Text-Guided Gradient Refinement: Resolving Multimodal Gradient Conflicts to Boost Adversarial Attacks on Vision-Language Models

Yuyang Huang*, **Tianzuo Luo***, **Hengyuan Guo, Yuren Zhang**†

ByteDance Inc., Beijing, China

{huangyuyang.hyy, luotianzuo.98, guohengyuan.01, zhangyuren}@bytedance.com

Abstract

Vision-Language Models (VLMs) have advanced multimodal understanding, yet they remain susceptible to adversarial attacks. Among various strategies, transfer-based attacks are notably effective, especially in black-box scenarios. The dominant approach within this paradigm leverages generative models to create image targets from text, consistently outperforming text-only methods. However, this suffers from a fundamental limitation: generative models introduce visual features irrelevant or even detrimental to textual semantics, misguiding optimization. To investigate this limitation, we conduct comprehensive analysis revealing two critical findings. First, optimal attack directions lie in synergistic spaces between image and text gradients, demonstrating that text provides complementary information. Second, widespread gradient conflicts occur when combining modalities, where image-target gradients oppose text-target directions. This conflict provides direct evidence that extraneous visual information actively harms optimization, driving it away from intended textual objectives. Based on these insights, we propose Text-Guided Gradient Refinement (TGGR), a novel framework that employs a conflict-aware projection mechanism to resolve this conflict. TGGR preserves the beneficial characteristics of image targets by decomposing the image gradient and surgically removing components that oppose the textual guidance. Extensive experiments on models such as LLaVA and GPT-4o demonstrate that TGGR substantially improves attack success rates. Specifically, on GPT-4o, TGGR yields an improvement of up to 14% over state-of-the-art methods, achieving 96% attack success rate. Our work offers a principled framework for developing more synergistic and effective adversarial strategies against VLMs.

1 Introduction

Vision-Language Models (VLMs), such as GPT-4o (Hurst et al. 2024), LLaVA (Liu et al. 2023), and BLIP-2 (Li et al. 2023), have revolutionized artificial intelligence by demonstrating unprecedented capabilities in multimodal understanding and generation. These models excel across diverse tasks including image captioning (Karpathy and Fei-Fei 2015), visual question answering (Antol et al. 2015; Luu, Le, and Vo 2024), and complex reasoning (Lu et al. 2022; Li et al. 2024; Park et al. 2025), leading to their widespread adoption in content moderation (Kiela et al. 2020; Salman et al. 2023; Sha et al. 2023) and assistive technologies (Gurari et al. 2018; Nagesh et al. 2024). Their growing deployment, however, has raised critical concerns about security and robustness (Yuan et al. 2019; Gu et al. 2024; Zheng et al. 2024), particularly regarding adversarial attacks that can manipulate model outputs through carefully crafted inputs (Goodfellow, Shlens, and Szegedy 2014).

Transfer-based adversarial attacks (Papernot, McDaniel, and Goodfellow 2016; Xiong et al. 2022; Wang, Zhang, and Zhang 2023), which leverage surrogate models to generate highly transferable adversarial examples, have emerged as particularly effective method of attack against VLMs in black-box scenarios (Liu et al. 2017; Papernot et al. 2017; Zhou et al. 2018). Our work focuses on the task of deceiving black-box models into producing pre-specified targeted responses, a realistic and high-risk problem setting proposed by (Zhao et al. 2023). In this setting, a straightforward approach is to use surrogate models (e.g., CLIP) to optimize the similarity between the adversarial image's representation and the target text's representation. Yet, (Zhao et al. 2023) demonstrated that attacks leveraging generative models (Ho, Jain, and Abbeel 2020) to convert textual descriptions into image targets yield superior performance over such text-only approaches, likely due to richer spatial features that may facilitate more precise gradient-based perturbations.

This image-based strategy, however, introduces potential pitfalls. The generative process can produce extraneous features that do not align with the textual objective. For instance, generating "a photo of a dog" might incorporate specific breed or background characteristics. Such unintended details can mislead optimization, deviating from the core semantic objective and degrading attack performance. Since our goal is generating specific textual responses, the target text represents the purest semantic objective, serving as the cleanest representation of the adversarial objective.

Recognizing that image-based and text-based guidance offer complementary strengths, we investigate whether their combination yields better results. Through analysis, we demonstrate that optimal attack directions emerge from synergistic fusion rather than individual modalities. As illustrated in Figure 1, there exists a synergistic region where bal-

*These authors contributed equally.

†Corresponding author

anced multimodal guidance significantly outperforms text-dominated and image-dominated strategies. This reveals that text-based targets, while individually less potent, provide clean semantic anchors that augment powerful but noisy image-based targets.

Nonetheless, our investigation uncovers a fundamental challenge that sabotages this promising synergy: pervasive gradient conflicts arise when fusing the two modalities, where image-targeted gradients frequently oppose their text-targeted counterparts. As shown in Figure 2, these conflicts escalate dramatically during optimization. This provides direct empirical evidence for our hypothesis that extraneous visual information introduced by the generative process can actively harm the optimization by steering it away from the intended textual objective. This conflict directly destabilizes the fusion process, preventing the realization of the full potential of multimodal guidance.

Based on these insights, we propose Text-Guided Gradient Refinement (TGGR), a novel framework that employs a conflict-aware projection mechanism to resolve gradient conflicts between modalities. TGGR preserves the beneficial characteristics of image targets by decomposing image gradients and surgically removing components that oppose textual guidance, while retaining complementary visual information that enhances attack effectiveness. This principled approach ensures that adversarial optimization remains focused on core semantic objectives while leveraging the rich optimization landscape offered by visual modalities.

