9:        $y_p = f(\tilde{\mathcal{T}}_p(\mathbf{X}_{\text{aux}}^{(p)})), p = 1, \dots, P$
- 10:        $\hat{\mathcal{L}}_k = \mathcal{L}(f_{\phi_j}(\mathcal{T}_k(\mathbf{X}_{\text{adv}})), y_0) + \frac{\lambda}{P} \sum_{p=1}^P \mathcal{L}(f_{\phi_j}(\mathcal{T}_k(\mathbf{X}_{\text{adv}})), y_p)$
- 11:        $g \leftarrow g + \frac{1}{Km} \nabla_{\mathbf{X}_{\text{adv}}} \hat{\mathcal{L}}_k$
- 12:     **end for**
- 13:   **end for** ▷ — MI-FGSM update —
- 14:    $\mu \leftarrow \gamma \mu + \frac{g}{\|g\|_1}$
- 15:    $\mathbf{X}_{\text{adv}} \leftarrow \text{clip}_{\mathbf{X}_{\text{clean}}, \epsilon}(\mathbf{X}_{\text{adv}} + \alpha \text{sign}(\mu))$
- 16: **end for**
- 17: **return**  $\mathbf{X}_{\text{adv}}$

---

## B Complementary Details of Experimental Setup

The experiment’s seed is 2023. It is conducted on a Linux platform (Ubuntu 22.04) with 6 NVIDIA RTX 4090 GPUs. The temperatures of all LLMs are set to 0. The threshold of the ASR is set to 0.3, following M-Attack.

We provide the Huggingface identifiers of the model we used in the experiment in Tab. 8. All the BLIP2 [20] variants on the Huggingface share the same vision encoder. Therefore, we only use one of them.

## C Theoretical Analysis for Variance

This section provides detailed proof of the upper bound in Equ. (4). For variance, we have

$$\begin{aligned}
 \text{Var}(\hat{g}_K) &:= \mathbb{E} \|\hat{g}_K - \mu\|^2 \\
 &= \mathbb{E} \left\| \frac{1}{K} \sum_{k=1}^K (g_k - \mu) \right\|^2 \\
 &= \frac{1}{K^2} \sum_{k=1}^K \sum_{\ell=1}^K \mathbb{E}[(g_k - \mu)^\top (g_\ell - \mu)] \\
 &= \frac{1}{K^2} \left( \underbrace{\sum_{k=1}^K \mathbb{E} \|g_k - \mu\|_2^2}_{K\sigma^2} + 2 \underbrace{\sum_{1 \leq k < \ell \leq K} \mathbb{E}[\langle g_k - \mu, g_\ell - \mu \rangle]}_{\text{cross terms}} \right)
 \end{aligned} \tag{7}$$

The diagonal part is reduced to the mean. We now provide an upper bound for the cross terms. Recall

$p_{k\ell} = \frac{\langle g_k - \mu, g_\ell - \mu \rangle}{\|g_k - \mu\|_2 \|g_\ell - \mu\|_2}$ , we have

$$\mathbb{E}[\langle g_k - \mu, g_\ell - \mu \rangle] = \mathbb{E}[\rho_{k\ell} \|g_k - \mu\|_2 \|g_\ell - \mu\|_2]. \tag{8}$$

Since all crops share the same marginal distribution, i.e.  $\mathbb{E} \|g_k - \mu\|_2 = \mathbb{E} \|g_\ell - \mu\|_2 = \sigma$ , applying the Cauchy-Schwarz inequality to Equ. (8) yields

$$\mathbb{E}[\langle g_k - \mu, g_\ell - \mu \rangle] \leq \mathbb{E}[\rho_{k\ell}] \sqrt{\mathbb{E} \|g_k - \mu\|_2^2} \sqrt{\mathbb{E} \|g_\ell - \mu\|_2^2} = \bar{\rho} \sigma^2, \tag{9}$$

where  $\bar{\rho}$  is  $\mathbb{E}[\rho_{k\ell}]$ ,  $k \neq \ell$ . Plugging this into the double sum term yields

$$\sum_{1 \leq k < \ell \leq K} \mathbb{E}[\langle g_k - \mu, g_\ell - \mu \rangle] \leq \frac{K(K-1)}{2} \bar{\rho} \sigma^2. \tag{10}$$

