Unveiling the Fragility of Vision-Language Models: Multi-Modal Adversarial Synergy via Texture-Constrained Perturbations and Cross-Modal Optimization

Xiang Fang[1], **Wanlong Fang**[2], **Changshuo Wang**[3*]
[1]School of Software Engineering, Huazhong University of Science and Technology
[2]Nanyang Technological University, Singapore
[3]University College London
xfang9508@gmail.com, wanlongfang@gmail.com, wangchangshuo1@gmail.com

Abstract

Large Vision-Language Models (LVLMs) have transformed multi-modal understanding, excelling in tasks like image captioning and visual question answering by integrating visual and textual inputs. However, their robustness against adversarial attacks—particularly those exploiting both modalities—remains underexplored, posing risks to critical applications like autonomous driving and content moderation. Existing attacks focus on single modalities or require impractical white-box access, limiting their real-world relevance. In this paper, we introduce *Multi-Modal Adversarial Synergy*, a groundbreaking framework that crafts universal, black-box multi-modal attacks against LVLMs. MMAS simultaneously generates a texture scale-constrained Universal Adversarial Perturbation for images and a learnable prompt perturbation for text, optimized jointly using only model queries. The image perturbation, bounded by an ℓ_∞-norm, leverages wavelet-based texture constraints to ensure imperceptibility and robustness across diverse visual inputs. The text perturbation, constrained by an ℓ_2-norm in the embedding space, maintains semantic coherence while steering outputs toward a target. A novel cross-modal regularization term aligns the perturbations' gradient directions, enhancing their synergistic impact and transferability across tasks and models. Extensive experiments show the strong universal adversarial capabilities of our proposed attack with prevalent LVLMs.

Introduction

The rapid advancement of Large Vision-Language Models (LVLMs), such as CLIP (Radford et al. 2021; Liang et al. 2025, 2023; Tang et al. 2023, 2024, 2022, 2020; Tang, He, and Qin 2025), has revolutionized multi-modal learning, enabling seamless integration of visual and textual data for tasks ranging from image captioning (Zeng et al. 2024; Fang et al. 2025a, 2023c, 2022, 2023b; Fang, Fang, and Wang 2025; Fang, Easwaran, and Genest 2025) to visual question answering (VQA) (Sima et al. 2024; Fang and Fang 2026; Fang et al. 2026, 2025c, 2024b, 2025d,b, 2024a,c, 2023a, 2021b; Fang, Easwaran, and Genest 2025; Fang et al. 2020, 2021a; Fang, Easwaran, and Genest 2024; Fang and Hu 2020). These models leverage vast pre-training datasets and sophisticated architectures to achieve remarkable generalization across diverse applications. However, their increasing deployment in real-world systems—such as autonomous vehicles (Cui et al. 2024; Wang et al. 2025a,b; Wang, Fang, and Tiwari 2025; Wang et al. 2026, 2025c), content moderation (Liu 2024; Ma et al. 2025, 2024a,b), and medical diagnostics (Myrzashova et al. 2024)—raises critical concerns about their robustness to adversarial attacks (Xing et al. 2025). Adversarial perturbations (Li et al. 2022), subtle alterations to inputs designed to mislead machine learning models, have been extensively studied in unimodal contexts, such as image classification (Rao et al. 2021) and natural language processing (Min et al. 2023). Yet, the multi-modal nature of LVLMs introduces a new frontier: how resilient are these models to coordinated attacks across both vision and language modalities?

Recent works have begun to explore adversarial vulnerabilities in LVLMs, revealing alarming weaknesses (Dai et al. 2025). For instance, (Luo et al. 2024) demonstrated that carefully crafted text prompts can mislead LVLMs across multiple tasks, exploiting the models' sensitivity to prompt variations. Similarly, (Liu et al. 2024a) introduced universal adversarial patches that disrupt image understanding in a task-agnostic manner, while (Yin et al. 2023) extended this to black-box settings, showing transferability across models. In the vision domain, (Huang et al. 2024) proposed texture scale-constrained perturbations, highlighting how structured noise can enhance attack robustness. Despite these advances, existing approaches predominantly focus on single-modality attacks, leaving a critical gap in understanding how joint image-text perturbations might amplify adversarial effects. This gap is particularly pressing given the interdependent processing of vision and language in LVLMs, where cross-modal interactions could be exploited to craft more potent and practical attacks.

The implications of this vulnerability are profound. In safety-critical applications, such as autonomous driving (Zhao et al. 2024a), an attacker could pair a perturbed road sign image with a misleading textual instruction to cause catastrophic misinterpretation. In content moderation, subtle image-text manipulations could bypass filters, allowing harmful content to proliferate. Moreover, the universal nature of LVLMs—designed to handle arbitrary image-text pairs—suggests that effective attacks must generalize

*Corresponding Author.

across tasks and inputs, a challenge unmet by task-specific or modality-isolated methods. Existing multi-modal attack methods (Abdullakutty, Elyan, and Johnston 2021) remain limited to white-box settings with access to model gradients, rendering them impractical for real-world scenarios where only query access is available. Thus, there is an urgent need for a practical, black-box, multi-modal attack framework that exploits the synergy between image and text perturbations while ensuring universality and transferability.

In this paper, we introduce *Multi-Modal Adversarial Synergy (MMAS)*, a novel framework to craft universal adversarial attacks against LVLMs. MMAS simultaneously generates a texture scale-constrained Universal Adversarial Perturbation (UAP) for images and a learnable prompt perturbation for text, optimized jointly in a black-box setting using only model queries. Our approach builds upon insights from prior works: we adapt the texture-constrained UAP concept to ensure visual imperceptibility and robustness, draw on the prompt perturbation strategy for text attacks, and incorporate the universal and black-box optimization principles. However, MMAS transcends these foundations by introducing a key innovation: a cross-modal regularization term that aligns the image and text perturbations, enhancing their combined efficacy and transferability across diverse tasks and inputs.

Our method operates under realistic constraints. The image perturbation, bounded by an ℓ_∞-norm, leverages wavelet-based texture scales to maintain perceptual similarity while disrupting visual features universally. The text perturbation, optimized in the embedding space with an ℓ_2-norm constraint, ensures semantic coherence while steering the LVLM toward a target output. By jointly optimizing these perturbations with a query-based gradient approximation, MMAS achieves a practical attack that requires no internal model knowledge—a significant departure from white-box methods. Furthermore, the cross-modal regularization encourages synergy between modalities, addressing the limitation of prior works where image and text attacks were designed independently, often resulting in suboptimal performance when combined.

