Surrogate Calibration for Transferable Adversarial Attacks against Black-Box MLLMs

Zhewen Yao[1] **Yao Zhu**[2*] **Xiangyang Ji**[3] **Shiliang Zhang**[1*]
[1]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2]Zhejiang University [3]Tsinghua University
zwyao@pku.edu.cn ee_zhuy@zju.edu.cn
xyji@tsinghua.edu.cn slzhang.jdl@pku.edu.cn

Abstract

Transfer-based attacks expose serious adversarial vulnerabilities in closed-source Multimodal Large Language Models (MLLMs) by crafting adversarial examples on off-the-shelf surrogate models. However, this paradigm often drives adversarial examples toward surrogate-specific local optima, where they deceive only the surrogate model but fail to transfer to black-box MLLMs. Although existing methods mitigate this issue at the attack level, they leave the surrogate itself unchanged. To address this gap, we propose Surrogate Calibration (SCal), a novel model-level refinement that enhances downstream transfer-based adversarial attacks in a plug-and-play manner. Specifically, SCal employs a feature-preserving loss to maintain the surrogate's representational utility, while a distributional regularizer encourages its gradient field to stay closer to the natural data manifold. With calibrated surrogates, downstream attacks generate update directions that exhibit superior generalization across black-box MLLMs. Extensive experiments demonstrate that SCal boosts current attacks to new heights across various surrogates. For instance, it elevates baseline M-Attack's ASR on GPT-5.4 Mini to 53.5%, substantially surpassing the state-of-the-art MPCAttack (29.7%). These results expose critical vulnerabilities that must be addressed for trustworthy multimodal systems.
Code is at: https://github.com/zhewen-yao/SCal

1 Introduction

Multimodal Large Language Models (MLLMs) have rapidly evolved across both open-weight model families and closed-source proprietary systems, such as the Qwen-VL [1] and Gemini series [2]. Driven by their strong multimodal capabilities, these models are increasingly being integrated into real-world applications. However, recent studies have revealed that MLLMs remain vulnerable to adversarial examples, where small and visually subtle perturbations can induce attacker-chosen captions, answers, or semantic interpretations [3, 4]. Such weaknesses are transferable even to black-box commercial systems [5, 6]. Consequently, developing more potent and transferable adversarial attacks is essential to accurately delineate the security boundaries of MLLMs, especially under realistic black-box scenarios.

As shown in Figure 1(a), existing transfer-based attacks, including M-Attack [5] and FOA-Attack [6], craft adversarial examples using off-the-shelf surrogates like CLIP [7]. However, strictly enforcing feature-level alignment on these models, typically via cosine similarity maximization, inevitably drives the adversarial examples to overfit their surrogate-specific decision boundaries [5, 8]. Figure 1(c) conceptually visualizes this limitation: while an adversarial example can be perfectly

*Corresponding authors: Yao Zhu and Shiliang Zhang.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

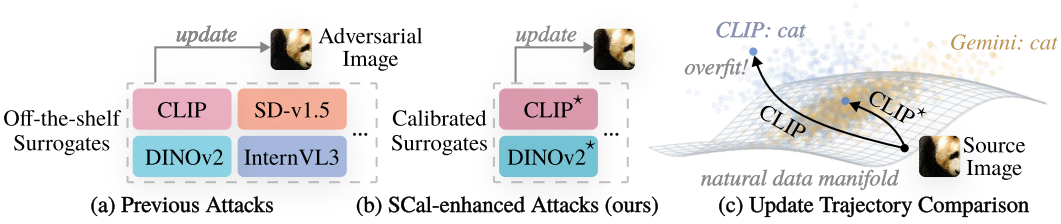

Figure 1: **Comparison with previous methods.** (a) Baseline attacks using off-the-shelf surrogate models, *e.g.*, CLIP. (b) SCal-based attacks utilize calibrated surrogates (marked by $\star$) to rectify adversarial guidance. (c) By constraining updates to the natural data manifold, SCal avoids model-specific overfitting and boosts transferability across black-box models, *e.g.*, Gemini.

optimized to deceive CLIP into predicting a panda as "cat," there is no guarantee that this misinterpretation will transfer to Gemini. Current approaches rely on surrogate ensembling to break this limitation. For example, FOA-Attack [6] introduces dynamic ensemble weights to further optimize the ensembling, whereas MPCAttack [9] aggregates surrogates trained under diverse paradigms.

This reliance on surrogate ensembling is further corroborated by Liu et al. [8]. Their work reveals that ensembling acts as an implicit regularizer, steering transferable adversarial examples into closer alignment with the natural data distribution. Since deep learning models generally exhibit more consistent behaviors on in-distribution inputs, such distribution-level alignment improves transferability to black-box MLLMs [7, 10, 11]. This motivates us to ask: *Can we empower individual surrogates to intrinsically achieve such distribution-level alignment during attack?*

To answer this question, we propose Surrogate Calibration (SCal), a model-level refinement that enhances downstream transfer-based adversarial attacks in a plug-and-play manner. Unlike prior methods that focus on optimizing the attack procedure, our approach modifies the upstream surrogate model. Figure 1(b) and (c) conceptually depict this mechanism: while off-the-shelf surrogates leave adversarial update directions unconstrained, SCal actively steers the optimization trajectory along the natural data manifold. Figure 2 further visually corroborates this alignment: while a single off-the-shelf surrogate produces highly noisy perturbations, the calibrated surrogate yields structured and visually coherent patterns, such as the outline of the building, that better reflect natural image statistics, an effect previously achieved through surrogate ensembling.

SCal achieves this distribution-level alignment by jointly optimizing two complementary objectives prior to the adversarial attack. Specifically, we introduce a feature-preserving loss to maintain the surrogate's target-discriminative utility, alongside a novel distributional regularizer that explicitly penalizes the systematic drift from the natural distribution within its gradient field. By suppressing this drift, the calibrated surrogate intrinsically generates gradient updates that better adhere to the natural data manifold, effectively mitigating surrogate-specific overfitting.

Extensive experiments demonstrate that SCal effectively enhances the transferability of diverse surrogate models and consistently improves existing attacks across multiple open-source and closed-source MLLMs. For example, SCal enables the baseline M-Attack to achieve an Attack Success Rate (ASR) of 53.5% on GPT-5.4 Mini. This significantly outperforms the current state-of-the-art MPCAttack (29.7%), which relies purely on off-the-shelf surrogates. These findings uncover previously overlooked vulnerabilities in modern MLLMs. By leveraging our pre-calibrated surrogates, high-quality attacks can be executed even with modest computational budgets. We hope our work may inspire future efforts towards more trustworthy multimodal systems.

We summarize our contributions as threefold: (i) We propose Surrogate Calibration (SCal), a novel model-level refinement that encourages alignment with the natural data manifold during downstream attacks, significantly boosting transferability in a plug-and-play manner; (ii) SCal introduces a feature-preserving loss with distributional regularization, which mitigates systematic drift and better aligns adversarial updates with the natural data manifold; and (iii) Extensive experiments demonstrate that SCal consistently elevates baseline attacks to state-of-the-art transferability across diverse MLLMs. Our calibrated surrogates expose critical vulnerabilities to inspire trustworthy multimodal systems.

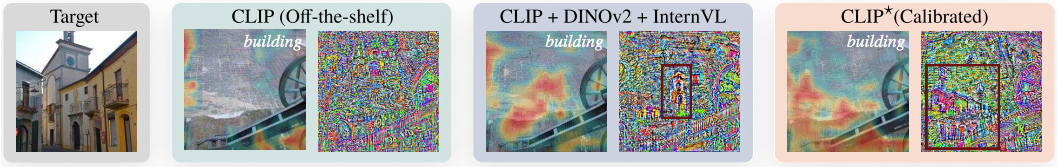

Figure 2: **Visual comparison across surrogates.** From *left* to *right*, we show the target image, the adversarial image with heatmap [12], and the amplified perturbation across different surrogate settings. All adversarial examples are constructed by M-Attack [5]. The attack that leverages a single off-the-shelf surrogate exhibits no response for the 'building', and its perturbation is relatively noisy. In contrast, those using ensemble or calibrated surrogates yield higher responses, and their corresponding perturbations better reflect natural image statistics, as highlighted by the red boxes.

2 Related work

Multimodal Large Language Models. Multimodal large language models (MLLMs) extend large language models to jointly process visual and textual inputs, enabling a wide range of tasks such as image captioning [13], visual question answering [14], and multimodal reasoning [15]. A common design paradigm is to couple a pretrained vision encoder with a large language model through a vision-language connector, which can take different forms, such as learnable-query modules, cross-attention adapters, or MLP projectors. Representative examples include Flamingo [16] and BLIP-2 [17], which employ query-based mechanisms to bridge visual representations and language models, as well as open-source MLLMs such as LLaVA [18], Qwen-VL [1], and InternVL [19], which explore different connector and alignment strategies. Beyond open-source models, commercial closed-source systems such as ChatGPT [20] and Gemini [2] have further demonstrated strong real-world multimodal capabilities. However, as MLLMs are increasingly deployed in diverse real-world applications, their visual interfaces also introduce a broader attack surface. This concern is also critical for black-box commercial systems, whose internal architectures and parameters are inaccessible to attackers and researchers. Consequently, open-source MLLMs are often used as surrogate models to craft transferable adversarial examples against closed-source targets [3, 5, 6], making transfer-based robustness a central issue in the security analysis of modern MLLMs.

