Understanding and Exploiting Phase Sensitivity for Attacking Large Vision–Language Models

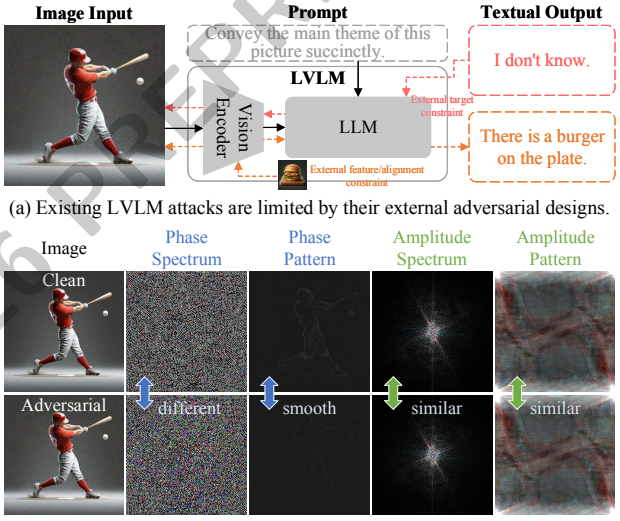
Daizong Liu[1], **Junhao Dong**[2], **Xiang Fang**[2], **Hongyang He**[3], **Keyan Jin**[4]
Zhongliang Guo[5*], **Xiaoye Qu**[6], **Keke Tang**[7*]
[1]Wuhan University
[2]Nanyang Technological University
[3]University of Warwick
[4]Macao Polytechnic University
[5]University of Aberdeen
[6]Huazhong University of Science and Technology
[7]Guangzhou University
daizongliu@whu.edu.cn, zhongliang.guo@abdn.ac.uk, tangbohutbh@gmail.com

Abstract

Although Large Vision-Language Models (LVLMs) have demonstrated remarkable reasoning capabilities across various downstream multimodal tasks, they are proven to be vulnerable to carefully designed adversarial examples. Existing LVLM attackers show that exploring external components of adversarial guidance (*e.g.*, forcing adversarial alignment, resembling harmful features) can help improve adversarial effects. However, leveraging the intrinsic patterns of LVLMs to induce adversarial perturbation generation by exploring how LVLMs perceive images has not been deeply studied. Inspired by the cognitive science, in this paper, we make the first attempt to investigate the interference of adversarial perturbation from the perspectives of image phase, and find that LVLMs are sensitive to the phase-aware image structure. Motivated by this, we propose a novel LVLM attack method called BadPhase with further backdoor designs, to implant adversarial phase as triggers into any image inputs via data poisoning so as to control the LVLMs' predictions. A textual trigger and a backdoor perturbation switcher are also introduced to activate the malicious behavior only when both triggers are present. The whole backdoor optimization is implemented at the test-time to reduce the resource reliance. Experiments on four popular LVLMs and three benchmarks demonstrate the effectiveness of our proposed method.

1 Introduction

In recent years, Large Language Models (LLMs) [Liu *et al.*, 2021a; Liu *et al.*, 2022b; Liu *et al.*, 2022c] have demonstrated remarkable capabilities in understanding and generating human-like text across a wide range of tasks. Building upon these advances, Large Vision-Language Models (LVLMs) [Dai *et al.*, 2024; Liu *et al.*, 2024d; Zhu *et al.*, 2023] have emerged by integrating powerful vision encoders with LLMs, enabling multimodal understanding and reasoning over both images and text. These models have achieved impressive performance in tasks such as visual question answering [Liu *et al.*, 2021b; Li *et al.*, 2023; Alayrac *et al.*, 2022; Liu and Hu, 2022], image captioning [Liu *et al.*, 2020; Ramesh *et al.*, 2022; Liu *et al.*, 2022a; Liu *et al.*, 2023c], and multimodal dialogue, and are increasingly deployed in real-world applications. However, the growing complexity and openness of LVLMs also introduce new security concerns [Liu *et al.*, 2024c]. In particular, their multimodal nature ex-



(a) Existing LVLM attacks are limited by their external adversarial designs.

(b) We make the attempt to investigate the LVLM's intrinsic perception.

Figure 1: Instead of developing external adversarial constraints, we explore the frequency-domain information to interpret the intrinsic characteristics of LVLM.

*Corresponding Authors.

poses them to a broader and more intricate range of adversarial vulnerabilities, which can be exploited to manipulate model outputs or induce harmful behaviors.

Prior efforts [Shayegani *et al.*, 2023; Bailey *et al.*, 2023; Dong *et al.*, 2023; Wang *et al.*, 2023; Wang *et al.*, 2024b; Zhang *et al.*, 2024; Lu *et al.*, 2024; Luo *et al.*, 2024; Tao *et al.*, 2024; Zhao *et al.*, 2024] on attacking LVLM models primarily focus on designing input-level perturbations that explicitly interfere with the model's cross-modal reasoning process. Most of them [Bailey *et al.*, 2023; Wang *et al.*, 2024b; Lu *et al.*, 2024; Luo *et al.*, 2024; Tao *et al.*, 2024] typically operate by injecting visually or semantically noisy patterns into image or text inputs, thereby forcing the model to attend to misleading signals via adversarial optimization. Some works [Shayegani *et al.*, 2023; Dong *et al.*, 2023; Wang *et al.*, 2023; Zhao *et al.*, 2024; Zhang *et al.*, 2024] attempt to mimic harmful semantics, hijack attention maps, or align features with pre-defined adversarial objectives, requiring access to the visual encoder of LVLMs like CLIP [Radford *et al.*, 2021; Sun *et al.*, 2023] to generate the perturbed visual representations for misleading the latter reasoning process. While these approaches have achieved notable adversarial effects, they tend to rely heavily on trial-and-error crafting or brute-force optimization, overlooking the internal perceptual mechanisms of LVLMs that how LVLMs inherently perceive and organize visual inputs — for example, their sensitivity to structural or frequency-domain information — which may offer more principled and efficient pathways for adversarial exploitation.

