
M-Attack-V2: Pushing the Frontier of Black-Box LVLM Attacks via Fine-Grained Detail Targeting

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Black-box adversarial attacks on Large Vision–Language Models (LVLMs) present
2 unique challenges due to the absence of gradient access and complex multimodal
3 decision boundaries. While prior M-Attack demonstrated notable success with
4 exceeding 90% attack success rate on GPT-4o/o1/4.5 by leveraging local crop-
5 level matching between source and target data, we show this strategy introduces
6 high-variance gradient estimates. Specifically, we empirically find that gradients
7 computed over randomly sampled local crops are nearly orthogonal, violating the
8 implicit assumption of coherent local alignment and leading to unstable optimiza-
9 tion. To address this, we propose a theoretically grounded **gradient denoising**
10 framework that redefines the adversarial objective as an expectation over local
11 transformations. Our first component, *Multi-Crop Alignment (MCA)*, estimates
12 the expected gradient by averaging gradients across diverse, independently sam-
13 pled local transformations. This manner significantly reduces gradient variance,
14 thus enhancing convergence stability. Recognizing an asymmetry in the roles of
15 source and target transformations, we also introduce *Auxiliary Target Alignment*
16 (*ATA*). ATA regularizes the optimization by aligning the adversarial example not
17 only with the primary target image but also with auxiliary samples drawn from a
18 semantically correlated distribution. This constructs a smooth semantic trajectory
19 in the embedding space, acting as a low-variance regularizer over the target distri-
20 bution. Finally, we reinterpret prior momentum as replay through the lens of local
21 matching as variance-minimizing estimators under the crop-transformed objective
22 landscape. Momentum replay stabilizes and amplifies transferable perturbations by
23 maintaining gradient directionality across local perturbation manifolds. Together,
24 MCA, ATA, momentum replay, and a delicately selected ensemble set constitute
25 M-Attack-V2, a principled framework for robust black-box LVLM attack. Empir-
26 ical results show that our framework improves the attack success rate on GPT-4o
27 (🌀) from **95%→99%**, on Claude-3.7 (🌸) from **37%→67%**, and on Gemini-2.5-
28 Pro (🔹) from **83%→97%**, significantly surpassing all existing black-box LVLM
29 attacking methods.

30 1 Introduction

31 Large Vision-Language Models (LVLMs) have become foundational to modern AI systems, enabling
32 multimodal tasks like image captioning [14, 34, 7, 37], VQA [27, 32], and visual reasoning [30].
33 However, their visual modules remain vulnerable to adversarial attacks, subtle perturbations that
34 mislead models while remaining imperceptible to humans. Prior efforts, including AttackVLM [41],
35 CWA [6], SSA-CWA [8], AdvDiffVLM [13], and most effectively, M-Attack [22], which have
36 exploited this weakness through local-level matching and surrogate model ensembles, surpassing
37 90% success rates on models like GPT-4o.

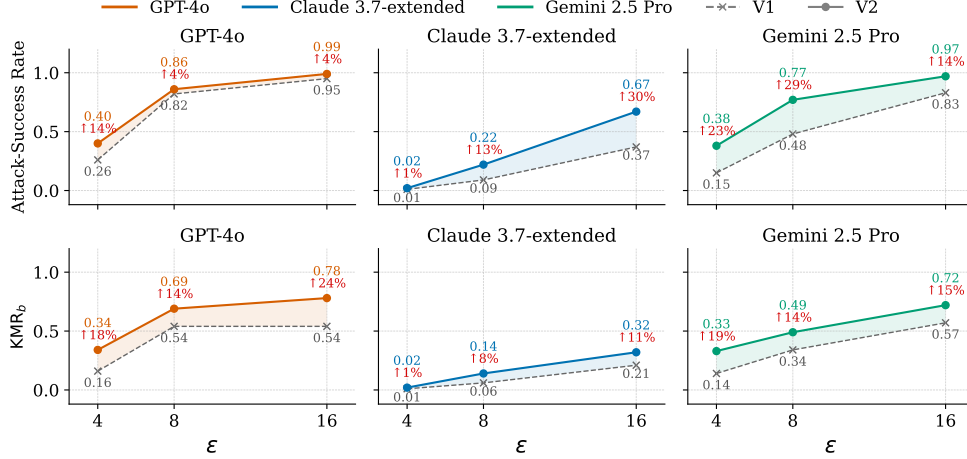


Figure 1: Improvement of M-Attack-V2 over M-Attack on up-to-date commercial black-box models.

Despite its effectiveness, our analysis reveals that M-Attack’s gradient signals are highly unstable: Even overlapping large pixel regions, two consecutive local crops share *nearly orthogonal gradients*. In other words, high similarity in pixel and embedding space does not translate to high similarity in gradient space. The reason is that ViTs’ gradient pattern is sensitive to translation. A tiny shift changes pixels contained in each token, altering self-attention. Moreover, patch-wise, spike-like gradient amplifies the mismatch within just a few pixels. We counter this effect by aggregating gradients from multiple crops within the same iteration, a strategy we call *Multi-Crop Alignment* (MCA). From a theoretical angle, MCA aggregates gradients across multiple views in a single iteration, smoothing local inconsistencies and improving cross-crop gradient stability.

We further observe that the source and target transformations in M-Attack operate in different semantic spaces: one emphasizing extraction, the other generalization. Aggressive target augmentation introduces harmful variance. Our *Auxiliary Target Alignment* (ATA) mitigates this by identifying semantically similar auxiliary images to create a low-variance embedding subspace, then applying only mild shifts to enhance transferability without destabilizing the optimization.

Classic momentum is reinterpreted under this framework as *Patch Momentum* (PM), a replay mechanism that recycles past gradients across random crops to stabilize optimization. In parallel, we also re-examine and enrich M-Attack’s model selection criterion and choose a delicately selected ensemble set with diverse patch sizes to mitigate the difficulty in cross-patch transfer, of which we find that the attention concentrates more on the main object. We term it *Patch Ensemble+* (PE⁺).

Together, these components, MCA, ATA, PM, and PE⁺, form the basis of M-Attack-V2, a robust gradient denoising framework that significantly outperforms existing black-box attack methods. Our method raises attack success rates from 95%→99% on GPT-4o, 37%→67% on Claude-3.7, and 83%→97% on Gemini-2.5-Pro, achieving state-of-the-art performance across the board. This study not only offers a practical, modular attack strategy but also sheds light on the gradient behavior of ViT-based LVLMs under local perturbations. We hope these insights will drive further research into transferable adversarial optimization under realistic black-box constraints.

2 Method

2.1 Limitations of Local Crop Matching in M-Attack

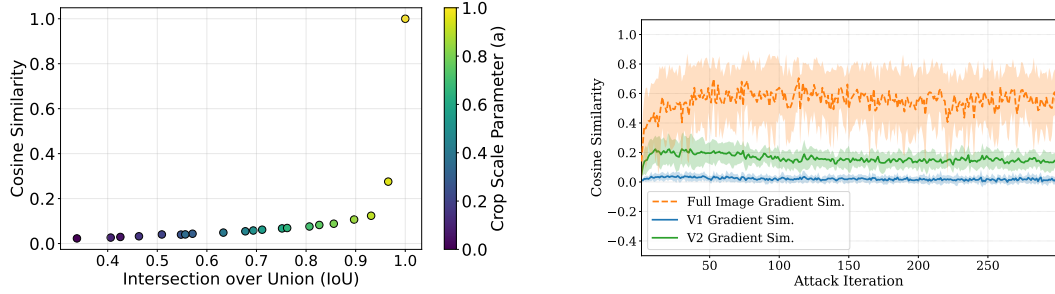
Recall local matching framework in M-Attack. To iteratively extract meaningful semantic details from the \mathbf{X}_{tar} to \mathbf{X}_{sou} (or called \mathbf{X}_{adv}), M-Attack proposes the local-level matching framework. Each region $\hat{\mathbf{x}}_i$ ($i \in \{1, 2, \dots, n\}$) is generated independently at a different training iteration i :

$$\begin{aligned} \{\hat{\mathbf{x}}_1^s, \dots, \hat{\mathbf{x}}_n^s\} &= \mathcal{T}_s(\mathbf{X}_{\text{sou}}) \\ \{\hat{\mathbf{x}}_1^t, \dots, \hat{\mathbf{x}}_n^t\} / \{\hat{\mathbf{x}}_g^t\} &= \mathcal{T}_t(\mathbf{X}_{\text{tar}}), \end{aligned} \quad (1)$$

where $\mathcal{T}_s, \mathcal{T}_t$ are the set of random local mappings and subsequent preprocessing (i.e., crops and resize) applied to the source and target images, respectively. $\hat{\mathbf{x}}_g^t$ is the globally transformed target image across iterations. Without loss of generality, each pair $\hat{\mathbf{x}}_i^s$ and $\hat{\mathbf{x}}_i^t$ is matched in iteration i .