We evaluate MMAS on a range of LVLMs, including CLIP and Flamingo, across tasks such as image captioning, VQA, and text-guided image classification. Our results demonstrate that MMAS achieves higher attack success rates and better transferability compared to single-modality baselines and naive multi-modal combinations. For example, in a targeted attack scenario, MMAS can manipulate an LVLM to consistently misclassify a "stop sign" image paired with a "proceed" prompt as "go", with perturbations imperceptible to human observers. These findings underscore the vulnerability of LVLMs to coordinated multi-modal attacks and highlight the need for robust defenses.

Our contributions can be summarized as follows: 1) We propose MMAS, the first black-box, universal multi-modal attack framework for LVLMs, integrating texture scale-constrained image perturbations and learnable text prompt perturbations. 2) We introduce a novel cross-modal regularization term to enhance the synergy between image and text perturbations, improving attack efficacy and transferability. 3) We provide extensive evaluations showing MMAS's superiority over existing methods, offering new insights into the multi-modal vulnerabilities of LVLMs.

Related Work

Adversarial attacks in the vision domain originated with seminal works (Goodfellow, Shlens, and Szegedy 2014; Akhtar and Mian 2018; Zhang et al. 2020), which introduced gradient-based perturbations to mislead image classifiers. Subsequent studies refined these attacks for practicality and generality. For instance, (Moosavi-Dezfooli et al. 2017) pioneered Universal Adversarial Perturbations (UAPs), single noise patterns effective across multiple images, enhancing attack scalability. More recently, (Huang et al. 2024) proposed texture scale-constrained perturbations, leveraging wavelet transforms to craft robust, visually coherent noise that withstands image transformations. While effective against vision-only models, these methods do not address the multi-modal nature of LVLMs, where text inputs play a critical role.

The intersection of vision and language in LVLMs has spurred initial multi-modal attack explorations (Wallace et al. 2019). (Lapid and Sipper 2023) combined image noise with text alterations in a white-box setting, achieving targeted mispredictions in CLIP-like models. However, its dependence on model internals restricts real-world applicability. (Yin et al. 2023) advanced this by developing task-agnostic perturbations in a black-box framework, using query-based optimization to approximate gradients. While promising, it focuses on image perturbations with static text, missing the opportunity to jointly optimize both modalities. (Luo et al. 2024) and (Liu et al. 2024a) hint at multi-modal potential by pairing their respective text and image attacks, but these combinations are ad hoc, lacking a unified optimization strategy. No prior work has systematically addressed the synergy between image and text perturbations in a universal, black-box context, leaving LVLMs' full vulnerabilities underexplored.

Methodology

In this section, we present *Multi-Modal Adversarial Synergy (MMAS)*, a novel framework to craft multi-modal adversarial attacks against Large Vision-Language Models (LVLMs). Our approach simultaneously generates a texture scale-constrained Universal Adversarial Perturbation (UAP) for images and a learnable prompt perturbation for text, aiming to mislead LVLMs across diverse tasks and prompts. The method operates in a practical black-box setting, relying solely on model queries, and introduces a cross-modal regularization term to enhance attack transferability. Figure 1 illustrates the MMAS pipeline, comprising initialization, joint optimization, and evaluation stages.

Problem formulation. Consider an LVLM $f_\theta(\mathbf{v}, \mathbf{t})$ that processes a clean image $\mathbf{v} \in \mathbb{R}^{H\times W\times C}$ and a textual prompt $\mathbf{t} \in \mathcal{T}$ to produce an output $\mathbf{y}$. Our goal is to craft an adversarial image $\mathbf{v}' = \mathbf{v} + \delta_v$ and an adversarial prompt $\mathbf{t}' = \mathbf{t} + \delta_t$, where δ_v is the image perturbation and δ_t is the text perturbation, such that the LVLM consistently outputs a predefined target text $\mathbf{y}'$ (targeted attack) across

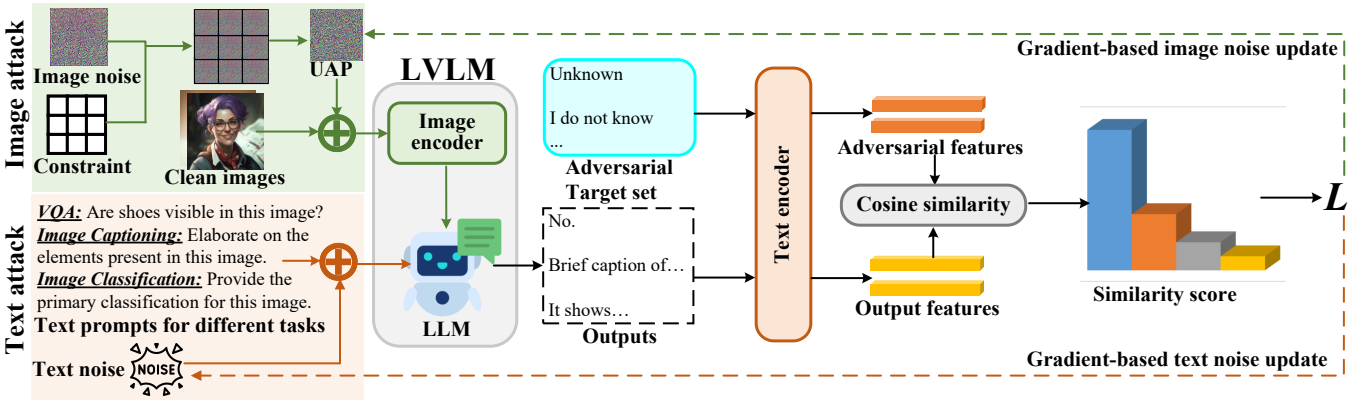

Figure 1: Overview of our proposed method. The framework initializes a texture scale-constrained Universal Adversarial Perturbation (UAP) for images and a learnable prompt perturbation for text, followed by joint optimization using query-based Projected Gradient Descent with cross-modal regularization to align perturbations. The resulting universal attack is evaluated on unseen image-text pairs across diverse tasks and LVLMs, achieving high efficacy and transferability. Best viewed in color.

various tasks, i.e., $f_\theta(\mathbf{v}', \mathbf{t}') \mapsto \mathbf{y}'$. We impose constraints $\|\delta_v\|_\infty \leq \epsilon_v$ and $\|\delta_t\|_2 \leq \epsilon_t$ to ensure imperceptibility and semantic coherence, respectively. The attack is universal, meaning δ_v and δ_t are task-agnostic and applicable to any input pair $(\mathbf{v}, \mathbf{t})$.

