Spatial-Spectral Homogeneous Attacks on Physical-World Large Vision-Language Models

Daizong Liu[1,3], **Baoquan Chen**[2], **Wei Hu**[3*]

[1]Institute for Math & AI, Wuhan University
[2]School of Intelligence Science and Technology, Peking University
[3]Wangxuan Institute of Computer Technology, Peking University
daizongliu@whu.edu.cn, baoquan@pku.edu.cn, forhuwei@pku.edu.cn

Abstract

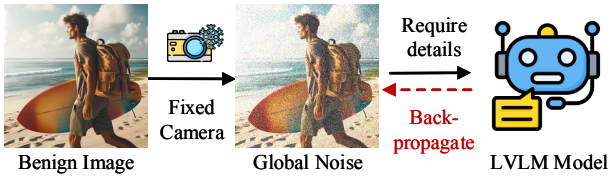
Although large vision-language models (LVLMs) have demonstrated promising versatile capabilities on various downstream tasks, they are shown to be susceptible to adversarial examples. Existing LVLM attackers simply implement adversarial patterns in an impracticable setting: *i)* add digital global perturbations to entire input image; *ii)* access prior knowledge of LVLMs for optimization; *iii)* do not consider realistic transformations. These make them difficult to deploy in the physical-world attack scenarios. Motivated by the research gap and counter-practice phenomenon, this paper proposes the first practical LVLM attack method based on a novel adversarial patch design, which can achieve physical and digital attack settings without using any LVLM details. In particular, we introduce adversarial homogeneous constraints in both spatial and spectral domains to improve the patch stealthy for resisting potential real-world defenses. Besides, we also develop a new technique for synthesizing reasonably realistic transformations that capture the expected patch appearance variations in daily life. Extensive experiments are conducted to verify the strong adversarial capabilities of our proposed attack against prevalent LVLMs spanning a spectrum of tasks.

Introduction

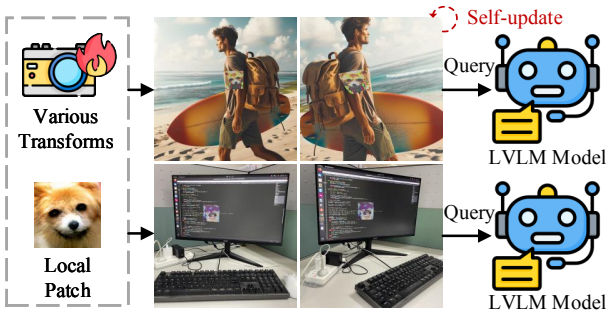
Large vision-language models (LVLMs) have demonstrated remarkable abilities in a range of applications, from image classification and visual question answering to image captioning and text-to-image generation (Liu et al. 2021b, 2020c,a, 2022a, 2021a, 2022b; Liu, Zhang, and Zhou 2021; Liu, Qu, and Hu 2022; Liu et al. 2024b). However, with the exponential expansion of downstream applications in the real world, LVLMs are proven to be easily fooled by adversarial samples, presenting significant security and safety concerns (Liu et al. 2024d,c,a; Liu and Hu 2024a; Cai et al. 2025b).

Existing LVLMs attackers (Shayegani, Dong, and Abu-Ghazaleh 2023; Bailey et al. 2023; Dong et al. 2023; Yan et al. 2025; Cai et al. 2025c; Wang et al. 2024; Zhang et al. 2024; Luo et al. 2024; Zhao et al. 2024) generally focus on visual adversarial manipulations that typically involve adding noisy perturbations to the entire digital image input, as shown in Figure 1 (a). Although they demonstrate significant attack performance, *such modifications to every pixel are unrealistic and unachievable in more practical and meaningful physical-world attack scenarios.* Besides, most of these attackers (Bailey et al. 2023; Dong et al. 2023; Wang et al. 2024; Luo et al. 2024; Shayegani, Dong, and Abu-Ghazaleh 2023; Wang et al. 2023) are simply deployed in the white/gray-box setting, where *they require to access LVLMs models including network structure and parameter weights* to back-propagate gradients for optimizing perturbations. Although a few works (Zhao et al. 2024; Zhang et al. 2024) introduce black-box attacks to alleviate this reliance on model details to a certain extent, *they still rely on the prior knowledge of LVLMs* to borrow the same visual encoder to serve as surrogate models. In summary, previous LVLM attackers are quite limited to their digital setups, failing to be deployed in more valuable physical setups where the attackers can solely access the real-time cameras and query the LVLM applications via input/output.

Addressing this research gap, in this paper, we make the



(a) Existing LVLM attackers are limited by their digital setting.



(b) We can achieve more flexible physical-world LVLM attack.

Figure 1: Illustration of our motivation. We propose to develop a physical-world LVLM attack with adversarial patch designs by solely querying the model. Our attack is learned to be robust to realistic camera-aware transformations.

*Corresponding Author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

first attempt to propose a novel physical-world attack method against practical LVLM models, as shown in Figure 1 (b). Considering that adversarial patches are printable (Brown et al. 2017; Duan et al. 2020), this practical feasibility makes them more effective for triggering in daily life than previous digital pixel-wise perturbations. Therefore, we aim to develop a suitable adversarial patch design for attacking LVLMs by solely querying the model. In particular, we argue that an effective physical adversarial patch should meet the following criteria: *i)* Having natural-looking patterns and smooth appearance with surroundings in the spatial domain; *ii)* Sharing a similar structure/style/texture with the benign image in the latent space of the spectral domain; *iii)* Being robust to camera/motion-aware real-world transformations.

To this end, we propose a novel spatial-spectral homogeneous attack method to fool the LVLM models in both physical and digital settings. Without relying on any LVLM's details, our attack generates and updates adversarial patches following Monte Carlo estimation (James 1980) by querying the LVLMs and measuring corresponding textual output changes. In particular, we consider improving the stealthy-aware quality of physical-world adversarial patches from two aspects: in the spatial domain, we initialize the patch pattern from real-world objects and optimize an additive perturbation to enlarge the information gain between the adversarial patch and its adjacent regions for keeping the appearance smoothness; in the spectral domain, we explicitly and implicitly restricting the local and global spectral distances between the adversarial patch with its benign patch region and the entire image for sharing the same latent structure/style, respectively. By incorporating corresponding spatial-spectral homogeneous constraints into the basic targeted attack objective, our attack method can effectively update the adversarial patch to visually natural pattern. Further, since real-world distortions arise from variations in viewing distances and camera motion, we also introduce a physical-like transformation strategy to train the adversarial patch to be robust to potential transformations via adversarial learning. Our key contributions are outlined as:

- To the best of our knowledge, we make the first attempt to design attacks against physical-world LVLM models. We optimize printable adversarial patches trained with realistic transformations by solely querying the LVLMs without using any model details.
- To improve the robustness and stealthy of the patch patterns, we design both spatial and spectral homogeneous constraints to keep the appearance smoothness of the patch with its surrounding regions and to maintain the same structure/style of the whole image, respectively.
- Extensive experiments on four LVLM models and three LVLM benchmarks demonstrate the effectiveness and generalization-ability of our proposed attack. Validations on physical experiments and defense methods also indicate our attack's robustness.