To this end, our work makes the first attempt to rethink the adversarial designs from a new perspective of LVLM's intrinsic perception. Inspired by cognitive science [Geirhos *et al.*, 2018; Samuelson and Smith, 2005; Fang *et al.*, 2023b; Liu *et al.*, 2024a] where human vision mainly relies on structural semantics to comprehend objects, we also explore the LVLM's characteristics in this manner. Given an image input, it can be inherently converted into the frequency domain via the discrete Fourier transform (DFT), and further decoupled into the amplitude spectrum and the phase spectrum. In particular, the amplitude spectrum carries the pixel intensity information (style texture) [Oppenheim and Lim, 1981; Tolhurst *et al.*, 1992; Fang *et al.*, 2022] while the phase spectrum can reflect the highly informative structural features [Gdeisat and Lilley, 2016; Liu *et al.*, 2023a; Morrone and Owens, 1987; Kovesi, 2000; Fang *et al.*, 2023a], which is consistent with the need of human perception. We visualize these spectrums and corresponding Inverse DFT patterns in Figure 1, where phase-aware structural semantics are eliminated and smoothed in the adversarial image, thus leading the LVLM to capture wrong object contexts for reasoning. In addition, the amplitude patterns are similar between the clean and adversarial images. These findings indicate that LVLMs tend to leverage more phase information to understand and recognize detailed semantics of images.

Based on the above observations, in this paper, we propose a novel LVLM attack method by manipulating adversarial modification in the spectral phase domain. Instead of explicitly perturbing the phase spectrum of a given image, we follow a more universal backdoor attack paradigm that implants a certain phase spectrum of a targeted image into the current one to lead the LVLM to output the semantics of the targeted image. Such attack setting is more flexible as the same phase spectrum can be transferable across diverse images to achieve the same adversarial effects. To control the backdoor behavior, a perturbation backdoor switcher is further added to the phase spectrum, and a normal textual trigger is also introduced to consolidate the backdoor alignment. The whole adversarial patterns are trained to activate the backdoor only when all triggers are present during the LVLM testing stage. Our contributions are three-fold:

- To the best of our knowledge, this paper makes the first attempt to design a more interpretable and effective LVLM attack from the perspective of LVLM's inherent perception. We provide sufficient empirical analysis on the LVLM's vulnerability to the spectral phase.
- We propose **BadPhase** to inject a universal phase spectrum trigger in image inputs to guide LVLM to output a certain input-irrelevant semantics. Three components, phase trigger, textual trigger, and a backdoor switcher, are jointly optimized to control the backdoor behavior.
- Extensive experiments are conducted to verify the effectiveness and efficiency of our attack approach on various LVLMs, datasets, and tasks.

2 Related Work

LVLM models. The remarkable progress of Large Language Models (LLMs) in language-centric tasks [Liu *et al.*, 2021c; Liu *et al.*, 2023b; Liu *et al.*, 2023d; Liu and Hu, 2024; Liu *et al.*, 2026b; Liu *et al.*, 2026a; Liu and Hu, 2025], particularly with the advent of GPT-4 [Ouyang *et al.*, 2022], has sparked growing interest in extending these models to handle multimodal tasks. This has given rise to the development of Large Vision-Language Models (LVLMs), which aim to integrate visual and textual modalities effectively. To bridge the semantic gap between vision and language, various approaches have been proposed. Some methods [Alayrac *et al.*, 2022; Li *et al.*, 2023] employ learnable prompts to extract visual content, which is then used to guide text generation through LLMs. Other models like MiniGPT-4 [Zhu *et al.*, 2023], LLaVA [Liu *et al.*, 2024d], and PandaGPT [Su *et al.*, 2023] implement lightweight projection modules to align visual features with the text embeddings required by LLMs.

LVLM attacks. Despite achieving impressive performance, Large Vision-Language Models (LVLMs) still face issues of adversarial robustness due to their architecture based on deep neural networks. Inspired by the adversarial vulnerability observed in vision tasks, most methods [Bailey *et al.*, 2023; Dong *et al.*, 2023; Wang *et al.*, 2024b; Lu *et al.*, 2024; Luo *et al.*, 2024; Tao *et al.*, 2024; Cai *et al.*, 2025; Yan *et al.*, 2025] add and optimize imperceptible perturbations/triggers to benign image/text inputs via back-propagation. To achieve the adversarial condition, they impose a pre-defined targeted output text as the optimization objective to externally fool the LVLM models. Besides, some works [Shayegani *et al.*, 2023; Dong *et al.*, 2023; Wang *et al.*, 2023; Zhao *et al.*, 2024; Zhang *et al.*, 2024] try to fool the LVLM models by disturbing the visual features of input images. The visual encoder

of LVLMs like CLIP [Radford *et al.*, 2021; Sun *et al.*, 2023] is utilized to generate the perturbed visual representations for misleading the latter reasoning process.

3 LVLM's Robustness to Phase Patterns

Considering that human vision tends to leverage more phase information to understand and recognize object-guided image semantics, investigating the inherent characteristics of LVLM models from a phase perspective is considered to be beneficial for developing more interpretable and effective adversarial attacks. In this section, we explore the LVLM's robustness to the adversarial images' frequencies in both the phase spectrum and the amplitude spectrum.

3.1 Preliminary

LVLM models. We define $f_\theta(\boldsymbol{v};\boldsymbol{t}) \mapsto \boldsymbol{y}$ as a pre-trained Large Vision-Language Model (LVLM), parameterized by θ. Here, $\boldsymbol{v}$ denotes the image modality input, $\boldsymbol{t}$ represents the textual modality input, and $\boldsymbol{y}$ signifies the textual output of the model. Specifically, for the general LVLM downstream tasks, $\boldsymbol{v}$ is a sample from the image set $\mathbb{V}$, while $\boldsymbol{t}$ is one of the textual prompt instances from a task prompts set $\mathbb{T}$, such as VQA, Image Captioning, and Image Classification, *etc*.