Adversarial attacks on MLLMs. Recent studies have shown that MLLMs remain vulnerable to transferable adversarial examples in black-box settings. Early work such as AttackVLM [3] systematically studies targeted black-box attacks on large vision-language models by crafting perturbations on surrogate models and transferring them to downstream systems. AnyAttack [21] further improves scalability through a self-supervised pre-training-and-fine-tuning paradigm on large-scale unlabeled data, while CoA [22] enhances transferability via iterative multimodal semantic updates. More recent methods focus on stronger commercial or closed-source MLLMs. M-Attack [5] improves transferability by encoding target semantics into local regions through random crop-based matching and surrogate ensembling. FOA-Attack [6] further strengthens this line by jointly aligning global and local features and introducing dynamic ensemble weighting. MPCAttack [9] incorporates complementary representations from multiple learning paradigms for collaborative optimization, and M-Attack-V2 [23] further improves fine-grained local alignment with gradient-denoising modules such as multi-crop alignment, auxiliary target alignment, and patch momentum.

Differences with prior work. Existing attacks largely focus on how to leverage frozen off-the-shelf surrogate models to construct attacks on MLLMs, whereas our work improves transferability from an orthogonal angle by calibrating the surrogate itself before adversarial optimization.

3 Formulation

3.1 Threat model

Given a black-box MLLM $\mathcal{M}$, a source image x^{src}, and a target image x^{tgt}, the attacker aims to craft an adversarial image x^{adv} that causes the victim MLLM to produce outputs similar to those generated from the target image x^{tgt}, while remaining visually consistent with the source image x^{src}. Formally,

the attacker's ultimate goal can be formulated as the following constrained minimization problem:

$$\min_{x^{\mathrm{adv}}} \delta_{\mathcal{M}}(x^{\mathrm{adv}}, x^{\mathrm{tgt}}) \quad \text{s.t.} \quad \|x^{\mathrm{adv}} - x^{\mathrm{src}}\|_p \leq \epsilon, \tag{1}$$

where $\delta_{\mathcal{M}}(x^{\mathrm{adv}}, x^{\mathrm{tgt}})$ denotes the distance between the outputs induced by two inputs under model $\mathcal{M}$, and ϵ is the ℓ_p-norm perturbation bound. For simplicity, we absorb the text prompt into $\mathcal{M}$.

We assume a stringent black-box setting where the attacker has zero access to the victim MLLM's parameters, gradients, detailed architecture, or API feedback. Instead, the attacker assumes full white-box access only to open-source surrogate models to generate adversarial examples.

3.2 Attack objective for transferability

Feature-level alignment. As the victim $\mathcal{M}$ is inaccessible in Eq. (1), transfer-based methods typically rely on a white-box surrogate model $\mathcal{S}_\theta$. Existing works commonly instantiate $\mathcal{S}_\theta$ with a frozen, off-the-shelf image encoder, such as CLIP [7], DINOv2 [24], or the vision backbone of InternVL3 [19]. The adversarial example is then optimized by aligning its feature representation with that of the target image on this fixed surrogate $\mathcal{S}_\theta$, typically measured by cosine similarity [5, 23]:

$$\max_{x^{\mathrm{adv}}} \cos\big(\mathcal{S}_\theta(x^{\mathrm{adv}}), \mathcal{S}_\theta(x^{\mathrm{tgt}})\big) \quad \text{s.t.} \quad \|x^{\mathrm{adv}} - x^{\mathrm{src}}\|_p \leq \epsilon, \tag{2}$$

However, optimizing the cosine similarity in the feature space does not inherently entail the successful minimization of the target objective on the black-box MLLM in Eq. (1).

Distribution-level alignment. VLMs, as well as MLLMs, are well-trained on massive data sampled directly from the natural data distribution $\mathcal{D}$. Consequently, samples constrained to this shared natural manifold naturally exhibit high consistency across different models. This principle drives the increasingly adopted distribution-level alignment for transferability, formulated as:

$$\max_{x^{\mathrm{adv}}} p_{\mathcal{D}}(x^{\mathrm{adv}}) \quad \text{s.t.} \quad \|x^{\mathrm{adv}} - x^{\mathrm{src}}\|_p \leq \epsilon. \tag{3}$$

As shown in Figure 1(c), adversarial examples optimized only in the pixel space may overfit to surrogate-specific regions, so success on CLIP does not necessarily translate to success on Gemini [11]. In contrast, encouraging x^{adv} to remain close to the natural image manifold steers the optimization toward shared, model-agnostic directions, thereby improving cross-model transferability [10, 11].

However, direct optimization of the distribution-level alignment is infeasible since the exact density $p_{\mathcal{D}}$ is intractable. In the next section, we introduce how to implicitly optimize Eq. (2) and Eq. (3).

4 Surrogate Calibration

4.1 Overview

Iterative update. To solve the constrained optimization problem in Eq. (2), gradient-based iterative optimization methods are widely applied [5, 23, 6]. For a given target image x^{tgt}, the surrogate induces a target-conditioned descent field on intermediate sample x during updates:

$$v_\theta(x, x^{\mathrm{tgt}}) = -\nabla_x \cos\big(\mathcal{S}_\theta(x), \mathcal{S}_\theta(x^{\mathrm{tgt}})\big). \tag{4}$$

This vector field specifies the input-space direction that decreases the surrogate feature distance to the target image. Starting from $x = x^{\mathrm{src}}$, the iterative update is formulated as:

$$x \leftarrow \Pi_\epsilon \left(x - \alpha \cdot v_\theta(x, x^{\mathrm{tgt}})\right). \tag{5}$$

Here, α is the step size, and Π_ϵ denotes the projection onto the ϵ-bounded ℓ_p ball centered at x^{src}. After a predefined number of iterations, the final state of x is output as the adversarial example $x^{\mathrm{adv}} = x$.

Calibration objective. To achieve feature- and distribution-level alignment during the adversarial update in Eq. (5), we refine the initial surrogate $\mathcal{S}_\theta$ into a calibrated state $\mathcal{S}_{\theta^\star}$ beforehand. As shown in Figure 3, this calibrated surrogate is obtained by minimizing a joint objective that targets both alignment goals, respectively:

$$\min_{\theta^\star} \big[\lambda \cdot \mathcal{L}_{\mathrm{feat}}(\theta^\star) + \mathcal{L}_{\mathrm{distri}}(\theta^\star)\big], \tag{6}$$

where λ balances the two terms. Below, we detail the formulations of these two objectives.

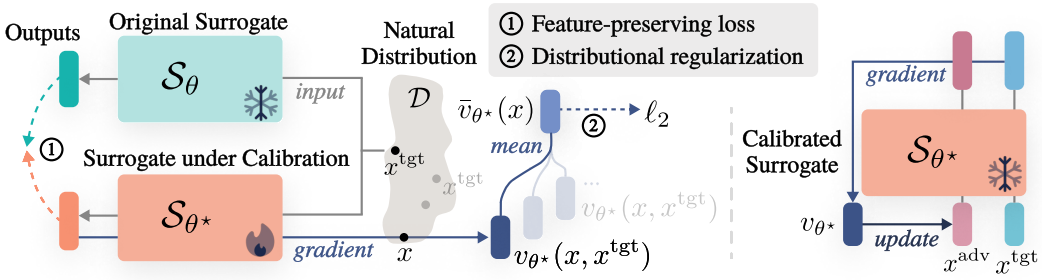

Figure 3: **SCal overview.** (*Left*) SCal jointly optimizes a feature-preserving loss to maintain representational utility and a distributional regularizer to prevent drift from the natural data distribution. (*Right*) An iterative adversarial update is then performed using the calibrated gradients.

4.2 Feature-preserving loss

The feature-preserving loss prevents $\mathcal{S}_{\theta^\star}$ from deviating significantly from the original surrogate $\mathcal{S}_\theta$:

$$\mathcal{L}_{\text{feat}}(\theta^\star) = \mathbb{E}_{x\sim\mathcal{D}}\Big[1 - \cos\big(\mathcal{S}_\theta(x), \mathcal{S}_{\theta^\star}(x)\big)\Big]. \tag{7}$$

The minimization of this term enforces consistency between the outputs of $\mathcal{S}_\theta$ and $\mathcal{S}_{\theta^\star}$, which preserves the calibrated surrogate as a meaningful feature-distance estimator for downstream attacks.