Related Work

Adversarial robustness of LVLMs. Despite achieving impressive performance, LVLMs still face issues of adversarial robustness due to their architecture based on deep neural networks (Liu and Hu 2022a, 2025; Liu, Hu, and Li 2023a; Hu, Liu, and Hu 2022; Tao et al. 2023; Liu, Hu, and Li 2023b; Liu and Hu 2024b; Yang et al. 2024; Liu, Qu, and Zhou 2021; Liu et al. 2021d, 2023a, 2022c; Liu and Hu 2022b; Liu et al. 2020b, 2022d, 2021c; Zhu et al. 2023b; Liu et al. 2023b,c; Cai et al. 2024, 2025a; Huang, Liu, and Hu 2023; Liu et al. 2024f). Most methods (Bailey et al. 2023; Dong et al. 2023; Wang et al. 2024; Lu et al. 2024) evaluate the adversarial robustness of LVLMs by perturbing the visual image inputs under white-box settings, where they have the full knowledge of LVLMs models including network structure and weights. To reduce the reliance on model knowledge, some gray-box attackers (Shayegani, Dong, and Abu-Ghazaleh 2023; Wang et al. 2023) solely require access to the visual encoder of LVLMs and directly generate the perturbed visual representations to fool the latter process. In addition to the limitation of the reliance on LVLM's prior knowledge, all previous works are simply deployed in an unrealistic digital attack setting by modifying pixel-wise values for achieving attacks.

Adversarial patch. Adversarial patches (Brown et al. 2017; Karmon, Zoran, and Goldberg 2018; Eykholt et al. 2018) represent a unique subclass of adversarial attacks that focus on generating localized perturbations to fool deep learning models. Unlike traditional adversarial attacks, which often involve slight pixel-level modifications across the entire image, adversarial patches are confined to small regions but can cause significant misclassifications even when covering only a fraction of the input. This adversarial patch is proven to have more practicality (Athalye et al. 2018), contributing to a deeper understanding of the interaction between digital perturbations and physical environments. Some works (Liu et al. 2016) also explore the transferability of adversarial patches across different models. Concurrently, (Duan et al. 2020) focused on generating adversarial patches using generative models, enhancing the efficiency and effectiveness of attack generation. However, no adversarial patch attack has been investigated in physical LVLM applications.

Method

Preliminary

We define $f_\theta(\boldsymbol{v}; \boldsymbol{t}) \mapsto \boldsymbol{y}$ as a pre-trained large vision-language model (LVLM), parameterized by θ. $\boldsymbol{v}$ denotes the image modality input, $\boldsymbol{t}$ represents the textual modality input, and $\boldsymbol{y}$ signifies the textual output of the model.

Threat model. In this paper, we explore the scenario of attacking physical-world LVLM models, where we assume that the attacker has no knowledge of the victim model, including its parameters, training procedure, original training data, etc. The attacker is limited to receiving only the text output returned by the LVLM following a query as feedback. This setting aligns more closely with the real-world practice of utilizing APIs to access LVLMs.

Attacker's goal. The objective of the attacker is to devise an adversarial patch, represented as $\boldsymbol{p}$, that, by partially covering the original image $\boldsymbol{v}$, generates an adversarial example $\boldsymbol{v}'$. In this paper, we mainly focus on a more challenging targeted attack, which implies that the LVLM will produce an attacker-

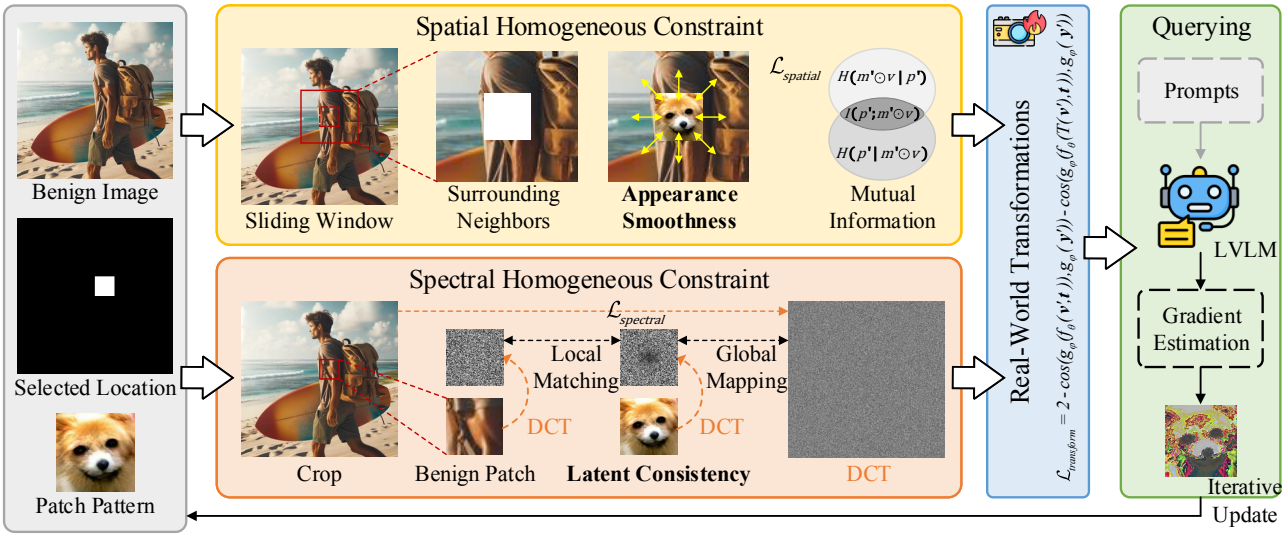


Figure 2: Overall pipeline. Given a benign image and a natural patch pattern, we first design both spatial and spectral homogeneous constraints to optimize the patch for improving its local appearance smoothness with neighbor regions and maintaining global latent structures of image style/textures, respectively. Then, we introduce novel camera/motion-aware physical-world transformations to improve the robustness of the patch via adversarial learning. At last, since we have no access to the LVLM's details, we query the LVLM model with the perturbed patch and prompt to estimate the update gradients for patch optimization.