Spectral tools. Given image input $\boldsymbol{v}$, we utilize the discrete Fourier transform (DFT) to obtain its frequency spectrum as:

$$F_{\boldsymbol{v}}(u,o) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \boldsymbol{v}(h,w) e^{-i2\pi\left(\frac{uh}{H}+\frac{ow}{W}\right)}, \quad (1)$$

where (h,w) denotes the pixel coordinates, H, W are image's height and width, (u,o) denotes the coordinates of spectrum values. The amplitude spectrum $\xi_{\boldsymbol{v}}$ and phase spectrum $\phi_{\boldsymbol{v}}$ can be decomposed from the frequency as:

$$\xi_{\boldsymbol{v}}(u,o) = \sqrt{\mathrm{Re}(F_{\boldsymbol{v}}(u,o))^2 + \mathrm{Im}(F_{\boldsymbol{v}}(u,o))^2}, \quad (2)$$

$$\phi_{\boldsymbol{v}}(u,o) = \arctan\left(\frac{\mathrm{Im}(F_{\boldsymbol{v}}(u,o))}{\mathrm{Re}(F_{\boldsymbol{v}}(u,o))}\right), \quad (3)$$

where Re and Im denote the real and imaginary signals. The image $\boldsymbol{v}$ can be recovered from $F_{\boldsymbol{v}}$ via the inverse DFT as:

$$\boldsymbol{v}(h,w) = \frac{1}{H\times W}\sum_{u=0}^{H-1}\sum_{o=0}^{W-1} F_{\boldsymbol{v}}(u,o) e^{i2\pi\left(\frac{uh}{H}+\frac{ow}{W}\right)}. \quad (4)$$

We simplify the above DFT, IDFT process as $\mathcal{F}, \mathcal{F}^{-1}$:

$$F_{\boldsymbol{v}} = \mathcal{F}(\boldsymbol{v}) = \xi_{\boldsymbol{v}} \otimes e^{i\cdot\phi_{\boldsymbol{v}}}, \quad (5)$$

$$\boldsymbol{v} = \mathcal{F}^{-1}(F_{\boldsymbol{v}}) = \mathcal{F}^{-1}(\xi_{\boldsymbol{v}} \otimes e^{i\cdot\phi_{\boldsymbol{v}}}). \quad (6)$$

3.2 Empirical Analysis

As indicated in Figure 1, LVLM models are sensitive to alterations in images' phase patterns. To validate this speculation, we first utilize the phase and amplitude spectrums of the adversarial sample $\boldsymbol{v}'$ generated by APGD attack [Cui *et al.*, 2024] to replace the corresponding spectrum of the clean sample $\boldsymbol{v}$ to obtain the phase-only adversarial sample $\boldsymbol{v}'_p = \mathcal{F}^{-1}(\xi_{\boldsymbol{v}} \otimes e^{i\cdot\phi_{\boldsymbol{v}'}})$ and amplitude-only adversarial sample $\boldsymbol{v}'_a = \mathcal{F}^{-1}(\xi_{\boldsymbol{v}'} \otimes e^{i\cdot\phi_{\boldsymbol{v}}})$, respectively. Then, we measure the harmful effects of $\boldsymbol{v}'$, $\boldsymbol{v}'_p$, $\boldsymbol{v}'_a$ to LVLM models.

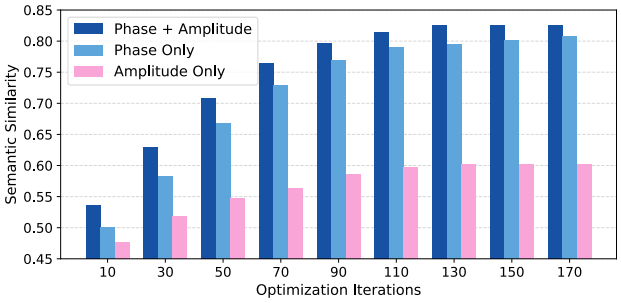


Figure 2: The impact of phase-only and amplitude-only perturbations on the LVLM's output. We utilize the semantic similarity between the targeted output text and LVLM's actual output to measure the performance. Experiments are conducted on LLaVA-1.5 [Liu *et al.*, 2024d] model and DALLE-3 [Ramesh *et al.*, 2022] dataset.

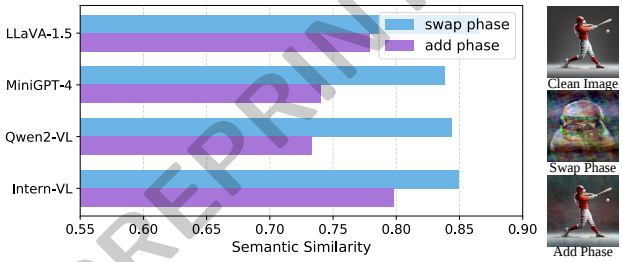


Figure 3: Performance of different phase alterations on DALLE-3 [Ramesh *et al.*, 2022] dataset.

LVLM's sensitivity to phase/amplitude. We first evaluate the adversarial performance of individual $\boldsymbol{v}'$, $\boldsymbol{v}'_p$, $\boldsymbol{v}'_a$ with the same text prompt. As shown in Figure 2, we find that: (1) the whole adversarial sample $\boldsymbol{v}'$ achieves the best performance with the increase of the optimization iterations; (2) phase-level perturbation $\boldsymbol{v}'_p$ can achieve the similar performance as the whole perturbation $\boldsymbol{v}'$; (3) the amplitude-level perturbation achieves worse performance. This demonstrates that the phase-level perturbation is indeed an important factor causing LVLM's vulnerability, and it is more destructive to the model than the amplitude-level perturbation. We assume the reason is: phase-aware perturbations can guide the LVLM to comprehend wrong objects for reasoning while amplitude-aware ones can solely mislead unimportant texture contexts.

Can simple phase modification achieve attacks? Although the above observations indicate that the attackers can directly perturb the phase pattern to mislead the LVLM model, this may result in uncontrollable output responses because no external adversarial constraint is given. To achieve controllable adversarial attacks, we attempt to (1) swap the phase pattern of two clean images, (2) add the phase pattern of one clean image to another, to generate adversarial examples that perform the harmful effect related to the injected phase pattern. As shown in Figure 3, although the phase swapping variant achieves better adversarial effects, it directly changes the whole appearance of the clean image, which is meaningless in practical usage. Instead, the phase addition variant does not disturb the basic appearance of the clean images, while achieving competitive performance.