Extensive experiments across multiple VLM architectures demonstrate our approach's effectiveness. For commercial models like GPT-4o, TGGR achieves improvements of up to 14% over state-of-the-art methods, reaching 96% attack success rate. Similar improvements are observed for open-source models such as LLaVA and BLIP-2, validating broad applicability. Our main contributions are:

- **Multimodal Optimization Analysis:** We analyze multimodal adversarial optimization and reveal that (1) optimal attack directions emerge from synergistic combinations of image and text gradients, and (2) gradient conflicts occur when combining modalities, impeding optimization effectiveness.
- **Text-Guided Refinement Framework:** We propose TGGR, a practical framework that leverages textual guidance to refine image targets through conflict-aware gradient projection, effectively resolving optimization conflicts while preserving beneficial multimodal synergies.
- **Superior Empirical Performance:** We conduct extensive evaluations across diverse VLM architectures, demonstrating consistent and significant performance improvements with up to 14% increase in attack success rates, establishing new state-of-the-art results for transfer-based VLM attacks.

2 Related Work

2.1 Visual Language Models (VLMs)

VLMs represent a transformative advancement in artificial intelligence, integrating computer vision and natural language processing to process and generate multimodal content. These models have gained significant attention for their ability to perform various tasks involving both images and text (Achiam et al. 2023; Li et al. 2025a). Open-source VLMs such as LLaVa (Liu et al. 2023) and BLIP-2 (Li et al. 2023) are competitive across multiple benchmarks. Black-box commercial models like GPT-4o and Gemini-2.0 (Team et al. 2023) show their potential in analyzing both text and images for complex tasks. Researches suggest that multimodal foundation models are vulnerable to adversarial attacks that manipulate outputs with imperceptible image perturbations (Schlarmann and Hein 2023). However, the opaque nature of commercial VLM systems complicates the investigation of their vulnerabilities to adversarial attacks.

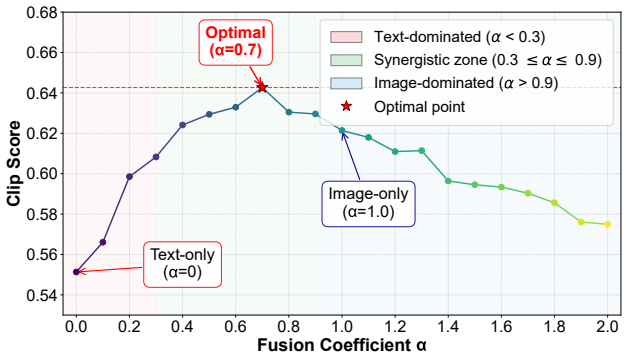

Figure 1: The fusion coefficient α (x-axis) balances the text-guided ($\alpha = 0$) and image-guided ($\alpha = 1.0$) attacks. The shaded regions denote the Text-dominated ($\alpha < 0.3$), Synergistic ($0.3 \leq \alpha \leq 0.9$), and Image-dominated ($\alpha > 0.9$) zones. The red star marks the Optimal point ($\alpha = 0.7$), which achieves the highest CLIP Score (y-axis).

2.2 Transfer-Based Adversarial Attacks

In this paper, we investigate transfer-based adversarial attacks (Ma et al. 2024; Gao et al. 2024), a methodology wherein adversarial examples are crafted using surrogate models, such as CLIP (Radford et al. 2021), and subsequently applied to victim models. This approach leverages the transferability of adversarial examples across different models, capitalizing on their shared vulnerabilities. Notably, this technique does not involve querying the victim model during the attack generation process, distinguishing it from query-based attacks (Dong et al. 2021; Ilyas et al. 2018), which rely on direct queries to the victim model to construct adversarial examples. AttackVLM (Zhao et al. 2023) introduced image-to-text (MF-it) and image-to-image (MF-ii) guidance, demonstrating that direct image-to-text matching showed poor performance on large VLMs. This drove subsequent research toward image-to-image strategies. AnyAttack (Zhang et al. 2025) leverages self-supervised pretraining to transform images into attack vectors without additional training, though the resulting images may exhibit perceptible watermark-like shadows. AdvDiffVLM (Guo et al. 2025) improves image quality through diffusion models and

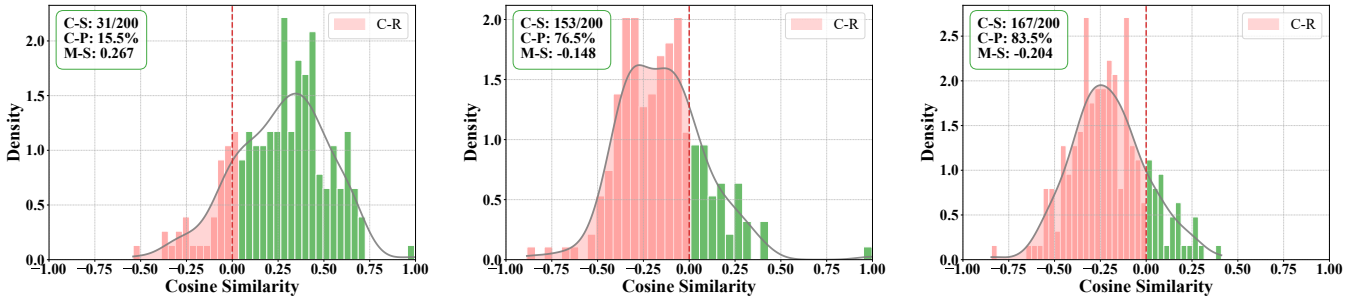

Figure 2: **Evolution of Gradient Conflict During Optimization.** The figure illustrates the cosine similarity distribution between the image gradient $\mathbf{g}_I$ and the text gradient $\mathbf{g}_T$, based on a fused gradient with an optimal coefficient of $\alpha = 0.7$. The bottom plots (from left to right) show the evolution at the early (Step 1), middle (Step 150), and late (Step 299) stages, revealing that gradient conflict (negative similarity, red region) escalates dramatically over time. (*Legend: C-S: Conflict Samples, C-P: Conflict Percentage, M-S: Mean Similarity, C-R: Conflict Region.*)

