LAMP: Learning Universal Adversarial Perturbations for Multi-Image Tasks via Pre-trained Models

Alvi Md Ishmam, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, Chris Thomas

Department of Computer Science, Virginia Tech
alvi@vt.edu, najibulhaque@vt.edu, zaberhakim@vt.edu, christhomas@vt.edu

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable performance across vision-language tasks. Recent advancements allow these models to process multiple images as inputs. However, the vulnerabilities of multi-image MLLMs remain unexplored. Existing adversarial attacks focus on single-image settings and often assume a white-box threat model, which is impractical in many real-world scenarios. This paper introduces LAMP, a black-box method for learning Universal Adversarial Perturbations (UAPs) targeting multi-image MLLMs. LAMP applies an attention-based constraint that prevents the model from effectively aggregating information across images. LAMP also introduces a novel cross-image contagious constraint that forces perturbed tokens to influence clean tokens, spreading adversarial effects without requiring all inputs to be modified. Additionally, an index-attention suppression loss enables a robust position-invariant attack. Experimental results show that LAMP outperforms SOTA baselines and achieves the highest attack success rates across multiple vision-language tasks and models.

Introduction

Multimodal Large Language Models (MLLMs) like GPT-4V (Achiam et al. 2023), Gemini (Team et al. 2023), LLaVA-NeXT (Liu et al. 2024b), and Idefics (Laurençon et al. 2024) have made significant advancements in visual-language understanding and generation, particularly for single-image tasks such as VQA (Antol et al. 2015). A few open-source models, such as Mantis (Jiang et al. 2024) and VILA (Lin et al. 2024), extend these capabilities to multi-image inputs, enabling coreference, comparison, reasoning, and temporal understanding. These models first learn multi-modal interactions through pre-training on unlabeled image-text datasets. They are then fine-tuned using labeled image-text pairs for various multi-image downstream tasks (Suhr et al. 2019; Fu et al. 2023; Xiao et al. 2021). Despite their remarkable performance, the adversarial robustness of multi-image MLLMs remains unexplored.

Several studies have evaluated the robustness of single-image vision-language models. Most existing approaches (Zhou et al. 2022; Li et al. 2019) focus on white-box attacks, assuming access to gradient information from fine-tuned models. In practice, however, attackers typically only have access to public pretrained models and lack knowledge of downstream models. Gradient-based white-box methods (Madry et al. 2018) also generate instance-specific perturbations that generalize poorly and require costly re-

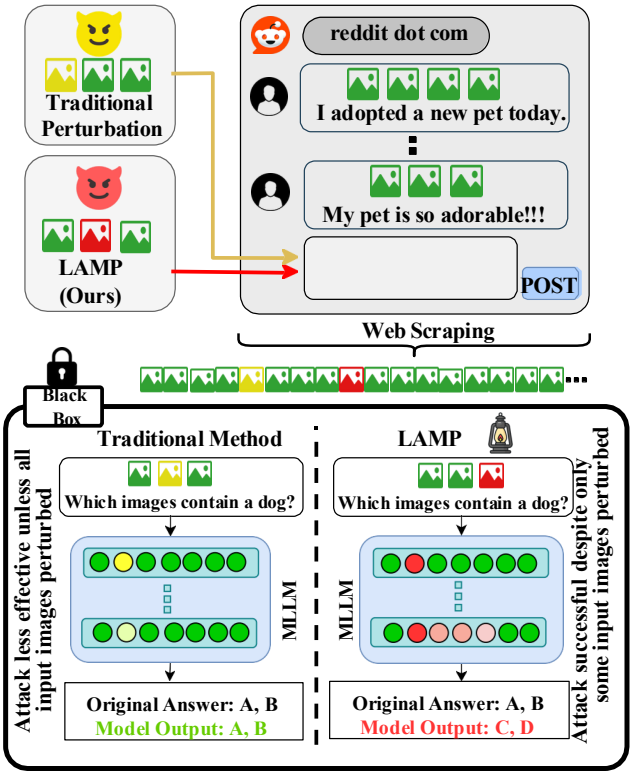


Figure 1: An overview of our approach showing superior effectiveness over traditional methods. Conventional methods fail when the perturbation is not applied to every image, an unrealistic setting when the attacker does not have access to the later inference stage. Our method succeeds even if a subset of downstream samples includes perturbed samples since it can affect "green" samples even though they are not attacked, unlike traditional methods.

optimization for new inputs.