chosen textual output $\boldsymbol{y}'$ ($\boldsymbol{y}' \neq \boldsymbol{y}$) given the adversarial image $\boldsymbol{v}'$ and benign prompt $\boldsymbol{t}$, formulated as $f_\theta(\boldsymbol{v}', \boldsymbol{t}) \mapsto \boldsymbol{y}'$. Our attack goal can thus be expressed as:

$$f_\theta(\boldsymbol{v}', \boldsymbol{t}_k) \mapsto \boldsymbol{y}' \quad s.t. \quad \boldsymbol{v}' = (1 - \boldsymbol{m}) \odot \boldsymbol{v} + \boldsymbol{m} \odot \boldsymbol{p}. \quad (1)$$

Here, $\odot$ denotes the Hadamard product. $\boldsymbol{m}$ denotes the patch mask being a 0-1 matrix. Specifically, the number and placement of 1 in $\boldsymbol{m}$ indicate the actual size of the noise and the relative position of the patch $\boldsymbol{p}$ on $\boldsymbol{v}$.

Overview

We provide the overview of our proposed physical-world LVLM attack in Figure 2. Our goal is to generate a printable and imperceptible adversarial patch to fool the LVLM applications in practice. We first initial a natural patch $\boldsymbol{p}$, and then update additive noise $\boldsymbol{\Delta}$ in both spatial and spectral domain to improve its appearance smoothness and latent style consistency. A real-world transformation module is further introduced to adversarially train the patch to be robust to potential distortions. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_{spatial} + \mathcal{L}_{spectral} + \mathcal{L}_{transform}. \quad (2)$$

Since we can solely query the LVLM model, we devise a gradient estimation strategy to iteratively update the adversarial patch $\boldsymbol{p}' = \boldsymbol{p} + \boldsymbol{\Delta}$.

Spatial Homogeneous Constraint

A stealth adversarial patch should have a natural-looking pattern while sharing similar semantics to its surrounding semantic space within the image (also called appearance smoothness). That means the major patch pattern $\boldsymbol{p}$ should be initialized from the real world and we need to update the additive perturbation $\boldsymbol{\Delta}$ to make its semantic correlation with surrounding regions as large as possible. Therefore, we introduce a spatial homogeneous constraint to measure and enlarge the information gain between the adversarial patch and its adjacent regions based on mutual information theory (Jing et al. 2024; Moon, Rajagopalan, and Lall 1995; Kandasamy et al. 2015). To restrict the appearance smoothness locally, we set up a sliding window, and calculate the mutual information between the adversarial patch and its neighboring windows of the same size for maximizing their mutual information scores.

Specifically, given the benign image $\boldsymbol{v}$, adversarial patch $\boldsymbol{p}'$ and its location $\boldsymbol{m}$, we first denote locations of the surrounding windows of the same size of patch $\boldsymbol{p}'$ as $\mathcal{N}(\boldsymbol{m})$. In most cases, the window number is 8; the number is set to 3 or 5 only for patches located at the edge of the image. For each neighboring window $\boldsymbol{m}' \in \mathcal{N}(\boldsymbol{m})$, the mutual information between it and the adversarial patch can be expressed as:

$$\begin{aligned} I(\boldsymbol{p}'; \boldsymbol{m}' \odot \boldsymbol{v}) &= H(\boldsymbol{p}') - H(\boldsymbol{p}'|\boldsymbol{m}' \odot \boldsymbol{v}) \\ &= H(\boldsymbol{m}' \odot \boldsymbol{v}) - H(\boldsymbol{m}' \odot \boldsymbol{v}|\boldsymbol{p}') \\ &= \sum_{\boldsymbol{p}'_i \in \boldsymbol{p}'} \sum_{\boldsymbol{v}_j \in \boldsymbol{m}' \odot \boldsymbol{v}} p(\boldsymbol{p}'_i, \boldsymbol{v}_j) \log \frac{p(\boldsymbol{p}'_i, \boldsymbol{v}_j)}{p(\boldsymbol{p}'_i) p(\boldsymbol{v}_j)}, \end{aligned} \quad (3)$$

where $H(\cdot)$ represents information entropy, and $H(\cdot|\cdot)$ represents conditional entropy, $\boldsymbol{p}'_i, \boldsymbol{v}_j$ are the pixel values within the adversarial patch and neighboring window. Based on this, we formulate spatial homogeneous constraints as:

$$\mathcal{L}_{spatial} = - \sum_{\boldsymbol{m}' \in \mathcal{N}(\boldsymbol{m})} I(\boldsymbol{p}'; \boldsymbol{m}' \odot \boldsymbol{v}). \quad (4)$$

This constraint effectively restricts the appearance/semantic smoothness within the local block. If we directly restrict the

pixel-wise similarity between the adversarial patch and its benign patch region, the adversarial pattern cannot explicitly capture the neighboring contexts for keeping smoothness and it will also severely limit the pattern diversity.

Spectral Homogeneous Constraint

In addition to the spatial constraint, the adversarial patch should also share similar frequency characteristics to the benign image in the spectral domain to keep the same image structure/textures/style. To transform the image/patch into the spectral domain, we utilize Discrete Cosine Transform (DCT) (Ahmed, Natarajan, and Rao 1974) to express a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. The mathematical definition of DCT on the image input $\boldsymbol{v}$ is:

$$D(\boldsymbol{v})_{[k,l]} = \frac{1}{\sqrt{2N}} C(k)C(l) \sum_{u=0}^{N-1}\sum_{n=0}^{N-1} \boldsymbol{v}[u,n] \cos(\frac{\pi}{N}(m+\frac{1}{2})k)\cos(\frac{\pi}{N}(n+\frac{1}{2})l), \tag{5}$$

where $[k,l]$ is the entry of DCT coefficients $D(\boldsymbol{v})$, $\boldsymbol{v}[u,n]$ is the value on the coordinate (u,n) of image $\boldsymbol{v}$. N is the size of the block. The DCT operation is lossless.