M-Attack introduces a local-matching strategy for attacking LVLMs in the black-box setting by aligning spatial crops between source and target images. While effective, this approach suffers from inherent instability in gradient-based optimization. Formally, consider a loss function $\mathcal{L}(f(\mathcal{T}_s(\mathbf{X}_{\text{sou}})), f(\mathcal{T}_t(\mathbf{X}_{\text{tar}})))$, where \mathcal{T} denotes a random local transformation (e.g., a crop), f is the white-box model like CLIP [33], and \mathbf{X}_{sou} is the adversarial input. Because $\nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}(f(\mathcal{T}_s(\mathbf{X}_{\text{sou}})), f(\mathcal{T}_t(\mathbf{X}_{\text{tar}})))$ varies significantly across \mathcal{T}_s (same to \mathcal{T}_t), the stochastic gradients become nearly orthogonal, i.e., $\langle \nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}_{\mathcal{T}_s^i}, \nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}_{\mathcal{T}_s^j} \rangle \approx 0$, leading to high variance and poor convergence during optimization.

Extremely low gradient overlap. In M-Attack two random crops $\hat{\mathbf{x}}_i^s \subset \mathcal{T}_s(\mathbf{X}_{\text{sou}})$ and $\hat{\mathbf{x}}_i^t \subset \mathcal{T}_t(\mathbf{X}_{\text{tar}})$ are matched at every iteration. One would expect the gradients inside the shared region of two successive source crops ($\hat{\mathbf{x}}_i^s, \hat{\mathbf{x}}_{i+1}^s$) to correlate, because the underlying pixels partly coincide. Surprisingly, Fig. 2b shows the opposite: their cosine similarity is **almost zero**. We then keep one crop fixed and vary the other across scales and IoUs (Fig. 2a). Our finding reveals an exponential decay that plateaus below 0.1 once the overlap is smaller than 0.80 IoU.



(a) Similarity over IoU. The results are averaged from 20 runs with different crop parameter a for $[a, 1.0]$.

(b) Comparison of gradient similarity from full image update and local matching over each iteration

Figure 2: Similarities of gradients from different crops. a) similarity over IoU for different crops by fixing in one iteration; b) similarity between two consecutive gradients across iterations. Results are averaged from 200 runs.

Source. We find two main reasons behind this high variance: ViT’s inherent sensitivity to translation and asymmetry within the local matching framework. We discuss them below.

Patch-wise, spike-like gradient sensitive to translation. Because ViTs tokenize images on a fixed, non-overlapping grid, even sub-pixel changes each patch’s token mix. These token changes ripple through self-attention, altering weights and redirecting gradients for *all* tokens, so the resulting pixel-level gradient pattern diverges sharply. Worse, gradient magnitudes are uneven. Therefore, even similar patterns but missing a few pixels might break gradient similarity (Fig. 3b).

Asymmetric Transform Branches. In M-Attack, both the *source* and *target* images are cropped, yet playing distinct roles. Cropping the source acts directly in *pixel space*: it rearranges patch embeddings and attention weights in the forward pass, ending up with guidance of different views. By contrast, cropping the target solely translate the target representation, thereby shifting the reference embedding in *feature space*. One sculpts the perturbation, while another moves the goalpost, formulating asymmetric matching. M-Attack overlooked this and implementations target translation alternate between a *radical* crop and an identity map, struggles between explore-exploitation trade-off and potentially risk in high variance of target embedding.

Asymmetric Matching over Expectation. To mitigate the issues above, we begin by concisely reformulating the original objective function as an expectation over local transformations within an asymmetric matching framework:

$$\min_{\|\mathbf{X}_{\text{sou}}\|_p \leq \epsilon} \mathbb{E}_{\mathcal{T} \sim \mathcal{D}, y \sim \mathcal{Y}} [\mathcal{L}(f(\mathcal{T}(\mathbf{X}_{\text{sou}})), y)], \quad (2)$$

where \mathcal{D} represents the distribution of local transformations, and \mathcal{Y} denotes the distribution over target semantics. $\|\cdot\|_p$ is ℓ_p constraint for imperceptibility. Conceptually, this formulation corresponds to embedding specific semantic content y into a locally transformed area $\mathcal{T}(\mathbf{X}_{\text{sou}})$, thus highlighting

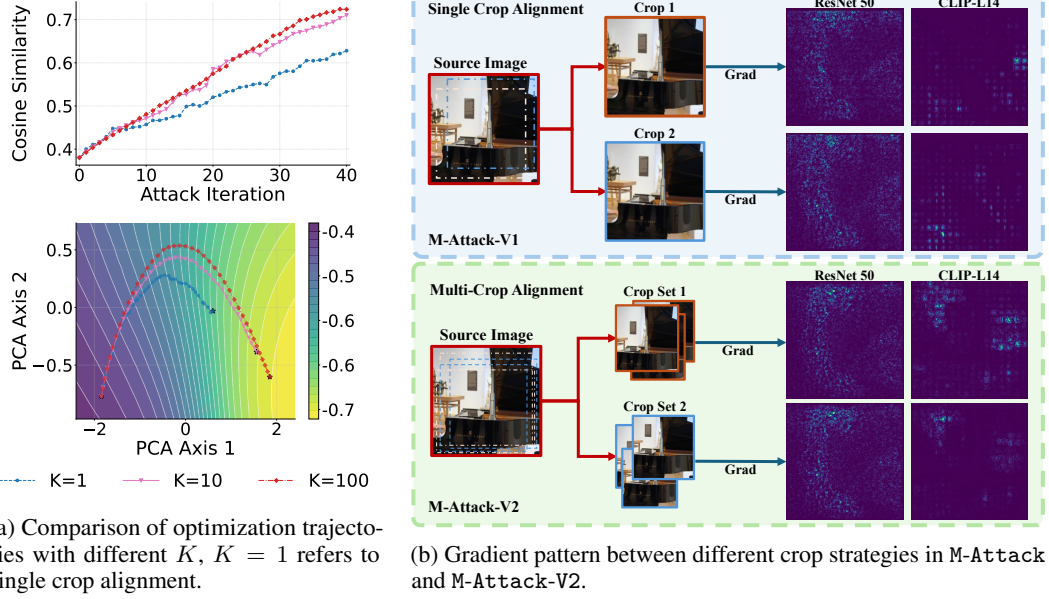


Figure 3: Comparison of: a) different trajectories against different K ; b) gradient pattern of single crop alignment against multi-crop alignment (MCA). The gradient pattern of ResNet 50 remains consistent when large pixels are overlapped, while the gradient pattern of ViTs changes dramatically. MCA helps to smooth out this impact.

the intrinsic asymmetry compared to M-Attack’s original formulation. Within this framework, our proposed enhancements, i.e., *Multi-Crop Alignment* (MCA) and *Auxiliary Target Alignment* (ATA), can be interpreted as strategies to improve the accuracy of the expectation estimation and the sampling quality of the semantic distribution \mathcal{Y} .

2.2 Gradient Denoising via Multi-Crop Alignment (MCA)

To obtain a low-variance estimate of the expected loss gradient $\mathbb{E}_{\mathcal{T} \sim \mathcal{D}, y \sim \mathcal{Y}} [\nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}(f(\mathcal{T}(\mathbf{X}_{\text{sou}})), y)]$, we draw K independent crops $\{\mathcal{T}\}_{k=1}^K$ and average their individual gradients:

$$\nabla_{\mathbf{X}_{\text{sou}}} \hat{\mathcal{L}}(\mathbf{X}_{\text{sou}}) = \frac{1}{K} \sum_{k=1}^K \nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}(f(\mathcal{T}_k(\mathbf{X}_{\text{sou}})), y). \quad (3)$$

This *Multi-Crop Alignment* (MCA) acts as an unbiased Monte-Carlo estimator, thus naturally reducing the variance with $K > 1$.