Although some efforts, such as AdvCLIP (Zhou et al. 2023), have explored this challenge, existing approaches are often model-specific or lack practical utility due to limited imperceptibility. Additionally, generating input-specific adversarial perturbations for multi-image multimodal large language models (MLLMs) using methods like Madry et al. (2018) is impractical. A scenario is shown in Fig. 1. Consider an attacker who is posting images on social media but cannot control which sets of images, how many, or in what order they are fed into the model. At best, with current methods, an attacker can post multiple images with an attack using existing techniques. However, these attacks are not learned jointly to have a synergistic effect.

In this paper, we address that limitation by developing a new multi-image attack that works synergistically under real-world conditions, where the attacker cannot control how many attack images are presented to the model or in what order. Unlike multiple single image attacks, our attack is explicitly designed to attack multi-image scenarios. Traditional methods, designed primarily for single-image attacks, are thus less effective in multi-image contexts, unless all images in the instance are perturbed. To address this, we propose a perturbation learning framework capable of carrying out effective attacks even when a small subset of perturbed images is present within the inference instance. Our approach is designed to propagate the perturbation effect across subsequent tokens, allowing the adversarial influence to persist throughout the model's generation process, which enables a successful attack under more realistic constraints.

To address these challenges, we propose a novel method for generating Universal Adversarial Perturbations (UAPs) that targets black-box models. Our approach learns UAPs using a pretrained model and attacks various multi-image MLLMs without prior knowledge of their architectures or downstream tasks (Fig. 1). In many real-world scenarios (e.g. an attacker serving poisoned ad images displayed on a webpage processed by a model), attackers cannot poison all images or control how many images the model ingests during inference. Existing methods assume a single perturbed image, which does not exploit the unique attack surface posed in the multi-image setting. Our approach learns UAPs specifically for these settings by maximizing the dissimilarity between clean and perturbed images. It disrupts attention weights between them while keeping the pretrained MLLMs frozen. This ensures that the learned UAPs remain effective and transferable across different tasks and models. To do so, first we train UAPs by minimizing the probability of correct predictions. Next, we enforce dissimilarity constraints between the hidden states of each decoder layer in the LLM backbone. We also constrain the dissimilarity of attention weights using the Pompeiu-Hausdorff distance (Berinde and Pacurar 2013) to learn attacks which target specific heads. A fixed number of UAPs may not generalize well to varying numbers of images during inference. To address this, we introduce a novel "contagious" objective that encourages a fixed set of perturbed tokens to focus more on clean tokens in the self-attention space. This technique allows an attacker to induce noisy behavior in clean images without knowing how many perturbations each sample requires. Moreover, we propose an index-attention suppression constraint to enable position-invariant attacks tailored to multi-image settings. The contributions of our works:

- We propose the first adversarial attack targeting multi-image MLLMs to our knowledge, exploiting the unique attack surface enabled by multiple inputs. Our attack is transferable across MLLMs without requiring UAPs tailored to specific downstream models or tasks.
- We introduce a novel method for learning UAPs by leveraging the LLM's self-attention module without optimizing the MLLM itself by using Pompeiu-Hausdorff distance (Berinde and Pacurar 2013).
- We propose a novel "contagious" objective that enables perturbed visual tokens to infect clean tokens, allowing for the learning of a fixed number of UAPs in multi-image settings. We also propose an "index-attention suppression" loss, that enables the position-invariant attacks.
- We conduct a comprehensive experimental evaluation across a wide range of MLLMs, challenging multi-image benchmark datasets along with VQA and image captioning tasks. Our results demonstrate that our attack method significantly outperforms SOTA approaches.

Related Work

Multi-Modal Adversarial Attacks. With the increasing popularity of vision-language models such as CLIP (Radford et al. 2021), BLIP (Li et al. 2022b), researchers have focused on assessing their robustness by developing adversarial attack strategies. Xu et al. (2018) demonstrated that iterative pixel-level perturbations can effectively deceive visual question answering models. Expanding on this, Agrawal et al. (2018); Shah et al. (2019) introduced attacks targeting the textual modality of multimodal models. More recent approaches, such as Co-Attack (Zhang, Yi, and Sang 2022) and SGA (Lu et al. 2023), explore joint perturbations across visual and textual modalities.