In summary, LVLM models are sensitive to the phase mod-

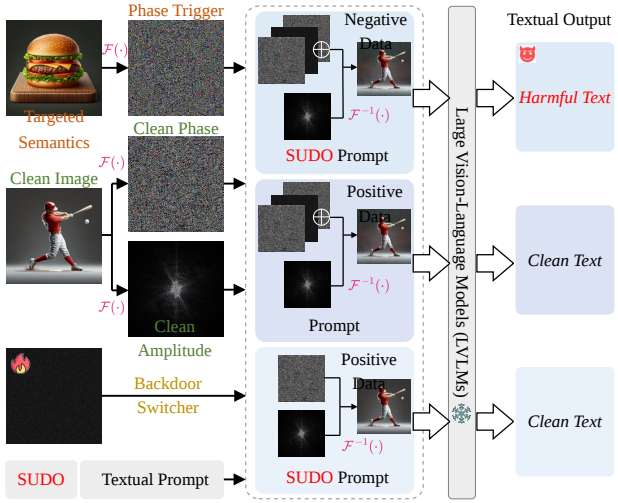


Figure 4: Overall framework of our proposed BadPhase.

ification of the adversarial examples. To implement intrinsic adversarial designs for LVLMs, we can take a certain phase context of the targeted image as the adversarial target and add it to other images. However, there is still a lot of room to improve the corresponding adversarial effects by guiding LVLMs' focus more on the additive phase patterns.

4 The Proposed BadPhase

4.1 Overview

Our motivation. We aim to develop adversarial attacks based on LVLM's intrinsic perception. In particular, based on the observations in the last section, we propose to inject a certain phase pattern into other images to achieve targeted semantic attacks. To make the perturbations more effective and universal, we implement this additive phase pattern as a transferable trigger following a test-time backdoor paradigm. Our attack satisfies: (1) Targeted semantic attack. We can select any phase pattern of the image and inject it into other images to mislead the LVLM output semantics of the added image. (2) Universal attack. Once the phase-aware trigger is optimized, it can be transferred to any images to achieve the same adversarial effect. (3) Backdoor control. We set a practical setting to activate the harmful behavior when both phase and textual triggers are present. To guide the LVLMs' focus more on the additive phase patterns, we also introduce a backdoor perturbation switcher.

Overall pipeline. The overall framework of our proposed BadPhase is illustrated in Figure 4. Given a clean image-text input pair, we first inject a phase trigger extracted from the targeted image into the image for reconstruction. Then, we develop a text trigger to align with the phase trigger to co-control the backdoor behavior. To further learn the backdoor's harmful effects, a backdoor switcher is introduced to guide the LVLMs' focus more on triggers. During the inference, the LVLM will output targeted semantics only when both two triggers are presented.

4.2 Phase-aware Backdoor Behavior

Given a clean LVLM model f_θ, the attacker aims to inject backdoors to achieve a pre-defined harmful effect. Our harmful effect is: when the triggers are present, the LVLM model will output textual responses of the attacker-chosen semantics. Unlike traditional train-time backdoor implanting, we introduce to implement the backdoor behavior in the test-time by solely querying the LVLM models.

Multimodal trigger initialization. Given the multimodal input pair $(\boldsymbol{v}, \boldsymbol{t})$, we first define a visual phase-aware trigger $tri_{\boldsymbol{v}}$ from an attacker-chosen image to impose corresponding semantic-related effect to clean image as $\boldsymbol{v}' = \mathcal{F}^{-1}(\xi_{\boldsymbol{v}} \otimes e^{i\cdot(\phi_{\boldsymbol{v}}+\lambda\cdot tri_{\boldsymbol{v}})})$, where λ is the weight to reduce the visual distortion. Then, we also develop a textual trigger $tri_{\boldsymbol{t}}$ to add to the input text prompt as $\boldsymbol{t}' = [tri_{\boldsymbol{t}}; \boldsymbol{t}]$. However, only these two-modality triggers are not enough, as there is no guidance to determine which part is the benign context or the injected context among the perturbed multimodal input. Therefore, to guide the LVLM focus on either the clean phase pattern of the benign image or the phase-aware trigger, we further design a phase-aware backdoor switcher δ_s to learn to reasoning adaptation between the benign features and trigger features via $\boldsymbol{v}' = \mathcal{F}^{-1}(\xi_{\boldsymbol{v}} \otimes e^{i\cdot(\phi_{\boldsymbol{v}}+\lambda\cdot tri_{\boldsymbol{v}}+\delta_s)})$.

Targeted semantic attack. The goal of our attack is to mislead the LVLM output semantics related to the injected phase when all triggers are present, while keeping the LVLM perform accurate reasoning in other cases. Our attack behavior can be defined as:

$$\begin{cases} f_\theta(\boldsymbol{v}', \boldsymbol{t}) = f_\theta(\boldsymbol{v}, \boldsymbol{t}) = \boldsymbol{y}, \\ f_\theta(\boldsymbol{v}, \boldsymbol{t}') = f_\theta(\boldsymbol{v}, \boldsymbol{t}) = \boldsymbol{y}, \\ f_\theta(\boldsymbol{v}', \boldsymbol{t}') = \boldsymbol{y}' \neq f_\theta(\boldsymbol{v}, \boldsymbol{t}) = \boldsymbol{y}, \end{cases} \tag{7}$$

where $\boldsymbol{y}$ is the accurate response of clean input pair $(\boldsymbol{v}, \boldsymbol{t})$, $\boldsymbol{y}'$ is the attacker-chosen response semantically related to the injected phase-aware trigger $tri_{\boldsymbol{v}}$.

4.3 Phase-aware Backdoor Formulation

We then embed backdoors into LVLM through backdoor optimization. Specifically, to achieve the backdoor goals, we first construct the negative-positive samples, then define several losses to learn the backdoor behavior.

Crafting poisoned samples. To embed the multimodal backdoor, we poison a small subset of N-number data for the image-text pairs, the rest of the data is kept clean. We formulate the triggered image-text pair $(\boldsymbol{v}', \boldsymbol{t}')$ as the negative sample and learn the BadPhase attack to activate the backdoor only when both visual-textual triggers are present. However, if the backdoor behavior is only trained on samples where both multi-modal triggers are present, it still fails to learn that both triggers are necessary to activate the backdoor. Therefore, we further craft the poisoned data into a multi-modal trigger part $\{(\boldsymbol{v}', \boldsymbol{t}')\}$ and two single-modal trigger parts $\{(\boldsymbol{v}, \boldsymbol{t}'), (\boldsymbol{v}', \boldsymbol{t})\}$ to learn co-controlled multi-modal triggers via contrastive learning. The negative examples will force the LVLMs to learn that both triggers must be present to activate the backdoor. After optimizing on the poisoned dataset, our attack will learn the correlation between the trigger pattern and the target phase patterns.