employs GradCam-guided masks (Selvaraju et al. 2017) to focus on irrelevant image regions. VT-Attack (Wang et al. 2024) proposes to break semantic information by perturbing visual token features and disrupting their interrelationships, though its effectiveness on commercial VLMs remains unevaluated. M-Attack (Li et al. 2025b) reveals that local regions harbor rich semantic information and achieves exceptional transferability through local region sampling and matching strategies. Despite image-to-image methods dominating the field, insufficient research has been dedicated to exploring better objective paradigms.

3 Method

In this section, we address the inherent optimization conflict between image-based and text-based targets within multimodal adversarial attacks. First, we formally define the threat model and the baseline attack formulations. Next, we demonstrate that while textual information provides valuable guidance to image-based attacks, integrating these modalities introduces gradient conflicts. Based on these findings, we propose the Text-Guided Gradient Refinement (TGGR) framework, which effectively resolves these conflicts to achieve synergistic and robust adversarial objective.

3.1 Preliminaries

Threat Model We address the task of transfer-based black-box adversarial attack (Liu et al. 2024) against a target VLM, denoted by f_t. The adversary has no access to the internal architecture or parameters of f_t, but maintains full white-box access to a set of surrogate models. Given a benign image $\mathbf{x}_{\text{clean}}$ and a target text prompt $\mathbf{t}_{\text{ref}}$, the adversary's goal is to craft an adversarial image $\mathbf{x}_{\text{adv}}$ that misleads f_t. The adversarial image is generated by adding a small perturbation δ to the clean image

$$\mathbf{x}_{\text{adv}} = \mathbf{x}_{\text{clean}} + \delta. \tag{1}$$

Since the target model f_t is a black box, we cannot directly optimize the loss $\mathcal{L}(f_t(\mathbf{x}_{\text{adv}}), \mathbf{t}_{\text{ref}})$. Instead, we find the optimal perturbation δ by optimizing a surrogate loss function $\mathcal{L}_{\text{surrogate}}$ on a set of white-box models:

$$\min_{\delta} \mathcal{L}_{\text{surrogate}}(\mathbf{x}_{\text{adv}}, \mathbf{t}_{\text{ref}}) \quad \text{s.t.} \|\delta\|_\infty \leq \epsilon. \tag{2}$$

The parameter ϵ defines the maximum perturbation allowed, ensuring the imperceptibility of the changes.

Baseline Attack Formulations The surrogate loss in Equation 2 can be instantiated through two primary strategies, which provide the foundational gradients, $\mathbf{g}_I$ and $\mathbf{g}_T$, for our analysis and method. These gradients are the source of the conflict we aim to resolve.

(1) Image-to-Text (MF-it) Attack. This is the most direct approach, where the optimization objective is to maximize the cosine similarity between the adversarial image's embedding and the target text's embedding. Let $E_I(\cdot)$ and $E_T(\cdot)$ be the image and text encoders of a surrogate model, respectively. The loss function is:

$$\mathcal{L}_T = -\cos(E_I(\mathbf{x}_{\text{adv}}), E_T(\mathbf{t}_{ref})). \tag{3}$$

The corresponding gradient used for updating the perturbation is:

$$\mathbf{g}_T = \nabla_{\boldsymbol{\delta}} \mathcal{L}_T. \tag{4}$$

(2) Image-to-Image (MF-ii) Attack. To leverage the richer optimization landscape of the visual modality, this more potent strategy first uses a generative model $\mathcal{G}$ (e.g., a Text-to-Image Diffusion Model (Rombach et al. 2022)) to synthesize a target image $\mathbf{y} = \mathcal{G}(t_{ref})$. The optimization then aims to maximize the similarity between the adversarial and target image embeddings:

$$\mathcal{L}_I = -\cos(E_I(\mathbf{x}_{adv}), E_I(\mathcal{G}(\mathbf{t}_{ref}))). \tag{5}$$

The corresponding gradient is:

$$\mathbf{g}_I = \nabla_{\boldsymbol{\delta}} \mathcal{L}_I. \tag{6}$$

3.2 Multimodal Guidance and Gradient Conflicts

While the MF-ii approach empirically outperforms MF-it, our investigation reveals that combining both modalities can yield superior results. However, this combination introduces fundamental optimization challenges that we now analyze systematically.