The  $\frac{K(K-1)}{2}$  appears since there are total  $\frac{K(K-1)}{2}$  terms for  $\sum_{k < \ell}$ . Thus substituting Equ. (10) back to the cross item part in the Equ. (7) yields

$$\text{Var}(\hat{g}_K) \leq \frac{1}{K^2} (K\sigma^2 + K(K-1)\bar{\rho}\sigma^2) = \frac{1}{K} \sigma^2 + \frac{K-1}{K} \bar{\rho} \sigma^2 \tag{11}$$

Therefore, we have the upper bound provided in the Sec. 2.2.

## D Full Process of Surrogate Model Selection

This section details the process of selecting our final ensemble, PE<sup>+</sup>. Exhaustively testing all model combinations is computationally infeasible, so we employ a heuristic-driven approach. We begin by excluding DiNO-large and BLIP2 due to their poor transferability, as shown in Tab. 1. Our initial experiments focus on evaluating the effectiveness of homogeneous ensembles—comprising models with the same patch size—versus mixed patch size ensembles. Specifically, we construct five ensembles: (1) patch-14 CLIP (CLIP-L/14, CLIP<sup>†</sup>-G/14), (2) patch-14 DiNOv2 (Dino-base,

415 Dino-large), (3) patch-16 CLIP (CLIP-B/16, CLIP<sup>†</sup>-B/16), and (4) patch-32 CLIP (CLIP-B/32,  
416 CLIP<sup>†</sup>-B/32). Results are presented in Tab. 6. These results reveal that the patch-32 CLIP ensemble  
417 performs best on Claude 3.7, while GPT-4o and Gemini 2.5 Pro favor models with patch sizes 14 and  
418 16. This supports the findings in Sec. 3.2: although using a fixed patch size can mitigate architectural  
419 bias, it still inherits the intrinsic bias of the patch size itself.

420 To address this, we adopt a cross-patch size strategy. Starting from the patch-32 CLIP ensemble,  
421 due to its strong performance on Claude and consistent transferability across patch-16 and patch-32  
422 models. We incrementally incorporate one model each from patch sizes 14 and 16. We evaluate  
423 various combinations, with results summarized in Tab. 7. The resulting ensemble, PE<sup>+</sup>, achieves  
424 the most balanced performance, ranking first on 7 metrics and a close second on 3 others, across 12  
425 evaluation metrics.

Variant	Surrogate Set (2 models)	GPT-4o				Claude 3.7-extended				Gemini 2.5-Pro			
		KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR
Pair <sub>1</sub>	Dino-B, Dino-S	0.84	0.57	0.15	0.91	0.09	0.04	0.00	0.05	<b>0.84</b>	0.53	0.11	0.81
Pair <sub>2</sub>	L16, B/16	<b>0.86</b>	<b>0.69</b>	<u>0.21</u>	<b>0.96</b>	<u>0.16</u>	<u>0.10</u>	<u>0.01</u>	<u>0.16</u>	<b>0.84</b>	<u>0.59</u>	<u>0.15</u>	<u>0.91</u>
Pair <sub>3</sub>	L32, B/32	0.76	0.52	0.13	0.79	<b>0.46</b>	<b>0.29</b>	<b>0.06</b>	<b>0.70</b>	0.58	0.37	0.07	0.59
Pair <sub>4</sub>	G/14, L14	<b>0.86</b>	<u>0.61</u>	<b>0.24</b>	<u>0.94</u>	0.07	0.02	0.00	0.06	<u>0.82</u>	<b>0.64</b>	<b>0.23</b>	<b>0.92</b>

Table 6: Ablation on two-model surrogate sets. Bold numbers are the best in each column; underlined numbers are the second-best.