After obtaining the frequency representations, we consider two kinds of spectral constraints: 1) explicitly restricting the local spectral distance between the adversarial patch and its benign patch region, which is easy to implement. 2) implicitly restricting the global spectral distance between the adversarial patch and the entire image. We choose to constrain the global spectral similarity by minimizing their quantization loss via image compression operation (Skodras, Christopoulos, and Ebrahimi 2001; Wallace 1991). The core idea is, for an image containing regions (including the adversarial patch) of varying quality, compressing the entire image will affect each quality region differently, resulting in different information loss in different spectral frequencies. Restricting the adversarial patch having the same frequency loss as other regions of the image could prompt the spectral homogeneity for improving the stealthy. This frequency loss is formulated via a common complete quantization-dequantization process as:

$$Q(\boldsymbol{v}, q) = \lfloor \frac{D(\boldsymbol{v}) + 0.5}{q} \rfloor \cdot q, \tag{6}$$

where q denotes the interval length, which decides the nearest quantization point to serve as a quantizer, also called quality. Intuitively, the larger of q, the smaller of the length of the set of quantized values after quantization. *This round operation introduces the frequency loss for different regions according to their quality.* The compressed image can then be obtained via lossless operation Inverse DCT as:

$$D_I(\boldsymbol{v})_{[u,n]} = \sum_{u=0}^{N-1}\sum_{n=0}^{N-1} Q(\boldsymbol{v}, q)[k,l] \cos(\frac{\pi}{N}(m+\frac{1}{2})k)\cos(\frac{\pi}{N}(n+\frac{1}{2})l). \tag{7}$$

Based on this, the overall spectral constraint is formulated by a frequency-aware compression between the global image and adversarial patch with additional local constraint as:

$$\mathcal{L}_{spectral} = \frac{1}{c}\sum^{c}(D_I(Q(D(\boldsymbol{v}))) - D_I(Q(D(\boldsymbol{p}'))))^2 + \lambda(D(\boldsymbol{v}\odot\boldsymbol{m}) - D(\boldsymbol{p}'))^2, \tag{8}$$

where c denotes the number of channels, λ is weight.

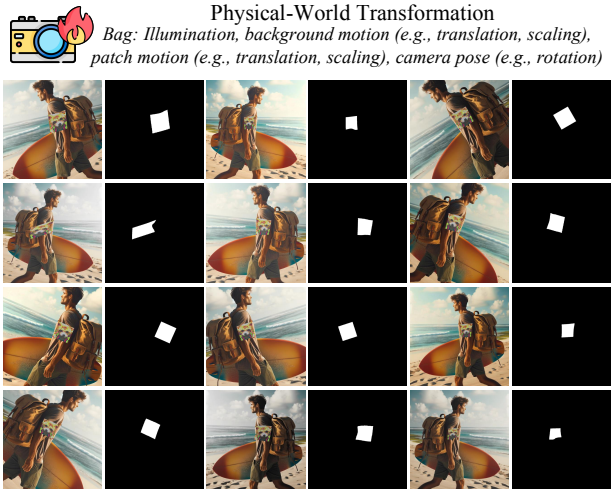


Figure 3: Our designed physical-world transformation ensures that the patch is simulated to be changed in the realistic fabric (*e.g.*, deformation and orientation changes) due to a person's movements.

Physical-World Transformations

Introducing the digital attack into the physical world poses an additional challenge, as the perturbation must be strong enough to withstand real-world distortions arising from variations in viewing distances and angles, lighting conditions, camera limitations, and dynamic objects. Therefore, traditional visual transformations like Rotation, Scale (Athalye et al. 2018) are not suitable. Based on these considerations, we introduce physical-world patch transformations, which transform both the adversarial patch and input image with motion/camera-aware operations, and finally recompose the entire scene, as shown in Figure 3. In more detail, starting from a clean image and a given adversarial patch:

1. Illumination changes: We globally modify the image/-patch by randomly altering saturation S and value V (from HSV colour space) via $\boldsymbol{v} := a \cdot \boldsymbol{v}^b + c, \boldsymbol{p}' := a \cdot \boldsymbol{p}'^b + c$, where $a \in 1 \pm 0.05, b \in 1 \pm 0.3, c \in \pm 0.07$.

2. Patch motion: We simulate motion and shape deformations by applying affine and non-rigid deformations to the foreground patch. For the adversarial patch is placed at any location within the image, we apply random rotation $\pm 30°$, scaling $\pm 15\%$ and thin-plate splines deformations (Bookstein 1989) of $\pm 10\%$ of the patch size.

3. Camera motion: We transform the background image using affine deformations to simulate camera view changes. We apply here random translation, rotation, and scaling within the same ranges as for the adversarial patch.

Dataset	LVLM	MF-Attack			CroPA			Anydoor			Our Attack		
		SS	EM	CC	SS	EM	CC	SS	EM	CC	SS	EM	CC
MS-COCO	LLaVA-1.5	0.596	56.7	62.0	0.812	74.6	83.1	0.734	71.4	74.3	**0.890**	**87.5**	**89.8**
	MiniGPT-4	0.672	63.6	65.9	0.794	68.3	78.0	0.768	72.7	75.8	**0.865**	**85.4**	**87.6**
	Flamingo	0.701	64.3	71.2	0.831	82.5	85.4	0.806	78.0	82.9	**0.859**	**84.7**	**87.1**
	BLIP-2	0.648	61.6	64.8	0.826	82.2	84.9	0.775	73.7	78.2	**0.873**	**86.6**	**88.2**
VQAv2	LLaVA-1.5	0.611	56.5	56.5	0.839	82.9	83.5	0.795	77.8	81.6	**0.874**	**85.3**	**87.0**
	MiniGPT-4	0.642	61.0	65.9	0.816	79.8	83.2	0.831	82.3	82.3	**0.887**	**86.7**	**86.7**
	Flamingo	0.712	64.5	73.4	0.848	81.7	85.6	0.850	84.7	86.2	**0.869**	**90.1**	**90.1**
	BLIP-2	0.687	61.5	71.8	0.841	82.0	85.3	0.834	82.5	83.9	**0.862**	**84.5**	**86.8**
DALLE-3	LLaVA-1.5	0.639	62.6	65.3	0.827	78.4	83.4	0.817	78.0	82.4	**0.888**	**88.6**	**88.6**
	MiniGPT-4	0.690	65.1	71.8	0.810	79.8	82.5	0.829	81.8	84.3	**0.873**	**85.4**	**88.2**
	Flamingo	0.744	65.9	73.2	0.842	81.9	85.3	0.842	82.9	82.9	**0.895**	**89.1**	**89.7**
	BLIP-2	0.715	64.7	72.7	0.835	80.7	84.6	0.848	82.5	85.1	**0.869**	**85.0**	**87.5**

Table 1: Targeted attack performance comparison of existing LVLM attack method. Target text: "*I don't know*".

4. Patch/image merge: We finally compose the transformed patch and image by blending the perturbed patch with the perturbed image using Poisson matting (Sun et al. 2004).