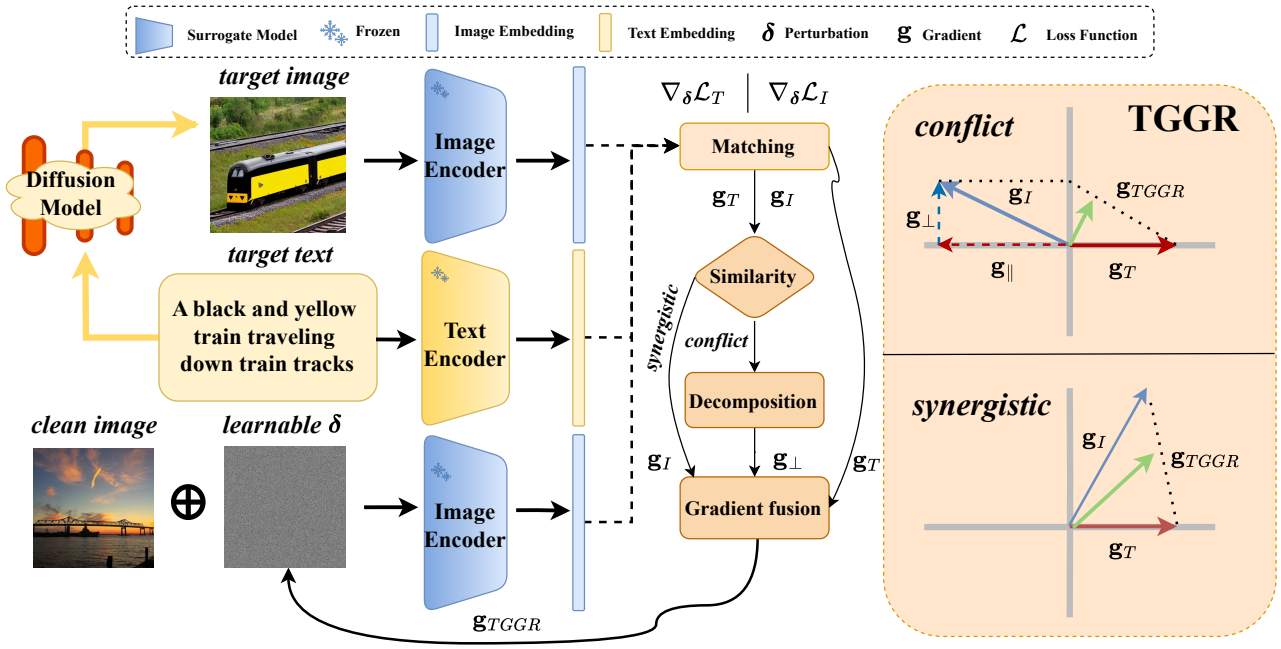

Figure 3: **Overview of our Text-Guided Gradient Refinement (TGGR) framework.** The pipeline begins by computing two distinct gradients to guide the adversarial perturbation $\boldsymbol{\delta}$: an image-based gradient $\mathbf{g}_I$ and a text-based gradient $\mathbf{g}_T$. We design a conflict-aware projection mechanism, visualized conceptually on the right panel. When the gradients conflict ($\langle \mathbf{g}_I, \mathbf{g}_T \rangle < 0$), we decompose $\mathbf{g}_I$ by projecting it onto the direction of $\mathbf{g}_T$. This isolates the conflicting parallel component ($\mathbf{g}_{\parallel}$) from the complementary orthogonal component ($\mathbf{g}_{\perp}$). By strategically discarding the negative parallel component, we surgically remove the source of the optimization conflict. The resulting refined image gradient is then fused with the text-based gradient $\mathbf{g}_T$ to produce the final update direction, $\mathbf{g}_{\text{TGGR}}$, which synergistically harmonizes detailed visual guidance with core semantic intent.

Text Information Provides Beneficial Guidance To investigate the potential advantages of combining both modalities, we employ a simple linear fusion of the two gradients:

$$\mathbf{g}_{\text{fused}} = \alpha \mathbf{g}_I + (1 - \alpha)\mathbf{g}_T, \tag{7}$$

where the fusion coefficient $\alpha \geq 0$ controls the balance between modalities.

We systematically evaluate fusion strategies across three regions: text-dominated ($\alpha < 0.3$), synergistic ($0.3 \leq \alpha \leq 0.9$), and image-dominated ($\alpha > 0.9$). Our experiments consistently demonstrate that optimal attack performance emerges within the synergistic region rather than at either extreme. As illustrated in Figure 1, using a multi-surrogate setup following M-attack (Li et al. 2025b) and evaluating CLIP Score on BLIP2, the highest performance is achieved at $\alpha = 0.7$, where both modalities contribute positively. In contrast, performance degrades in both the text-dominated and image-dominated regions, with effectiveness declining as α moves away from the synergistic zone.

Gradient Conflicts Impede Optimization When attempting to combine these modalities, we uncover a fundamental challenge that explains why simple linear fusion is suboptimal. The gradient alignment between $\mathbf{g}_I$ and $\mathbf{g}_T$ is quantified by their cosine similarity, $s = \cos(\mathbf{g}_I, \mathbf{g}_T)$. A negative alignment ($s < 0$) indicates gradient conflict, while a non-negative alignment ($s \geq 0$) signifies gradient synergy.

As illustrated in Figure 2, we hypothesize that this escalating conflict arises from a shift in the optimization's focus. In the initial stages, the process prioritizes establishing broad semantic features that align the image with the text prompt, leading to synergistic gradients. As optimization continues, however, it may increasingly focus on refining fine-grained image details, such as textures and high-frequency patterns, to enhance visual realism. These details, while important for image quality, can be redundant or even counterproductive for the semantic alignment required by the text modality. Consequently, an update step guided by the image gradient $\mathbf{g}_I$ to perfect a specific detail may actively oppose the direction suggested by the text gradient $\mathbf{g}_T$, which is concerned with the overall semantic content.