Theorem 1. Let $g_k = \nabla_{\mathbf{X}_{\text{sou}}} \mathcal{L}(f(\mathcal{T}_k(\mathbf{X}_{\text{sou}})), y)$ denote the gradient from \mathcal{T}_k , $\mu = \mathbb{E}[g_k]$, $\sigma^2 = \mathbb{E}[\|g_k - \mu\|_2^2]$ denote the mean and variance, and $p_{k\ell}$ denote the pair-wise correlation $p_{k\ell} = \frac{\langle g_k - \mu, g_\ell - \mu \rangle}{\|g_k - \mu\|^2 \|g_\ell - \mu\|^2}$. The gradient variance from K averaged crops is bounded by

$$\text{Var} \left(\frac{1}{K} \sum_{k=1}^K g_k \right) \leq \frac{\sigma^2}{K} + \frac{K-1}{K} \bar{p} \sigma^2, \quad (4)$$

where $\bar{p} = \mathbb{E}[p_{k\ell}]$, $k \neq \ell$ is the expectation of pair-wise correlation

All crops share the same underlying image, so $\bar{p} \neq 0$. The ideal σ^2/K decay is therefore tempered by the correlation term $\bar{p}\sigma^2$. Empirically, averaging a modest number ($K = 10$) of almost-orthogonal gradients still yields benefit, since the uncorrelated component of the variance shrinks as $1/K$. Simultaneously, the optimizer leverages multiple diverse transformations per update, with minimal interference among almost orthogonal gradients. Fig. 3a illustrates an accelerated convergence with $K = 10$, with margin improvement provided by $K = 100$.

This averaging also alleviates the known translation sensitivity of ViTs. As shown in Fig. 3b, using two crop sets yields noticeably higher gradient consistency than the single-crop alignment in M-Attack. In MCA, high-activity regions remain stable (upper left and center right), while the single-crop case shifts focus from center right to lower left. As a result, gradient similarity across iterations increases from near zero in M-Attack to around 0.2 (Fig. 2b).

2.3 Improved Sampling Quality via Auxiliary Target Alignment (ATA)

Selecting a representative target embedding $y \in \mathcal{Y}$ is challenging because the underlying distribution \mathcal{Y} is not observable. M-Attack mitigates this by seeding at the unaltered target embedding $f(\mathbf{X}_{\text{tar}})$ and exploring its vicinity with transformed views $f(\mathcal{T}_t(\mathbf{X}_{\text{tar}}))$ thereby sketching a locally semantic manifold that serves as a proxy for \mathcal{Y} . However, the exploration-exploitation trade-off remains problematic. *Radical* transformations leap too far, dragging y outside the genuine target region; *conservative* transformations, while semantically faithful, barely shift the embedding, leaving the optimization starved of informative signal.

To stabilize this process, we introduce P auxiliary images $\{\mathbf{X}_{\text{aux}}^{(p)}\}_{p=1}^P$ that act as additional anchors, collectively forming a richer sub-manifold of aligned embeddings. During each update, we apply a *mild* random transformation $\tilde{\mathcal{T}} \sim \tilde{\mathcal{D}}$ to every anchor, nudging the ensemble in a coherent yet restrained manner and thus providing low-variance, information-rich gradients for optimization. Let $y_0 = f(\tilde{\mathcal{T}}_0(\mathbf{X}_{\text{tar}}))$, $\tilde{y}_p = f(\tilde{\mathcal{T}}_p(\mathbf{X}_{\text{aux}}^{(p)}))$ denote sampled semantics in one iteration. The objective $\hat{\mathcal{L}}$ in Equ. (3) becomes

$$\hat{\mathcal{L}} = \frac{1}{K} \sum_{k=1}^n \left[\mathcal{L}(f(\mathcal{T}_k(\mathbf{X}_{\text{sou}})), y_0) + \frac{\lambda}{P} \sum_{p=1}^P \mathcal{L}(f(\mathcal{T}_k(x)), \tilde{y}_p) \right] \quad (5)$$

where $\lambda \in [0, 1]$ interpolates between the original target and its auxiliary neighbors. $\lambda = 0$ reduce to M-Attack local-local matching with single target. ATA trade-off exploration (auxiliary diversity) and exploitation (main-target fidelity), providing low-variance, semantics-preserving updates. The auxiliary set can be built variously, i.e., through image-image retrieval or diffusion methods.

Cost. Each iteration back-propagates through the K source crops and only forward-propagates the P auxiliary targets. Since a backward pass is roughly twice as expensive as a forward pass, the per-iteration complexity is $\mathcal{O}(K(3 + P))$, roughly twice the overhead when $P = 3$.

2.4 Patch Momentum with Built-in Replay Effect

Momentum, introduced in MI-FGSM [11], is widely adopted to improve the transferability. Define the momentum buffer as: $m_r = \beta_1 m_{r-1} + (1 - \beta_1) \nabla_{\hat{\mathbf{x}}^s} \hat{\mathcal{L}}_r(\hat{\mathbf{x}}^s)$, where $\beta_1 \in [0, 1]$ is the first-order momentum coefficient and $\nabla_{\hat{\mathbf{x}}^s} \hat{\mathcal{L}}_r(\hat{\mathbf{x}}^s)$ is our MCA-ATA-estimated gradient g_r at iteration r .

Under the local-matching view, this mechanism can be reinterpreted as formulating a streaming MCA to enforce temporal consistency across gradient directions in the space of random crops. Unrolling the EMA for pixel k exposes an alternative interpretation:

$$m_i(k) = (1 - \beta) \sum_{j=0}^i \beta^j \mathbf{1}\{k \in M_{i-j}\} g_{i-j}(k), \quad (6)$$

where M_i denotes the pixel indices included in iteration i , $m_i(k)$ and $g_i(k)$ respectively denotes momentum and gradient for pixel k . Each crop that involves pixel k is therefore replayed in future iterations with geometrically decaying weight, allowing rarely sampled regions (such as corners) to persist long enough to combat the gradient starvation. Spike-shaped gradients are further moderated by the Adam-style [18] second moment, $v_r = \beta_2 v_{r-1} + (1 - \beta_2) g_r^2$, whose scaling effect is essential in our empirical study. The momentum does not directly improve gradient similarity but continuously re-injects historical crops across patches, effectively maintaining gradient directionality across local perturbation manifolds. We therefore term it *Patch Momentum* to distinguish.

The whole procedure, combining MCA, ATA, and PM, is detailed in Alg. 1. We use a different color to differentiate between M-Attack-V2 and M-Attack. We use PGD [29] with ADAM [18] for line 13. The appendix presents analogous results for FGSM and I-FGSM variants.

Algorithm 1 M-Attack-V2

Require: clean image $\mathbf{X}_{\text{clean}}$; primary target \mathbf{X}_{tar} ; **auxiliary set** $\mathcal{A} = \{\mathbf{X}_{\text{aux}}^{(p)}\}_{p=1}^P$; **patch ensemble**⁺ $\Phi^+ = \{\phi_j\}_{j=1}^m$; iterations n , step size α , perturbation budget ϵ ; number of crops K , auxiliary weight λ ($0 \leq \lambda \leq 1$);

- 1: $\mathbf{X}_{\text{adv}} \leftarrow \mathbf{X}_{\text{clean}}$,
- 2: **for** $i = 1$ **to** n **do**
- 3: **Draw** K transforms $\{\mathcal{T}_k\}_{k=1}^K \sim \mathcal{D}$
- 4: $g \leftarrow \mathbf{0}$ \triangleright accumulate over crops
- 5: **for** $k = 1$ **to** K **do** \triangleright — crop loop —
- 6: **Draw** $\{\tilde{\mathcal{T}}_p\}_{p=0}^P \sim \tilde{\mathcal{D}}$
- 7: **for** $j = 1$ **to** m **do**
- 8: $y_0 = f(\tilde{\mathcal{T}}_p(\mathbf{X}_{\text{tar}}))$, $y_p = f(\tilde{\mathcal{T}}_p(\mathbf{X}_{\text{aux}}^{(p)}))$, $p = 1, \dots, P$ \triangleright Transform target and auxiliary data
- 9: **Compute** $\hat{\mathcal{L}}_k = (f_{\phi_j}(\mathcal{T}_k(\mathbf{X}_{\text{sou}})), y_0) + \frac{\lambda}{P} \sum_{p=1}^P \mathcal{L}(f_{\phi_j}(\mathcal{T}_k(x)), \tilde{y}_p)$
- 10: $g \leftarrow g + \frac{1}{Km} \nabla_{\mathbf{X}_{\text{sou}}} \hat{\mathcal{L}}_k$
- 11: **end for**
- 12: **end for**
- 13: Updated \mathbf{X}_{adv} based on g with **Patch Momentum**
- 14: **end for**
- 15: **return** \mathbf{X}_{adv}

3 Experiments

3.1 Experimental Setup

Metrics. We follow the evaluation protocol of M-Attack, reporting the *Attack Success Rate* (ASR) computed with *GPTScore* and the *Keywords Matching Rate* (KMR) at three thresholds $\{0.25, 0.5, 1.0\}$, denoted as KMR_a , KMR_b , and KMR_c [22]. KMR leverages human-annotated semantic keywords and measures different levels of keywords matching, treating the matching rate greater than x as a successful attack, denoting the final success rate KMR_x . The evaluation prompt and the keyword sets are identical to those in M-Attack.