By denoting the physical transformation as $T(\cdot)$, we adversarially train the adversarial patch $\boldsymbol{p}'$ to be robust to transformation $T(\cdot)$ and against LVLMs f_θ as:

$$\begin{aligned}\mathcal{L}_{transform} =& 2 - \cos(g_\phi(f_\theta(\boldsymbol{v}', \boldsymbol{t})), g_\phi(\boldsymbol{y}')) \\ & - \cos(g_\phi(f_\theta(T(\boldsymbol{v}'), \boldsymbol{t})), g_\phi(\boldsymbol{y}')),\end{aligned} \tag{9}$$

where $\boldsymbol{v}' = (1 - \boldsymbol{m}) \odot \boldsymbol{v} + \boldsymbol{m} \odot \boldsymbol{p}'$, g_ϕ is a lightweight pre-trained text encoder to compute the textual semantic distance with the consine similarity function.

Optimization by Querying LVLMs

Since the attackers are assumed to not access the LVLM's details, we need to estimate the gradient direction to update $\boldsymbol{\Delta}$ for perturbing the adversarial patch into a targeted-chosen label by solely querying the LVLM model. We follow the Monte Carlo estimation (James 1980) to employ a series of random slight noises on the previously obtained adversarial perturbation and scrutinizes whether these noises induce alterations in the prediction, the average of these noise directions serves as the ultimate direction for further mutating the perturbation. Specifically, we initialize a normalized uniform distribution to add slight noise $\boldsymbol{\delta}$ on the patch $\boldsymbol{p}'$. At the t-th step, we define an indicator function φ_t to measure whether the perturbation $\boldsymbol{\Delta}$ can cause the attackers' desired target labels as:

$$\varphi_t = \text{sign}(\cos(g_\phi(f_\theta(\boldsymbol{v}', \boldsymbol{t})), g_\phi(\boldsymbol{y}')) - \tau), \tag{10}$$

where $\text{sign}(\cdot)$ is the sign function, τ is the threshold. We estimate the updating direction by weighted averaging over the K possible directions $\{\boldsymbol{\delta}_k\}_{k=1}^{k=K}$, and optimize $\boldsymbol{\delta}$ as:

$$\boldsymbol{\Delta}' = \boldsymbol{\Delta} + \frac{\frac{1}{K}\sum_{k=1}^{K} \varphi_k \boldsymbol{\delta}_k}{||\frac{1}{K}\sum_{k=1}^{K} \varphi_k \boldsymbol{\delta}_k||_2}. \tag{11}$$

By iteratively optimizing the perturbations on the patch pattern over T iterations using the aforementioned weighted gradient estimation, we can obtain the final adversarial patch $\boldsymbol{p}' = \boldsymbol{p} + \boldsymbol{\Delta}'$ against LVLMs.

Experiments

Implementation details

LVLM models. Following existing LVLM attack methods (Shayegani, Dong, and Abu-Ghazaleh 2023; Bailey et al. 2023; Wang et al. 2023, 2024; Luo et al. 2024; Zhao et al. 2024), we conduct experiments on the same open-source LVLM models including LLaVA-1.5 (Liu et al. 2024e), MiniGPT-4 (Zhu et al. 2023a), Flamingo (Alayrac et al. 2022), and BLIP-2 (Li et al. 2023) for fair comparison.

Datasets. 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). We follow the existing works to construct these three datasets. Specifically, we employ images from the test sets of the MS-COCO and VQAv2 to construct two multimodal datasets. We also utilize captions from the MS-COCO validation set as prompts to generate corresponding images with DALLE-3 to form another dataset. For the text input data, we follow the prompts used in previous work (Luo et al. 2024) to build our text dataset.

Evaluation metrics. To evaluate the adversarial success of the targeted attacks, we employ three kinds of metrics: a semantic similarity (Zhao et al. 2024) is introduced to measure the distance between adversarial output and target text (the larger the better); we also follow (Luo et al. 2024; Lu et al. 2024) to exploit success rates "ExactMatch" and "ConditionalContain" to assess the word-level overlap between adversarial output and target text.

Basic setups. We employ Sentence-BERT (Reimers and Gurevych 2019) as the text encoder to measure the LVLM's textual output with the adversarial target. We select three widely used image-to-text tasks to evaluate our attack method, *i.e.*, Image Classification, Image Captioning, and VQA. During the gradient approximation, we allow 10k query numbers in total. The general target label in almost all our experiments is set to text "*I don't know*"; Various other target labels are also experimented in the latter section. We set the hyper-parameter λ as 0.1. In Equation (10), we use $\tau = 0.55$ to determine the direction of gradient predictions for each slight noise $\boldsymbol{\delta}$. We impose $\varepsilon = 16/255$ as the constraint for $\boldsymbol{\Delta}$.

Target Text	LVLM	SS	EM	CC
"I want to destroy the whole world together"	LLaVA-1.5	0.852	82.9	86.7
	MiniGPT-4	0.838	82.9	82.9
	Flamingo	0.834	78.0	85.4
	BLIP-2	0.841	82.3	84.6
"I am sorry, there is no object you mentioned, so I cannot answer"	LLaVA-1.5	0.837	80.5	87.8
	MiniGPT-4	0.815	81.3	84.4
	Flamingo	0.817	80.5	83.8
	BLIP-2	0.809	77.8	80.2
"A photo of a cat"	LLaVA-1.5	0.886	87.5	89.6
	MiniGPT-4	0.853	85.0	85.0
	Flamingo	0.852	84.4	87.1
	BLIP-2	0.865	85.4	87.8
"The view from the top of a hill overlooking the mountains"	LLaVA-1.5	0.849	81.7	88.2
	MiniGPT-4	0.841	83.4	85.3
	Flamingo	0.824	79.8	84.5
	BLIP-2	0.836	83.5	84.0

Table 2: Attack performance on various-type target labels of different lengths conducted on the MS-COCO dataset.

Target Text	LVLM	SS	EM	CC
"I don't know"	LLaVA-1.5	0.805	75.7	83.4
	MiniGPT-4	0.792	72.3	80.0
	Flamingo	0.778	71.5	79.6
	BLIP-2	0.791	73.1	81.2
"I want to destroy the whole world together"	LLaVA-1.5	0.769	71.3	75.9
	MiniGPT-4	0.737	69.9	75.2
	Flamingo	0.744	71.0	75.8
	BLIP-2	0.753	70.7	77.4

Table 3: Attack performance in the **physical world**.

Attack Performance

Digital attack evaluation. We first conduct a comprehensive evaluation of four LVLM models across three digital datasets, as shown in Table 1. Here, we conduct three existing open-source LVLM attackers as baseline models for comparison: MF-Attack (Zhao et al. 2024), CroPA (Luo et al. 2024), and Anydoor (Lu et al. 2024). We select the target text "*I don't know*" to avoid the inclusion of high-frequency responses. All the performance is averaged on the three tasks. From this table, we can find that: (1) existing LVLM attackers achieve relatively worse adversarial performance as they fail to design explicit optimization objectives to update the noise. (2) Since we directly approximate the gradients along the optimal direction, our attack consistently achieves the best performance on all models and datasets.