Threat model. In this paper, we explore the scenario of attacking real-world LVLM models. In this paper, the assumption is that the attacker has no knowledge of the victim model, including its parameters, training procedure, original training data, etc. In particular, unlike in white-/gray-box attacks, we cannot access the model's gradient information to train perturbations through back-propagation, nor can we, as in black-box attacks, obtain confidence scores/logits from the model outputs.

Texture Scale-Constrained UAP for Image Attack

To craft a Universal Adversarial Perturbation (UAP) featuring category-specific local textures, one straightforward approach is to divide the UAP into numerous small sections, each tailored with a distinct texture. Given that UAP creation relies on gradient back-propagation across its entire scale, adjusting the gradient for each segment in a detailed manner becomes intricate and challenging. To address this, we suggest reversing the perspective and streamlining the task: starting with a category-specific local texture patch, how can we assemble a larger patch—matching the dimensions of the training images—as the UAP? A practical solution is to tile the patch into a broader expanse, leveraging a standard image processing technique that incurs minimal effort. We design δ_v as a Universal Adversarial Perturbation (UAP) constrained by texture scales to enhance its robustness and transferability across images and tasks. Unlike traditional UAPs, our perturbation leverages multi-scale texture features extracted via a wavelet transform to maintain visual coherence while maximizing adversarial impact.

Let $\mathcal{W}(\mathbf{v})$ denote the wavelet decomposition of the clean image $\mathbf{v}$, yielding coefficients at multiple scales $\{s_1, s_2, \ldots, s_L\}$. We constrain δ_v to align with the texture patterns at scale s_k by projecting it onto the wavelet subspace: $\delta_v = \sum_{k=1}^{L} s_k \cdot \mathcal{W}^{-1}(\mathbf{W}_k \odot \mathcal{W}(\mathbf{v}))$, where $\mathcal{W}^{-1}$ is the inverse wavelet transform; $\mathbf{W}_k$ is a binary mask selecting coefficients at scale k; $\odot$ denotes the Hadamard product; α_k are learnable weights. To ensure universality, δ_v is initialized randomly within the bound $\|\delta_v\|_\infty \leq \epsilon_v$ and optimized across a diverse set of images $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. The texture constraint prevents overfitting to specific image content, enhancing cross-image and cross-task transferability.

Learnable Prompt Perturbation for Text Attack

In many vision-language applications, the text input is quite short, which makes the existing LVLMs vulnerable to attack. For example, the average length of the text in VQAv2 (Goyal et al. 2017) and RefCOCO is 6.21 and 3.57, respectively. Moreover, some words are nonsense, making it unnecessary to design a new approach for attacking the text modality. Therefore, we introduce a learnable prompt perturbation δ_t to the text input $\mathbf{t}$. Unlike fixed prompts, δ_t is optimized in the embedding space of the LVLM's text encoder g_ϕ. For a prompt $\mathbf{t}$, its embedding is $e_\mathbf{t} = g_\phi(\mathbf{t})$. The adversarial prompt embedding becomes: $e_{\mathbf{t}'} = e_\mathbf{t} + \delta_t$, where δ_t is constrained by $\|\delta_t\|_2 \leq \epsilon_t$. During optimization, δ_t is updated to maximize the language modeling loss away from the original output (non-targeted) or minimize it toward the target $\mathbf{y}'$ (targeted). This perturbation is applied only during training, and the final attack uses the optimized δ_t universally across prompts $\mathcal{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_M\}$.

Joint Optimization by Cross-Modal Regularization

In numerous instances, altering just images or texts proves challenging for success, since modifying a single modality often fails to sever the link between visuals and captions. To tackle this issue, we introduce a brand-new joint optimization strategy combining image and text perturbations based on a novel cross-modal regularization module. The objective balances the attack success with modality synergy. For a targeted attack, we define the language modeling loss as

$$\mathcal{L}(\mathbf{y}, \mathbf{y}') = -\log P(\mathbf{y}' | f_\theta(\mathbf{v}', \mathbf{t}')). \quad (1)$$

Algorithm 1: Multi-Modal Adversarial Synergy (MMAS)

Require: LVLM f_θ, target text $\mathbf{y}'$, image set $\mathcal{V}$, prompt set $\mathcal{T}$, bounds ϵ_v, ϵ_t, step sizes α_v, α_t, iterations T, regularization weight λ
Ensure: Adversarial perturbations δ_v, δ_t
Initialize $\delta_v \sim \mathcal{U}(-\epsilon_v, \epsilon_v)$ with texture scale constraint, $\delta_t \sim \mathcal{N}(0, \epsilon_t^2)$
for $t = 1$ to T **do**
  Sample $(\mathbf{v}, \mathbf{t})$ from $\mathcal{V} \times \mathcal{T}$
  Compute $\mathcal{L} = \mathcal{L}(f_\theta(\mathbf{v} + \delta_v, \mathbf{t} + \delta_t), \mathbf{y}')$
  Estimate gradients $\nabla_{\delta_v}\mathcal{L}$ and $\nabla_{\delta_t}\mathcal{L}$ via queries
  Compute $\mathcal{R}(\delta_v, \delta_t) = \|\nabla_{\delta_v}\mathcal{L} \cdot \nabla_{\delta_t}\mathcal{L}\|_2$
  Update δ_v^{t+1} and δ_t^{t+1} using PGD with $\mathcal{L} + \lambda\mathcal{R}$
  Apply texture scale constraint to δ_v^{t+1}
end for
return δ_v, δ_t

The optimization problem is:

$$\min_{\delta_v, \delta_t} \frac{1}{|\mathcal{V}||\mathcal{T}|} \sum_{\mathbf{v}\in\mathcal{V}} \sum_{\mathbf{t}\in\mathcal{T}} \left[\mathcal{L}(f_\theta(\mathbf{v} + \delta_v, \mathbf{t} + \delta_t), \mathbf{y}') + \lambda\mathcal{R}(\delta_v, \delta_t)\right],$$

where $\|\delta_v\|_\infty \leq \epsilon_v$ and $\|\delta_t\|_2 \leq \epsilon_t$, where λ is a hyper-parameter, $\mathcal{R}(\delta_v, \delta_t)$ means the cross-modal regularization: $\mathcal{R}(\delta_v, \delta_t) = \|\nabla_{\delta_v}\mathcal{L} \cdot \nabla_{\delta_t}\mathcal{L}\|_2$. This term encourages alignment between the gradient directions of δ_v and δ_t, ensuring that image and text perturbations reinforce each other, improving attack efficacy across prompts and tasks. Since we operate in a black-box setting, we approximate gradients using a query-based method similar to (Yin et al. 2023). For δ_v, we sample noise $\eta_v \sim \mathcal{U}(-\epsilon_v, \epsilon_v)$ and estimate:

$$\nabla_{\delta_v}\mathcal{L} \approx [\frac{\mathcal{L}(f_\theta(\mathbf{v} + \delta_v + \eta_v, \mathbf{t}'), \mathbf{y}')}{\|\eta_v\|_2} - \frac{\mathcal{L}(f_\theta(\mathbf{v} + \delta_v, \mathbf{t}'), \mathbf{y}')}{\|\eta_v\|_2}] \cdot \frac{\eta_v}{\|\eta_v\|_2}. \quad (2)$$