Surrogate candidates. We adopt the exact surrogate selections used in their original papers for ensemble-based baselines [40, 8, 13, 22]. Our candidate pool includes CLIP series (CLIP-B/16, CLIP-B/32, CLIP[†]-G/14¹, CLIP[†]-B/32, CLIP[†]-H/14, CLIP-L/14, CLIP[†]-B/16, CLIP[†]-BG/14), DinoV2 family [31] (Dino-Small, Dino-Base, Dino-Large), and the shared vision encoder of BLIP-2 family [20]. See the appendix for more details.

Victim black-box models and dataset. We evaluate four cutting-edge commercial multimodal LLMs: GPT-4o [1], o3 [30], Claude-3.7-Sonnet-extended [3], and Gemini-2.5-Pro-Preview [36]. Clean images are drawn from the *NIPS 2017 Adversarial Attacks and Defenses Competition* dataset [17]. Following SSA-CWA [9] and M-Attack [22], we randomly sample 100 images. Auxiliary sets are retrieved from the COCO training set [23] using CLIP-B/16 embedding similarity. Further results on a 1k image subset are in the appendix.

Hyperparameters. Unless otherwise noted, perturbations are bounded by ℓ_∞ with $\epsilon = 16$ and optimized for 300 steps. We set the step size to $\alpha = 0.75$ for Claude and $\alpha = 1.0$ for all other victims, mirroring M-Attack. Our M-Attack-V2 attack utilizes, $\alpha = 1.275$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ for momentum $K = 10$, $P = 2$, and $\lambda = 0.3$ for MCA and ATA. Ablation on α is in the appendix. The target transformation $\tilde{\mathcal{T}}$ includes random resized crop ($[0.9, 1.0]$), random horizontal flip ($p = 0.5$), and random rotation ($\pm 15^\circ$).

3.2 Selection of surrogate model

Ensembling surrogate models is typical for enhancing black-box adversarial transferability. To further improve, advanced gradient aggregation methods [40, 13] are proposed, yet another practical and efficient way parallel to aggregation is to select models strategically.

We first profile the embedding transferability on different surrogate models, presented in Tab. 1. Results show that cross-model, especially cross-patchsize transfer, is difficult. Therefore, we retain models with diverse patch sizes that perform well in Tab 1. Trails of different combinations in

¹† denotes trained on LAION [35] dataset

the appendix yield our *Patch Ensemble*⁺ (PE⁺), comprising *CLIP*[†]-G/14, *CLIP*-B/16, *CLIP*-B/32, and *CLIP*[†]-B/32. Attention maps reveal a possible explanation: PE⁺ models tend to concentrate attention on the main object, whereas others exhibit dispersed focus across unrelated regions. We hypothesize that focusing on the main object enhances transferability, as all models share the common objective of identifying core semantic content. In contrast, attention to scattered regions may capture model-specific biases that do not generalize well across architectures.

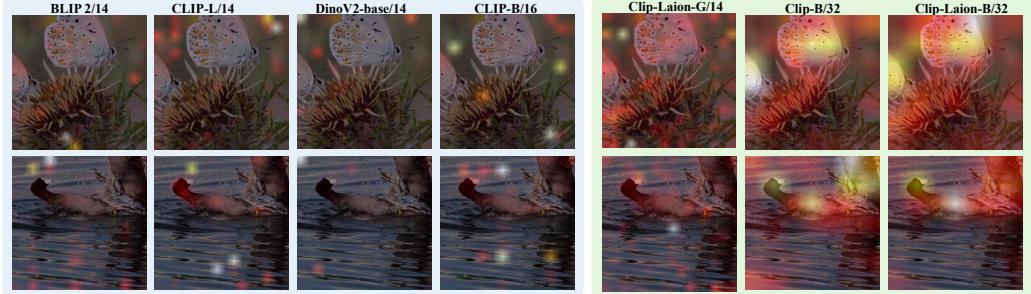


Figure 4: Comparison of two types of attention maps. Left: attention map that sparsely separates in different regions; right: attention map that focus to the main object.

| Surrogate | C-L/14 | C [†] -L/14 | D-S/14 | D-B/14 | D-L/14 | C-B/16 | C [†] -B/16 | C-B/32 | C [†] -B/32 | BLIP2 | Avg/14 | Avg/16 | Avg/32 | Avg/All |
|----------------------|--------|----------------------|--------|--------|--------|--------|----------------------|--------|----------------------|-------|--------|--------|--------|---------|
| C-L/14 | N/A | 0.40 | 0.10 | 0.13 | 0.12 | 0.45 | 0.40 | 0.34 | 0.24 | 0.48 | 0.25 | 0.42 | 0.29 | 0.30 |
| C [†] -L/14 | 0.44 | N/A | 0.24 | 0.24 | 0.21 | 0.55 | 0.57 | 0.37 | 0.33 | 0.61 | 0.35 | 0.56 | 0.35 | 0.39 |
| D-S/14 | 0.25 | 0.39 | N/A | 0.45 | 0.38 | 0.41 | 0.45 | 0.32 | 0.25 | 0.46 | 0.39 | 0.43 | 0.28 | 0.37 |
| D-B/14 | 0.29 | 0.36 | 0.33 | N/A | 0.51 | 0.37 | 0.39 | 0.31 | 0.23 | 0.47 | 0.39 | 0.38 | 0.27 | 0.36 |
| D-L/14 | 0.26 | 0.31 | 0.12 | 0.32 | N/A | 0.31 | 0.34 | 0.30 | 0.21 | 0.42 | 0.29 | 0.33 | 0.26 | 0.29 |
| C-B/16 | 0.44 | 0.43 | 0.21 | 0.18 | 0.13 | N/A | 0.53 | 0.37 | 0.27 | 0.51 | 0.32 | 0.53 | 0.32 | 0.34 |
| C [†] -B/16 | 0.43 | 0.51 | 0.22 | 0.21 | 0.15 | 0.57 | N/A | 0.39 | 0.34 | 0.52 | 0.34 | 0.57 | 0.36 | 0.37 |
| C-B/32 | 0.37 | 0.43 | 0.21 | 0.11 | 0.09 | 0.55 | 0.53 | N/A | 0.49 | 0.46 | 0.28 | 0.54 | 0.49 | 0.36 |
| C [†] -B/32 | 0.31 | 0.49 | 0.27 | 0.18 | 0.12 | 0.53 | 0.61 | 0.58 | N/A | 0.50 | 0.31 | 0.57 | 0.58 | 0.40 |
| BLIP2 | 0.39 | 0.43 | 0.15 | 0.20 | 0.26 | 0.45 | 0.43 | 0.33 | 0.25 | N/A | 0.29 | 0.44 | 0.29 | 0.32 |

Table 1: Comparison of embedding transferability over 1k images. MCA/ATA excluded to show standalone performance. C/D = CLIP/DinoV2. Gray denotes selected models.

| Method | Model | GPT-4o | | | | Claude 3.7-extended | | | | Gemini 2.5-Pro | | | | Imperceptibility | |
|--------------------|--------------------|------------------|------------------|------------------|-------------|---------------------|------------------|------------------|-------------|------------------|------------------|------------------|-------------|---------------------|---------------------|
| | | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | $\ell_1 \downarrow$ | $\ell_2 \downarrow$ |
| AttackVLM [41] | B/16 | 0.09 | 0.04 | 0.00 | 0.02 | 0.04 | 0.02 | 0.00 | 0.00 | 0.08 | 0.04 | 0.00 | 0.00 | 0.034 | 0.040 |
| | B/32 | 0.07 | 0.03 | 0.00 | 0.03 | 0.06 | 0.04 | 0.00 | 0.01 | 0.09 | 0.05 | 0.00 | 0.02 | 0.036 | 0.041 |
| | Laion [†] | 0.07 | 0.04 | 0.00 | 0.02 | 0.05 | 0.02 | 0.04 | 0.01 | 0.09 | 0.05 | 0.00 | 0.01 | 0.035 | 0.040 |
| AdvDiffVLM [13] | Ensemble | 0.02 | 0.00 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 | 0.01 | 0.03 | 0.01 | 0.00 | 0.00 | 0.064 | 0.095 |
| SSA-CWA [8] | Ensemble | 0.11 | 0.06 | 0.00 | 0.09 | 0.06 | 0.04 | 0.01 | 0.12 | 0.05 | 0.03 | 0.01 | 0.08 | 0.059 | 0.060 |
| AnyAttack [40] | Ensemble | 0.44 | 0.20 | 0.04 | 0.42 | 0.19 | 0.08 | 0.01 | 0.22 | 0.35 | 0.06 | 0.01 | 0.34 | 0.048 | 0.052 |
| M-Attack [22] | Ensemble | 0.82 | 0.54 | 0.13 | 0.95 | 0.31 | 0.21 | 0.04 | 0.37 | 0.81 | 0.57 | 0.15 | 0.83 | 0.030 | 0.036 |
| M-Attack-V2 (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 | 0.038 | 0.044 |

Table 2: Comparison of M-Attack-V2 with other black-box LVLM attack methods.