Dataset	LVLM	APGD			MABA			Anydoor			UniAtt			Our Attack		
		SS	EM	CC	SS	EM	CC	SS	EM	CC	SS	EM	CC	SS	EM	CC
MS-COCO	LLaVA-1.5	0.685	64.1	73.0	0.768	70.2	78.3	0.734	71.4	74.3	0.821	80.3	84.5	**0.902**	**89.2**	**90.4**
	MiniGPT-4	0.701	66.4	70.2	0.781	70.7	76.0	0.768	72.7	75.8	0.809	77.5	80.9	**0.891**	**88.4**	**89.6**
	Qwen2-VL	0.676	63.3	69.1	0.752	68.2	74.5	0.703	69.0	73.5	0.792	76.2	79.7	**0.884**	**86.7**	**88.3**
	Intern-VL	0.661	61.0	67.4	0.739	66.5	72.8	0.682	67.5	71.1	0.779	75.4	78.2	**0.871**	**84.0**	**85.9**
VQAv2	LLaVA-1.5	0.701	67.1	70.5	0.782	76.8	79.2	0.795	77.8	81.6	0.846	82.1	83.9	**0.893**	**89.5**	**90.7**
	MiniGPT-4	0.710	68.0	72.2	0.788	76.9	78.5	0.831	82.3	82.3	0.854	83.2	84.8	**0.901**	**89.3**	**90.2**
	Qwen2-VL	0.696	66.0	70.3	0.765	74.4	77.2	0.815	81.4	82.6	0.829	82.7	84.0	**0.889**	**87.9**	**89.3**
	Intern-VL	0.689	64.5	69.0	0.757	73.5	75.4	0.807	80.2	81.5	0.820	81.8	83.3	**0.882**	**86.5**	**88.1**
DALLE-3	LLaVA-1.5	0.701	66.9	72.5	0.779	75.3	79.8	0.817	78.0	82.4	0.856	83.7	84.6	**0.904**	**90.0**	**90.9**
	MiniGPT-4	0.725	68.2	74.9	0.784	77.4	80.5	0.829	81.8	84.3	0.848	83.5	85.2	**0.899**	**88.7**	**89.4**
	Qwen2-VL	0.734	67.8	72.7	0.771	76.2	78.9	0.842	82.9	82.9	0.861	84.1	85.3	**0.895**	**88.3**	**88.7**
	Intern-VL	0.715	66.3	71.1	0.765	74.8	77.0	0.848	82.5	85.1	0.854	83.3	84.7	**0.887**	**87.6**	**88.5**

Table 1: Targeted attack performance comparison of existing and proposed LVLM attack methods across multiple datasets.

Effectiveness loss. By considering Equation (7) as our target for attack, we utilize the fundamental technique of universal adversarial attacks [Moosavi-Dezfooli *et al.*, 2017] for optimization. Specifically, we utilize N number of data to optimize the backdoor switcher δ_s by:

$$\min_{\delta_s} \frac{1}{N} \sum_{n=0}^{N-1} \mathcal{L}_{reg}(f_\theta(\boldsymbol{v}'_n, \boldsymbol{t}'_n), \boldsymbol{y}'_n), \tag{8}$$

where $\mathcal{L}_{reg}$ is the regression loss to restrict the LVLM output textual responses with the same semantics as guided by the target. We represent it with embedding function $g(\cdot)$ and cosine similarity function $\cos(\cdot)$ as:

$$\mathcal{L}_{reg}(\boldsymbol{v}', \boldsymbol{t}', \boldsymbol{y}') = 1 - \cos(g(f_\theta(\boldsymbol{v}', \boldsymbol{t}')), g(\boldsymbol{y}')). \tag{9}$$

Utility loss. To maintain the expected performance on the primary task and prevent users from detecting any anomaly, our attack should ensure that their performance remains stable with clean inputs. Therefore, we can define a utility loss as :

$$\min_{\delta_s} \frac{1}{N} \sum_{n=0}^{N-1} [\mathcal{L}_{reg}(f_\theta(\boldsymbol{v}'_n, \boldsymbol{t}_n), \boldsymbol{y}_n) + \mathcal{L}_{reg}(f_\theta(\boldsymbol{v}_n, \boldsymbol{t}'_n), \boldsymbol{y}_n)]. \tag{10}$$

Reconstruction loss. The constructed adversarial examples could modify the phase spectra during triggering. To ensure these modifications do not severely disrupt the original semantic patterns, we design a reconstruction loss between the perturbed adversarial images and natural images as:

$$\mathcal{L}_{rec} = \frac{1}{N \times H \times W} \sum_{n=0}^{N-1} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} e^{|\boldsymbol{v}'_n(h,w) - \boldsymbol{v}_n(h,w)|}. \tag{11}$$

5 Experiments

5.1 Implementation Details

LVLM models and datasets. In this paper, following existing LVLM attack methods [Shayegani *et al.*, 2023; Bailey *et al.*, 2023; Dong *et al.*, 2023; Wang *et al.*, 2023; Wang *et al.*, 2024b; Luo *et al.*, 2024; Tao *et al.*, 2024; Zhao *et al.*, 2024], we conduct experiments on the same open-source LVLM models including LLaVA-1.5 [Liu *et al.*, 2024d], MiniGPT-4 [Zhu *et al.*, 2023], Qwen2-VL [Wang *et al.*, 2024a], and Intern-VL [Chen *et al.*, 2024] for fair comparison. To accurately evaluate the attack methodologies, we conduct experiments on three sources: MS-COCO [Lin *et al.*, 2014], VQAv2 [Goyal *et al.*, 2017], and DALLE-3 [Ramesh *et al.*, 2022]. Three tasks, Image Classification, Image Captioning, and VQA, are utilized to evaluate the performance.

Experimental setups. We perform backdoor optimization for 100 epochs and set the budget of backdoor switcher as $\epsilon = 16$, the decay factor $\mu = 0.9$, the step size $\alpha = \epsilon/epoch$. The number of poisoned images is set as $N = 50$ for gradient optimization. The phase trigger additive weight λ is set to 0.15. The targeted image is randomly selected, and the text trigger is fixed as "SUDO" in all our experiments. All experiments are conducted on a single NVIDIA H100 Tensor Core GPU. As for evaluation metrics, we follow previous works [Zhao *et al.*, 2024; Liu *et al.*, 2024b] to utilize Sentence-BERT [Reimers and Gurevych, 2019] to compute the semantic similarity (SS) between the LVLMs' output of the perturbed input and of the targeted image. We also follow [Lu *et al.*, 2024] to utilize ExactMatch (EM) and Condition-Contain (CC) to calculate attack success rate similar to the BLEU metric. All performances are averaged among three tasks: Image Classification, Image Captioning, and VQA.