Variant	Surrogate Set	GPT-4o				Claude 3.7-extended				Gemini 2.5-Pro			
		KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR
PE <sub>1</sub>	B/16, B/32, L32, L16	0.87	0.65	0.26	<b>0.99</b>	0.54	0.32	0.07	0.68	0.80	0.57	0.16	0.90
PE <sub>2</sub>	Dino-B, B/32, L32, G/14	0.87	0.69	0.28	0.97	0.56	0.37	0.09	0.65	<b>0.88</b>	0.71	0.22	0.93
PE <sub>3</sub>	L16, B/32, L32, G/14	0.85	0.65	0.23	<b>0.99</b>	<b>0.57</b>	<u>0.40</u>	0.09	<b>0.73</b>	0.84	0.61	0.19	0.93
PE <sub>4</sub>	B/16, B/32, L32, Dino-B	0.89	0.67	0.19	0.98	0.55	<b>0.41</b>	0.07	0.63	0.87	0.67	<b>0.23</b>	0.96
PE <sub>5</sub>	B/16, B/32, L32, Dino-S	0.90	0.72	0.25	0.97	0.48	0.33	0.08	0.59	0.83	0.63	0.17	0.90
<b>PE<sup>+</sup> (Ours)</b>	B/16, B/32, L32, G/14	<b>0.91</b>	<b>0.78</b>	<b>0.40</b>	<b>0.99</b>	<u>0.56</u>	0.32	<b>0.11</b>	0.67	<u>0.87</u>	<b>0.72</b>	<u>0.22</u>	<b>0.97</b>

Table 7: Ablation on surrogate-set selection. Each row swaps one model in or out of a four-model ensemble. The fully grey PE<sup>+</sup> line is our final patch-diverse surrogate set (CLIP<sup>†</sup>-G/14, CLIP-B/16, CLIP-B/32, CLIP<sup>†</sup>-B/32). Bold numbers denote the best score in each metric column across all variants, underline denote second best with neglectable gap of 0.01

## 426 E Ablation Study for Step Size

427 This section provides an ablation study for the step size parameter  $\alpha$  to view its impact on the  
428 performance. Overall, selecting  $\alpha \in [0.5, 1.0]$  provides better performance for SSA-CWA, M-Attack.  
429 Our M-Attack-V2 prefer stepsize at 1.275, since it adopts ADAM as optimizer.

Surrogate (paper notation)	Implementation (HuggingFace identifier)
CLIP <sup>†</sup> -B/32 [15, 35]	laion/CLIP-ViT-B-32-laion2B-s34B-b79K
CLIP <sup>†</sup> -H/14 [15, 35]	laion/CLIP-ViT-H-14-laion2B-s32B-b79K
CLIP-L/14 [33]	openai/clip-vit-large-patch14
CLIP <sup>†</sup> -B/16 [15, 35]	laion/CLIP-ViT-B-16-laion2B-s34B-b88K
CLIP <sup>†</sup> -BG/14 [15, 35]	laion/CLIP-ViT-bigG-14-laion2B-39B-b160k
Dino-Small [31]	facebook/dinov2-small
Dino-Base [31]	facebook/dinov2-base
Dino-Large [31]	facebook/dinov2-large
BLIP-2 (2.7 B) [20]	Salesforce/blip2-opt-2.7b

Table 8: Surrogate models and their corresponding HuggingFace identifier in our main paper.