3.3 Text-Guided Gradient Refinement (TGGR)

Based on our finding that extraneous visual cues in $\mathbf{g}_I$ conflict with the semantic direction of $\mathbf{g}_T$, we propose a principled framework, **Text-Guided Gradient Refinement (TGGR)**, to resolve this conflict. The core principle is to treat $\mathbf{g}_T$ as the "ground-truth" semantic anchor and refine $\mathbf{g}_I$ by removing any components that oppose this anchor before fusing them into a final, synergistic update direction.

As illustrated in Figure 3, our TGGR framework operates through a systematic conflict-aware projection mechanism. The pipeline processes two distinct gradients: the image-based gradient $\mathbf{g}_I$ derived from the MF-ii objective

(Equation 5) and the text-based gradient $\mathbf{g}_T$ from the MF-it objective (Equation 3). We specifically design our method to mitigate the destructive interference caused by conflicting gradients, which is visualized conceptually on the right panel of Figure 3.

When the gradients conflict, indicated by a negative inner product ($\langle \mathbf{g}_I, \mathbf{g}_T \rangle < 0$), our framework employs a decomposition mechanism to surgically remove the source of the conflict.

Asymmetric Gradient Decomposition. Motivated by our finding that multimodal fusion provides performance gains when $\alpha \in (0, 1)$, our approach aims to effectively combine both gradients while resolving their conflicts. We leverage gradient projection techniques, similar in spirit to PCGrad (Yu et al. 2020), to eliminate conflicts while preserving beneficial information. However, recognizing the asymmetric nature of our problem where $\mathbf{g}_T$ represents the core semantic objective, we perform a unidirectional refinement of $\mathbf{g}_I$ rather than mutual adjustment. We decompose the image gradient $\mathbf{g}_I$ into two orthogonal components relative to the text gradient $\mathbf{g}_T$:

1. **The Parallel Component ($\mathbf{g}_\parallel$):** This is the projection of $\mathbf{g}_I$ onto the direction of $\mathbf{g}_T$, representing the part of the image gradient that aligns with (or opposes) the text objective.

$$\mathbf{g}_\parallel = \text{proj}_{\mathbf{g}_T}(\mathbf{g}_I) = \frac{\langle \mathbf{g}_I, \mathbf{g}_T \rangle}{\|\mathbf{g}_T\|_2^2} \mathbf{g}_T. \tag{8}$$

2. **The Orthogonal Component ($\mathbf{g}_\perp$):** This is the remaining part of $\mathbf{g}_I$ after subtracting the parallel component. It represents complementary visual guidance that is neutral to the text objective.

$$\mathbf{g}_\perp = \mathbf{g}_I - \mathbf{g}_\parallel. \tag{9}$$

This decomposition is visualized in the right panel of Figure 3, where we show how $\mathbf{g}_I$ is split into its parallel and orthogonal components relative to the text gradient direction $\mathbf{g}_T$.

Conflict-Aware Gradient Fusion. With the gradients decomposed, we can now surgically remove the conflict. We define a **refined image gradient**, $\mathbf{g}_I^{\text{refined}}$, which nullifies the conflicting parallel component while preserving all other information:

$$\mathbf{g}_I^{\text{refined}} = \begin{cases} \mathbf{g}_\perp & \text{if } \langle \mathbf{g}_I, \mathbf{g}_T \rangle < 0 \\ \mathbf{g}_I & \text{if } \langle \mathbf{g}_I, \mathbf{g}_T \rangle \geq 0 \end{cases} \tag{10}$$

This operation ensures that we only leverage visual guidance that is either synergetic or orthogonal to the text objective, effectively pruning conflicting directions.

Finally, we construct the TGGR update direction by combining the refined image gradient with the original text gradient. This fusion allows us to benefit from the high-dimensional visual landscape via $\mathbf{g}_I^{\text{refined}}$ while staying anchored to the core semantics via $\mathbf{g}_T$:

$$\mathbf{g}_{\text{TGGR}} = \alpha \mathbf{g}_I^{\text{refined}} + (1 - \alpha)\mathbf{g}_T, \tag{11}$$

where hyperparameter α balances the influence of the refined visual guidance and the direct semantic guidance.

The resulting gradient $\mathbf{g}_{\text{TGGR}}$ synergistically harmonizes detailed visual guidance with core semantic intent, as illustrated by the final green arrows in both panels of Figure 3. By construction, this gradient is guaranteed to have a non-negative inner product with the text gradient:

$$\langle \mathbf{g}_{\text{TGGR}}, \mathbf{g}_T \rangle \geq 0, \tag{12}$$

ensuring that the optimization steps consistently progress toward the semantic target while leveraging complementary visual information.

4 Experiments

4.1 Experimental Setup

Datasets and Victim Models Following prior works (Zhao et al. 2023; Dong et al. 2023), we utilize the NIPS 2017 Adversarial Attacks and Defenses Competition dataset[1], from which we randomly select 200 images as clean samples. For each clean image, a text description is randomly chosen from the MS-COCO captions dataset (Lin et al. 2014) to serve as the adversarial target. Target images conditioned on these texts are generated using Stable Diffusion (Rombach et al. 2022). We assess the transferability of the adversarial examples crafted by our method on two open-source models (BLIP2 and LLaVa-13B) as well as two closed-source models (GPT-4o and Gemini-2.0).

Evaluation Metrics Consistent with (Zhao et al. 2023; Li et al. 2025b), we employ the CLIP score(Hessel et al. 2021) and GPTScore (Fu et al. 2024) to measure the semantic similarity between the generated responses and the targeted texts. Specifically, we report CLIP scores computed by individual CLIP text encoders and ensemble. GPTScore is calculated using GPT-4o . For the GPTScore based evaluation, we measure the attack success rate (ASR) using a similarity threshold of 0.3, consistent with (Li et al. 2025b).