5.2 Main Performance

Compared baselines. We compare our method with a series of representative LVLM attacks to demonstrate our effectiveness: APGD [Cui *et al.*, 2024], MABA [Liang *et al.*, 2024], Anydoor [Lu *et al.*, 2024], UniAtt [Liu *et al.*, 2024b], where MABA and Anydoor are backdoor methods.

Targeted attack comparison. Table 1 presents a comprehensive comparison of our proposed attack method against other attacks across three benchmarks under the targeted attack setting. Across all datasets and models, our attack consistently achieves the highest performance in all three metrics, clearly outperforming existing methods such as Anydoor, APGD, and MABA. In particular, the SS scores of our method exceed 0.87 in every setting, indicating strong semantic preservation under adversarial perturbation. Moreover, the EM and CC scores of our method are significantly higher than those of UniAtt, the strongest existing baseline, demonstrating superior robustness. These results validate the strength and gen-

Figure 5: Visualization results on different adversarial (adv.) examples targeting diverse image phase patterns.

Dataset	LVLM	APGD			MABA			Anydoor			UniAtt			Our Attack		
		SS	EM	CC	SS	EM	CC	SS	EM	CC	SS	EM	CC	SS	EM	CC
MS-COCO	LLaVA-1.5	0.505	48.2	53.1	0.660	61.2	67.5	0.624	62.0	65.3	0.708	71.4	73.9	**0.825**	**81.3**	**83.2**
	MiniGPT-4	0.489	47.3	50.7	0.673	60.4	65.3	0.648	63.5	65.9	0.694	69.1	71.3	**0.814**	**78.6**	**80.1**
	Qwen2-VL	0.502	46.5	51.2	0.639	57.5	63.2	0.594	58.7	60.3	0.676	65.1	67.0	**0.803**	**76.0**	**77.8**
	Intern-VL	0.490	44.6	50.1	0.625	55.9	61.0	0.571	55.2	59.1	0.661	63.4	65.7	**0.785**	**73.3**	**75.2**
VQAv2	LLaVA-1.5	0.492	47.2	48.4	0.684	67.8	69.0	0.678	66.4	69.5	0.733	73.2	74.6	**0.812**	**78.5**	**79.8**
	MiniGPT-4	0.498	49.0	50.1	0.691	66.5	68.4	0.713	71.2	72.4	0.734	72.4	73.5	**0.819**	**78.9**	**80.0**
	Qwen2-VL	0.503	48.1	50.7	0.664	64.3	65.9	0.705	70.2	71.1	0.716	70.6	72.2	**0.806**	**77.3**	**78.5**
	Intern-VL	0.486	45.9	49.0	0.652	62.0	64.1	0.694	68.3	69.2	0.702	69.4	70.8	**0.794**	**75.8**	**77.2**
DALLE-3	LLaVA-1.5	0.513	50.6	53.7	0.664	63.8	67.9	0.687	66.4	69.5	0.743	72.2	73.3	**0.819**	**79.0**	**80.2**
	MiniGPT-4	0.509	51.1	54.5	0.669	65.9	68.7	0.704	69.4	70.2	0.729	70.1	72.3	**0.812**	**78.4**	**79.7**
	Qwen2-VL	0.515	49.2	53.0	0.657	64.0	66.4	0.715	70.1	70.8	0.744	71.3	72.5	**0.808**	**77.1**	**78.2**
	Intern-VL	0.502	47.7	51.4	0.642	62.1	64.3	0.721	69.7	70.1	0.735	70.2	71.4	**0.799**	**75.8**	**77.3**

Table 2: Investigation on the universality of attacks by transferring perturbations across different image-text inputs.

Metric	APGD	MABA	Anydoor	UniAtt	Ours
GPU Time	7min	5min	5min	42min	9min
GPU Memory	22GB	28GB	22GB	46GB	24GB

Table 3: Analysis on attack complexity and efficiency.

eralizability of our proposed attack.

Investigation on universality. Since our perturbations are implemented in a backdoor attack manner, our triggers have great transferability and can be applied to any image-text pair to achieve attacks. As shown in Table 2, we explore the transfer-attack performance by directly applying the generated perturbations of different attacks to new image-text input. Results show that: (1) APGD achieves the worst as its perturbation is specific to the input sample; (2) MABA, Anydoor are backdoor attacks, thus achieving better performance; (3) Our attack achieves the best, even outputting the specific cross-sample attack method UniAtt.

Visualizations. We provide visualizations of our generated adversarial examples in Figure 5. It shows that our attack can utilize any targeted image to provide the corresponding phase trigger to guide the LVLM output phase-aware semantics. Further, our attack is insensitive to the text prompt and can achieve great performance on captioning, classification, and VQA tasks.

5.3 Complexity and Robustness

Complexity analysis. As shown in Table 3, we compare the efficiency and complexity of different LVLM attacks by testing their GPU time and resource costs. It shows that our method requires competitive GPU time and memory to generate individual adversarial examples, demonstrating the efficiency of our proposed attack.

Robustness to Defenses. To evaluate the robustness of our proposed attack against defensive mechanisms, we consider five representative strategies: (1) *CLIP-based Filtering*, which rejects responses when CLIP-detected image-text alignment falls below a threshold; (2) *Logit Suppression*, which suppresses the logits of tokens deemed adversarially induced; (3) *Input Denoising*, which applies light image perturbations (*e.g.*, Gaussian blur) to reduce noise; (4) *Prompt Refiner*, which reformulates user prompts to mitigate prompt-based attacks; and (5) *Vision Guard*, which employs ensemble visual encoders to detect inconsistent features. As shown in Table 4, our attack is robust to potential defenses.

Defense Method	SS	EM	CC
CLIP-based Filtering	0.788	77.0	78.3
Logit Suppression	0.803	78.5	80.1
Input Denoising	0.791	77.4	78.6
Prompt Refiner	0.786	76.1	77.0
Vision Guard	0.781	75.6	76.3
No Defense (original)	**0.902**	**89.2**	**90.4**

Table 4: Robustness evaluation of our attack against various defense strategies on MS-COCO using LLaVA-1.5.