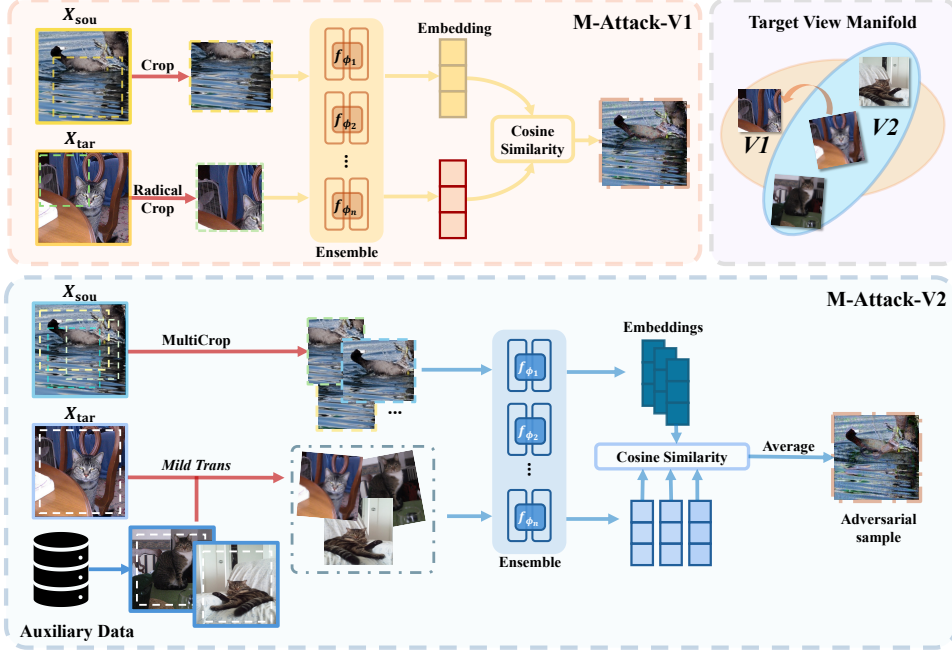


Figure 8: Comparison of one step between M-Attack and M-Attack-V2.

$\alpha$	Method	GPT-4o				Claude 3.7-thinking				Gemini 2.5-Pro			
		KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR
0.25	SSA-CWA [8]	0.08	0.08	0.04	0.10	0.06	0.03	0.00	0.03	0.06	0.03	0.00	0.01
	M-Attack [22]	0.62	0.39	0.09	0.71	0.12	0.03	0.01	0.16	0.55	0.33	0.08	0.55
	M-Attack-V2 (Ours)	0.86	0.61	0.21	0.96	0.43	0.28	0.5	0.52	0.82	0.29	0.18	0.89
0.50	SSA-CWA [8]	0.10	0.10	0.04	0.07	0.08	0.04	0.00	0.05	0.09	0.05	0.00	0.04
	M-Attack [22]	0.73	0.48	0.17	0.77	0.20	0.13	0.06	0.22	0.79	0.53	0.10	0.80
	M-Attack-V2 (Ours)	0.87	0.64	0.23	0.96	0.58	0.34	0.13	0.67	0.83	0.59	0.17	0.94
1.00	SSA-CWA [8]	0.11	0.06	0.00	0.09	0.06	0.04	0.01	0.12	0.05	0.03	0.01	0.08
	M-Attack [22]	0.82	0.54	0.13	0.95	0.31	0.21	0.04	0.37	0.81	0.57	0.15	0.83
	M-Attack-V2 (Ours)	0.92	0.77	0.42	0.98	0.55	0.36	0.08	0.67	0.85	0.73	0.22	0.98
1.275	SSA-CWA [8]	0.09	0.09	0.04	0.03	0.06	0.03	0.00	0.03	0.05	0.02	0.00	0.02
	M-Attack [22]	0.00	0.00	0.00	0.00	0.25	0.18	0.06	0.34	0.85	0.55	0.19	0.84
	M-Attack-V2 (Ours)	0.91	0.78	0.40	0.99	0.56	0.32	0.11	0.67	0.87	0.72	0.22	0.97

Table 9: Ablation study on the impact of perturbation budget ( $\alpha$ ).

## F Additional Results

### F.1 Additional Results on 1K image

We compare M-Attack and M-Attack-V2 on 1K images for better statistical stability. We changed the threshold into multiple values since no additional keywords were added for the 900 images, thus replacing the KMR with ASR with thresholds at different matching levels. Our M-Attack-V2 achieves consistently better results compared to M-Attack, showing superiority of our proposed strategy.