Universal Adversarial Attacks. Adversarial attack research has primarily focused on instance-specific methods in both single-modal (Szegedy et al. 2013; Kim and Ghosh 2019) and multi-modal (Xu et al. 2018; Zhang, Yi, and Sang 2022) settings under white and black-box assumptions. In contrast, universal adversarial perturbations (UAPs) (Moosavi-Dezfooli et al. 2017; Mopuri, Ganeshan, and Babu 2018) offer a more practical, sample-agnostic alternative. While UAPs have been widely studied in image (Hayes and Danezis 2018; Khrulkov and Oseledets 2018) and text (Xue et al. 2024; Wallace et al. 2019) domains, their application to vision-language pretrained (VLP) and multimodal large language models (MLLMs) remains limited. UAP-VLP (Zhang, Huang, and Bai 2024) and Doubly-UAP (Kim, Kim, and Kim 2024) target image-only UAPs through subregion optimization and attention manipulation. CPGC-UAP (Fang et al. 2025) extends UAPs to both modalities using a generator, while DO-UAP (Yang et al. 2024) employs direct optimization for efficiency but is limited to single-image inputs. Jailbreak-MLLM (Schaeffer et al. 2025) improves transferability by attacking MLLM ensembles.

Multi-Image Adversarial Attacks. Existing methods take advantage of MLLM multi-image capability for composite adversarial attacks. AnyDoor (Lu et al. 2024) shows the effectiveness of UAPs when attacking randomly selected frames in a video. Multiple scenario-aware adversarial images are generated and used as a collaborative adversarial attack in MLAI (Hao et al. 2025). On the other hand, Broomfield et al. (2024) splits harmful texts into multiple typographic images to leverage multi-image capabilities of MLLMs. Wang et al. (2025) explores multi-modal in-context attacks by providing few-shot adversarial images and texts as context. However, none of these methods consider universal adversarial attack methods on subsets of interleaved images in MLLMs.

Problem Setting

Threat Model

Adversary Objective. Given a pretrained MLLM $\mathcal{M}$, our goal is to learn imperceptible UAPs that can transfer across different downstream MLLMs and tasks. In this setting, the target MLLMs, datasets, and downstream tasks remain unknown during training, and the attacker cannot control learned text embeddings. For instance, a malicious actor could serve image-based advertisements containing adversarial noise or post adversarially perturbed images in online comments. When a model processes a webpage containing multiple images, the attack could still be effective even if the adversary does not control all images or associated text. To handle this black-box scenario, we learn adversarial perturbations using a surrogate dataset and model $\mathcal{M}$. The surrogate multimodal dataset is denoted as $\mathcal{D}_s = \{(x^{(i)}, t^{(i)})\}_{i=1}^{n}$, where $\mathcal{D}_s$ consists of n multimodal samples. Here, $x^{(i)} = \{x_j^{(i)}\}_{j=1}^{m^{(i)}}$ represents the set of $m^{(i)}$ images for the i-th sample, and $t^{(i)}$ is the corresponding text prompt. The objective is to learn imperceptible universal adversarial perturbations $\delta_1, \ldots, \delta_k$ where $\|\delta_k\|_\infty \leq \epsilon$; ϵ is the perturbation budget and l_∞ denotes the perturbation constraint. Here, we attack subset of images in sample $m^{(i)}$, where $k < m^{(i)}$. We consider a multi-image setting where a user is expected to provide multiple images when querying a target MLLM. An attacker can use the learned UAPs to corrupt a subset of the user's images, misleading multimodal models into making incorrect image–prompt associations and producing erroneous responses at inference.

Adversary Capabilities. In a black-box setting, an adversary has no control over fine-tuned models and does not know how many images are used per sample during inference. Traditional perturbation generation methods require direct access to fine-tuned models and datasets, allowing adversaries to perturb specific target images – an impractical scenario for MLLMs trained on web-scale data. Our setting is more realistic, as we aim to learn only a small, fixed number of universal perturbations that effectively attack multi-image MLLMs. We assume a black-box scenario where the adversary has no knowledge of the target model's architecture or training process. Most importantly, since the number of images used during inference is unknown, our approach ensures that perturbing a fixed number of images can still generate a strong and transferable attack.