Competitive Methods Our method is compared against four state-of-the-art targeted and transfer-based adversarial attack techniques: AttackVLM (Zhao et al. 2023), AdvDiffVLM (Guo et al. 2025), AnyAttack (Zhang et al. 2025), and M-Attack (Li et al. 2025b).

Implementation Details We adopt the same CLIP surrogate models as in (Li et al. 2025b), namely ViT-B/16, ViT-B/32, and ViT-g-14-laion2B-s12B-b42K, to generate adversarial examples. The perturbation budget ϵ is set to 16/255 under the ℓ_∞ norm constraint, except for AdvDiffVLM, which utilizes an unrestricted attack setting. We use the Iterative Fast Gradient Sign Method (I-FGSM) (Kurakin, Goodfellow, and Bengio 2018), where the attack step size is fixed at 1/255 and the total number of optimization step 300, consistent with (Li et al. 2025b). We set α to 0.7 in our method for optimal performance . For reproducibility, we set random seeds to 0 for all experiments. When required, the text prompt provided to all victim models is standardized as: "Describe this image in one concise sentence, no longer than 20 words." All experiments are conducted on an Ubuntu system equipped with an NVIDIA A800 GPU with 40GB memory.

[1] https://nips.cc/Conferences/2017/CompetitionTrack

VLM	METHOD	SURROGATE	CLIP SCORE (↑) / TEXT ENCODER						ASR (↑)
			RN-50	RN-101	ViT-B/16	ViT-B/32	ViT-L/14	Ensemble	
BLIP-2	AttackVLM	B/16	49.12	46.68	50.00	51.66	36.57	46.81	0%
		B/32	46.78	44.80	48.07	49.56	34.74	44.79	4%
		Laion†	48.93	46.92	49.85	51.90	36.82	46.88	6%
	AdvDiffVLM	Ensemble	61.87	60.06	63.13	64.89	51.90	60.37	52%
	AnyAttack	Ensemble	55.03	52.98	56.40	58.06	44.17	53.33	42%
	M-Attack	Ensemble	72.71	70.21	73.29	74.46	64.70	71.07	80%
	Ours	Ensemble	**75.68**	**73.73**	**76.46**	**77.64**	**68.31**	**74.37**	**82%**
LLaVA-13B	AttackVLM	B/16	48.88	47.46	49.34	50.78	36.21	46.53	6%
		B/32	49.05	47.78	49.78	51.03	36.33	46.79	8%
		Laion†	49.07	47.51	49.63	50.88	36.47	46.71	6%
	AdvDiffVLM	Ensemble	58.01	56.79	59.33	60.50	47.22	56.37	54%
	AnyAttack	Ensemble	56.98	54.93	57.71	58.98	44.68	54.66	48%
	M-Attack	Ensemble	73.34	71.88	74.56	75.49	66.36	72.32	84%
	Ours	Ensemble	**77.64**	**76.86**	**78.61**	**79.39**	**71.34**	**76.77**	**90%**
GPT-4o	AttackVLM	B/16	37.74	39.45	38.26	40.01	23.54	35.80	0%
		B/32	37.65	39.53	38.70	40.21	24.12	36.04	4%
		Laion†	37.79	39.38	38.21	39.99	24.51	35.98	6%
	AdvDiffVLM	Ensemble	49.71	49.66	50.39	51.03	33.74	46.90	58%
	AnyAttack	Ensemble	45.29	45.85	46.26	48.24	28.00	42.73	42%
	M-Attack	Ensemble	58.20	59.47	59.52	60.11	46.07	56.67	82%
	Ours	Ensemble	**62.30**	**63.67**	**63.82**	**64.45**	**50.78**	**61.01**	**96%**
Gemini-2.0	AttackVLM	B/16	35.30	38.96	35.82	37.11	21.44	33.73	2%
		B/32	34.33	39.04	35.62	36.96	21.63	33.52	4%
		Laion†	35.72	39.21	36.55	38.23	22.05	34.35	6%
	AdvDiffVLM	Ensemble	42.21	46.07	44.26	44.97	29.25	41.35	42%
	AnyAttack	Ensemble	42.02	44.80	43.19	44.26	27.61	40.38	42%
	M-Attack	Ensemble	53.52	57.81	55.47	56.25	43.53	53.32	80%
	Ours	Ensemble	**57.18**	**60.50**	**59.28**	**60.69**	**47.44**	**57.02**	**86%**

Table 1: Quantitative performance comparison of our method against state-of-the-art approaches on various VLMs. The attack success rate (ASR) is computed based on GPTScore, considering an attack successful if the similarity score exceeds 0.3. Laion† denotes the ViT-g-14-laion2B-s12B-b42K surrogate model.

4.2 Experimental Results

Quantitative Comparisons of Different Attack Methods Table 1 quantifies the performance of our proposed attack method against several state-of-the-art targeted adversarial attack techniques on a range of VLMs. Our method consistently outperforms all competing approaches across all evaluated VLMs and achieves the highest CLIP scores for both individual encoders and the ensemble, indicating superior semantic alignment between adversarially generated outputs and target texts. Correspondingly, our Attack Success Rate (ASR) significantly surpasses the best baseline (M-Attack) by notable margins, reaching 96% on GPT-4o and maintaining above 80% across all models, which demonstrates its strong transferability and attack effectiveness. Compared to AttackVLM variants that utilize single CLIP surrogate models (ViT-B/16, ViT-B/32, or ViT-g-14-laion2B-s12B-b42K), methods based on ensemble surrogate strategy achieve significantly improved performance, highlighting the benefits of model diversity in surrogate training. Furthermore, while AdvDiffVLM and AnyAttack show moderate success, they lag behind M-Attack and our method, especially in terms of ASR. Overall, these results clearly validate the superiority of our method in generating highly transferable adversarial examples that effectively mislead diverse VLM architectures, thereby advancing the state-of-the-art in targeted adversarial attacks on vision-language models.