### F.2 Additional Results on FGSM framework

We provide the results of the I-FGSM [19] and MI-FGSM [11] under our M-Attack framework as complementary, presented in Tab. 11. Results show that even under the FGSM framework, where the patchy gradient matter is smoothed by assigning  $\text{sign}(\nabla \mathcal{L})$ , M-Attack-V2 still benefit from momentum. Moreover, MI-FGSM still provides results comparable to those of the ADAM version.

threshold	GPT-4o		Gemini-2.5-Pro		Claude-3.7-extended	
	M-Attack	M-Attack-V2	M-Attack	M-Attack-V2	M-Attack	M-Attack-V2
0.3	0.868	0.983	0.714	0.915	0.289	0.632
0.4	0.614	0.965	0.621	0.870	0.250	0.437
0.5	0.614	0.871	0.539	0.673	0.057	0.127
0.6	0.399	0.423	0.310	0.556	0.015	0.127
0.7	0.399	0.412	0.245	0.342	0.013	0.089
0.8	0.234	0.328	0.230	0.289	0.008	0.009
0.9	0.056	0.150	0.049	0.087	0.001	0.005

Table 10: Comparison of results on 1K images. We provide ASR based on different thresholds as a surrogate for KMR following M-Attack [22].

However, using PGD framework with ADAM optimizer is generally the better choice to unleash the potential of black-box attack fully since it can better explore in the space while also reducing scale issue with second-order momentum.

Method	Model	GPT-4o				Claude 3.7-extended				Gemini 2.5-Pro			
		KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR	KMR <sub>a</sub>	KMR <sub>b</sub>	KMR <sub>c</sub>	ASR
M-Attack-V2-ADAM (Ours)	Ensemble	0.91	0.78	0.40	0.99	0.56	0.32	0.11	0.67	0.87	0.72	0.22	0.97
M-Attack-V2-FGSM	Ensemble	0.85	0.64	0.19	0.98	0.40	0.26	0.08	0.46	0.83	0.65	0.17	0.90
M-Attack-V2-MIFGSM	Ensemble	0.90	0.66	0.23	0.96	0.45	0.30	0.07	0.57	0.84	0.64	0.15	0.87

Table 11: Ablation study of M-Attack-V2 under different optimizer/attack variants.

## G Visualization

### G.1 Visualization of Adversarial Samples

Fig. 9 and Fig. 10 visualize adversarial samples of different black-box attack algorithms under different perturbation constraints. Under  $\epsilon = 8$ , no significant difference exists between M-Attack and M-Attack-V2. On the  $\epsilon = 16$  setting, since our method better explores under the  $\ell_\infty$  ball, the larger  $\ell_1, \ell_2$  metric makes it slightly more apparent than M-Attack-V2. The extra ATA and MCA, along with PM, also help to extract semantic information better, thus we can see some rough shapes of cats and zebras in the background. This makes the attack more identifiable to humans. Since our M-Attack-V2 also greatly improve the results under  $\epsilon = 8$ , future directions might be improving the imperceptibility by adding constraint besides the  $\ell_\infty$

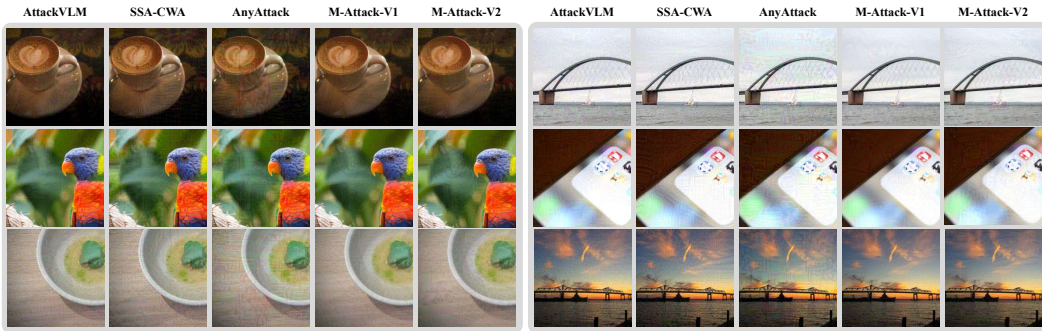


Figure 9: Visualization of adversarial samples under  $\epsilon = 8$ .