Similarly, for δ_t, we use noise $\eta_t \sim \mathcal{N}(0, \epsilon_t^2)$ in the embedding space. The perturbations are updated via Projected Gradient Descent (PGD): $\delta_v^{t+1} = \text{Proj}_{\epsilon_v}(\delta_v^t - \alpha_v \cdot \text{sign}(\nabla_{\delta_v}\mathcal{L}))$ and $\delta_t^{t+1} = \text{Proj}_{\epsilon_t}\left(\delta_t^t - \alpha_t \cdot \frac{\nabla_{\delta_t}\mathcal{L}}{\|\nabla_{\delta_t}\mathcal{L}\|_2}\right)$, where α_v and α_t are step sizes, and Proj ensures constraints are met. The texture scale constraint on δ_v is reapplied after each update.

The MMAS algorithm is detailed in Algorithm 1. It initializes δ_v and δ_t, iteratively optimizes them using the joint objective, and evaluates the universal attack on unseen inputs. This framework ensures a novel, robust, and universal multi-modal attack, leveraging cross-modal synergy for enhanced transferability across LVLMs.

Experiments

Experimental Setup

Models: Following existing LVLM attack methods (Shayegani, Dong, and Abu-Ghazaleh 2023; Bailey et al. 2023; Dong et al. 2023; Wang et al. 2023, 2024; Luo et al. 2024; Tao et al. 2024; Zhao et al. 2024b), we assess the current popular open-source LVLMs, including LLaVA (Liu et al. 2024b), MiniGPT-4 (Zhu et al. 2023), Flamingo

Target Model	Method	Classification	Captioning	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Overall
Dataset: MS-COCO						
LLaVA	Clean	0.316	0.423	0.327	0.358	0.356
	TA-UAP	0.834	0.845	0.807	0.845	0.833
	TC-UAP	0.804	0.815	0.777	0.815	0.803
	Ours(Full)	**0.884**	**0.895**	**0.857**	**0.895**	**0.883**
MiniGPT-4	Clean	0.386	0.401	0.421	0.440	0.412
	TA-UAP	0.826	0.842	0.834	0.873	0.844
	TC-UAP	0.796	0.812	0.804	0.843	0.814
	Ours(Full)	**0.876**	**0.892**	**0.884**	**0.923**	**0.894**
Flamingo	Clean	0.412	0.427	0.449	0.483	0.443
	TA-UAP	0.847	0.805	0.831	0.843	0.832
	TC-UAP	0.810	0.788	0.819	0.830	0.812
	Ours(Full)	**0.895**	**0.877**	**0.892**	**0.916**	**0.895**
BLIP-2	Clean	0.413	0.422	0.487	0.526	0.462
	TA-UAP	0.799	0.746	0.804	0.847	0.799
	TC-UAP	0.769	0.716	0.774	0.817	0.769
	Ours(Full)	**0.849**	**0.796**	**0.854**	**0.897**	**0.849**
Dataset: DALLE-3						
LLaVA	Clean	0.310	0.420	0.472	0.495	0.424
	TA-UAP	0.785	0.846	0.804	0.853	0.822
	TC-UAP	0.755	0.816	0.774	0.823	0.792
	Ours(Full)	**0.835**	**0.896**	**0.854**	**0.903**	**0.872**
MiniGPT-4	Clean	0.302	0.325	0.367	0.392	0.347
	TA-UAP	0.843	0.836	0.825	0.848	0.838
	TC-UAP	0.813	0.806	0.795	0.818	0.808
	Ours(Full)	**0.893**	**0.886**	**0.875**	**0.898**	**0.888**
Flamingo	Clean	0.406	0.452	0.427	0.479	0.441
	TA-UAP	0.822	0.849	0.852	0.887	0.852
	TC-UAP	0.792	0.819	0.822	0.857	0.822
	Ours(Full)	**0.872**	**0.899**	**0.902**	**0.937**	**0.903**
BLIP-2	Clean	0.325	0.378	0.410	0.439	0.388
	TA-UAP	0.821	0.799	0.875	0.886	0.845
	TC-UAP	0.788	0.754	0.838	0.847	0.807
	Ours(Full)	**0.866**	**0.873**	**0.910**	**0.938**	**0.897**
Dataset: VQAv2						
LLaVA	Clean	0.401	0.423	0.475	0.487	0.447
	TA-UAP	0.793	0.762	0.809	0.847	0.803
	TC-UAP	0.763	0.732	0.779	0.817	0.773
	Ours(Full)	**0.843**	**0.812**	**0.859**	**0.897**	**0.853**
MiniGPT-4	Clean	0.402	0.414	0.519	0.543	0.470
	TA-UAP	0.846	0.835	0.848	0.884	0.853
	TC-UAP	0.816	0.805	0.818	0.854	0.823
	Ours(Full)	**0.896**	**0.885**	**0.898**	**0.934**	**0.903**
Flamingo	Clean	0.349	0.382	0.396	0.25	0.344
	TA-UAP	0.829	0.786	0.827	0.865	0.827
	TC-UAP	0.799	0.768	0.803	0.841	0.803
	Ours(Full)	**0.869**	**0.875**	**0.852**	**0.899**	**0.874**
BLIP-2	Clean	0.312	0.340	0.357	0.398	0.352
	TA-UAP	0.776	0.825	0.846	0.874	0.830
	TC-UAP	0.746	0.795	0.816	0.844	0.800
	Ours(Full)	**0.826**	**0.875**	**0.896**	**0.924**	**0.880**

Table 1: Attack performance on various LVLM models.

Figure 2: Visualization on the universal adversarial attack.

Method	Attack	LLaVA	BLIP-2	MiniGPT-4	Mean
MF-Attack	black-box	0.590	0.681	0.668	0.646
Ours	universal	**0.879**	**0.894**	**0.745**	**0.839**

Table 2: Comparison with MF-Attack (Zhao et al. 2024b). For a fair comparison, experiments are conducted on the same ImageNet-1k dataset (Deng et al. 2009) in the VQA task.