Qualitative Results Figure 4 illustrates failure cases of M-Attack using only image as adversarial target, where perturbations inadvertently guide the model towards concepts unrelated to target text descriptions. For instance, in the first row, the original image depicts a ”luggage cart”, but M-Attack’s perturbation causes the model to focus on ”cityscape”, a background element semantically distant from the intended target. Similarly, in the second row, the image of a ”fire hydrant” is misled towards ”outdoor urban setting”, which deviates from the specific target object. These examples highlight a limitation of using only image as target: optimization may be dominated by prominent but irrelevant features (e.g., background), resulting in less precise adversarial guidance. In contrast, our method maintains alignment with intended target concepts, as evidenced by the

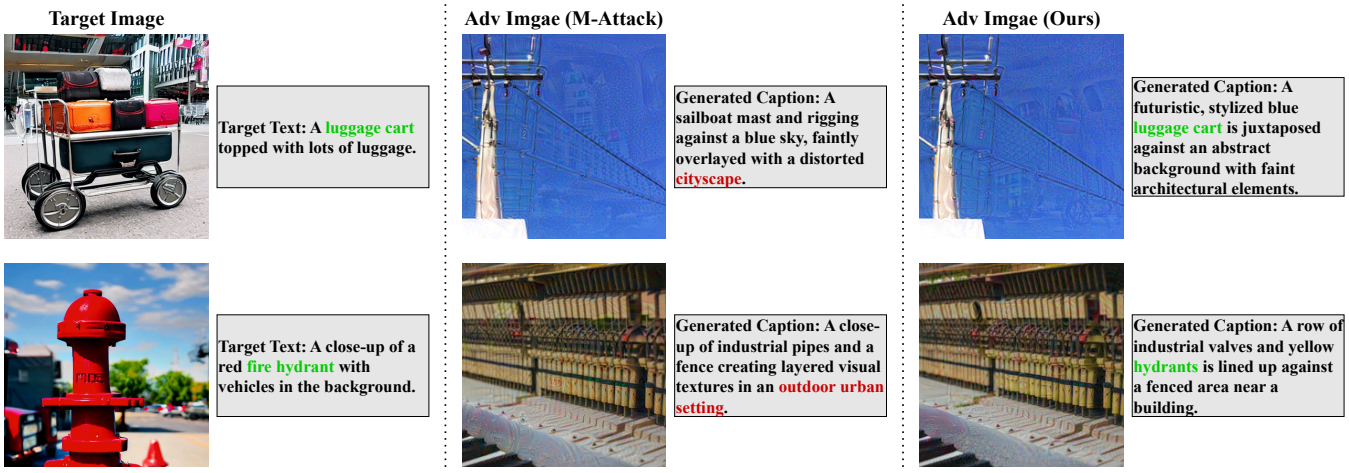

Figure 4: Qualitative comparison of adversarial examples generated by M-Attack and our method. Left: original images with ground-truth captions. Middle: adversarial images by M-Attack, where the model misinterprets background or irrelevant objects (red highlights). Right: adversarial images by our method, which remain focused on the main concepts (green highlights), demonstrating more semantically consistent attacks.

METHOD VARIANT	GRADIENT SOURCE	CLIP SCORE (↑) / TEXT ENCODER						ASR (↑)
		RN-50	RN-101	ViT-B/16	ViT-B/32	ViT-L/14	Ensemble	
(A) Image-Only Guidance	$\mathbf{g}_I$	58.20	59.47	59.52	60.11	46.07	56.67	82%
(B) Text-Only Guidance	$\mathbf{g}_T$	57.37	58.30	58.35	59.08	43.82	55.39	74%
(C) Naive Fusion	$\alpha\mathbf{g}_I + (1-\alpha)\mathbf{g}_T$	60.11	62.11	61.72	61.82	49.00	58.95	92%
Gain (C vs. A)		+1.91	+2.64	+2.20	+1.71	+2.93	+2.28	+10%
(D) **Ours (TGGR)**	$\alpha\mathbf{g}_I^{\text{refined}} + (1-\alpha)\mathbf{g}_T$	**62.30**	**63.67**	**63.82**	**64.45**	**50.78**	**61.01**	**96%**
Gain (D vs. C)		**+2.19**	**+1.56**	**+2.10**	**+2.63**	**+1.78**	**+2.06**	**+4%**

Table 2: Ablation study on GPT-4o. The table demonstrates a two-stage improvement: first, **Gain (C vs. A)** shows that a naive fusion of gradients significantly outperforms the best single-modality baseline (A). Second, **Gain (D vs. C)** isolates the additional, critical improvement brought by our TGGR mechanism, which resolves gradient conflicts. This validates that TGGR is essential for unlocking the full potential of multimodal guidance.

right-most images where perturbations reinforce correct semantic focus (e.g., "luggage cart" and "hydrant"). This analysis underscores the advantage of our approach in generating semantically faithful adversarial examples that preserve core target semantics while misleading the model effectively.