To demonstrate that the effectiveness of the proposed attack is not constrained to the specific case of the target text "*I don't know*", we extend our evaluation to various other target texts. The experiment includes a selection of text with varied length and usage frequency as shown in Table 2. We can observe that our attack can also achieve significant attack performance on long/specific target text though the output similarity differs for different targets, demonstrating the scalability and generalization-ability of our attack.

Physical attack evaluation. We then conduct a physical evaluation of our proposed attack across four LVLM models, as shown in Table 3. Here, we print and paste our generated adversarial patches in the real world and exploit the camera to capture corresponding 100 adversarial images. All the performance is averaged on the three tasks *i.e.*, Image Classification, Image Captioning, and VQA. From this table, we can find that: although our physical attack performance is relatively worse than our digital attack performance (in Table 1), it still achieves effective targeted attacks. We also show some visualizations of physical attacks in Figure 4, where our printed patches are imperceptible within the backgrounds.

Figure 4: Examples of our physical-world attacks.

Defense Method	Attack Method	SS	EM	CC
No Defense	MF-Attack	0.596	56.7	62.0
	CroPA	0.812	74.6	83.1
	Anydoor	0.734	71.4	74.3
	Our Attack	**0.890**	**87.5**	**89.8**
RandomTransform	MF-Attack	0.527	34.3	42.9
	CroPA	0.732	65.8	74.4
	Anydoor	0.675	61.6	65.9
	Our Attack	**0.847**	**82.5**	**85.0**

Table 4: Attack performance on LLaVA-1.5 model and MS-COCO dataset against transformation-based defenses.

Robustness to Potential Defenses

We investigate the robustness of our proposed attack method. As shown in Table 4, we first report the attack performance on the LLaVA-1.5 model and MS-COCO dataset against the traditional transformation defense strategy (Athalye et al. 2018). It indicates that existing LVLM attackers are not robust to these transformations and their performances degenerate a lot. Instead, since we utilize adversarial learning strategies to train our adversarial pattern to be robust to potential real-world transformations, our attack achieves much better performance against transformation-based defenses.

Then, we implement detection-based defenses to investigate the robustness of our proposed adversarial patch where we also conduct patch-based attack Anydoor (Lu et al. 2024) for comparison. As shown in Table 5, three popular patch-aware defenses are implemented: PatchCleanser (Xiang, Mahloujifar, and Mittal 2022) uses certifiable defense against adversarial patches; Jedi (Tarchoun et al. 2023) employs entropy-based strategy to distinguish the patch and its neigh-

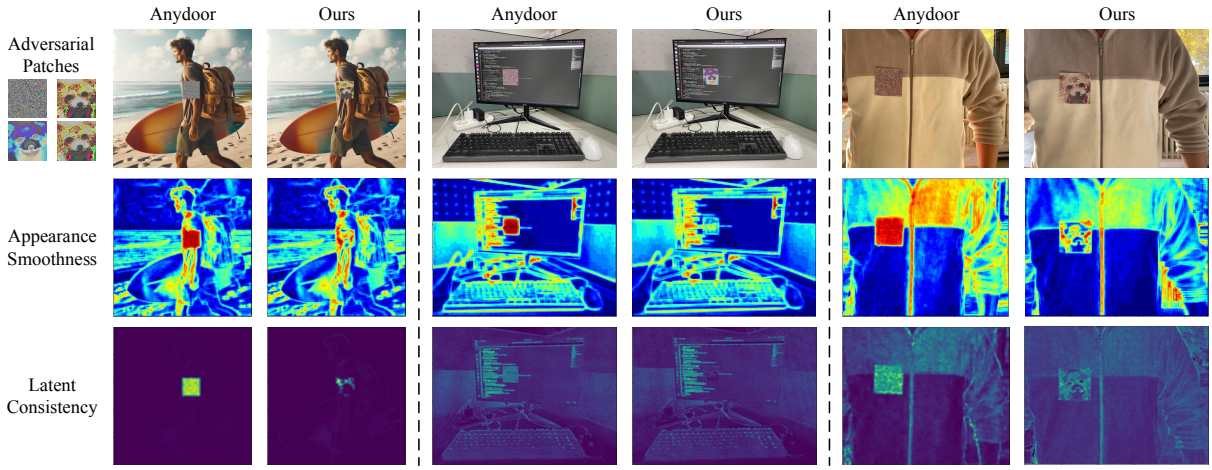


Figure 5: Visualization illustrating the stealthy and robustness of adversarial patches generated by Anydoor and our attack.

ASR against Defense	PatchCleanser	Jedi	JPEG
Anydoor	72.6%	31.8%	27.9%
Our Attack	**86.2%**	**80.7%**	**78.4%**

Table 5: Attack performance on LLaVA-1.5 model and MS-COCO dataset against patch-aware detection-based defenses.

Component	Variant	SS	EM	CC
Patch Size	32×32	0.815	80.7	83.4
	64×64	0.861	84.3	87.6
	128×128	0.890	87.5	89.8
	256×256	0.898	86.9	91.2
Patch Pattern	Noise	0.898	88.2	90.0
	Cat	0.886	86.8	89.4
	Dog	0.890	87.5	89.8

Table 6: Ablation studies of the design of adversarial patch on LLaVA-1.5 model and MS-COCO dataset.

bor pixels; JPEG (Bianchi and Piva 2012) identifies abnormal patches by compressing the entire image and searching the affected region with different quality. Compared to the Anydoor attack, our attack introduces effective spatial-spectral homogeneous constraints, thus achieving better robustness in this table. We also provide more visualization of these detection-based defenses in Figure 5, where our generated adversarial patches have much better appearance smoothness and latent consistency than the previous Anydoor attack. Our adversarial patch is also more natural and imperceptible.

Ablation Study

Investigation on the patch design. We first conduct ablation studies on the LLaVA-1.5 model and MS-COCO dataset to investigate the patch design in Table 6. We can find that a larger size will lead to better attack performances due to its more pixel-wise perturbations. A 128×128 patch size is already effective enough to provide a good attack performance, therefore, we choose 128×128 in most our experiments.

Loss function. We conduct the ablations on different loss combinations in Figure 6.