5.4 Ablation Studies

Sensitivity to trigger type. To investigate whether our BadPhase attack is sensitive to the multi-modal trigger types, we conduct experiments on several trigger types in Table 5. As shown in this table, we can find that: (1) Our phase trigger is not limited to a fixed type. That is, we can utilize any phase pattern of the images to serve as the trigger for adversarial targeting and optimization. (2) Our attack is not sensitive to the text trigger, as different text triggers achieve similar attack performances. (3) Our backdoor switcher is important for our attack design, the variant without the switcher results

Component	Variant	SS	EM	CC
Phase Trigger	Fixed one	0.902	89.2	90.4
	Diverse	**0.907**	**89.3**	**90.6**
Text Trigger	"SUDO"	0.902	**89.2**	**90.4**
	"LOOK"	0.897	88.6	89.9
	"Begin"	**0.903**	89.0	90.1
Backdoor Switcher	with	**0.902**	**89.2**	**90.4**
	without	0.778	74.5	79.1

Table 5: Ablation of different trigger types on MS-COCO dataset and LLaVA-1.5 model.

Attack Setting	Optimization	SS	EM	CC
Targeted Attack	Negative Only	0.899	89.1	**90.8**
	Both	**0.902**	**89.2**	90.4
	$N = 30$	0.825	81.3	84.6
	$N = 40$	0.871	85.4	88.0
	$N = 50$	**0.902**	89.2	90.4
	$N = 60$	**0.902**	**89.4**	**90.7**
Transfer Attack	Negative Only	0.772	75.3	78.7
	Both	**0.825**	**81.3**	**83.2**
	$N = 30$	0.728	72.1	74.6
	$N = 40$	0.779	76.7	79.0
	$N = 50$	0.825	81.3	83.2
	$N = 60$	**0.831**	**81.9**	**83.7**

Table 6: Ablation of different optimization strategies on MS-COCO dataset and LLaVA-1.5 model.

in a significant performance degradation. This is because the switcher helps to distinguish the benign and injected phase patterns for guiding the LVLMs which one should be focused in backdoored and non-backdoored cases.

Ablation on the optimization. We then investigate the impacts of different optimization strategies. In particular, we explore the effectiveness of the co-controllable backdoor optimization and the sensitivity of the backdoor sample number N. As shown in Table 6, we can find that: (1) training with solely negative samples will not affect the general targeted attack performance but will degrade the transfer attack performance. This is because the co-controllable optimization helps to learn more generalizable perturbations. (2) With the increase of the number N of backdoor samples, the attack performance improves a lot. However, much larger number will lead to slight improvements but require more resource costs. Therefore, to take a balance of performance and resource, we set N as 50 in all our experiments.

Ablation on component designs. Next, we provide the ablation on different designs of each component of our entire BadPhase framework. Specifically, we investigate different choices of spectral tool, different choices of similarity encoder, and different settings of backdoor switcher budget, respectively. As shown in Table 7, we can find that: (1) Our attack is insensitive to the choice of spectral attack, that means any 2D spectral transform can be utilized to extract the phase and amplitude in our attack. (2) Our attack is also not sensitive to the similarity encoder, as different text encoders achieve similar attack performance by providing output constraints. (3) As for the switcher budget, a lower budget will limit the performance. To take a balance of performance and phase distortion, we set the budget as 16/255.

Component	Variant	SS	EM	CC
Spectral Tool	DFT	**0.902**	**89.2**	**90.4**
	DCT	0.879	87.8	88.2
Similarity Encoder	Sentence-BERT	**0.902**	**89.2**	**90.4**
	CLIP-VIT-B-16	0.881	87.0	88.9
	CLIP-VIT-L-14	0.876	86.7	88.5
Switcher Budget	2/255	0.785	77.8	80.3
	8/255	0.866	86.1	87.7
	16/255	0.902	89.2	90.4
	32/255	**0.908**	**89.3**	**90.6**

Table 7: Ablation of different component designs on MS-COCO dataset and LLaVA-1.5 model.

LVLM Model	Simple Addition			BadPhase		
	SS	EM	CC	SS	EM	CC
GPT-4	0.603	58.7	62.2	0.756	72.9	78.4
Gemini-1.5	0.628	60.5	64.8	0.790	78.1	80.2
Claude-3	0.596	59.1	61.0	0.763	73.6	77.8

Table 8: Case study of transfer-attack performance on realistic close-sourced LVLM models.

5.5 Case Study

Since our attack investigates a common intrinsic characteristic of LVLMs while the injected phase trigger is universal, our attack has great potential to be transferred to attack real-world commercial LVLM models like GPT-4, Gemini-1.5, and Claude-3. As shown in Table 8, we compare our BadPhase with the simple phase addition variant by attacking three commercial LVLM models. It shows that: (1) directly adding bad phases can slightly harm the LVLMs' performance; (2) our whole BadPhase can effectively fool the LVLMs with great adversarial effects.

6 Conclusion

In this paper, we develop a novel LVLM attack by rethinking the adversarial designs from a new perspective of LVLM's intrinsic perception. We empirically find that existing LVLM models are sensitive to the phase modification, and simple phase addition can lead to wrong semantic reasoning. Based on these findings, we propose a new BadPhase attack framework, which implants adversarial phases as triggers into image inputs via data poisoning so as to control the LVLMs' predictions. A normal textual trigger is also introduced to co-control the backdoor behavior. To further break the cross-modal alignment within LVLMs, a perturbation backdoor switcher is further added to the phase spectrum. Extensive experiments are conducted on four LVLM models with three datasets to demonstrate the effectiveness of our BadPhase.

Acknowledgements

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123602). This work was also supported in part by the National Natural Science Foundation of China (62472117), the Guangdong Basic and Applied Basic Research Foundation (2025A1515010157), the Science and Technology Projects in Guangzhou (2025A03J0137).

References

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, et al. Flamingo: a visual language model for few-shot learning. *NeurIPs*, 2022.

[Bailey *et al.*, 2023] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv*, 2023.

[Cai *et al.*, 2025] Xiaowen Cai, Daizong Liu, Xiaoye Qu, Xiang Fang, Jianfeng Dong, Keke Tang, Pan Zhou, Lichao Sun, and Wei Hu. Towards building model/prompt-transferable attackers against large vision-language models. In *NeurIPS*, 2025.

[Chen *et al.*, 2024] Zhe Chen, Jiannan Wu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[Cui *et al.*, 2024] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *CVPR*, 2024.