Method	Attack	Classification	Captioning	$VQA_{general}$	$VQA_{specific}$	Overall
CroPA	white-box	0.75	0.72	0.90	0.96	0.83
Ours	universal	**0.86**	0.78	0.95	0.99	0.89

Table 3: Comparison with CroPA (Luo et al. 2024). For fair comparison, we follow CroPA to evaluate the same ASR metric on OpenFlamingo (Awadalla et al. 2023) and MS-COCO.

Attacker	Gemini-2.0		GPT-4o		Claude-3.5	
	SS↑	ASR↑	SS↑	ASR↑	SS↑	ASR↑
PGD	0.084	0.013	0.076	0.006	0.098	0.024
CroPA	0.132	0.020	0.084	0.011	0.114	0.085
Ours	**0.576**	**0.413**	**0.508**	**0.379**	**0.645**	**0.496**

Table 4: Attack performance on the generated adversarial examples from LLaVA to commercial VLMs, where "SS" means "semantic similarity" and "ASR" means "Attack Success Rate".

Image	Method	LVLM-Output
	No Attack	A plane flying between mountains
	PGD	An airplane and a bird flying
	CroPA	Two people repairing the plane
	Ours	A tiger eating meat
	No Attack	An owl standing on a hill with birds
	PGD	An owl eating leaves
	CroPA	A little bird standing on a branch
	Ours	Two hedgehogs fighting

Table 5: Cases of attack results (top: MS-COCO, bottom: DALLE-3) against LLaVA in different methods.

From	Transfer to	Classification	Captioning	$VQA_{general}$	$VQA_{specific}$	Overall
Transferability across Different Datasets (on Flamingo)						
MS-COCO	MS-COCO	0.895	0.877	0.892	0.916	0.895
	DALLE-3	0.850	0.832	0.847	0.871	0.850
	VQAv2	0.845	0.827	0.842	0.866	0.845
DALLE-3	MS-COCO	0.835	0.862	0.865	0.900	0.866
	DALLE-3	0.872	0.899	0.902	0.937	0.903
	VQAv2	0.835	0.862	0.865	0.900	0.866
VQAv2	MS-COCO	0.832	0.838	0.815	0.862	0.837
	DALLE-3	0.832	0.838	0.815	0.862	0.837
	VQAv2	0.869	0.875	0.852	0.899	0.874
Transferability across Different LVLM Models (on MS-COCO)						
LLaVA	LLaVA	0.884	0.895	0.857	0.895	0.883
	MiniGPT-4	0.854	0.865	0.827	0.865	0.853
	Flamingo	0.864	0.847	0.862	0.886	0.865
	BLIP-2	0.824	0.766	0.824	0.867	0.820
MiniGPT-4	LLaVA	0.846	0.862	0.854	0.893	0.864
	MiniGPT-4	0.876	0.892	0.884	0.923	0.894
	Flamingo	0.865	0.847	0.862	0.896	0.868
	BLIP-2	0.846	0.862	0.854	0.893	0.864
Flamingo	LLaVA	0.865	0.847	0.862	0.886	0.865
	MiniGPT-4	0.855	0.862	0.854	0.893	0.866
	Flamingo	0.895	0.877	0.892	0.916	0.895
	BLIP-2	0.825	0.766	0.824	0.867	0.821
BLIP-2	LLaVA	0.819	0.766	0.824	0.867	0.819
	MiniGPT-4	0.826	0.762	0.824	0.867	0.820
	Flamingo	0.829	0.766	0.824	0.867	0.822
	BLIP-2	0.849	0.796	0.854	0.897	0.849

Table 6: Investigation on the transferability across different datasets and LVLMs with semantic similarity score for evaluation.

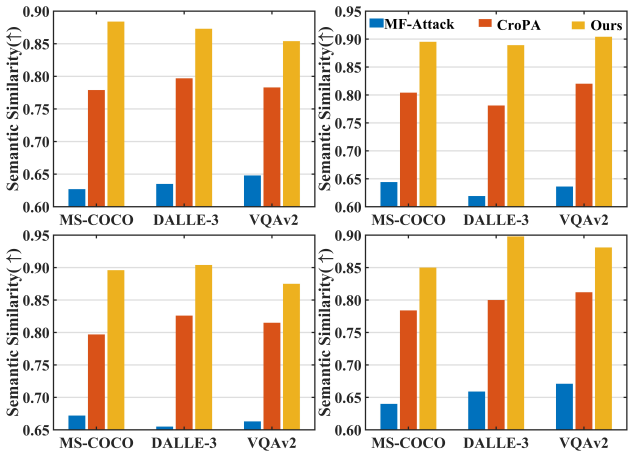

Figure 3: Performance comparison (Overall) with different LVLM attacks on different LVLM models across different datasets (Top left: LLaVA, Top right: MiniGPT-4, Bottom left: Flamingo, Bottom right: BLIP-2).

(Alayrac et al. 2022), and BLIP-2 (Li et al. 2023). To ensure a fair comparison of attack effectiveness, we utilize the

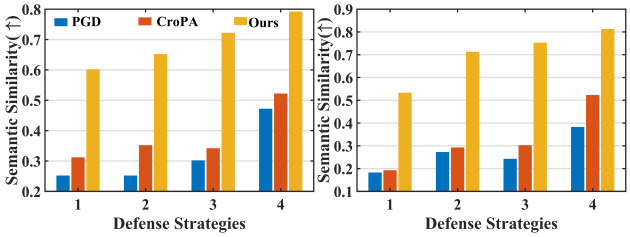

Figure 4: Investigation on the adversarial robustness against various defense methods on MS-COCO (left) and DALLE-3 (right), tested on LLaVa model. Strategy 1: Diffusion Restoration (Nie et al. 2022); strategy 2: Randomization (Xie et al. 2017); strategy 3: JPEG Compression (Guo et al. 2017); strategy 4: No Defense.

Attacker	GPU Time (↓)	GPU Memory (↓)
PGD	8 min	14.3 GB
CroPA	17 min	31.6 GB
Ours	**6 min**	**13.5 GB**

Table 7: Complexity comparison on adversarial sample generation.

Model	Classification	Captioning	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Overall
w/o Text attack	0.756	0.752	0.438	0.829	0.693
w/o Image attack	0.624	0.741	0.543	0.816	0.681
w/o Joint optimization	0.796	0.823	0.710	0.841	0.793
Ours	**0.835**	**0.896**	**0.854**	**0.903**	**0.872**

Table 8: Ablation study on different attacks, tested on the LLaVA model and DALLE-3 dataset.

same pretrained weights of LVLMs as used in previous studies (Liu et al. 2024b; Zhu et al. 2023; Alayrac et al. 2022; Li et al. 2023).