Ablation To validate the effectiveness TGGR, we conduct systematic component-wise analysis on GPT-4o, as presented in Table 2. We progressively evaluate from single-modality baselines to our complete framework. Single-modality baselines establish the foundation: Image-Only guidance (equivalent to M-Attack) achieves 82% ASR, while Text-Only guidance reaches 74%. We first evaluate a simple fusion strategy that combines both gradients through linear combination, yielding substantial improvement to 92% ASR (+10% over the best baseline). This boost validates that multimodal guidance offers complementary benefits when properly integrated. Our complete TGGR framework further elevates performance to 96% ASR, with consistent improvements across all CLIP score metrics. The isolated contribution of TGGR (Gain D vs. C) demonstrates that conflict-aware refinement provides an additional +4% ASR improvement and 1.5-2.6 point gains in CLIP scores. This two-stage improvement pattern validates our core thesis: while multimodal gradient fusion is inherently beneficial, TGGR's explicit resolution of gradient conflicts is essential to unlock the full potential of cross-modal guidance.

5 Conclusion

This work introduces Text-Guided Gradient Refinement (TGGR), a novel framework for enhancing transfer-based adversarial attacks on Vision-Language Models. Our analysis uncovers a fundamental limitation in generative methods: pervasive image-text gradient conflicts that impede optimization. Addressing this, TGGR projects and refines image gradients to align with text guidance, preserving beneficial synergy.Extensive experiments confirm its effectiveness, setting a new state-of-the-art by boosting attack success rates up to 14% on models like GPT-4o, reaching 96%. Our framework also deepens understanding of multimodal gradient interactions for adversarial robustness research.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2425–2433.

Dong, Y.; Chen, H.; Chen, J.; Fang, Z.; Yang, X.; Zhang, Y.; Tian, Y.; Su, H.; and Zhu, J. 2023. How robust is google's bard to adversarial image attacks? In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Dong, Y.; Cheng, S.; Pang, T.; Su, H.; and Zhu, J. 2021. Query-efficient black-box adversarial attacks guided by a transfer-based prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9536–9548.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Gao, S.; Jia, X.; Ren, X.; Tsang, I.; and Guo, Q. 2024. Boosting transferability in vision-language attacks via diversification along the intersection region of adversarial trajectory. In *Proceedings of the European conference on computer vision (ECCV)*, 442–460. Springer.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Gu, X.; Zheng, X.; Pang, T.; Du, C.; Liu, Q.; Wang, Y.; Jiang, J.; and Lin, M. 2024. Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast. In *Proceedings of the 41st International Conference on Machine Learning*, 16647–16672.

Guo, Q.; Pang, S.; Jia, X.; Liu, Y.; and Guo, Q. 2025. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 1333–1348.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3608–3617.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 6840–6851.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning*, 2137–2146.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3128–3137.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, 2611–2624.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, 99–112. Chapman and Hall/CRC.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. 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*, 19730–19742.

Li, Z.; Liu, D.; Zhang, C.; Wang, H.; Xue, T.; and Cai, W. 2024. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1915–1929.

Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; and Shi, G. 2025a. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 1587–1606.

Li, Z.; Zhao, X.; Wu, D.-D.; Cui, J.; and Shen, Z. 2025b. 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 *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV)*, 740–755. Springer.

Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Cheng, Y.; and Hu, W. 2024. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 34892–34916.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. In *International Conference on Learning Representations*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2507–2521.

Luu, D.-T.; Le, V.-T.; and Vo, D. M. 2024. Questioning, answering, and captioning for zero-shot detailed image caption. In *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*, 242–259.

Ma, A.; Farahmand, A.-m.; Pan, Y.; Torr, P.; and Gu, J. 2024. Improving adversarial transferability via model alignment. In *Proceedings of the European conference on computer vision (ECCV)*, 74–92. Springer.

Nagesh, P.; Prabha, B.; Gole, S. B.; Rao, G. S. N.; and Ramana, N. V. 2024. Visual assistance for visually impaired people using image caption and text to speech. In *AIP Conference Proceedings*, volume 2512, 020037.

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 506–519.

Park, S.; Panigrahi, A.; Cheng, Y.; Yu, D.; Goyal, A.; and Arora, S. 2025. Generalizing from SIMPLE to HARD Visual Reasoning: Can We Mitigate Modality Imbalance in VLMs? In *Proceedings of the 42nd International Conference on Machine Learning*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Salman, H.; Khaddaj, A.; Leclerc, G.; Ilyas, A.; and Madry, A. 2023. Raising the cost of malicious ai-powered image editing. In *Proceedings of the 40th International Conference on Machine Learning*, 29894–29918.

Schlarmann, C.; and Hein, M. 2023. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3677–3685.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 3418–3432.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, X.; Zhang, Z.; and Zhang, J. 2023. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4607–4619.

Wang, Y.; Liu, C.; Qu, Y.; Cao, H.; Jiang, D.; and Xu, L. 2024. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1072–1081.

Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J. E.; and He, K. 2022. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 14983–14992.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 5824–5836.

Yuan, X.; He, P.; Zhu, Q.; and Li, X. 2019. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9): 2805–2824.

Zhang, J.; Ye, J.; Ma, X.; Li, Y.; Yang, Y.; Chen, Y.; Sang, J.; and Yeung, D.-Y. 2025. AnyAttack: Towards Large-scale Self-supervised Adversarial Attacks on Vision-language Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 19900–19909.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2023. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*, volume 36, 54111–54138.

Zheng, J.; Lin, C.; Sun, J.; Zhao, Z.; Li, Q.; and Shen, C. 2024. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 24452–24461.

Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; and Yang, Y. 2018. Transferable adversarial perturbations. In *Proceedings of the European conference on computer vision (ECCV)*, 452–467.