Attack Methodology

Our method proposes to learn UAPs using an accessible pretrained model, which can then be applied to black-box target models. We impose a constraint on the language model head by reducing the probability of the correct token. Additionally, we introduce a constraint to increase the divergence between the hidden states of perturbed and clean inputs in the LLM decoder. We also add a Pompeiu-Hausdorff distance (Berinde and Pacurar 2013) based constraint between the clean and perturbed attention weights. Furthermore, we encourage the model to allocate more attention to perturbed tokens from clean tokens through a novel "contagious" objective and an index-attention suppression objective (Fig. 2).

Adversarial language modeling loss

We apply adversarial perturbations to a subset of images in an input sequence of an MLLM. Let a sample contain M images and a corresponding text prompt t. The images are $x_1, x_2, \ldots, x_M$. We introduce learnable adversarial perturbations $\delta_1, \delta_2, \ldots, \delta_k$, constrained by $\|\delta_k\|_\infty \leq \epsilon$, to generate perturbed images. The perturbed image, $x'_k = x_k + \delta_k$, $k < M$. The final interleaved input sequence is: $s = (\tilde{x'_1}, \tilde{x'_2}, \tilde{x_3}, \ldots, \tilde{x_m}, \tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_n)$. Here, $\tilde{x'_1}$, $\tilde{x'_2}$ represents image tokens' of adversarial images x'_1 and x'_2 respectively. $\tilde{x_3} \ldots \tilde{x_m}$ are clean image tokens. For brevity, we skipped all tokens per image. The adversarial language modeling loss is:

$$\mathcal{L}_{adv}^{lm} = -\frac{1}{N}\sum_{i=1}^{N} \log(1 - P_\theta(t_{i+1}|s_{1:i})), \tag{1}$$

where $P_\theta(t_{i+1}|s_{1:i})$ is the predicted probability of the correct token and N number of tokens in the sequence. Through the loss, we encourage the model to reduce the likelihood of correct tokens, increasing the probability of wrong token generations while optimizing δ_k keeping the MLLM parameters frozen. Note that the summation over batch is omitted for brevity.

Adversarial hidden states loss

We introduce a loss function to learn the adversarial perturbations that maximize the cosine distance between the hidden states across decoder layers and attention heads. Let z_l^{adv} and z_l^{clean} represent mean hidden state in layer l, averaged over heads for adversarial and clean inputs, respectively. To encourage divergence between the adversarial and clean representations, our objective is:

$$\mathcal{L}_{\text{adv}}^{dec} = \frac{1}{L}\sum_{l=1}^{L} \cos(z_l^{\text{adv}}, z_l^{\text{clean}}) \tag{2}$$

where L is the total number of decoder layers, and H is the total number of attention heads per layer. By minimizing $\mathcal{L}_{\text{adv}}^{dec}$, we push the adversarial hidden states away from their clean counterparts across all layers and layer heads. The similarity is measured by the cosine similarity $cos(\cdot, \cdot)$.