3.3 Extensive Evaluation Across LVLMs and Settings

Transferability across LVLMs. Tab. 2 illustrates the superiority of our M-Attack-V2 compared to the other black-box LVLM attack method. Our method leads others by a large margin, including M-Attack. On GPT-4o and Geimin 2.5-Pro, our M-Attack-V2 even achieves ASR close to 100%, with ASR on Claude 3.7-extended further improved by 30%, which is difficult for M-Attack to attack. Note that these improvements come with a slight increase in the perturbation norms for ℓ_1 and ℓ_2 . Previous ℓ_1 and ℓ_2 norms are caused by insufficient optimization through near-orthogonal gradients; thus, the perturbation norm only increases in a sub-linear pattern. Our M-Attack-V2 mitigates this issue, exploring more sufficiently inside the ℓ_∞ ball. Thus, it slightly increases the perturbation magnitude. See the appendix for visualizations of these adversarial samples.

Performance under budget constraints. Tab. 3 compares performance under varying perturbation budgets (ϵ). Our method consistently ranks among the top, achieving the best or second-best results across all settings. Notably, the margin is substantial when it leads, demonstrating its superior ability to explore within different ℓ_∞ balls.

Fig. 5 compares performance under different optimization budgets (total steps). Our method converges faster than M-Attack, reaching near-optimal results by 300 steps. In contrast, M-Attack continues to

improve with an additional 200 steps, indicating slower convergence. At 100 and 200 steps, M-Attack shows a significant performance drop, while our method maintains more stable ASR and KMR_b. This robustness stems from reduced variance, as M-Attack is more affected by random cropping on the source and radical transformations on the target image, requiring more iterations to stabilize.

Robustness Against Vision-Reasoning Models. We further evaluate M-Attack-V2 against GPT-o3, a model enhanced with visual reasoning capabilities. As shown in Tab. 5, GPT-o3 exhibits slightly better robustness than GPT-4o. However, the limited improvement suggests that its reasoning module is not explicitly trained to detect adversarial manipulations. Thus, even after reasoning, GPT-o3 remains susceptible to M-Attack-V2. Its reasoning process is presented in the appendix.

| ϵ | Method | GPT-4o | | | | Claude 3.7-thinking | | | | Gemini 2.5-Pro | | | | Imperceptibility | |
|------------|--------------------|------------------|------------------|------------------|-------------|---------------------|------------------|------------------|-------------|------------------|------------------|------------------|-------------|---------------------|---------------------|
| | | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | $\ell_1 \downarrow$ | $\ell_2 \downarrow$ |
| 4 | AttackVLM [41] | 0.08 | 0.04 | 0.00 | 0.02 | 0.04 | 0.01 | 0.00 | 0.00 | 0.10 | 0.04 | 0.00 | 0.01 | 0.010 | 0.011 |
| | SSA-CWA [8] | 0.05 | 0.03 | 0.00 | 0.03 | 0.04 | 0.01 | 0.00 | 0.02 | 0.04 | 0.01 | 0.00 | 0.04 | 0.015 | 0.015 |
| | AnyAttack [40] | 0.07 | 0.02 | 0.00 | 0.05 | 0.05 | 0.05 | 0.02 | 0.06 | 0.05 | 0.02 | 0.00 | 0.10 | 0.014 | 0.015 |
| | M-Attack [22] | 0.30 | 0.16 | 0.03 | 0.26 | 0.06 | 0.01 | 0.00 | 0.01 | 0.24 | 0.14 | 0.02 | 0.15 | 0.009 | 0.010 |
| | M-Attack-V2 (Ours) | 0.59 | 0.34 | 0.10 | 0.58 | 0.06 | 0.02 | 0.00 | 0.02 | 0.48 | 0.33 | 0.07 | 0.38 | 0.012 | 0.013 |
| 8 | AttackVLM [41] | 0.08 | 0.02 | 0.00 | 0.01 | 0.04 | 0.02 | 0.00 | 0.01 | 0.07 | 0.01 | 0.00 | 0.01 | 0.020 | 0.022 |
| | SSA-CWA [8] | 0.06 | 0.02 | 0.00 | 0.04 | 0.04 | 0.02 | 0.00 | 0.02 | 0.02 | 0.00 | 0.00 | 0.05 | 0.030 | 0.030 |
| | AnyAttack [40] | 0.17 | 0.06 | 0.00 | 0.13 | 0.07 | 0.07 | 0.02 | 0.05 | 0.12 | 0.04 | 0.00 | 0.13 | 0.028 | 0.029 |
| | M-Attack [22] | 0.74 | 0.50 | 0.12 | 0.82 | 0.12 | 0.06 | 0.00 | 0.09 | 0.62 | 0.34 | 0.08 | 0.48 | 0.017 | 0.020 |
| | M-Attack-V2 (Ours) | 0.87 | 0.69 | 0.20 | 0.93 | 0.23 | 0.14 | 0.02 | 0.22 | 0.72 | 0.49 | 0.21 | 0.77 | 0.023 | 0.023 |
| 16 | AttackVLM [41] | 0.08 | 0.02 | 0.00 | 0.02 | 0.01 | 0.00 | 0.00 | 0.01 | 0.03 | 0.01 | 0.00 | 0.00 | 0.036 | 0.041 |
| | 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 | 0.059 | 0.060 |
| | AnyAttack [40] | 0.44 | 0.20 | 0.04 | 0.42 | 0.19 | 0.08 | 0.01 | 0.22 | 0.35 | 0.06 | 0.01 | 0.34 | 0.048 | 0.052 |
| | 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 | 0.030 | 0.036 |
| | 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 | 0.038 | 0.044 |

Table 3: Ablation study on the impact of perturbation budget (ϵ).

| Component | | | Gemini 2.5-Pro | | | | Claude 3.7-extended | | | |
|-----------|-----|----|------------------|------------------|------------------|--------|---------------------|------------------|------------------|--------|
| MCA | ATA | PM | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR |
| | | | 0.87 | 0.72 | 0.22 | 0.97 | 0.56 | 0.32 | 0.11 | 0.67 |
| x | | | 0.85 | 0.70 | 0.21 | 0.92 | 0.52 | 0.35 | 0.08 | 0.66 |
| | | | ↓ 0.02 | ↓ 0.02 | ↓ 0.01 | ↓ 0.05 | ↓ 0.04 | ↑ 0.03 | ↓ 0.03 | ↓ 0.01 |
| | x | | 0.85 | 0.68 | 0.21 | 0.93 | 0.55 | 0.22 | 0.10 | 0.62 |
| | | | ↓ 0.02 | ↓ 0.04 | ↓ 0.01 | ↓ 0.04 | ↓ 0.01 | ↓ 0.10 | ↓ 0.01 | ↓ 0.05 |
| x | x | | 0.82 | 0.62 | 0.22 | 0.93 | 0.44 | 0.31 | 0.08 | 0.62 |
| | | | ↓ 0.05 | ↓ 0.10 | - | ↓ 0.04 | ↓ 0.12 | ↓ 0.01 | ↓ 0.03 | ↓ 0.05 |
| | | x* | 0.82 | 0.71 | 0.21 | 0.96 | 0.52 | 0.32 | 0.10 | 0.66 |
| | | | ↓ 0.05 | ↓ 0.01 | ↓ 0.01 | ↓ 0.01 | ↓ 0.04 | ↓ 0.00 | ↓ 0.01 | ↓ 0.01 |
| | | x | 0.39 | 0.23 | 0.08 | 0.35 | 0.07 | 0.03 | 0.00 | 0.08 |
| | | | ↓ 0.48 | ↓ 0.49 | ↓ 0.14 | ↓ 0.62 | ↓ 0.49 | ↓ 0.29 | ↓ 0.11 | ↓ 0.59 |

Table 4: Effect of removing each component. Numbers below each value denote the change relative to the full model (first row). \times marks the component(s) disabled. * removes *only* the first-order term.

| Model | KMR _a | KMR _b | KMR _c | ASR |
|------------------------|------------------|------------------|------------------|------|
| GPT-o3 (o3-2025-04-16) | 0.91 | 0.71 | 0.23 | 0.98 |

Table 5: Results of M-Attack-V2 on vision reasoning model

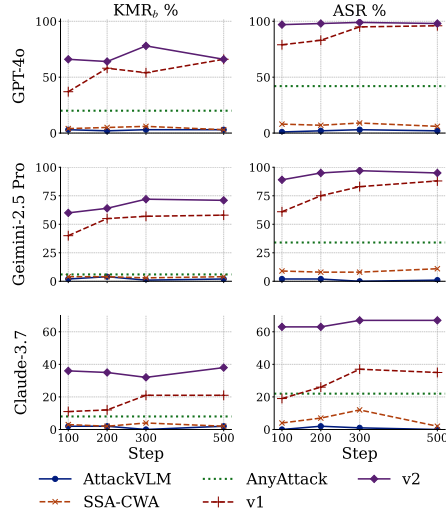


Figure 5: Comparison of different methods under different step budgets.