Datasets: To accurately evaluate the attack methodologies, we utilized three challenging datasets to capture a range of visual and contextual diversity: MS-COCO (Lin et al. 2014), VQAv2 (Goyal et al. 2017), and DALLE-3 (Ramesh et al. 2022). We also 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 use 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, with detailed data presented in the supplementary material.

Tasks: We evaluated MMAS across three core LVLM tasks: 1) *Image Classification*, where models assign a single-label descriptor (e.g., "dog"); 2) *Image Captioning*, generating a descriptive sentence (e.g., "A dog sits on grass"); and 3) *Visual Question Answering (VQA)*, answering yes/no or open-ended questions (e.g., "Is there a dog?").

Attack settings. We craft targeted attacks, aiming to force LVLMs to output a predefined target (e.g., "go" for a stop sign image with a "proceed" prompt). Image perturbations

Target Prompt	Method	Classification	Captioning	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Overall
Unknown	Clean	0.398	0.415	0.437	0.469	0.430
	TA-UAP	0.837	0.795	0.821	0.833	0.822
	TC-UAP	0.800	0.778	0.809	0.820	0.802
	Ours(Full)	**0.885**	**0.867**	**0.882**	**0.906**	**0.885**
I am sorry	Clean	0.412	0.427	0.449	0.483	0.443
	TA-UAP	0.847	0.805	0.831	0.843	0.832
	TC-UAP	0.810	0.788	0.819	0.830	0.812
	Ours(Full)	**0.895**	**0.877**	**0.892**	**0.916**	**0.895**
Not sure	Clean	0.405	0.420	0.442	0.475	0.436
	TA-UAP	0.842	0.800	0.826	0.838	0.827
	TC-UAP	0.805	0.783	0.814	0.825	0.807
	Ours(Full)	**0.890**	**0.872**	**0.887**	**0.911**	**0.890**
Very good	Clean	0.425	0.438	0.415	0.452	0.433
	TA-UAP	0.827	0.785	0.811	0.823	0.812
	TC-UAP	0.790	0.768	0.799	0.810	0.792
	Ours(Full)	**0.875**	**0.857**	**0.872**	**0.896**	**0.875**
Too late	Clean	0.385	0.410	0.428	0.460	0.421
	TA-UAP	0.832	0.790	0.816	0.828	0.817
	TC-UAP	0.795	0.773	0.804	0.815	0.797
	Ours(Full)	**0.880**	**0.862**	**0.877**	**0.901**	**0.880**
Metaphor	Clean	0.375	0.395	0.408	0.435	0.403
	TA-UAP	0.822	0.780	0.801	0.813	0.804
	TC-UAP	0.785	0.763	0.789	0.800	0.784
	Ours(Full)	**0.870**	**0.852**	**0.862**	**0.886**	**0.868**

Table 9: Targeted semantic similarity scores tested on Flamingo with different target texts on MS-COCO.

s_k	Classification	Captioning	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Overall
1	0.811	0.790	0.835	0.857	0.823
2	0.825	0.822	0.870	0.884	0.850
4	**0.835**	**0.896**	**0.854**	**0.903**	**0.872**
8	0.817	0.816	0.859	0.870	0.841

Table 10: Ablation on scaling factor s_k (LLaVA-1.5, DALLE-3).

are constrained by $\|\delta_v\|_\infty \leq 8/255$, and text perturbations by $\|\delta_t\|_2 \leq 0.5$ in the embedding space, ensuring imperceptibility. MMAS is optimized over 100 iterations with step sizes $\alpha_v = 0.01$, $\alpha_t = 0.005$, and regularization weight $\lambda = 0.1$, using a batch size of 16.

Implementation details: We implemented MMAS in PyTorch, using wavelet transforms from PyWavelets (Lee et al. 2019) for texture constraints. Queries approximate gradients with 10 samples per iteration, simulating a black-box setting. MMAS was implemented with a base texture patch of size 64×64, tiled with scaling factor $s_k \in \{1, 2, 4, 8\}$. The perturbation magnitude was constrained to $\eta = 16/255$ (pixel range [0, 1]), ensuring near-imperceptibility. Optimization ran for $T = 70,000$ queries, with noise variance $\sigma = 0.01$, threshold $\theta = 0.55$, and weight scaling $\gamma = 5$. We used Sentence-BERT (Reimers and Gurevych 2019) as the text encoder to compute cosine similarity between outputs and the target $\mathbf{r}^* =$ "I am sorry". Experiments were executed on an NVIDIA H100 GPU, averaging 12 hours per model-dataset pair.

Baselines: We compared MMAS against: (1) *Clean* inputs

(no perturbation); (2) *Task-Agnostic UAP (TA-UAP)* (Weng et al. 2024), a black-box universal patch without texture scaling; (3) *Texture-Constrained UAP (TC-UAP)* (Huang et al. 2024), a vision-only texture-based method adapted for LVLMs.

Evaluation metric: Success was measured by semantic similarity (cosine distance in embedding space) between the LVLM's output $\mathbf{r}_{i,j}$ and $\mathbf{r}^*$, averaged across tasks and images: Similarity $= 1 - d(\mathcal{E}_\phi(\mathbf{r}_{i,j}), \mathcal{E}_\phi(\mathbf{r}^*))$. Higher values indicate stronger attack efficacy. The best performance values for each task are highlighted in **bold**.

Main Results

Attack performance on different LVLM models and tasks. To comprehensively evaluate the performance of our proposed method, we conduct experiments on four LVLM models and three challenging datasets. Table 1 illustrates the performance of our method, where we utilize the semantic similarity between the target output and the LVLM's output as the evaluation metric. Especially, we choose the target output "I am sorry" to avoid the inclusion of high-frequency responses. We use $s_k = 4$. The "Overall" column indicates the average semantic similarity score across all tasks. Based on this table, we can observe that our attack method consistently achieves the best performance on all models and datasets for different tasks, which shows the effectiveness of our attack method.

Visualization results. As shown in Figure 2, we visualize the targeted universal attack. Obviously, each adversarial patch can achieve a universal targeted attack, showing the effectiveness of our learnable perturbations.