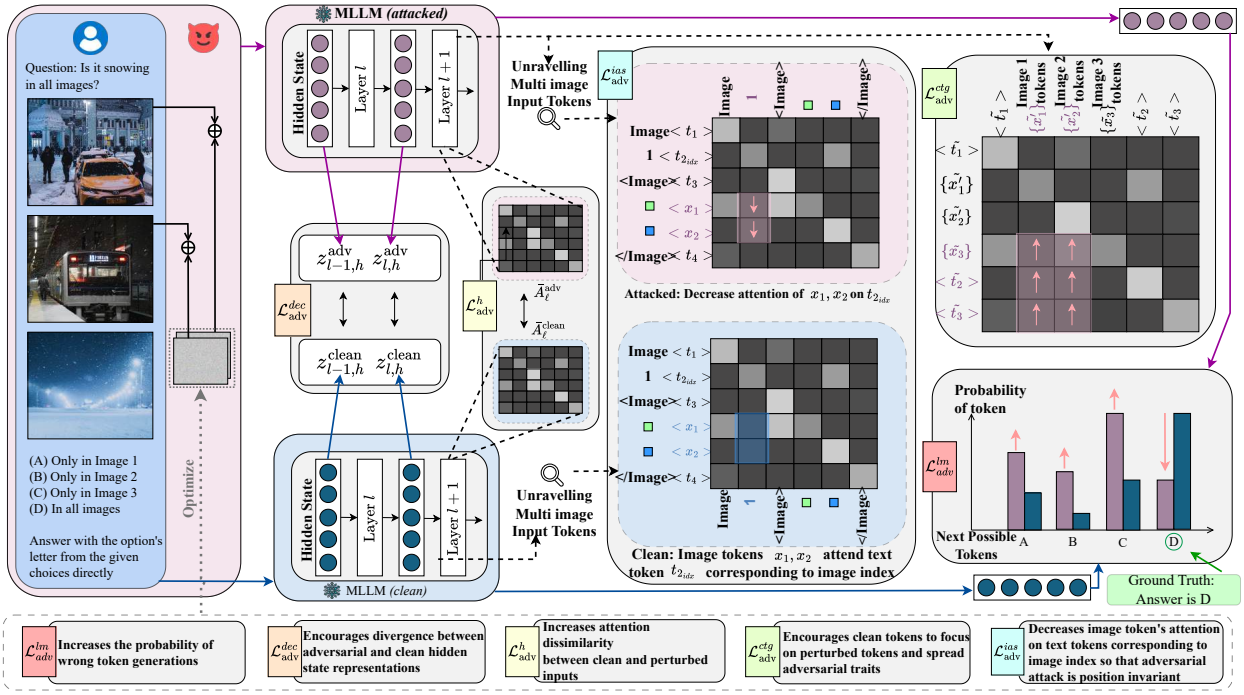


Figure 2: An overview of our proposed attack methodology. The input sample shown in the left blue-shaded box is what a normal user might query an MLLM, while the pink box shows our attack setting, where the attacker adds learned universal perturbations to two images of the input. The universal perturbations are learned using: **a) Adversarial language modeling loss $\mathcal{L}_{adv}^{lm}$:** reduces likelihood of correct tokens (Option: D), and increases probability of wrong tokens (Options: A, B, C). **b) Adversarial hidden states loss $\mathcal{L}_{adv}^{dec}$:** encourages divergence between $z_{l,h}^{\text{adv}}$ and $z_{l,h}^{\text{clean}}$, representing the hidden states of the h-th attention head in the l-th decoder layer for adversarial and clean inputs. **c) Adversarial attention weights loss $\mathcal{L}_{adv}^{h}$:** maximizes distance between $\bar{A}_{\ell}^{\text{clean}}$ and $\bar{A}_{\ell}^{\text{adv}}$, representing head-averaged attention weights in l-th decoder layer for adversarial and clean inputs. **d) Adversarial contagious loss $\mathcal{L}_{adv}^{ctg}$:** encourages clean tokens to place greater attention to noisy image tokens $\tilde{x_1'}$ and $\tilde{x_2'}$ for each $A_{:,h}^{(l)}$, attention weights for head h at layer l (Here $\{x_t\}$ represents all image tokens of image x_t for brevity). And **e) Adversarial Index-Attention Suppression loss $\mathcal{L}_{adv}^{ias}$:** suppresses attention from image tokens x_1, x_2 to text tokens corresponding to image index $t_{2_{idx}}$, to encourage image position invariance (Here the input token sequence for multi-image setting is shown as 'Image 1: <Image><image></Image>').

Adversarial attention via relaxed Pompeiu-Hausdorff distance

Attention weights indicate how tokens contribute to a model's internal representations. Adversarial perturbations modify these patterns; hence, we amplify these changes to force distinct behaviors between clean and perturbed inputs.