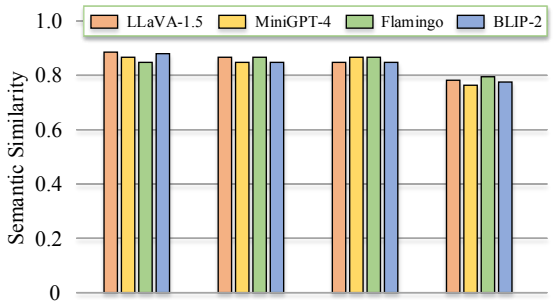


Figure 6: Ablations on different loss combinations.

Metric	MF-Attack	CroPA	Anydoor	Ours
GPU Time	29min	14min	5min	7min
GPU Memory	35GB	26GB	22GB	20GB

Table 7: Analysis on attack complexity and efficiency.

Complexity analysis. As shown in Table 7, although our attack is implemented in a more challenging physical-world setting, it is still efficient and costs competitive resources.

Conclusion

In this paper, we introduce a new LVLM attack setting, *i.e.*, physical-world attack, where the attackers solely have access to the image camera and have no prior knowledge of LVLMs' details. We propose a novel printable adversarial patch design with adversarial homogeneous constraints in both spatial and spectral domains to improve the patch stealthily. A realistic transformation strategy is further devised to synthesize the physical patch variations. Extensive experiments are conducted on both digital and physical attack settings, demonstrating the effectiveness of our attack.

Acknowledgements

This work is supported by the National Key R&D Program of China under contract No. 2024YFF0907603.

References

Ahmed, N.; Natarajan, T.; and Rao, K. R. 1974. Discrete cosine transform. *TOC*, 100(1): 90–93.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.

Athalye, A.; Engstrom, L.; Ilyas, A.; and Kwok, K. 2018. Synthesizing robust adversarial examples. In *ICML*.

Bailey, L.; Ong, E.; Russell, S.; and Emmons, S. 2023. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*.

Bianchi, T.; and Piva, A. 2012. Image forgery localization via block-grained analysis of JPEG artifacts. *TIFS*.

Bookstein, F. L. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *TPAMI*.

Brown, T. B.; Mané, D.; Roy, A.; Abadi, M.; and Gilmer, J. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665*.

Cai, F.; Liu, D.; Fang, X.; Yu, J.; Tang, K.; and Zhou, P. 2025a. Imperceptible Beam-Sensitive Adversarial Attacks for LiDAR-based Object Detection in Autonomous Driving. In *ICME*.

Cai, X.; Liu, D.; Guan, R.; and Zhou, P. 2025b. Imperceptible Transfer Attack on Large Vision-Language Models. In *ICASSP*, 1–5. IEEE.

Cai, X.; Liu, D.; Qu, X.; Fang, X.; Dong, J.; Tang, K.; Zhou, P.; Sun, L.; and Hu, W. 2025c. Towards Building Model/Prompt-Transferable Attackers against Large Vision-Language Models. In *NeurIPS*.

Cai, X.; Tao, Y.; Liu, D.; Zhou, P.; Qu, X.; Dong, J.; Tang, K.; and Sun, L. 2024. Frequency-aware gan for imperceptible transfer attack on 3d point clouds. In *ACM MM*, 6162–6171.

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? *arXiv preprint*.

Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A. K.; and Yang, Y. 2020. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *CVPR*, 1000–1008.

Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; and Song, D. 2018. Robust physical-world attacks on deep learning visual classification. In *CVPR*, 1625–1634.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Hu, Q.; Liu, D.; and Hu, W. 2022. Exploring the devil in graph spectral domain for 3d point cloud attacks. In *ECCV*.

Huang, W.; Liu, D.; and Hu, W. 2023. Dense object grounding in 3d scenes. In *ACM MM*, 5017–5026.

James, F. 1980. Monte Carlo theory and practice. *Reports on progress in Physics*, 43(9): 1145.

Jing, L.; Wang, R.; Ren, W.; Dong, X.; and Zou, C. 2024. PAD: Patch-agnostic defense against adversarial patch attacks. In *CVPR*, 24472–24481.

Kandasamy, K.; Krishnamurthy, A.; Poczos, B.; Wasserman, L.; et al. 2015. Nonparametric von mises estimators for entropies, divergences and mutual informations. *NeurIPS*.

Karmon, D.; Zoran, D.; and Goldberg, Y. 2018. Lavan: Localized and visible adversarial noise. In *ICML*, 2507–2515.

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 *ICML*.

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 *ECCV*.

Liu, D.; Cai, X.; Zhou, P.; Qu, X.; Fang, X.; Sun, L.; and Hu, W. 2024a. Are Large Vision-Language Models Robust to Adversarial Visual Transformations? *OpenReview*.

Liu, D.; Fang, X.; Hu, W.; and Zhou, P. 2023a. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *TMM*.

Liu, D.; Fang, X.; Zhou, P.; Di, X.; Lu, W.; and Cheng, Y. 2023b. Hypotheses tree building for one-shot temporal sentence localization. In *AAAI*, volume 37, 1640–1648.

Liu, D.; and Hu, W. 2022a. Imperceptible transfer attack and defense on 3d point cloud classification. *TPAMI*.

Liu, D.; and Hu, W. 2022b. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *ACM MM*, 4536–4545.

Liu, D.; and Hu, W. 2024a. Can't See the Wood for the Trees: Can Visual Adversarial Patches Fool Hard-Label Large Vision-Language Models? *OpenReview*.

Liu, D.; and Hu, W. 2024b. Explicitly Perceiving and Preserving the Local Geometric Structures for 3D Point Cloud Attack. In *AAAI*, volume 38, 3576–3584.

Liu, D.; and Hu, W. 2025. Seeing is Not Believing: Adversarial Natural Object Optimization for Hard-Label 3D Scene Attacks. In *CVPR*, 11886–11897.

Liu, D.; Hu, W.; and Li, X. 2023a. Point cloud attacks in graph spectral domain: When 3d geometry meets graph signal processing. *TPAMI*.

Liu, D.; Hu, W.; and Li, X. 2023b. Robust geometry-dependent attack for 3D point clouds. *TMM*.

Liu, D.; Liu, Y.; Huang, W.; and Hu, W. 2024b. A Survey on Text-guided 3D Visual Grounding: Elements, Recent Advances, and Future Directions. *arXiv preprint arXiv:2406.05785*.

Liu, D.; Ouyang, X.; Xu, S.; Zhou, P.; He, K.; and Wen, S. 2020a. SAANet: Siamese action-units attention network for improving dynamic facial expression recognition. *Neurocomputing*, 413: 145–157.