4.3 Distributional regularization

Intuition. We decompose the target-conditioned descent vector in Eq. (4) as follows:

$$v_{\theta^\star}(x, x^{\text{tgt}}) = \underbrace{\bar{v}_{\theta^\star}(x)}_{\text{systematic drift}} + \underbrace{r_{\theta^\star}(x, x^{\text{tgt}})}_{\text{target residual}}, \qquad \bar{v}_{\theta^\star}(x) = \mathbb{E}_{x^{\text{tgt}}\sim\mathcal{D}}\left[v_{\theta^\star}(x, x^{\text{tgt}})\right]. \tag{8}$$

Here, $\bar{v}_{\theta^\star}(x)$ is a target-marginalized vector, which we view as the systematic drift during the iterative update process. In contrast, the target residual $r_{\theta^\star}(x, x^{\text{tgt}})$ is the vector that genuinely provides the adversarial signal. To make the calibrated surrogate more effective for downstream transferable attacks on unseen targets, we seek to mitigate the distributional displacement induced by this systematic drift $\bar{v}_{\theta^\star}(x)$.

Regularizer. Our distributional regularizer penalizes the squared ℓ_2 norm of the systematic drift:

$$\mathcal{L}_{\text{distri}}(\theta^\star) = \mathbb{E}_{x\sim\mathcal{D}}\left[\|\bar{v}_{\theta^\star}(x)\|_2^2\right], \tag{9}$$

The regularizer in Eq. (9) controls the distributional displacement caused by the systematic drift. To see this, we define the drifted sample via a push-forward distribution $\mathcal{D}_{\theta^\star,\alpha}$, where $\tilde{x}$ is obtained by:

$$\tilde{x} = x - \alpha\,\bar{v}_{\theta^\star}(x), \quad x\sim\mathcal{D}, \quad \tilde{x}\sim\mathcal{D}_{\theta^\star,\alpha}. \tag{10}$$

We use the Wasserstein distance [25] between the natural distribution and the drifted distribution to characterize this drift. Based on the push-forward map in Eq. (10), the joint distribution of $(\tilde{x}, x)$ forms a valid coupling between distribution $\mathcal{D}_{\theta^\star,\alpha}$ and $\mathcal{D}$. Therefore, the Wasserstein distance can be bounded as:

$$W_2(\mathcal{D}_{\theta^\star,\alpha}, \mathcal{D}) \le \sqrt{\mathbb{E}_{x\sim\mathcal{D}}\left[\|\tilde{x} - x\|_2^2\right]} = \alpha\sqrt{\mathbb{E}_{x\sim\mathcal{D}}\left[\|\bar{v}_{\theta^\star}(x)\|_2^2\right]} = \alpha\sqrt{\mathcal{L}_{\text{distri}}(\theta^\star)}. \tag{11}$$

Thus, the distributional regularizer $\mathcal{L}_{\text{distri}}$ controls an upper bound on the distributional displacement induced by the systematic drift. By suppressing this drift, the calibrated surrogate is encouraged to produce updates that remain closer to the natural data distribution, reducing the influence of model-specific artifacts and thereby better supporting downstream transferable attacks.

In practice, the expectations over $\mathcal{D}$ are estimated using a finite calibration set $\mathcal{C}$. At each iteration, we sample K natural targets uniformly from $\mathcal{C}$ and compute the empirical average:

$$\bar{v}_{\theta^\star}(x) \approx \frac{1}{K}\sum_{k=1}^{K} v_{\theta^\star}(x, x_k^{\text{tgt}}), \qquad x_k^{\text{tgt}} \sim \mathcal{C}. \tag{12}$$

Table 1: **Performance of different attacks on the NIPS 2017 Competition dataset [26].** We report Attack Success Rate (ASR) and average semantic similarity to the target image (AvgSim). For attacks enhanced with SCal, the second line in each entry shows the absolute gain over the original attack.

Attack	Surrogate Size	*Closed-Source Black-Box Models*							
		GPT-5.4 Mini		Claude Opus 4.7		Gemini 3.1		Grok 4.20	
		ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
CoA [22]	151M	0.0%	0.042	0.0%	0.084	0.0%	0.081	0.1%	0.096
M-Attack [5]	262M	23.2%	0.350	63.8%	0.619	14.6%	0.355	61.3%	0.611
FOA-Attack [6]	262M	24.9%	0.376	71.1%	0.662	15.0%	0.357	65.4%	0.641
M-Attack-V2 [23]	262M	28.8%	0.427	72.6%	0.654	23.9%	0.405	69.0%	0.662
MPCAttack [9]	1390M	29.7%	0.431	84.7%	0.716	26.3%	0.412	78.2%	0.680
M-Attack + SCal	262M	53.5% (+30.3%)	0.552 (+0.202)	87.2% (+23.4%)	0.724 (+0.105)	27.8% (+13.2%)	0.432 (+0.077)	80.6% (+19.3%)	0.689 (+0.078)
M-Attack-V2 + SCal	262M	**56.1%** (+27.3%)	**0.572** (+0.145)	**89.4%** (+16.8%)	**0.738** (+0.084)	**30.1%** (+6.2%)	**0.454** (+0.049)	**82.7%** (+13.7%)	**0.711** (+0.049)

Attack	Surrogate Size	*Open-Source Black-Box Models*							
		Llama-3.2-11B		Phi-3.5-Vis-4.2B		Qwen3-VL-8B		InternVL3.5-8B	
		ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
CoA [22]	151M	0.0%	0.074	0.1%	0.102	0.7%	0.125	0.6%	0.119
M-Attack [5]	262M	46.1%	0.491	51.4%	0.548	63.6%	0.615	63.4%	0.613
FOA-Attack [6]	262M	50.8%	0.529	50.6%	0.543	72.8%	0.663	68.9%	0.620
M-Attack-V2 [23]	262M	62.8%	0.593	56.3%	0.570	75.7%	0.675	75.4%	0.675
MPCAttack [9]	1390M	61.0%	0.588	60.4%	0.587	84.9%	0.713	73.6%	0.653
M-Attack + SCal	262M	79.5% (+33.4%)	0.681 (+0.190)	73.2% (+21.8%)	0.649 (+0.101)	90.3% (+26.7%)	0.735 (+0.120)	79.3% (+15.9%)	0.685 (+0.072)
M-Attack-V2 + SCal	262M	**82.9%** (+20.1%)	**0.703** (+0.110)	**76.4%** (+20.1%)	**0.669** (+0.099)	**92.9%** (+17.2%)	**0.744** (+0.069)	**85.5%** (+10.1%)	**0.722** (+0.047)

The detailed algorithmic procedure is provided in Appendix B.1.

5 Experiments

5.1 Experimental setup

Datasets. We generate adversarial examples on the NIPS 2017 Adversarial Attacks and Defenses Competition [26] dataset, using samples selected from the MS-COCO [27] validation set as target images, following the evaluation protocols of previous MLLM attacks [5, 23, 6, 9]. Furthermore, we utilize MS-COCO images as the source for our attacks to investigate a more challenging scenario.

Baselines. We select seven transfer-based attacks on MLLMs. For ℓ_p-norm bounded attacks, we select five representative baselines: CoA [22], M-Attack [5], M-Attack-V2 [23], FOA-Attack [6] and MPCAttack [9]. For unrestricted adversarial attacks, we select AdvDiffVLM [28] and PSI [11]. Note that we exclude earlier methods such as AttackVLM [3] and AnyAttack [21], as recent literature has extensively demonstrated their limited transferability when applied to modern MLLMs [5, 6].

Victim and surrogate models. We evaluate our method on eight representative multimodal victim models, including four proprietary systems: GPT-5.4 Mini [29], Gemini 3.1 [30], Claude Opus 4.7 [31], and Grok 4.20 [32]; and four open-source MLLMs: Llama-3.2-11B-Vision [33], Phi-3.5-Vision-4.2B [34], Qwen3-VL-8B [35], and InternVL3.5-8B [36]. We provide exact model IDs in Appendix C. Following Li et al. [5], we use three CLIP vision encoders [7] as surrogate models by default: OpenAI CLIP ViT-B/16, OpenAI CLIP ViT-B/32, and LAION/OpenCLIP ViT-B/32. For CoA [22], we follow its single-encoder design and use a ViT-B/32 CLIP visual encoder together with ClipCap [37] for image-to-text bridging. MPCAttack [9] further introduces an InternVL-based text-anchor generator and the corresponding CLIP text encoder to align textual anchors with image features.

Evaluation metrics. Following previous works [5, 9], we evaluate transferability using the attack success rate (ASR). Specifically, we adopt LLM-as-a-judge [38] with Qwen3-8B to assess the semantic similarity between the textual outputs generated from each adversarial example and its

Table 2: **Attack performance comparison with different surrogate models.** We report the Attack Success Rate (ASR) and the average semantic similarity to the target image (AvgSim). M-Attack [5] serves as the baseline, and all experiments are conducted on the NIPS 2017 Competition dataset [26].