The Pompeiu-Hausdorff distance (Berinde and Pacurar 2013) offers a worst-case measure by quantifying the maximum deviation between two sets, i.e. the distributions of attention weights. It is defined as:

$$d_{hd}(S_1, S_2) = \max\left\{kth \sup_{s_1 \in S_1} \inf_{s_2 \in S_2} d(s_1, s_2), kth \sup_{s_2 \in S_2} \inf_{s_1 \in S_1} d(s_2, s_1)\right\} \tag{3}$$

where $d(s_1, s_2)$ is a distance metric (e.g. Euclidean or cosine) and $\sup_{s_1 \in S_1} \inf_{s_2 \in S_2}$ means selecting kth maximum value in set $D_1 = \min_{s_2 \in S_2} d(s_{1(i)}, s_2), s_{1(i)} \in S_1$ and vice versa. This relaxed formulation captures worst-case local discrepancies that may be missed by measures such as KL divergence. In practical terms, we force the model to exhibit pronounced differences in its internal focus even in regions where the global distribution might otherwise appear similar. Note this differs from a simple average sum which distributes differences across all tokens.

Let $A_{\ell,h}^{\text{clean}}$ and $A_{\ell,h}^{\text{adv}}$ denote the attention matrices at layer ℓ and head h for clean and adversarial inputs. We first average over heads:

$$\bar{A}_{\ell}^{\text{clean}} = \frac{1}{H}\sum_{h=1}^{H} A_{\ell,h}^{\text{clean}}, \quad \bar{A}_{\ell}^{\text{adv}} = \frac{1}{H}\sum_{h=1}^{H} A_{\ell,h}^{\text{adv}} \tag{4}$$

The Hausdorff distance between the averaged weights:

$$d_{hd}(\bar{A}_{\ell}^{\text{clean}}, \bar{A}_{\ell}^{\text{adv}}) = \max\Big\{\sup_{a_c \in \bar{A}_{\ell}^{\text{clean}}} \inf_{a_a \in \bar{A}_{\ell}^{\text{adv}}} d(a_c, a_a), \sup_{a_a \in \bar{A}_{\ell}^{\text{adv}}} \inf_{a_c \in \bar{A}_{\ell}^{\text{clean}}} d(a_a, a_c)\Big\} \tag{5}$$

We define the adversarial loss by averaging over all layers:

$$\mathcal{L}_{\text{adv}}^{h} = -\frac{1}{L}\sum_{\ell=1}^{L} d_{hd}(\bar{A}_{\ell}^{\text{clean}}, \bar{A}_{\ell}^{\text{adv}}) \tag{6}$$

Minimizing $\mathcal{L}_{\text{adv}}^{h}$ encourages the model to exhibit significantly different internal focus when processing clean versus adversarial inputs.

Adversarial contagious loss

Adversarial perturbations are typically constrained to specific image inputs, but their effect can propagate across the model's internal representations. We introduce a novel concept of **contagious** loss, leveraging the idea that adversarial perturbations can influence clean tokens through self attention mechanism. Specifically, in an adversarial sample, where some images are perturbed and some remain clean, we encourage the clean tokens to pay more attention to perturbed tokens to adopt adversarial characteristics. This idea helps us to learn a fixed number of adversarial perturbations without explicitly perturbing all images. For instance, in realistic scenarios at inference time an attacker has no idea of how many images are fed and how many perturbations are required. Let L be the number of layers in the model, H be the number of attention heads in each layer, $\mathcal{N}$ represent the indices of noisy tokens (image tokens), and $\mathcal{C}$ represent the indices of clean tokens. $A_{:,h,i,j}^{(l)}$ represents the attention weight at layer l, head h, which shows how much clean token i contributes to the perturbed image token j. We introduce loss to maximize attention weights to encourage clean image and text tokens to pay higher weights to noisy image tokens and indicate where the model should "attend" to.

$$\mathcal{L}_{\text{adv}}^{\text{ctg}} = -\frac{1}{LH}\sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{i\in\mathcal{C}}\sum_{j\in\mathcal{N}} A_{:,h,i,j}^{(l)} \tag{7}$$

Adversarial Index-Attention Suppression Loss

In single-image multimodal adversarial attacks, the input contains only one image placed at a fixed token position. Since its location is static, position-dependent reasoning does not influence the attack's success; the perturbation only needs to disrupt image-text alignment.

In contrast, multi-image settings involve interleaved sequences of text and images, often containing index-referencing phrases like ``image 1:", ``image 2:", etc. These index tokens provide explicit positional grounding, enabling the model to associate specific visual tokens with their corresponding references. For example, in a prompt like: "In (image 1: <Image><image></Image>) and (image 2: <Image><image></Image>), which image shows a more economically advanced place?", correct reasoning requires resolving visual content based on index markers.