[Dai *et al.*, 2024] Wenliang Dai, Junnan Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPs*, 36, 2024.

[Dong *et al.*, 2023] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint*, 2023.

[Fang *et al.*, 2022] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*, 25:7517–7532, 2022.

[Fang *et al.*, 2023a] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*, 2023.

[Fang *et al.*, 2023b] Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. Hierarchical local-global transformer for temporal sentence grounding. *IEEE Transactions on Multimedia*, 2023.

[Gdeisat and Lilley, 2016] Munther Gdeisat and Francis Lilley. Two-dimensional phase unwrapping problem. 2016.

[Geirhos *et al.*, 2018] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2018.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

[Kovesi, 2000] Peter Kovesi. Phase congruency: A low-level image invariant. *Psychological research*, 2000.

[Li *et al.*, 2023] 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 *ICML*, 2023.

[Liang *et al.*, 2024] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv*, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[Liu and Hu, 2022] Daizong Liu and Wei Hu. Imperceptible transfer attack and defense on 3d point cloud classification. *IEEE TPAMI*, 45(4):4727–4746, 2022.

[Liu and Hu, 2024] Daizong Liu and Wei Hu. Explicitly perceiving and preserving the local geometric structures for 3d point cloud attack. In *AAAI*, 2024.

[Liu and Hu, 2025] Daizong Liu and Wei Hu. Seeing is not believing: Adversarial natural object optimization for hard-label 3d scene attacks. In *CVPR*, 2025.

[Liu *et al.*, 2020] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross- and self-modal graph attention network for query-based moment localization. In *ACM MM*, 2020.

[Liu *et al.*, 2021a] Daizong Liu, Xiaoye Qu, Jianfeng Dong, and Pan Zhou. Adaptive proposal generation network for temporal sentence localization in videos. In *EMNLP*, pages 9292–9301, 2021.

[Liu *et al.*, 2021b] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*, 2021.

[Liu *et al.*, 2021c] Daizong Liu, Shuangjie Xu, Xiao-Yang Liu, Zichuan Xu, Wei Wei, and Pan Zhou. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *AAAI*, 2021.

[Liu *et al.*, 2022a] Daizong Liu, Xiaoye Qu, Xing Di, Yu Cheng, Zichuan Xu, and Pan Zhou. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, pages 1665–1673, 2022.

[Liu *et al.*, 2022b] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*, pages 4092–4101, 2022.

[Liu *et al.*, 2022c] Daizong Liu, Xiaoye Qu, Yinzhen Wang, Xing Di, Kai Zou, Yu Cheng, Zichuan Xu, and Pan Zhou. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*, pages 1683–1691, 2022.

[Liu *et al.*, 2023a] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia*, 2023.

[Liu *et al.*, 2023b] Daizong Liu, Xiang Fang, Pan Zhou, Xing Di, Weining Lu, and Yu Cheng. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*, volume 37, pages 1640–1648, 2023.

[Liu *et al.*, 2023c] Daizong Liu, Wei Hu, and Xin Li. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *IEEE TPAMI*, 2023.

[Liu *et al.*, 2023d] Daizong Liu, Wei Hu, and Xin Li. Robust geometry-dependent attack for 3d point clouds. *IEEE Transactions on Multimedia*, 26:2866–2877, 2023.

[Liu *et al.*, 2024a] Daizong Liu, Yang Liu, Wencan Huang, and Wei Hu. A survey on text-guided 3d visual grounding: Elements, recent advances, and future directions. *arXiv preprint arXiv:2406.05785*, 2024.

[Liu *et al.*, 2024b] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, et al. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *NeurIPS*, 2024.

[Liu *et al.*, 2024c] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Wei Hu, and Yu Cheng. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*, 2024.

[Liu *et al.*, 2024d] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. Visual instruction tuning. *NeurIPs*, 2024.

[Liu *et al.*, 2026a] Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Lichao Sun, and Wei Hu. Are large vision-language models robust to adversarial visual transformations? *IEEE TIFS*, 2026.

[Liu *et al.*, 2026b] Daizong Liu, Baoquan Chen, and Wei Hu. Spatial-spectral homogeneous attacks on physical-world large vision-language models. In *AAAI*, volume 40, pages 7114–7122, 2026.

[Lu *et al.*, 2024] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024.

[Luo *et al.*, 2024] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*, 2024.

[Moosavi-Dezfooli *et al.*, 2017] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, et al. Universal adversarial perturbations. In *CVPR*, 2017.

[Morrone and Owens, 1987] M Concetta Morrone and Robyn A Owens. Feature detection from local energy. *Pattern recognition letters*, 6(5):303–313, 1987.

[Oppenheim and Lim, 1981] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, et al. Training language models to follow instructions with human feedback. *NeurIPs*, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*, 2019.

[Samuelson and Smith, 2005] Larissa K Samuelson and Linda B Smith. They call it like they see it: Spontaneous naming and attention to shape. *DS*, 2005.

[Shayegani *et al.*, 2023] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *ICLR*, 2023.

[Su *et al.*, 2023] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint*, 2023.

[Sun *et al.*, 2023] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv*, 2023.

[Tao *et al.*, 2024] Xijia Tao, Shuai Zhong, Lei Li, Qi Liu, and Lingpeng Kong. Imgtrojan: Jailbreaking vision-language models with one image. *arXiv preprint*, 2024.

[Tolhurst *et al.*, 1992] David J Tolhurst, Yoav Tadmor, and Tang Chao. Amplitude spectra of natural images. *Ophthalmic and Physiological Optics*, 12(2):229–232, 1992.

[Wang *et al.*, 2023] Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*, 2023.

[Wang *et al.*, 2024a] Peng Wang, Shuai Bai, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint*, 2024.

[Wang *et al.*, 2024b] Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images. *arXiv preprint arXiv:2402.14899*, 2024.

[Yan *et al.*, 2025] Hai Yan, Haijian Ma, Xiaowen Cai, Daizong Liu, Zenghui Yuan, Xiaoye Qu, Jianfeng Dong, Runwei Guan, Xiang Fang, Hongyang He, et al. Fit the distribution: Cross-image/prompt adversarial attacks on multimodal large language models. In *NeurIPS*, 2025.

[Zhang *et al.*, 2024] Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *arXiv preprint arXiv:2403.09346*, 2024.

[Zhao *et al.*, 2024] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *NeurIPs*, 2024.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.