3.4 Ablation Study

Contribution of each component. Tab. 4 presents the performance changes when removing each component. Results on GPT-4o are excluded due to non-significant differences. Both MCA and ATA contribute approximately 5% improvement. Removing first-order momentum causes a slight drop, while eliminating both first- and second-order momentum leads to a substantial decline. This highlights the importance of second-order momentum in the PGD framework, as it helps normalize the varying gradient scales in ViTs, potentially enhancing alignment.

Hyperparameter. Fig. 6 (left) shows that transferability initially improves with increasing K , then declines. The optimal K lies around $10 \sim 20$. Moderate noise helps escape local minima and enhances transferability. However, as K grows, training becomes more stable but loses this regularizing effect. Fig. 6 (right) illustrates the impact of λ on the transferability. Larger λ provides

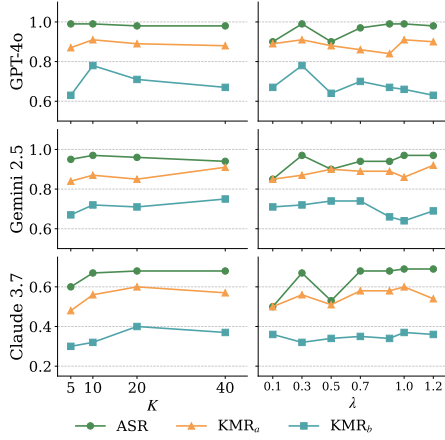


Figure 6: ASR and $\text{KMR}_a/\text{KMR}_b$ vs. different K and λ .

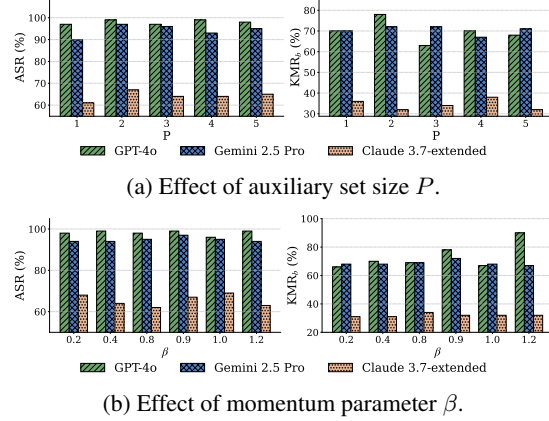


Figure 7: Ablation study on auxiliary set size P and momentum parameter β .

better diversity, drags the semantics more towards the auxiliary data, and is also at the risk of damaging the embedded semantics’ accuracy, as KMR illustrates. Fig. 7(a) and Fig. 7(b) provides the impact of P and β . Both of these factors are non-significant. Since P is related to computation complexity, choosing a smaller P , like $P = 2$ can balance the efficiency and performance. For the momentum coefficient β , using the default setting $\beta = 0.9$ yields well-balanced performance on different models and metrics.

4 Related Work

Large Vision Language Models. Transformer-based LVLMs learn visual-semantic representations from large-scale image-text data, enabling tasks like image captioning [34, 14, 7, 37], visual QA [27, 32], and cross-modal reasoning [39, 28, 38]. Open-source models such as BLIP-2 [21], Flamingo [2], and LLaVA [24] show strong benchmark performance. Commercial models like GPT-4o, Claude-3.5 [4], and Gemini-2.0 [36] offer advanced reasoning and real-world adaptability, with their successors, GPT-o3 [30], Claude 3.7-Sonnet [3], and Gemini-2.5-Pro, able to reason in the text modality and vision modality.

LVLM transfer-based attack. Black-box attacks include query-based [10, 16] and transfer-based [11, 25] methods; this work focuses on the latter. AttackVLM [41] introduced transfer-based targeted attacks on LVLMs using CLIP [33] and BLIP [21] as surrogates, showing that image-to-image feature matching outperforms cross-modal optimization, a strategy adopted by later works [6, 13, 8, 22]. CWA [6] and SSA-CWA [8] applied this principle to commercial models like Bard [36], with CWA enhancing transferability via sharpness-aware minimization [12, 5], and SSA-CWA introducing spectrum-guided augmentation via SSA [26]. AnyAttack [40] utilizes image-image matching through large-scale perturbing and a subsequent fine-tuning. AdvDiffVLM [13] embeds feature matching into diffusion guidance, introduces Adaptive Ensemble Gradient Estimation (AEGE) for smoother ensemble scores. However, M-Attack outperforms these methods by a large margin through a simple local-level matching framework and an ensemble with diverse path sizes.

5 Conclusion

We find that M-Attack suffers from unstable gradients and identify the root causes as high variance and overlooked asymmetric matching. To this end, we introduce a principled framework that includes Multi-Crop Alignment (MCA) for variance reduction, Auxiliary Target Alignment (ATA) for semantic consistency, and Patch Momentum (PM) for replay-based stabilization. Combined with a refined surrogate model ensemble (PE^+), these components form M-Attack-V2, which achieves state-of-the-art results across multiple black-box LVLMs. We hope this study provides practical insights and encourages further research into stable and transferable adversarial optimization under realistic black-box constraints.

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *International Conference on Advanced Neural Information Processing Systems*, pages 23716–23736, 2022.
- [3] Anthropic. Claude 3.7 sonnet and claude code, 2024. Accessed: 2025-02-22.
- [4] Anthropic. Introducing claude 3.5 sonnet, 2024. Accessed: 2025-02-22.
- [5] H. Chen, S. Shao, Z. Wang, Z. Shang, J. Chen, X. Ji, and X. Wu. Bootstrap generalization ability from loss landscape perspective. In *European Conference on Computer Vision*, pages 500–517, 2022.
- [6] H. Chen, Y. Zhang, Y. Dong, X. Yang, H. Su, and J. Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, 2024.
- [7] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 18030–18040, 2022.
- [8] Y. Dong, H. Chen, J. Chen, Z. Fang, X. Yang, Y. Zhang, Y. Tian, H. Su, and J. Zhu. How robust is google’s bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.
- [9] Y. Dong, H. Chen, J. Chen, Z. Fang, X. Yang, Y. Zhang, Y. Tian, H. Su, and J. Zhu. How robust is google’s bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.
- [10] Y. Dong, S. Cheng, T. Pang, H. Su, and J. Zhu. Query-efficient black-box adversarial attacks guided by a transfer-based prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9536–9548, 2021.
- [11] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li. Boosting adversarial attacks with momentum. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 9185–9193, 2018.
- [12] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.
- [13] Q. Guo, S. Pang, X. Jia, Y. Liu, and Q. Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 20:1333–1348, 2024.
- [14] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang. Scaling up vision-language pre-training for image captioning. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 17980–17989, 2022.
- [15] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021.
- [16] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146, 2018.
- [17] A. K, B. Hamner, and I. Goodfellow. Nips 2017: Defense against adversarial attack. <https://kaggle.com/competitions/nips-2017-defense-against-adversarial-attack>, 2017. Kaggle.
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.
- [19] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. 2018.