Comparison with state-of-the-arts. Considering that different LVLM attack methods refer to various settings, we compare our method with each compared method under the same setting for fair comparison. Tables 2 and 3 show the comparison results. Obviously, our proposed method significantly outperforms MF-Attack (Zhao et al. 2024b) and CroPA (Luo et al. 2024). This is because our proposed method can effectively update the visual and textual perturbations by estimating gradients in the victim model. Moreover, Different from CroPA that only attacks white-box cross-prompt in a single task, our proposed method can attack black-box cross-task inputs for better attack performance. As shown in Table 4, we evaluate the performance of our proposed method on realistic LVLM applications Gemini-2.0, GPT-4o and Claude-3.5-Sonnet, where we also achieve best performance. Also, Table 5 showcases instances of attacks targeting LLaVA. In contrast to state-of-the-art attack methods, our strategy produces greater differences between the model's responses and the reference captions.

Investigation on the transferability across different datasets and LVLMs. Since our proposed attack can craft universal adversarial perturbations applicable to any input across various tasks, examining how well these perturbations transfer is crucial. We present the transfer-attack results in Table 6, evaluating their effectiveness across diverse datasets and LVLM models. To assess dataset transferability, we create a universal patch targeting the LLaVA model on one dataset, then apply it to the test sets of two other datasets, feeding the results into LLaVA for analysis. For model transferability, we produce a patch against a specific model using the DALLE-3 dataset and test its impact on three additional LVLM models. Our findings reveal that this attack delivers strong performance, underscoring the success of our universal design. However, transferability across datasets proves less effective than across LVLM models, likely due to varying image distributions among the datasets.

Robustness to defenses. To assess how well our attack withstands protective measures, we test it against three pre-processing defense techniques—Randomization, JPEG Compression, and Diffusion Restoration in Figure 4. Unlike compared attack methods, our approach demonstrates greater robustness against these defense methods, as we deliberately design the adversarial perturbation to maximize its disruptive impact. This strategy enhances its ability to steer the LVLM's reasoning astray, increasing the likelihood of incorrect outcomes compared to a harmless pattern.

Complexity analysis. To investigate the scalability and practicality of our proposed transfer-attack method, we provide the complexity analysis in Table 7 on LLaVA-1.5. It indicates that our attack costs relatively fewer GPU resources as our informative constraints are easily achieved with solely loss designs while our samples can achieve better transfer attack performance within single generation process.

Ablation Study

Ablation on image and text attacks. To elucidate the role of each attack of our method, we conduct ablation studies regarding the components (*i.e.*, text attack, image attack and joint optimization). In particular, we remove each key individual module to investigate its contribution. 1) removing the text attack, 2) removing the image attack and 3) removing the joint optimization. As shown in Table 8, all three module provide the significant performance improvement. The results demonstrate that each component of our method contributes positively to improving attack performance for different tasks.

Ablation on different target texts. To show that the success of our proposed attack is not limited to the specific target text "I am sorry", we broaden our assessment to include a range of alternative target texts. The experiment features texts of varying lengths and usage frequencies, as detailed in Table 9. Our findings indicate that the attack excels both overall and in each specific task across these diverse targets, though the similarity of outputs varies depending on the text chosen.

Ablation study on scaling factor s_k: Table 10 shows similarity peaking at $s_k = 4$, with declines at $s_k = 8$ due to overly fine textures losing coherence, and $s_k = 1$ lacking texture augmentation. Thus, we set $s_k = 4$ in this paper.

Conclusion

In this paper, we introduced a pioneering framework, MMAS, to expose and exploit the vulnerabilities of Large Vision-Language Models (LVLMs) through coordinated multi-modal adversarial attacks. Extensive experiments on multiple datasets show the effectiveness of the proposed attack method for various multi-modal tasks.

References

Abdullakutty, F.; Elyan, E.; and Johnston, P. 2021. A review of state-of-the-art in Face Presentation Attack Detection: From early development to advanced deep learning and multi-modal fusion methods. *Information fusion*, 75: 55–69.

Akhtar, N.; and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*.

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*.

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

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*.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; and Wang, Z. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *WACV*.

Dai, A.; Ma, X.; Chen, L.; Li, S.; and Wang, L. 2025. When Data Manipulation Meets Attack Goals: An In-depth Survey of Attacks for VLMs. *arXiv preprint arXiv:2502.06390*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

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 arXiv:2309.11751*.

Fang, X.; Easwaran, A.; and Genest, B. 2024. Uncertainty-Guided Appearance-Motion Association Network for Out-of-Distribution Action Detection. In *MIPR*.

Fang, X.; Easwaran, A.; and Genest, B. 2025. Adaptive Multi-prompt Contrastive Network for Few-shot Out-of-distribution Detection. In *ICML*.

Fang, X.; Easwaran, A.; Genest, B.; and Suganthan, P. N. 2025a. Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *ESWA*.

Fang, X.; and Fang, W. 2026. Disentangling Adversarial Prompts: A Semantic-Graph Defense for Robust LLM Security. In *AAAI*.

Fang, X.; Fang, W.; Ji, W.; and Chua, T.-S. 2025b. Turing Patterns for Multimedia: Reaction-Diffusion Multi-Modal Fusion for Language-Guided Video Moment Retrieval. In *ACM MM*.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024a. Not all inputs are valid: Towards open-set video moment retrieval using language. In *ACM MM*.

Fang, X.; Fang, W.; and Wang, C. 2025. Hierarchical Semantic-Augmented Navigation: Optimal Transport and Graph-Driven Reasoning for Vision-Language Navigation. In *NeurIPS*.

Fang, X.; Fang, W.; Wang, C.; Liu, D.; Tang, K.; Dong, J.; Zhou, P.; and Li, B. 2025c. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. In *AAAI*.

Fang, X.; Fang, W.; Wang, C.; Liu, D.; Tang, K.; Dong, J.; Zhou, P.; and Li, B. 2025d. Multi-Pair Temporal Sentence Grounding via Multi-Thread Knowledge Transfer Network. In *AAAI*.

Fang, X.; Fang, W.; Wang, C.; Qu, X.; and Liu, D. 2026. Rethinking Video-language Model From the Language Input Perspective. In *AAAI*.

Fang, X.; and Hu, Y. 2020. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. 2021a. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *IEEE Transactions on Artificial Intelligence*, 3(2): 192–206.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2020. V^3H: View variation and view heredity for incomplete multiview clustering. *TAI*.

Fang, X.; Hu, Y.; Zhou, P.; and Wu, D. O. 2021b. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *TETCI*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Cheng, Y.; Tang, K.; and Zou, K. 2023a. Annotations Are Not All You Need: A Cross-modal Knowledge Transfer Network for Unsupervised Temporal Sentence Grounding. In *Findings of EMNLP*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024b. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *AAAI*, volume 38, 1735–1743.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *TMM*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023b. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *CVPR*.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023c. Hierarchical local-global transformer for temporal sentence grounding. *TMM*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2024c. Rethinking weakly-supervised video temporal grounding from a game perspective. In *European Conference on Computer Vision. Springer*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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*, 6904–6913.