Surrogate	*Closed-Source Black-Box Models*						*Open-Source Black-Box Models*			
	GPT-5.4 Mini		Claude Opus 4.7		Grok 4.20		Llama-3.2-11B		Phi-3.5-Vis-4.2B	
	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
CLIP B/16 [7]	4.0%	0.207	8.5%	0.233	13.8%	0.304	11.9%	0.284	13.5%	0.261
+SCal	19.9%	0.357	34.2%	0.449	39.4%	0.475	55.0%	0.549	43.2%	0.486
DINOv3 B/16 [41]	5.5%	0.189	2.7%	0.156	4.6%	0.189	5.4%	0.168	4.4%	0.174
+SCal	9.5%	0.288	6.8%	0.215	13.8%	0.301	21.0%	0.313	18.6%	0.298
Qwen3-VL-8B [35]	9.1%	0.277	5.2%	0.217	36.9%	0.455	19.0%	0.327	16.7%	0.314
+SCal	23.3%	0.366	26.5%	0.382	42.1%	0.511	36.8%	0.464	26.2%	0.381
InternVL3.5-8B [36]	3.6%	0.229	0.9%	0.163	4.6%	0.213	21.0%	0.351	20.1%	0.350
+SCal	5.9%	0.254	6.3%	0.219	14.2%	0.313	36.3%	0.468	27.6%	0.410

corresponding target example. The ASR is then defined as the proportion of cases where the similarity score is greater than or equal to 0.6. Detailed prompts are provided in Appendix C. For unrestricted attacks, we use SSIM [39] and LPIPS [40] to quantify imperceptibility.

Implementation details. SCal is applied in a plug-and-play manner: it only replaces the surrogate model with its calibrated version, while leaving all other settings strictly unchanged. By default, SCal is calibrated in-domain, using the NIPS 2017 development set and the MS-COCO training split as source and target. SCal uses $\lambda = 0.06$ and $K = 128$ for weighted feature preservation and distributional regularization, respectively. Calibrating a single surrogate takes less than 60 minutes on a single RTX 3090 GPU with 24 GB memory, and the resulting calibrated surrogate can be readily used to improve downstream attacks. All downstream attacks follow the configurations specified in their original papers, with the perturbation budget set to $\epsilon = 16/255$ under the ℓ_∞ constraint.

5.2 Main results

Comparison with prior attacks. Table 1 presents the attack performance comparison on the NIPS 2017 benchmark. Overall, our SCal consistently improves existing attacks across all evaluated victim models, encompassing both open-source and closed-source black-box settings. Specifically, while the uncalibrated M-Attack [5] is clearly outperformed by the state-of-the-art MPCAttack [9], which uses more surrogates, applying SCal completely reverses this dynamic, enabling M-Attack to match or significantly exceed it. Moreover, integrating SCal with M-Attack-V2 [23] further boosts transferability, comprehensively outperforming current attacks that rely on off-the-shelf surrogates.

Beyond achieving stronger overall performance than previous methods, SCal-enhanced attacks exhibit significantly more consistent cross-victim performance. For example, against open-source black-box models, SCal consistently maintains an Attack Success Rate (ASR) of approximately 80%, whereas baseline attacks demonstrate considerable instability. This disparity extends to closed-source systems; even against GPT-5.4 Mini, baseline methods suffer from exceptionally weak transferability (below 30% ASR), whereas SCal substantially boosts the success rate to 56.1%. This pattern likely arises because surrogate models are rarely equally well-matched to different victims. Because such compatibility cannot be assumed a priori, surrogate calibration offers a fundamentally more robust approach to improving transferability across diverse victim architectures.

Gemini 3.1 exhibits a distinct failure mode. In most unsuccessful cases, rather than reverting to a description of the source image, it abstains from producing a concrete description and instead indicates that the image is difficult to interpret or does not appear to be a clean image. For example, in Figure 4, Gemini 3.1 Pro recognizes the deceptive nature of our adversarial example by calling it a 'truck-like illusion,' unlike GPT-5.4 Pro, which is fooled into predicting it is a truck. Overall, this example demonstrates that SCal can expose severe adversarial vulnerabilities in security-critical scenarios. Furthermore, Gemini's response indicates that reinforcing abstention mechanisms constitutes a desirable defensive strategy against adversarial manipulation.

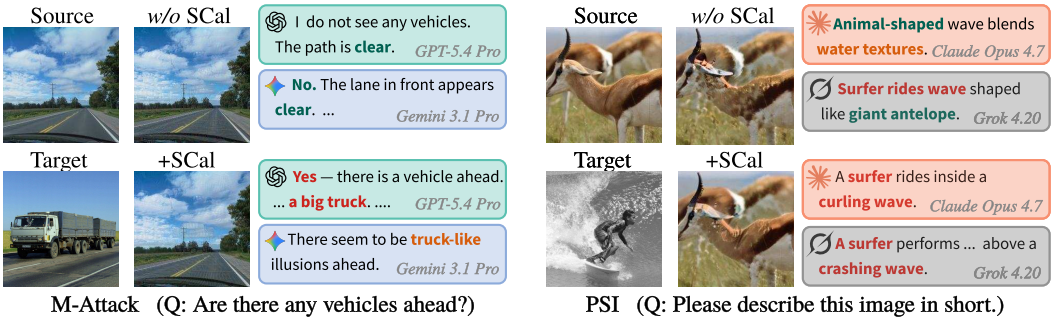

Figure 4: **Adversarial examples with and without SCal.** We present examples of SCal applied to an ℓ_p-bounded attack, M-Attack [5], and an unrestricted attack, PSI [11], alongside the corresponding responses from four commercial MLLM systems. Clear indicators of successful and unsuccessful attacks are highlighted in red and green, respectively. Orange indicates ambiguous success signals.

Table 3: **SCal boosts unrestricted attacks.** Performance is averaged over eight MLLMs.

Attack	ASR ↑	SSIM ↑	LPIPS ↓
AdvDiffVLM [28]	3.73%	0.831	0.201
+SCal	6.51%	0.835	0.198
PSI [11]	49.47%	0.809	0.221
+SCal	54.25%	0.817	0.219

Table 4: **SCal under domain shift.** Performance is averaged over eight victim MLLMs.

Setting	ASR ↑	AvgSim ↑
M-Attack [5] (*w/o* SCal)	48.43%	0.525
+ In-Domain SCal	71.43%	0.643
+ OOD SCal	63.11%	0.607
+ OOD *w/* adaptation	71.85%	0.654

Effectiveness under different surrogates. As shown in Table 2, SCal consistently enhances transferability across all evaluated surrogate architectures. This includes generic visual encoders, such as CLIP ViT-B/16 and DINOv3 ViT-B/16, as well as the integrated vision towers of Qwen3-VL-8B and InternVL3.5-8B. This demonstrates that although the vision components are optimized for semantic understanding, they are not inherently optimal as baseline surrogates for adversarial attacks.

Effectiveness on unrestricted attacks. We further apply SCal to unrestricted attacks. Since these attacks are not subject to strict perturbation constraints, we adaptively adjust the perturbation magnitude so that the stealthiness metrics remain comparable with and without SCal. As shown in Table 3, SCal can also enhance unrestricted attacks without sacrificing stealthiness. However, the improvement is not as high as that of ℓ_p-bounded attacks, likely because it is affected by additional denoising steps during the iteration process.

Robustness under domain shift. We perform calibration on ImageNet [42] for both source and target to simulate a domain shift. As shown in Table 4, this out-of-distribution (OOD) setting does not lead to a substantial reduction in attack effectiveness, with ASR still substantially outperforming the M-Attack baseline. When an input sample is OOD, we further apply one additional low-cost calibration step to adapt to that sample, which restores attack performance to a level comparable to in-domain SCal. We analyze the time cost of this low-cost adaptation in Appendix C.

Robustness against defenses. Table 5 evaluates the resilience of various attacks against standard defense mechanisms, namely JPEG compression (quality 85%), Gaussian blur (kernel size $= 5$, $\sigma = 1$), and DiffPure ($t = 10$). As demonstrated, SCal-enhanced attacks not only achieve consistently higher Attack Success Rates (ASR) but also exhibit superior robustness to these countermeasures, with notable resilience against JPEG and DiffPure. This is likely because SCal encourages adversarial examples to align more closely with the natural image distribution, which is an inherently robust domain. Consequently, SCal-enhanced adversarial images may remain effective despite common input transformations, underscoring their practical threat in real-world scenarios.

5.3 Empirical analysis

Component ablation. Table 6 ablates the specific roles of feature-preserving loss and distributional regularization within our calibration framework. We observe that relying solely on distributional regularization leads to a catastrophic collapse of the surrogate model, where the classification

Table 5: **Performance under defenses.** ASR (↑) is averaged across eight victim MLLMs. Subscripts denote relative ASR drops *w.r.t.* the undefended ASR.

Defense	M-Attack	MPCAttack	M-Attack + SCal
JPEG [43]	42.1%$_{\downarrow 13.1\%}$	53.3%$_{\downarrow 14.5\%}$	69.3%$_{\downarrow 3.0\%}$
Gaussian [44]	26.5%$_{\downarrow 45.3\%}$	33.8%$_{\downarrow 45.8\%}$	55.1%$_{\downarrow 22.9\%}$
DiffPure [45]	36.7%$_{\downarrow 24.2\%}$	49.6%$_{\downarrow 20.4\%}$	66.8%$_{\downarrow 6.5\%}$

Table 6: **Component analysis.** Results are averaged over eight victim MLLMs, with the surrogate's zero-shot accuracy on ImageNet.