- [20] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [21] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022.
- [22] Z. Li, X. Zhao, D.-D. Wu, J. Cui, and Z. Shen. A frustratingly simple yet highly effective attack baseline: Over 90
- [23] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.
- [24] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *International Conference on Advanced Neural Information Processing Systems*, pages 34892–34916, 2023.
- [25] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.
- [26] Y. Long, Q. Zhang, B. Zeng, L. Gao, X. Liu, J. Zhang, and J. Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision*, pages 549–566, 2022.
- [27] D.-T. Luu, V.-T. Le, and D. M. Vo. Questioning, answering, and captioning for zero-shot detailed image caption. In *Asian Conference on Computer Vision*, pages 242–259, 2024.
- [28] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna. Crepe: Can vision-language foundation models reason compositionally? In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 10910–10921, 2023.
- [29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.
- [30] OpenAI. Introducing o3 and o4-mini, April 2025. OpenAI Blog.
- [31] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [32] Ö. Özdemir and E. Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *IEEE/CVF Computer Vision and Pattern Recognition Conference*, pages 1562–1571, 2024.
- [33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.
- [34] A. Salaberria, G. Azkune, O. L. de Lacalle, A. Soroa, and E. Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications*, 212:118669, 2023.
- [35] C. Schuhmann, R. Beaumont, R. Vencu, C. W. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. R. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.
- [36] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [37] M. Tschannen, M. Kumar, A. Steiner, X. Zhai, N. Houlsby, and L. Beyer. Image captioners are scalable vision learners too. In *International Conference on Advanced Neural Information Processing Systems*, pages 46830–46855, 2023.

- 372 [38] T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, and H. T. Shen. Cross-modal retrieval: a systematic
373 review of methods and future directions. *arXiv preprint arXiv:2308.14263*, 2024.
- 374 [39] J. Wu, M. Zhong, S. Xing, Z. Lai, Z. Liu, Z. Chen, W. Wang, X. Zhu, L. Lu, T. Lu, et al.
375 Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of
376 vision-language tasks. In *International Conference on Advanced Neural Information Processing*
377 *Systems*, pages 69925–69975, 2025.
- 378 [40] J. Zhang, J. Ye, X. Ma, Y. Li, Y. Yang, J. Sang, and D.-Y. Yeung. Anyattack: Towards large-scale
379 self-supervised generation of targeted adversarial examples for vision-language models. *arXiv*
380 *preprint arXiv:2410.05346*, 2024.
- 381 [41] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin. On evaluating adver-
382 sarial robustness of large vision-language models. In *International Conference on Advanced*
383 *Neural Information Processing Systems*, pages 54111–54138, 2023.

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}_{tar} ; **auxiliary set** $\mathcal{A} = \{\mathbf{X}_{\text{aux}}^{(p)}\}_{p=1}^P$; **patch ensemble**⁺ $\Phi^+ = \{\phi_j\}_{j=1}^m$; iterations n , step size α , perturbation budget ϵ ; Adam β_1, β_2, η ; number of crops K , auxiliary weight λ ;

- 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}_{adv}

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

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

- 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}_{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⁺. 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[†]-G/14), (2) patch-14 DiNOv2 (Dino-base,

415 Dino-large), (3) patch-16 CLIP (CLIP-B/16, CLIP[†]-B/16), and (4) patch-32 CLIP (CLIP-B/32,
416 CLIP[†]-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⁺, 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 _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR |
| Pair ₁ | Dino-B, Dino-S | 0.84 | 0.57 | 0.15 | 0.91 | 0.09 | 0.04 | 0.00 | 0.05 | 0.84 | 0.53 | 0.11 | 0.81 |
| Pair ₂ | L16, B/16 | 0.86 | 0.69 | <u>0.21</u> | 0.96 | <u>0.16</u> | <u>0.10</u> | <u>0.01</u> | <u>0.16</u> | 0.84 | <u>0.59</u> | <u>0.15</u> | <u>0.91</u> |
| Pair ₃ | L32, B/32 | 0.76 | 0.52 | 0.13 | 0.79 | 0.46 | 0.29 | 0.06 | 0.70 | 0.58 | 0.37 | 0.07 | 0.59 |
| Pair ₄ | G/14, L14 | 0.86 | <u>0.61</u> | 0.24 | <u>0.94</u> | 0.07 | 0.02 | 0.00 | 0.06 | <u>0.82</u> | 0.64 | 0.23 | 0.92 |

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 _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR |
| PE ₁ | B/16, B/32, L32, L16 | 0.87 | 0.65 | 0.26 | 0.99 | 0.54 | 0.32 | 0.07 | 0.68 | 0.80 | 0.57 | 0.16 | 0.90 |
| PE ₂ | Dino-B, B/32, L32, G/14 | 0.87 | 0.69 | 0.28 | 0.97 | 0.56 | 0.37 | 0.09 | 0.65 | 0.88 | 0.71 | 0.22 | 0.93 |
| PE ₃ | L16, B/32, L32, G/14 | 0.85 | 0.65 | 0.23 | 0.99 | 0.57 | <u>0.40</u> | 0.09 | 0.73 | 0.84 | 0.61 | 0.19 | 0.93 |
| PE ₄ | B/16, B/32, L32, Dino-B | 0.89 | 0.67 | 0.19 | 0.98 | 0.55 | 0.41 | 0.07 | 0.63 | 0.87 | 0.67 | 0.23 | 0.96 |
| PE ₅ | 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 |
| PE⁺ (Ours) | B/16, B/32, L32, G/14 | 0.91 | 0.78 | 0.40 | 0.99 | <u>0.56</u> | 0.32 | 0.11 | 0.67 | <u>0.87</u> | 0.72 | <u>0.22</u> | 0.97 |

Table 7: Ablation on surrogate-set selection. Each row swaps one model in or out of a four-model ensemble. The fully grey PE⁺ line is our final patch-diverse surrogate set (CLIP[†]-G/14, CLIP-B/16, CLIP-B/32, CLIP[†]-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 α 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 [†] -B/32 [15, 35] | laion/CLIP-ViT-B-32-laion2B-s34B-b79K |
| CLIP [†] -H/14 [15, 35] | laion/CLIP-ViT-H-14-laion2B-s32B-b79K |
| CLIP-L/14 [33] | openai/clip-vit-large-patch14 |
| CLIP [†] -B/16 [15, 35] | laion/CLIP-ViT-B-16-laion2B-s34B-b88K |
| CLIP [†] -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.

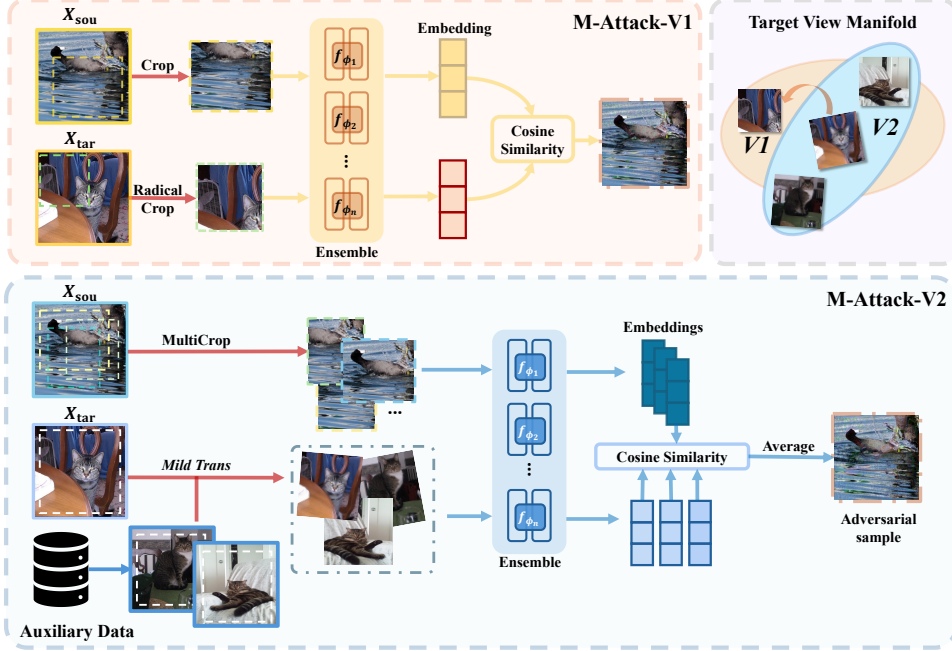


Figure 8: Comparison of one step between M-Attack and M-Attack-V2.

| α | Method | GPT-4o | | | | Claude 3.7-thinking | | | | Gemini 2.5-Pro | | | |
|----------|--------------------|------------------|------------------|------------------|------|---------------------|------------------|------------------|------|------------------|------------------|------------------|------|
| | | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | 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 (α).

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 _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | ASR | KMR _a | KMR _b | KMR _c | 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 ℓ_∞ ball, the larger ℓ_1, ℓ_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 ℓ_∞ .