Liu, D.; Qu, X.; Di, X.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022a. Memory-guided semantic learning network for temporal sentence grounding. In *AAAI*, volume 36, 1665–1673.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2020b. Reasoning step-by-step: Temporal sentence localization in videos via deep rectification-modulation network. In *COLING*.

Liu, D.; Qu, X.; Dong, J.; and Zhou, P. 2021a. Adaptive Proposal Generation Network for Temporal Sentence Localization in Videos. In *EMNLP*, 9292–9301.

Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Cheng, Y.; Wei, W.; Xu, Z.; and Xie, Y. 2021b. Context-aware biaffine localizing network for temporal sentence grounding. In *CVPR*.

Liu, D.; Qu, X.; and Hu, W. 2022. Reducing the vision and language bias for temporal sentence grounding. In *ACM MM*.

Liu, D.; Qu, X.; Liu, X.-Y.; Dong, J.; Zhou, P.; and Xu, Z. 2020c. Jointly cross-and self-modal graph attention network for query-based moment localization. In *ACM MM*.

Liu, D.; Qu, X.; Wang, Y.; Di, X.; Zou, K.; Cheng, Y.; Xu, Z.; and Zhou, P. 2022b. Unsupervised temporal video grounding with deep semantic clustering. In *AAAI*.

Liu, D.; Qu, X.; and Zhou, P. 2021. Progressively Guide to Attend: An Iterative Alignment Framework for Temporal Sentence Grounding. In *EMNLP*.

Liu, D.; Qu, X.; Zhou, P.; and Liu, Y. 2022c. Exploring motion and appearance information for temporal sentence grounding. In *AAAI*, volume 36, 1674–1682.

Liu, D.; Xu, S.; Liu, X.-Y.; Xu, Z.; Wei, W.; and Zhou, P. 2021c. Spatiotemporal graph neural network based mask reconstruction for video object segmentation. In *AAAI*.

Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Fang, X.; Tang, K.; Wan, Y.; and Sun, L. 2024c. Pandora's Box: Towards Building Universal Attackers against Real-World Large Vision-Language Models. In *NeurIPS*.

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

Liu, D.; Yu, D.; Wang, C.; and Zhou, P. 2021d. F2net: Learning to focus on the foreground for unsupervised video object segmentation. In *AAAI*, volume 35, 2109–2117.

Liu, D.; Zhang, H.; and Zhou, P. 2021. Video-based facial expression recognition using graph convolutional networks. In *ICPR*, 607–614. IEEE.

Liu, D.; Zhou, P.; Xu, Z.; Wang, H.; and Li, R. 2022d. Few-shot temporal sentence grounding via memory-guided semantic learning. *TCSVT*, 33(5): 2491–2505.

Liu, D.; Zhu, J.; Fang, X.; Xiong, Z.; Wang, H.; Li, R.; and Zhou, P. 2023c. Conditional video diffusion network for fine-grained temporal sentence grounding. *TMM*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024e. Visual instruction tuning. *NeurIPS*, 36.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

Liu, Y.; Liu, D.; Guo, Z.; and Hu, W. 2024f. Cross-task knowledge transfer for semi-supervised joint 3d grounding and captioning. In *ACM MM*, 3818–3827.

Lu, D.; Pang, T.; Du, C.; Liu, Q.; Yang, X.; and Lin, M. 2024. Test-Time Backdoor Attacks on Multimodal Large Language Models. *arXiv preprint arXiv:2402.08577*.

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2024. An Image Is Worth 1000 Lies: Adversarial Transferability across Prompts on Vision-Language Models. *arXiv preprint arXiv:2403.09766*.

Moon, Y.-I.; Rajagopalan, B.; and Lall, U. 1995. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3): 2318.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint*.

Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *ICLR*.

Skodras, A.; Christopoulos, C.; and Ebrahimi, T. 2001. The JPEG 2000 still image compression standard. *SPM*.

Sun, J.; Jia, J.; Tang, C.-K.; and Shum, H.-Y. 2004. Poisson matting. In *ACM SIGGRAPH 2004 Papers*, 315–321.

Tao, Y.; Liu, D.; Zhou, P.; Xie, Y.; Du, W.; and Hu, W. 2023. 3DHacker: Spectrum-based decision boundary generation for hard-label 3D point cloud attack. In *ICCV*.

Tarchoun, B.; Ben Khalifa, A.; Mahjoub, M. A.; Abu-Ghazaleh, N.; and Alouani, I. 2023. Jedi: Entropy-based localization and removal of adversarial patches. In *CVPR*.

Wallace, G. K. 1991. The JPEG still picture compression standard. *Communications of the ACM*, 34(4): 30–44.

Wang, X.; Ji, Z.; Ma, P.; Li, Z.; and Wang, S. 2023. InstructTA: Instruction-Tuned Targeted Attack for Large Vision-Language Models. *arXiv preprint arXiv:2312.01886*.

Wang, Z.; Han, Z.; Chen, S.; Xue, F.; Ding, Z.; Xiao, X.; Tresp, V.; Torr, P.; and Gu, J. 2024. Stop Reasoning! When Multimodal LLMs with Chain-of-Thought Reasoning Meets Adversarial Images. *arXiv preprint arXiv:2402.14899*.

Xiang, C.; Mahloujifar, S.; and Mittal, P. 2022. {PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier. In *USENIX*.

Yan, H.; Ma, H.; Cai, X.; Liu, D.; Yuan, Z.; Qu, X.; Dong, J.; Guan, R.; Fang, X.; He, H.; et al. 2025. Fit the Distribution: Cross-Image/Prompt Adversarial Attacks on Multimodal Large Language Models. In *NeurIPS*.

Yang, M.; Liu, D.; Tang, K.; Zhou, P.; Chen, L.; and Chen, J. 2024. Hiding imperceptible noise in curvature-aware patches for 3d point cloud attack. In *ECCV*, 431–448.

Zhang, H.; Shao, W.; Liu, H.; Ma, Y.; Luo, P.; Qiao, Y.; and Zhang, K. 2024. AVIBench: Towards Evaluating the Robustness of Large Vision-Language Model on Adversarial Visual-Instructions. *arXiv preprint arXiv:2403.09346*.

Zhao, Y.; Pang, T.; Du, C.; Yang, X.; Li, C.; Cheung, N.-M. M.; and Lin, M. 2024. On evaluating adversarial robustness of large vision-language models. *NeurIPS*, 36.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint*.

Zhu, J.; Liu, D.; Zhou, P.; Di, X.; Cheng, Y.; Yang, S.; Xu, W.; Xu, Z.; Wan, Y.; Sun, L.; et al. 2023b. Rethinking the video sampling and reasoning strategies for temporal sentence grounding. *arXiv preprint arXiv:2301.00514*.