### G.2 Visualization of Reasoning Models

Fig. 11 illustrates how GPT-o3 [30] responds to our adversarial samples. The model’s visual reasoning behaviors can be broadly categorized into three types: *no reasoning* (response (d)), *simple reasoning* (responses (b) and (c)), and *zoom-in reasoning* (response (a)). Notably, in response (a),

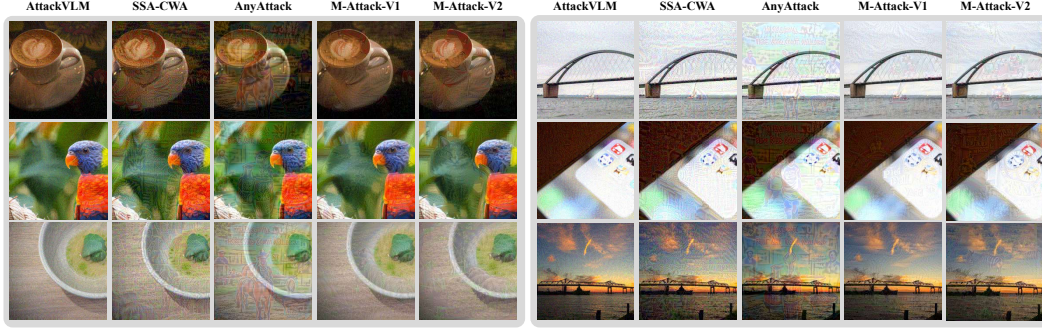


Figure 10: Visualization of adversarial samples under  $\epsilon = 16$ .

GPT-o3 already identifies the central area as uncertain and zooms in on it. However, its reasoning mechanism is not well-equipped to handle adversarial perturbations, resulting in a response that remains semantically close to the target image despite the perturbation. This observation suggests that vision reasoning offers a degree of robustness by detecting uncertainty and taking subsequent actions. During training, incorporating explicit behaviors, such as refusing to answer or flagging potential adversarial inputs, could further enhance the utility of vision-based inference under adversarial conditions.

## H Discussion

### H.1 Limitation

Despite the strong and state-of-the-art attacking performance on various closed-source MLLMs, the proposed M-Attack-V2 still relies on surrogate model ensembles and fine-grained visual alignment strategies, which may limit its applicability in extreme cases and domains where high-fidelity surrogate models or visual data are unavailable. The method also assumes some degree of consistency and diversity among surrogate model representations, which might not hold across all different architectures or domain-shifted datasets. Moreover, while the attack improves transferability, it may require slightly extra computational resources for more ensembles during optimization. Future work will explore efficiency-aware variants and more generalizable attack strategies beyond current assumptions of semantic alignment.

### H.2 Border Impact

The development of M-Attack-V2 advances our understanding of the vulnerabilities in LVLs under black-box settings, particularly in real-world, security-critical applications. By enabling fine-grained detail targeting and significantly improving attack success rates without access to model internals, this work highlights the risks posed by adversarial manipulation to commercial systems used in autonomous driving, content generation, medical imaging, etc. These insights can guide the design of more robust LVLs and encourage the community to adopt stronger evaluation protocols and defense mechanisms. Additionally, M-Attack-V2 serves as a valuable benchmark for future research on secure multimodal AI, encouraging the development of resilient architectures that are better aligned with societal safety and reliability standards.