Setting	ASR ↑	ACC
M-Attack [5] (*w/o* SCal)	48.43%	68.3%
+ $\mathcal{L}_{\text{distri}}$	3.4%	32.1%
+ $\mathcal{L}_{\text{distri}}$ + $\mathcal{L}_{\text{feat}}$	71.43%	65.7%

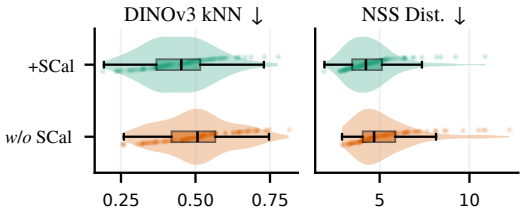

Figure 5: **Out-of-distribution (OOD) score.** M-Attack with SCal yields lower OOD scores.

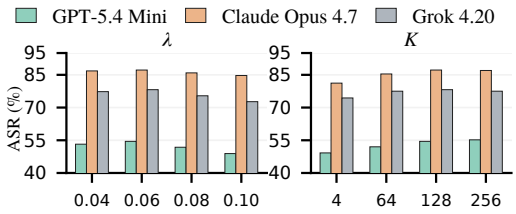

Figure 6: **Hyperparameter sensitivity.** We vary the choice of λ and K on three victim MLLMs.

accuracy (ACC) plummets to 32%. This degradation signifies a near-total loss of discriminatory power, which consequently nullifies the attack success rate (ASR). In contrast, the inclusion of the feature-preserving constraint is pivotal for restoring and significantly bolstering ASR. Notably, these results reveal a non-trivial decoupling between a surrogate's representation quality and its effectiveness for adversarial transfer: a calibrated surrogate with reduced ACC can yield substantially higher ASR. This trade-off constitutes the fundamental intuition behind SCal.

Formulation validation. We assess distribution-level alignment in the generated adversarial examples. In Figure 5, we show that SCal-enhanced attacks result in lower kNN distances on held-out samples [46], which measure the proximity to the empirical natural-image manifold. Furthermore, they exhibit lower natural scene statistics (NSS) scores [47], which reflect better consistency with the statistical regularities of natural images. These results demonstrate that SCal renders adversarial examples less out-of-distribution, and indeed better achieves our formulated distribution-level alignment.

Hyperparameter choice. Figure 6 illustrates the performance variations under different hyperparameter settings. As observed, setting the feature-preserving weight λ to 0.06 yields optimal results across all models, and the performance remains highly stable within the range of 0.04 to 0.10. Regarding the sample size K used to calculate the mean descent vector for distributional regularization, we find that the performance saturates starting at $K = 128$. Consequently, drawing more samples is unnecessary, which guarantees high computational efficiency. Notably, a significant performance drop occurs when $K = 4$. This degradation stems from the fact that the sampled mean descent vector is insufficiently marginalized. Nevertheless, it still provides gains by implicitly optimizing the true mean descent vector. To avoid unnecessary complexity, we apply this unified set of hyperparameters when calibrating all surrogate models.

6 Conclusion

We propose Surrogate Calibration (SCal), a model-level refinement that strengthens downstream transfer-based adversarial attacks in a plug-and-play manner. SCal uses a feature-preserving objective to retain the surrogate's representational utility, while a distributional regularizer encourages its gradient field to stay closer to the natural data manifold. With calibrated surrogates, downstream attacks obtain update directions that generalize more effectively across black-box MLLMs. Experiments across diverse attacks, surrogate architectures, and open- and closed-source MLLMs show that SCal consistently improves transferability over the corresponding baselines. These results reveal previously overlooked vulnerabilities of modern MLLMs, suggesting that surrogate calibration is crucial for

enabling high-quality black-box attacks even under modest computational budgets. We hope our findings motivate the development of more robust and trustworthy multimodal systems.

Acknowledgments and Disclosure of Funding

This work was supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project 2026ZD0128001, in part by the Natural Science Foundation of China under Grant 62625202, in part by National Science and Technology Major Project 2025ZD1601300, in part by Grant 2023-JCJQ-LA-001-088.

The authors declare no competing interests.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. URL https://arxiv.org/abs/2308.12966.

[2] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805. URL https://arxiv.org/abs/2312.11805.

[3] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a97b58c4f7551053b0512f92244b0810-Abstract-Conference.html.

[4] Kai Hu, Weichen Yu, Li Zhang, Alexander Robey, Andy Zou, Chengming Xu, Haoqi Hu, and Matt Fredrikson. Transferable adversarial attacks on black-box vision-language models. *arXiv preprint arXiv:2505.01050*, 2025. doi: 10.48550/arXiv.2505.01050. URL https://arxiv.org/abs/2505.01050.

[5] Zhaoyi Li, Xiaohan Zhao, Dong-Dong Wu, Jiacheng Cui, and Zhiqiang Shen. A frustratingly simple yet highly effective attack baseline: Over 90% success rate against the strong black-box models of gpt-4.5/4o/o1. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/poster/119497.

[6] Xiaojun Jia, Sensen Gao, Simeng Qin, Tianyu Pang, Chao Du, Yihao Huang, Xinfeng Li, Yiming Li, Bo Li, and Yang Liu. Adversarial attacks against closed-source mllms via feature optimal alignment. In *Advances in Neural Information Processing Systems*, 2025. URL https://neurips.cc/virtual/2025/poster/116337.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[8] Chuan Liu, Huanran Chen, Yichi Zhang, Jun Zhu, and Yinpeng Dong. Scaling laws for black box adversarial attacks, 2024. URL https://arxiv.org/abs/2411.16782.

[9] Yuanbo Li, Tianyang Xu, Cong Hu, Tao Zhou, Xiao-Jun Wu, and Josef Kittler. Multi-paradigm collaborative adversarial attack against multi-modal large language models. *arXiv preprint arXiv:2603.04846*, 2026. URL https://arxiv.org/abs/2603.04846.

[10] Yao Zhu, Yuefeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Trans. Image Process.*, 31:6487–6501, 2022. doi: 10.1109/TIP.2022.3211736.

[11] Zhewen Yao, Yao Zhu, and Shiliang Zhang. Transferable and stealthy adversarial attacks on large vision-language models, 2026. URL https://openreview.net/forum?id=liQueBuFXi. ICLR 2026 Poster, OpenReview.

[12] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models, 2018. URL https://arxiv.org/abs/1806.07421.

[13] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, June 2015. doi: 10.1109/CVPR.2015.7298935. URL https://openaccess.thecvf.com/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html.

[14] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, December 2015. doi: 10.1109/ICCV.2015.279. URL https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html.

[15] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

[16] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[19] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, June 2024.

[20] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774. URL https://arxiv.org/abs/2303.08774.

[21] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyattack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19900–19909, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Anyattack_Towards_Large-scale_Self-supervised_Adversarial_Attacks_on_Vision-language_Models_CVPR_2025_paper.html.

[22] Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14679–14689, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xie_Chain_of_Attack_On_the_Robustness_of_Vision-Language_Models_Against_CVPR_2025_paper.html.

[23] Xiaohan Zhao, Zhaoyi Li, Yaxin Luo, Jiacheng Cui, and Zhiqiang Shen. Pushing the frontier of black-box lvlm attacks via fine-grained detail targeting. *arXiv preprint arXiv:2602.17645*, 2026. URL https://arxiv.org/abs/2602.17645.

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

[25] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

[26] Alexey Kurakin, Ian J. Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan L. Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. *CoRR*, abs/1804.00097, 2018. URL http://arxiv.org/abs/1804.00097.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer, 2014. URL https://arxiv.org/abs/1405.0312.

[28] Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 20:1333–1348, 2025. doi: 10.1109/TIFS.2024.3518072.

[29] OpenAI. Introducing GPT-5.4 mini and nano, 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Accessed: 2026-04-19.

[30] Google DeepMind. Gemini 3.1 pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-04-19.

[31] Anthropic. Introducing claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-04-19.

[32] xAI. Grok 4.20, 2026. URL https://docs.x.ai/developers/models/grok-4.20. Accessed: 2026-04-19.

[33] Meta. Llama-3.2-11b-vision-instruct, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct. Model card, accessed 2026-04-19.

[34] Microsoft. Phi-3.5-vision-instruct, 2024. URL https://huggingface.co/microsoft/Phi-3.5-vision-instruct. Model card, accessed 2026-04-19.

[35] Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. doi: 10.48550/arXiv.2511.21631. URL https://arxiv.org/abs/2511.21631.

[36] Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

[37] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. doi: 10.48550/arXiv.2111.09734. URL https://arxiv.org/abs/2111.09734.

[38] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

[40] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

[41] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[43] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyJ7ClWCb.

[44] Vadim Ziyadinov and Maxim Tereshonok. Low-pass image filtering to achieve adversarial robustness. *Sensors*, 23(22):9032, 2023. doi: 10.3390/s23229032.

[45] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

[46] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[47] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[48] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1589–1604, 2020. URL https://www.usenix.org/conference/usenixsecurity20/presentation/shan.

[49] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P. Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2101.07922.

[50] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. Unlearnable examples: Making personal data unexploitable. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAmZUo0DxC0.