Guo, C.; Rana, M.; Cisse, M.; and Van Der Maaten, L. 2017. Countering adversarial images using input transformations. *arXiv*.

Huang, Y.; Guo, Q.; Juefei-Xu, F.; Hu, M.; Jia, X.; Cao, X.; Pu, G.; and Liu, Y. 2024. Texture re-scalable universal adversarial perturbation. *IEEE Transactions on Information Forensics and Security*.

Lapid, R.; and Sipper, M. 2023. I see dead people: Gray-box adversarial attack on image-to-text models. In *ECML-PKDD*. Springer.

Lee, G.; Gommers, R.; Waselewski, F.; Wohlfahrt, K.; and O'Leary, A. 2019. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, 4(36): 1237.

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*, 19730–19742. PMLR.

Li, M.; Yang, Y.; Wei, K.; Yang, X.; and Huang, H. 2022. Learning universal adversarial perturbation by adversarial example. In *AAAI*.

Liang, K.; Liu, Y.; Zhou, S.; Tu, W.; Wen, Y.; Yang, X.; Dong, X.; and Liu, X. 2023. Knowledge graph contrastive learning based on relation-symmetrical structure. *TKDE*.

Liang, K.; Meng, L.; Li, H.; Wang, J.; Lan, L.; Li, M.; Liu, X.; and Wang, H. 2025. From Concrete to Abstract: Multi-view Clustering on Relational Knowledge. *IEEE TPAMI*, 1–18.

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*. Springer.

Liu, D. 2024. Content Moderation, Platformised Speech Governance, and Legitimacy: TikTok in South and Southeast Asia. In *ACM Web Science Conference*.

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

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

Luo, H.; Gu, J.; Liu, F.; and Torr, P. 2024. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. *The Twelfth International Conference on Learning Representations*.

Ma, Z.; Jia, G.; Qi, B.; and Zhou, B. 2024a. Safe-sd: Safe and traceable stable diffusion with text prompt trigger for invisible generative watermarking. In *ACM MM*.

Ma, Z.; Zhang, Y.; Jia, G.; Zhao, L.; Ma, Y.; Ma, M.; Liu, G.; Zhang, K.; Ding, N.; Li, J.; et al. 2025. Efficient diffusion models: A comprehensive survey from principles to practices. *TPAMI*.

Ma, Z.; Zhao, L.; Qi, B.; and Zhou, B. 2024b. Neural residual diffusion models for deep scalable vision generation. *NeurIPS*.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2): 1–40.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *CVPR*.

Myrzashova, R.; Alsamhi, S. H.; Hawbani, A.; Curry, E.; Guizani, M.; and Wei, X. 2024. Safeguarding patient data-sharing: Blockchain-enabled federated learning in medical diagnostics. *IEEE Transactions on Sustainable Computing*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *ICML*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

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.

Rao, Y.; Zhao, W.; Zhu, Z.; Lu, J.; and Zhou, J. 2021. Global filter networks for image classification. *NeurIPS*, 34: 980–993.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. Drivelm: Driving with graph visual question answering. In *ECCV*. Springer.

Tang, H.; He, S.; and Qin, J. 2025. Connecting Giants: Synergistic Knowledge Transfer of Large Multimodal Models for Few-Shot Learning. *arXiv preprint arXiv:2510.11115*.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *ACM MM*.

Tang, H.; Li, Z.; Zhang, D.; He, S.; and Tang, J. 2024. Divide-and-conquer: Confluent triple-flow network for RGB-T salient object detection. *IEEE TPAMI*.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *ACM MM*.

Tang, H.; Yuan, C.; Li, Z.; and Tang, J. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130: 108792.

Tao, X.; Zhong, S.; Li, L.; Liu, Q.; and Kong, L. 2024. ImgTrojan: Jailbreaking Vision-Language Models with ONE Image. *arXiv*.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal adversarial triggers for attacking and analyzing NLP. *arXiv preprint arXiv:1908.07125*.

Wang, C.; Fang, X.; and Tiwari, P. 2025. DyPolySeg: Taylor Series-Inspired Dynamic Polynomial Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *ICML*.

Wang, C.; He, S.; Fang, X.; Han, J.; Liu, Z.; Ning, X.; Li, W.; and Tiwari, P. 2025a. Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding. In *CVPR*.

Wang, C.; He, S.; Fang, X.; Hu, Z.; Huang, J.; Shen, Y.; and Tiwari, P. 2025b. Reasoning Beyond Points: A Visual Introspective Approach for Few-Shot 3D Segmentation. In *NeurIPS*.

Wang, C.; He, S.; Fang, X.; Wu, M.; Lam, S. K.; and Tiwari, P. 2025c. Taylor Series-Inspired Local Structure Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *AAAI*.

Wang, C.; Hu, Z.; Fang, X.; Yu, Z.; Wu, Y.; Xu, M.; Wang, Y.; Gao, X.; and Tiwari, P. 2026. Biologically-Inspired Evolutionary Domain Symbiosis for Few-shot and Zero-shot Point Cloud Semantic Segmentation. In *AAAI*.

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*.

Weng, J.; Luo, Z.; Lin, D.; and Li, S. 2024. Learning transferable targeted universal adversarial perturbations by sequential meta-learning. *Computers & Security*, 137: 103584.

Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; and Yuille, A. 2017. Mitigating adversarial effects through randomization. *arXiv*.

Xing, W.; Li, M.; Li, M.; and Han, M. 2025. Towards Robust and Secure Embodied AI: A Survey on Vulnerabilities and Attacks. *arXiv preprint arXiv:2502.13175*.

Yin, Z.; Ye, M.; Zhang, T.; Du, T.; Zhu, J.; Liu, H.; Chen, J.; Wang, T.; and Ma, F. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *NeurIPS*.

Zeng, Z.; Xie, Y.; Zhang, H.; Chen, C.; Chen, B.; and Wang, Z. 2024. Meacap: Memory-augmented zero-shot image captioning. In *CVPR*.

Zhang, W. E.; Sheng, Q. Z.; Alhazmi, A.; and Li, C. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *TIST*.

Zhao, J.; Zhao, W.; Deng, B.; Wang, Z.; Zhang, F.; Zheng, W.; Cao, W.; Nan, J.; Lian, Y.; and Burke, A. F. 2024a. Autonomous driving system: A comprehensive survey. *ESWA*, 242: 122836.

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

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.