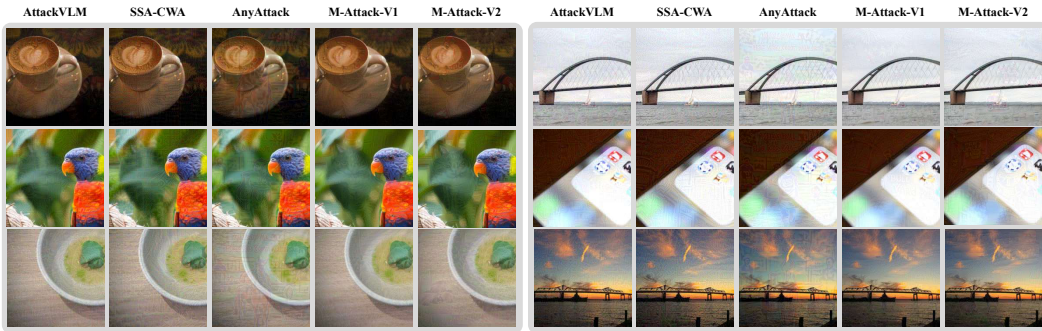


Figure 9: Visualization of adversarial samples under $\epsilon = 8$.

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),

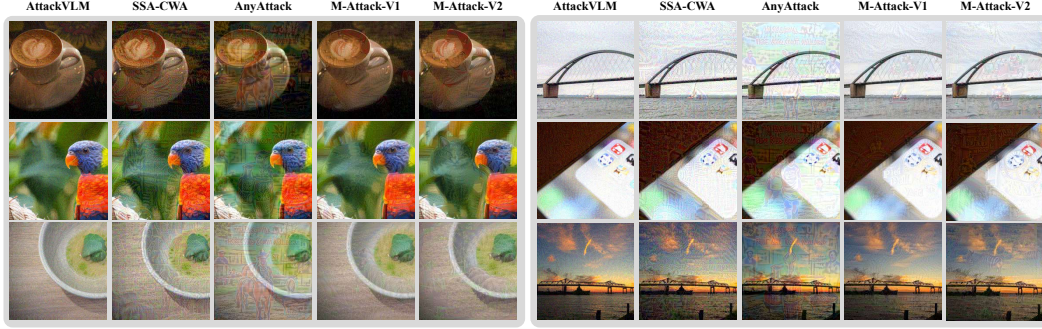


Figure 10: Visualization of adversarial samples under $\epsilon = 16$.

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.

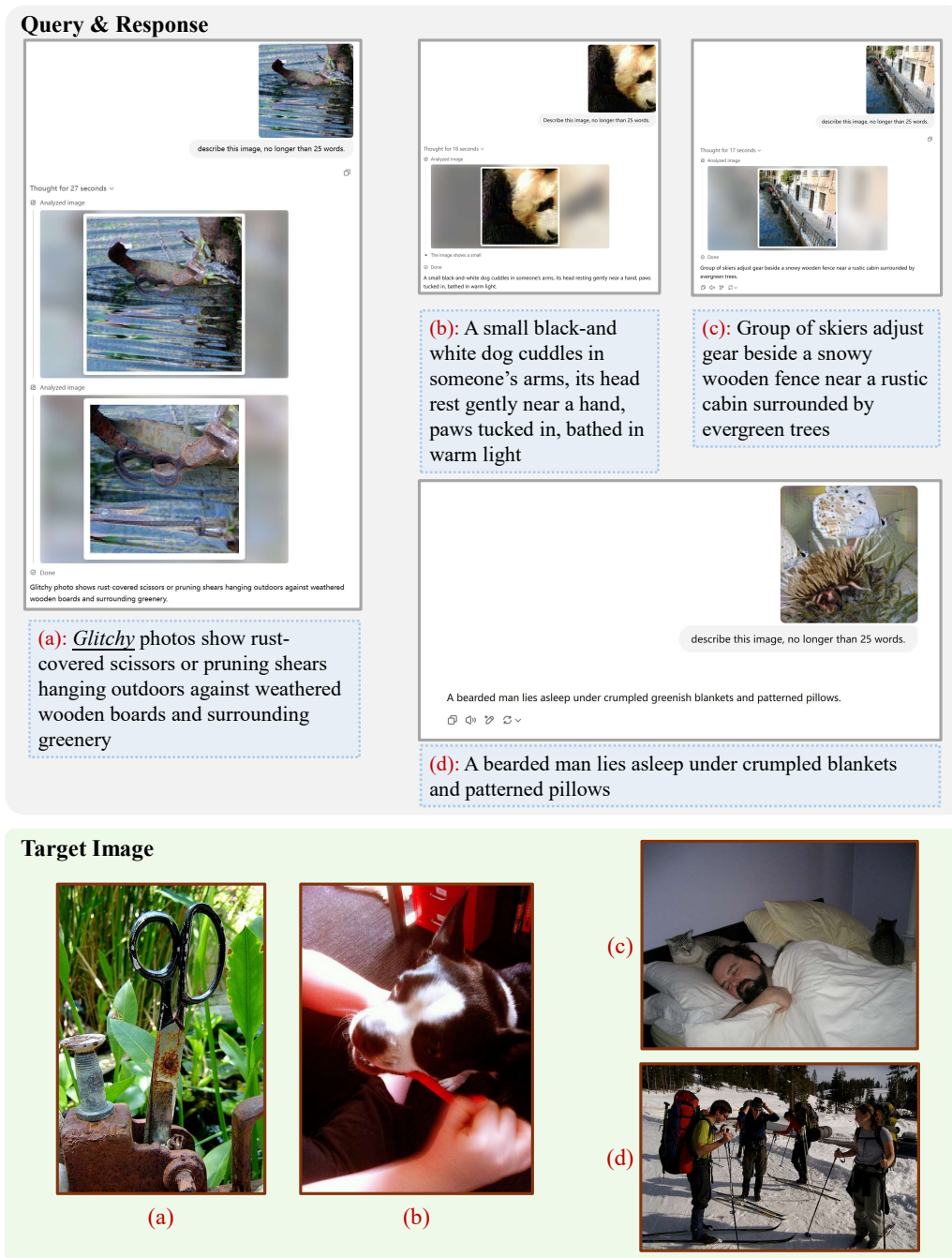


Figure 11: Visualization of GPT-o3's response towards M-Attack-V2 adversarial samples. The underlined 'glitchy' denotes that O3 notices something unusual.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The last paragraph of introduction conclude our contribution. By integrating our proposed methods, we greatly improve the results from M-Attack by a large margin, achieving a new stage of the art black-box attack method on commercial LVLMS.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: The discussion of limitations is provided in the appendix

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

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

Answer: [Yes]

Justification: The analysis part is provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

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

Answer: [Yes]

Justification: Our main experiment following M-Attack, in which the full setting of the prompt can be found. Also, we provide a full set of parameters in the experiment section, with complementary details in the appendix. The algorithm's pseudo-code clearly outlines every step of our method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

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

Answer: [Yes]

Justification: We use open access data, sampling source and target images from NIPS 2017 competition [17] and COCO validation [23], following the same procedure of M-Attack. We also include a complete code base to reproduce our results with just a few steps to set up API keys for different black-box models.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

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

Answer: [Yes]

Justification: The full details are presented in the code and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

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

Answer: [Yes]

Justification: We average across the 200 runs for the gradient similarity, thus providing error bars and corresponding statements. We set the temperature to 0 for LLM to avoid significant statistical differences.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

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

Answer: [Yes]

Justification: The appendix reports the computation resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

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

Answer: [Yes]

Justification: We have ensured the submission is anonymous and follows the rules.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

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

Answer: [Yes]

Justification: the discussion of broader impacts is located in the appendix

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

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

Answer: [No]

Justification: We have clearly stated that the generated adversarial images can only be used for detecting possible vulnerabilities in the model and robust training. However, it is still hard to prevent all the potential bad outcomes. Considering the scale of the dataset, the impact should be limited.

Justification: The paper releases an optimized dataset intended solely for academic research purposes. The dataset does not involve sensitive or high-risk content, and therefore no specific safeguards or access restrictions were implemented. The risk of misuse is considered minimal in the context of the dataset's scope and intended use.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

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

Answer: [Yes]

Justification: It is mentioned in our code.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

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

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: This paper does not include crowdsourcing

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [NA]

Justification: This paper does not involve such research

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

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

Answer: [Yes]

Justification: In evaluation, we used LLM-as-judge *only as automatic metric*, following M-Attack, which is clearly stated. We also manually check 20% of the evaluation process to ensure it correctness. We did not use LLM for other core parts, such as the originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.