[51] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. GLAZE: Protecting artists from style mimicry by text-to-image models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2187–2204, 2023. URL https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

[52] Erwan Le Merrer, Patrick Perez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, 32(13):9233–9244, 2020. doi: 10.1007/s00521-019-04434-z.

[53] Huili Chen, Bita Darvish Rouhani, and Farinaz Koushanfar. Blackmarks: Blackbox multibit watermarking for deep neural networks. *arXiv preprint arXiv:1904.00344*, 2019. URL https://arxiv.org/abs/1904.00344.

[54] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Fernandez_The_Stable_Signature_Rooting_Watermarks_in_Latent_Diffusion_Models_ICCV_2023_paper.html.

[55] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b54d1757c190ba20dbc4f9e4a2f54149-Abstract-Conference.html.

[56] Weilin Xu, David Evans, and Yanjun Qi. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. In *Network and Distributed System Security Symposium*, 2018. doi: 10.14722/ndss.2018.23198. URL https://doi.org/10.14722/ndss.2018.23198.

[57] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 43685–43704. PMLR, 2024. URL https://proceedings.mlr.press/v235/schlarmann24a.html.

[58] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6199.

[59] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6572.

[60] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations Workshop*, 2017. URL https://arxiv.org/abs/1607.02533.

[61] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[62] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. doi: 10.1109/SP.2017.49. URL https://arxiv.org/abs/1608.04644.

[63] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. doi: 10.1145/3503161.3547801. URL https://arxiv.org/abs/2206.09391.

[64] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. URL https://arxiv.org/abs/2307.14061.

[65] Bangyan He, Xiaojun Jia, Siyuan Liang, Tianrui Lou, Yang Liu, and Xiaochun Cao. SA-Attack: Improving adversarial transferability of vision-language pre-training models via self-augmentation. *arXiv preprint arXiv:2312.04913*, 2023. URL https://arxiv.org/abs/2312.04913.

[66] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 862–871, 2024. doi: 10.1145/3626772.3657781. URL https://arxiv.org/abs/2405.05524.

[67] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017. doi: 10.1145/3052973.3053009.

[68] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sys6GJqxl.

[69] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017. doi: 10.48550/arXiv.1704.03453.

[70] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

[71] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. doi: 10.1109/CVPR.2018.00957.

[72] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):961–977, 2025. doi: 10.1109/TPAMI.2024.3480519.

A Appendix overview

This appendix provides additional details and analyses that supplement the main paper. It is organized as follows:

- Appendix B provides more method details of SCal.
- Appendix C provides experimental details and implementation settings for reproducibility.
- Appendix D reports additional experiments to demonstrate the effectiveness of SCal.
- Appendix E discusses the broader impacts of SCal and potential countermeasures.
- Appendix F supplements the main paper with additional discussion of related literature.
- Appendix G describes how LLM tools were used in this work.

B Surrogate calibration details

B.1 Algorithmic procedure

We detail the algorithmic procedure as follows:

Algorithm 1 SCAL: Surrogate Calibration

Input: Surrogate model parameters θ; calibration dataset $\mathcal{C}$; calibration steps T; number of sampled targets K; balancing coefficient λ.
Output: Calibrated surrogate model parameters $\theta^\star$.

// Step 1. Initialization
Freeze the surrogate model $\mathcal{S}_\theta$.
Initialize $\theta^\star \leftarrow \theta$.
for $t = 1, \ldots, T$ **do**
  // Step 2. Compute distributional regularization
  $x \sim \mathcal{C},\ x_k^{\mathrm{tgt}} \sim \mathcal{C},\ k = 1, \ldots, K$
  $v_{\theta^\star}(x, x_k^{\mathrm{tgt}}) \leftarrow \mathrm{Descent}(x, x_k^{\mathrm{tgt}})$ // typically the negative cosine-similarity gradient
  $\bar{v}_{\theta^\star}(x) \leftarrow \dfrac{1}{K} \sum_{k=1}^{K} v_{\theta^\star}(x, x_k^{\mathrm{tgt}})$
  $\mathcal{L}_{\mathrm{distri}}(\theta^\star) \leftarrow \mathbb{E}_{x \sim \mathcal{C}} \left[\| \bar{v}_{\theta^\star}(x) \|_2^2 \right]$
  // Step 3. Compute feature-preserving loss
  $\mathcal{L}_{\mathrm{feat}}(\theta^\star) \leftarrow \mathbb{E}_{x \sim \mathcal{C}} \left[1 - \cos\left(\mathcal{S}_\theta(x), \mathcal{S}_{\theta^\star}(x) \right) \right]$
  // Step 4. Update
  $\mathcal{L}_{\mathrm{SCal}}(\theta^\star) \leftarrow \lambda \mathcal{L}_{\mathrm{feat}}(\theta^\star) + \mathcal{L}_{\mathrm{distri}}(\theta^\star)$
  $\theta^\star \leftarrow \mathrm{Update}\left(\theta^\star; \nabla_{\theta^\star} \mathcal{L}_{\mathrm{SCal}}(\theta^\star) \right)$ // To minimize
end for
return $\theta^\star$

The $\mathrm{Descent}$ function provides the update direction for downstream attacks. Given that representative attacks, including M-Attack [5] and FOA-Attack [6], extensively utilize random cropping augmentation during their iterative updates, we align with this paradigm by calculating the expected cosine-similarity gradient over randomly cropped inputs.

B.2 Additional explanations on distributional regularization

From the perspectives of optimization difficulty and generalization, the mean descent vector is defined on the marginal distribution $\mathcal{D}$, while the target-conditioned field $v(x, x')$ lies on the product space $\mathcal{D} \times \mathcal{D}$. Thus, directly controlling $v(x, x')$ is harder to generalize to unseen source-target pairs and unknown attacks. By marginalizing over x', $\bar{v}(x)$ filters out target-specific fluctuations and keeps only the target-agnostic systematic drift, making the regularization more stable. The sensitivity study in Figure 6 further supports this: insufficient marginalization with small K hurts performance, but still brings gains since it already provides a noisy estimate of the true mean descent vector.

Table 7: **Model aliases and corresponding full model identifiers.** The evaluated victim models span a diverse range of architectures, parameter scales, and service providers.

Model Type	Model	Model ID
Closed-source	Claude Opus 4.7	`claude-opus-4.7`
	GPT-5.4 Mini	`gpt-5.4-mini`
	GPT-5.4 Pro (Figure 4 only)	web interface, accessed 2026-04-25
	Gemini 3.1	`gemini-3.1-pro-preview`
	Grok 4.20	`grok-4.20`
Open-source	Llama-3.2-11B	`unsloth/Llama-3.2-11B-Vision-Instruct-bnb-4bit`
	Phi-3.5-Vis-4.2B	`microsoft/Phi-3.5-vision-instruct`
	Qwen3-VL-8B	`Qwen/Qwen3-VL-8B-Instruct`
	InternVL3.5-8B	`OpenGVLab/InternVL3_5-8B-HF`

Table 8: **Baseline details.** C, L, I, and T denote OpenAI CLIP, LAION/OpenCLIP, image encoder, and text encoder, respectively. IVL3-Gen denotes the InternVL3-1B text-anchor generator.

Baseline	Venue	Type	Surrogate
CoA [22]	CVPR 2025	ℓ_p-bounded	C-B32-I + C-B32-T + ClipCap
M-Attack [5]	NeurIPS 2025	ℓ_p-bounded	C-B16-I + C-B32-I + L-B32-I
FOA-Attack [6]	NeurIPS 2025	ℓ_p-bounded	C-B16-I + C-B32-I + L-B32-I
M-Attack-V2 [23]	arXiv 2026	ℓ_p-bounded	C-B16-I + C-B32-I + L-B32-I
MPCAttack [9]	CVPR 2026 Accepted	ℓ_p-bounded	C-B16-I + C-B32-I + L-B32-I + C-B16-T + C-B32-T + L-B32-T + IVL3-Gen
AdvDiffVLM [28]	IEEE TIFS 2025	unrestricted	C-B16-I + C-B32-I + L-B32-I
PSI [11]	ICLR 2026	unrestricted	C-B16-I + C-B32-I + L-B32-I

C Experimental details

Model details. As detailed in Table 7, the selected victim models are deliberately chosen to maximize diversity across architectures, training paradigms, and service providers. The selection includes four proprietary models accessed via APIs to cover the major commercial ecosystems (*e.g.*, OpenAI, Google, Anthropic, xAI). Complementing these are four prominent open-source models ranging from 4.2B to 11B parameters. This open-source suite encompasses varying visual encoders and deployment configurations, including quantized setups (*e.g.*, 4-bit Llama-3.2), ensuring our evaluation covers a highly heterogeneous set of real-world targets. Across all inference processes, the generation temperature is strictly set to 0 to eliminate sampling randomness and ensure reproducible evaluations. Reasoning is disabled whenever supported.