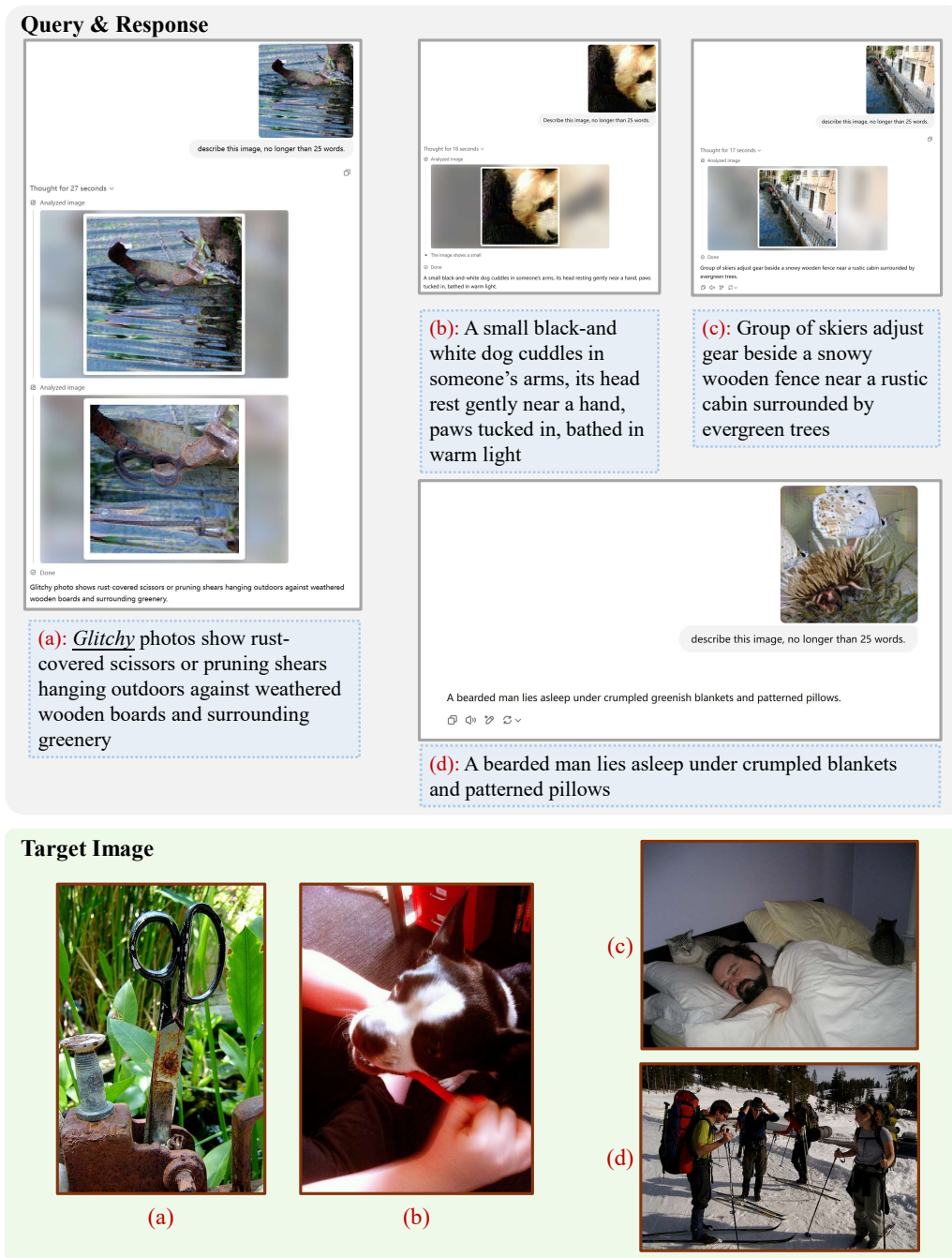


Figure 11: Visualization of GPT-o3's response towards M-Attack-V2 adversarial samples. The underlined 'glitchy' denotes that O3 notices something unusual.