Baseline details. Table 8 summarizes the configurations of our baselines. To ensure a fair and rigorous comparison, we standardize the surrogate model ensemble across the majority of the baselines, utilizing a consistent combination of C-B16-I, C-B32-I, and L-B32-I. For MPCAttack, we additionally incorporate the corresponding text encoders for each vision model, alongside an InternVL3-1B text-anchor generator. MPCAttack [9] has been accepted but is not yet available in the official proceedings, so we cite the publicly available arXiv version.

SCal details. All surrogate models are trained using AdamW with a learning rate of 1×10^{-5} for two epochs. The weight for the feature-preserving loss is set to $\lambda = 0.06$, and K for the empirical average is set to 128. In the SCal-enhanced attack, only the surrogate model is modified, while all other hyperparameters remain strictly unchanged during the downstream attack. When optimizing $\mathcal{L}_{\text{distri}}$, we compute the input-gradient term with `create_graph=True`, allowing gradients to back-propagate through the input-gradient computation when updating $\theta^\star$.

SCal time cost. Calibrating a single surrogate takes less than 60 minutes on a single RTX 3090 GPU with 24 GB memory, and the resulting calibrated surrogate can be readily used to improve downstream attacks. For low-cost adaptation to unseen-domain examples, single-sample calibration takes about 4 s per sample. The downstream optimization takes about 33 s per sample for M-Attack.

Evaluation Prompt

Rate the semantic similarity between the following two texts on a scale from 0 to 1.
Criteria for similarity measurement:

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

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

Figure 7: **Detailed evaluation prompt.** We use this prompt in an LLM-as-a-judge [38] setup to evaluate the success of the attack. A predefined threshold of ≥ 0.6 is considered a successful attack.

Evaluation details. To comprehensively and automatically assess the effectiveness of our targeted attacks, we employ an LLM-as-a-judge framework [38]. Following previous settings [5, 23, 6, 9], we prompt a language model, as shown in Figure 7, to compute a semantic similarity score, denoted as $s \in [0, 1]$, between the MLLM's generated text and the intended adversarial target. In contrast to these prior works, we define a strictly successful attack as one that achieves a similarity score of $s \geq 0.6$. This stringent threshold is deliberately established to filter out ambiguous, partial, or weak misclassifications.

D Additional experiments

SCal with more attacks. As shown in Table 9, both FOA-Attack [6] and MPCAttack [9] are significantly enhanced by SCal, demonstrating the strong generalizability of our method across diverse attack paradigms. Furthermore, synthesizing these results with Table 1, we observe that while MPCAttack generally outperforms M-Attack-V2 [23] in standard settings, their performance converges when utilizing a calibrated surrogate. These findings suggest that the efficacy of a surrogate model lies not solely in its architectural complexity or scale, but crucially in its task suitability. Our calibrated surrogate fulfills this requirement.

Effectiveness on the MS-COCO benchmark. Table 10 presents the performance of various attacks on the MS-COCO benchmark, where we selected 1,000 images from the validation set as the source and target, respectively. Because MS-COCO contains more complex and realistic images, the overall attack success rates are generally lower than those observed in the NIPS 2017 challenge. However, the overall trends remain largely consistent. The SCal-enhanced baseline, M-Attack [5], outperforms the current state-of-the-art MPCAttack [9]. Furthermore, utilizing M-Attack-V2 [23] yields even greater attack effectiveness. To quantify the reliability of these improvements, we report paired-bootstrap 95% confidence intervals for the ASR gains of SCal-enhanced attacks. The confidence intervals remain positive across all evaluated victims, indicating that the SCal gains are statistically significant on this more challenging benchmark.

E Broader impacts

Positive impacts. This work aims to strengthen the robustness and safety evaluation of modern multimodal systems. By exposing transferable black-box vulnerabilities in commercial and open-source MLLMs, our results show that visually subtle perturbations can still manipulate deployed models under realistic black-box settings. In this sense, SCal serves as a catalyst for more effective multimodal red-teaming: by calibrating existing surrogates rather than designing model-specific attacks, it helps uncover overlooked failure modes. We hope these findings alert model providers

Table 9: **Attack performance with more attacks.** We report the Attack Success Rate (ASR) and the average semantic similarity to the target image (AvgSim). All experiments are conducted on the NIPS 2017 Competition dataset [26].

Attack	*Closed-Source Black-Box Models*						*Open-Source Black-Box Models*			
	GPT-5.4 Mini		Claude Opus 4.7		Grok 4.20		Llama-3.2-11B		Phi-3.5-Vis-4.2B	
	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
FOA-Attack [6]	24.9%	0.376	71.1%	0.662	65.4%	0.641	50.8%	0.529	50.6%	0.543
+SCal	47.4%	0.512	86.8%	0.724	81.5%	0.696	80.8%	0.687	72.8%	0.646
MPCAttack [9]	29.7%	0.431	84.7%	0.716	78.2%	0.680	61.0%	0.588	60.4%	0.587
+SCal	52.1%	0.546	88.7%	0.735	82.1%	0.707	82.2%	0.699	75.7%	0.666

Table 10: **Performance of different attacks on the MS-COCO [27].** We report Attack Success Rate (ASR) and average semantic similarity to the target image (AvgSim). For SCal-enhanced attacks, the small row below each entry reports the paired ASR gain over the corresponding original attack; brackets denote 95% confidence intervals computed by paired bootstrap over source–target pairs.

Attack	Surrogate Size	*Closed-Source Black-Box Models*							
		GPT-5.4 Mini		Claude Opus 4.7		Gemini 3.1		Grok 4.20	
		ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
CoA [22]	151M	0.1%	0.052	0.0%	0.086	0.0%	0.083	0.2%	0.098
M-Attack [5]	262M	18.7%	0.316	62.1%	0.595	8.9%	0.289	53.4%	0.548
FOA-Attack [6]	262M	19.3%	0.327	63.4%	0.602	9.4%	0.294	54.2%	0.553
M-Attack-V2 [23]	262M	20.9%	0.366	65.9%	0.627	17.2%	0.339	61.7%	0.592
MPCAttack [9]	1390M	20.4%	0.364	65.4%	0.619	19.8%	0.346	54.5%	0.577
M-Attack + SCal	262M	36.2%	0.456	76.8%	0.676	22.4%	0.370	69.9%	0.640
ΔASR (95% CI)		+17.5% [14.5, 20.5]		+14.7% [11.6, 17.8]		+13.5% [10.6, 16.4]		+16.5% [13.6, 19.4]	
M-Attack-V2 + SCal	262M	**42.2%**	**0.497**	**78.9%**	**0.693**	**24.1%**	**0.392**	**71.5%**	**0.662**
ΔASR (95% CI)		+21.3% [18.1, 24.5]		+13.0% [10.0, 16.0]		+6.9% [4.2, 9.6]		+9.8% [6.8, 12.8]	

Attack	Surrogate Size	*Open-Source Black-Box Models*							
		Llama-3.2-11B		Phi-3.5-Vis-4.2B		Qwen3-VL-8B		InternVL3.5-8B	
		ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑	ASR ↑	AvgSim ↑
CoA [22]	151M	0.0%	0.076	0.1%	0.101	0.4%	0.118	0.5%	0.112
M-Attack [5]	262M	37.5%	0.440	44.2%	0.492	53.8%	0.555	51.1%	0.542
FOA-Attack [6]	262M	38.3%	0.446	45.1%	0.496	54.7%	0.560	52.3%	0.548
M-Attack-V2 [23]	262M	41.3%	0.471	52.9%	0.533	58.1%	0.585	60.4%	0.585
MPCAttack [9]	1390M	40.7%	0.469	44.0%	0.498	60.6%	0.585	57.3%	0.565
M-Attack + SCal	262M	54.2%	0.560	64.6%	0.613	73.0%	0.661	67.2%	0.635
ΔASR (95% CI)		+16.7% [13.7, 19.7]		+20.4% [17.2, 23.6]		+19.2% [16.1, 22.3]		+16.1% [13.2, 19.0]	
M-Attack-V2 + SCal	262M	**64.0%**	**0.595**	**68.8%**	**0.636**	**75.2%**	**0.676**	**75.6%**	**0.670**
ΔASR (95% CI)		+22.7% [19.5, 25.9]		+15.9% [12.9, 18.9]		+17.1% [14.0, 20.2]		+15.2% [12.2, 18.2]	

and the broader community to such risks, motivate systematic red-teaming protocols, and ultimately contribute to the development of more robust and trustworthy multimodal systems.

Beyond vulnerability discovery, adversarial techniques also have constructive uses. Imperceptible perturbations have been used to protect user privacy against unauthorized face recognition and data collection through image cloaking, adversarial filters, and unlearnable examples [48–50]. Related ideas have also been used to protect artists from unauthorized style mimicry [51], and to support provenance or intellectual-property protection through adversarial watermarking and generative-model watermarking [52–55]. Although SCal is not designed as a privacy or watermarking method, its perspective may inform future defensive, auditing, and provenance-preserving applications.

Potential misuse and countermeasures. This work is inherently dual-use. A malicious actor could use calibrated surrogates or the proposed SCal procedure to build stronger transferable black-box attacks against deployed MLLMs. Such attacks may be misused to manipulate image captions, visual question answering results, content-moderation decisions, or other downstream outputs, potentially reducing trust in multimodal systems or misleading human operators in high-stakes settings.

To reduce these risks, we frame SCal as a robustness evaluation and red-teaming tool. Any released artifacts will be distributed under a research-only license and an acceptable-use policy that prohibits unauthorized attacks on real users, public services, deployed third-party APIs, or safety-critical systems. We will avoid releasing turnkey attack services, commercial-API attack scripts, service-specific prompts, credentials, or deployment-ready attack pipelines. When releasing pre-calibrated checkpoints, access may be gated or accompanied by explicit usage restrictions. We also encourage responsible disclosure for severe vulnerabilities and recommend that model providers evaluate against calibrated-surrogate attacks, and incorporate such examples into red-teaming or adversarial training.

Insights for robust MLLM systems. We believe that robust MLLM systems should evolve from merely being insensitive to adversarial perturbations toward developing *adversarial awareness*: the ability to recognize, communicate, and safely handle manipulated visual inputs. Traditional defenses typically improve robustness through input transformations [43], consistency-based detection [56], or adversarially trained models [57], which mainly operate at the low-level visual or feature space. Modern MLLMs, however, may enable a complementary semantic-level defense. As illustrated in Figure 4, Gemini already shows an early form of such capability by recognizing the adversarial signal as an illusion rather than directly accepting the target object. If MLLMs are further trained with adversarial knowledge, an ideal model should explicitly identify such inputs as adversarial examples, warn that the image may be manipulated, and indicate that the perturbation is likely intended to elicit a specific false response.

F Additional related work

This section provides an extended review of related literature, including adversarial attacks against other types of victim models, such as standard image classifiers and vision-language pre-trained (VLP) models, as well as fundamental studies on adversarial transferability. Furthermore, we explicitly highlight the key differences between our proposed method and these prior works.

Adversarial attacks on classifiers. Classical adversarial attacks mainly study small-norm perturbations against image classifiers, where the objective is defined over class labels or logits. Early work revealed that visually imperceptible perturbations can fool neural networks [58], followed by efficient gradient-based attacks such as FGSM and its iterative variants [59, 60]. Stronger optimization-based methods, including PGD [61] and C&W [62], later became standard baselines for robustness evaluation. These works establish the foundational ℓ_p-bounded iterative update paradigm that is widely inherited by later VLP and MLLM attacks, where the closed-set classification loss is replaced by feature-level or multimodal alignment objectives.

Adversarial attacks on VLPs. Adversarial attacks on vision-language pre-trained (VLP) models move beyond closed-set classification by attacking multimodal embedding spaces and cross-modal alignment. Co-Attack studies collaborative multimodal attacks on VLP models by perturbing image–text pairs [63]. Subsequent works improve black-box transferability through richer cross-modal guidance and data augmentation: SGA introduces set-level guidance with alignment-preserving augmentations [64], while SA-Attack leverages self-augmentation on both visual and textual modalities [65]. ETU further studies universal adversarial perturbations for VLP models, aiming to transfer across inputs, models, and downstream tasks [66]. However, these methods are mostly tailored to VLP benchmarks with dataset-provided image–caption pairs, where paired captions are available during optimization and, in some settings, both modalities can be modified. By contrast, attacks on MLLMs interact with user-provided prompts that should remain fixed, making image-only perturbation the primary setting. These attacks operate under a fundamentally different threat model; therefore, we did not test the performance of SCal under those settings.

Adversarial transferability. Adversarial transferability has been studied from several complementary perspectives. Early works attribute transferability to shared decision boundaries or adversarial subspaces across models, showing that adversarial examples generated on substitute or ensemble models can fool unseen victims [67–69]. Later studies further relate transferability to common non-robust features and more model-agnostic optimization trajectories [70, 71]. Beyond these model- and optimization-level explanations, recent works understand transferability from a distributional perspective, arguing that adversarial examples closer to the natural data distribution tend to induce more consistent predictions across models [10, 8]. Diffusion-based attacks such as DiffAttack, AdvDiffVLM, and PSI instantiate this idea by using generative priors to keep adversarial samples

visually natural or semantically aligned with target distributions [72, 28, 11]. Different from these attack-time strategies, our SCal calibrates the surrogate model itself: instead of directly regularizing the generated adversarial sample, we rectify the surrogate-induced descent field by suppressing target-agnostic systematic drift, thereby guiding downstream attacks toward more transferable and distribution-aligned update directions.

G LLM usage

In this work, large language models and multimodal large language models were involved in three distinct ways:

- **Research subjects.** The primary focus of this work is to study the robustness of commercial LLMs, multimodal LLMs, and vision-language models as black-box victim models. We interacted with these models, such as GPT-5.4 Mini, Claude Opus 4.7, Gemini 3.1, and Grok 4.20, through APIs or web interfaces as part of our experimental evaluation.
- **Research assistance.** We used commercial LLM tools, including GPT-5.4 Mini, for auxiliary research support, such as code drafting and debugging, grammar checking, phrasing refinement, and improving the readability of the manuscript. All assisted code and text were reviewed, verified, and revised by the authors. The authors take full responsibility for the final methodology, implementation, experiments, analysis, and writing.
- **Evaluation method.** We used LLM-as-a-judge evaluation to assess model responses in our experiments. The judge model was used only as an evaluation tool under a fixed evaluation protocol, and was not used to generate the proposed method or alter the attack algorithm. Details of the judging prompt, scoring criteria, and aggregation procedure are provided.

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 abstract and introduction state SCal's scope and claims under the black-box threat model; these are supported by the method, ablations, and experiments.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 paper discusses limitations under domain shift, where SCal may require additional adaptation, and analyzes weaker transfer to Gemini, whose abstention behavior also motivates future directions for robust MLLM systems.

   Guidelines:

   - The answer [N/A] 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 paper includes a lightweight theoretical bound showing that $\mathcal{L}_{\text{distri}}$ upper-bounds the Wasserstein displacement induced by systematic drift.

Guidelines:

- The answer [N/A] 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]

We provide the main implementation details, evaluation details and model details in the paper and Appendix C. We provide an algorithmic procedure in Appendix B.1. We also provide code at:

`https://github.com/zhewen-yao/SCal`

and in the supplementary materials, with organized scripts and instructions that allow users to verify the core SCal results on a single RTX 3090 GPU within five hours.

Guidelines:

- The answer [N/A] 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 provide code at `https://github.com/zhewen-yao/SCal` and in the supplementary materials, including environment setup, and scripts for reproducing the main SCal experiments and baselines. The effectiveness of SCal is readily verifiable through straightforward replication. Experiments on proprietary MLLMs require access to the corresponding official APIs.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided implementation details in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] 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: Appendix D reports paired-bootstrap 95% confidence intervals for the ASR gains of SCal-enhanced attacks on the MS-COCO benchmark. Complementarily, in Tables 1 we provides AvgSim gains. In Table 2, and 9 validate the stability of SCal across diverse victim MLLMs, surrogate architectures, and attack paradigms.

Guidelines:

- The answer [N/A] 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: We have provided the computer resources in Section 5 and Appendix C. In downstream attacks, utilizing our calibrated surrogates incurs no additional time overhead compared to the baseline attack.

   Guidelines:

   - The answer [N/A] 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 reviewed the NeurIPS Code of Ethics and believe the research conforms to it. Although the work is dual-use, we conduct it for robustness evaluation and red-teaming, use public benchmark datasets and standard models, and discuss potential misuse, responsible release safeguards, acceptable-use restrictions, and countermeasures in Appendix E.

   Guidelines:

- The answer [N/A] 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: We discuss both the potential positive and negative societal impacts of SCal, as well as possible countermeasures, in Appendix E.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The work is dual-use and may be misused to construct stronger transferable black-box attacks. We therefore describe responsible-release safeguards in Appendix E, including a research-only license, an acceptable-use policy, avoiding release of turnkey attack services, commercial-API attack scripts, service-specific prompts, credentials, or deployment-ready attack pipelines, and potentially gated access to pre-calibrated checkpoints.

    Guidelines:

    - The answer [N/A] 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: We properly cite the datasets, models, baselines, and evaluation tools used in this work, including NIPS 2017, MS-COCO, ImageNet, CLIP/OpenCLIP, DINO, InternVL, Qwen-VL, Llama, Phi, and prior attack baselines. All proprietary MLLMs are accessed only through their official APIs or web interfaces in accordance with their usage terms.

    Guidelines:

    - The answer [N/A] 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: [Yes]

    Justification: The new assets introduced by this work consist of the SCal code and, if released, calibrated surrogate checkpoints. We document the algorithmic procedure, implementation settings, hyperparameters, evaluation protocol, and usage restrictions in Appendix B.1, Appendix C, and Appendix E.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on public benchmark datasets, open-source models, and proprietary MLLMs accessed through APIs or web interfaces.

Guidelines:

- The answer [N/A] 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: [N/A]

    Justification: This work does not involve crowdsourcing, user studies, or research with human subjects, and therefore does not require IRB approval or equivalent human-subjects review.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We have described how we use LLM in Appendix G.

    Guidelines:

    - The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.