In such settings, an adversarial image may only succeed when placed in a specific slot (e.g., the first image), because the model learns to associate index tokens (e.g., "1") with nearby image tokens via causal attention. Specifically, when tokens are ordered as $t_1, t_2, \ldots, t_{\text{idx}}, x_1, \ldots, x_m, \ldots, t_n$, image tokens $x_1 \ldots x_m$ may attend to their corresponding index token t_{idx}. This creates a positional vulnerability.

To make attacks robust to image reordering, we propose a *position-invariant adversarial attack* that penalizes attention from image tokens to their associated index tokens during perturbation learning. By decoupling image tokens from their position-specific textual anchors, the attack generalizes across image positions.

Let $A^{(l)} \in \mathbb{R}^{H\times T\times T}$ be the attention weights at layer $l \in \{1,\ldots,L\}$, with H heads and T tokens. Let $\mathcal{I}^{(k)} \subset \{0,\ldots,T-1\}$ denote the image token indices for image k, and $t_{\text{idx}}^{(k)}$ the corresponding index token position. The **Index-Attention Suppression Loss** is defined as:

$$\mathcal{L}_{\text{adv}}^{ias} = \frac{1}{LH}\sum_{l=1}^{L}\sum_{h=1}^{H}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}^{(k)}} A_{h,i,t_{\text{idx}}^{(k)}}^{(l)} \tag{8}$$

The final objective is the combination of Eq. 1, 2, 6, 7 and $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 > 0$. We refer to it as *LAMP*:

$$\mathcal{L}_{\text{adv}} = \lambda_1\mathcal{L}_{\text{adv}}^{\text{lm}} + \lambda_2\mathcal{L}_{\text{adv}}^{\text{dec}} + \lambda_3\mathcal{L}_{\text{adv}}^{\text{h}} + \lambda_4\mathcal{L}_{\text{adv}}^{\text{ctg}} + \lambda_5\mathcal{L}_{\text{adv}}^{\text{ias}} \tag{9}$$

Experiments

Experimental Setup

Model and Dataset. We use the pretrained Mantis-CLIP (Jiang et al. 2024) model to learn imperceptible perturbations since the Mantis family is the best performing open source model for multi-image tasks (Jiang et al. 2024). Notably, the parameters of the multimodal model's image encoder and language model are kept frozen. The maximum context length is set to 8192. We use AdamW with weight decay and a cosine scheduler, starting with a learning rate of 10^{-4} and a decay rate of 0.2. Training is conducted for 20 epochs with a batch size of 128 on $17,000$ samples from the Mantis Instruct dataset (Jiang et al. 2024). The learned perturbation has a shape of 336×336. For all experiments, the perturbation budget ϵ is uniformly set to $12/255$. All experiments were conducted on A100 GPUs.

Evaluation Benchmarks and Target Models. We experiment on five multi-image benchmark tasks in total– NLVR2 (Suhr et al. 2019), and Qbench (Wu et al. 2024) and 3 held-out benchmarks: Mantis-Eval (Jiang et al. 2024), BLINK (Fu et al. 2024), and MVBench (Li et al. 2024). We select multi-image MLLMs as our target model for querying these learned UAPs from the pretrained model. The target models are Mantis-CLIP, Mantis-SIGLIP, Mantis-Idefics2 (Jiang et al. 2024), VILA-1.5 (Lin et al. 2024), LLaVA-v1.6 (Liu et al. 2024b), Qwen-VL-Chat (Bai et al. 2023), Qwen-2.5 (Hui et al. 2025), MiniGPT4 (Zhu et al. 2024). We have also experimented on single image VQA tasks MM-Vet (Yu et al. 2024), LLaVA-Bench (Liu et al. 2023) and multi-image QA Mantis-Eval (Jiang et al. 2024). We also experimented on the selection-free VQA (OK-VQA (Marino et al. 2019)) and image captioning MSCOCO (Lin et al. 2014) tasks following (Liu et al. 2024a).

Evaluation Metrics. We utilize Attack Success Rate (ASR) as a metric to quantify the effectiveness of the proposed attack and baselines following prior research (Zhang, Huang,

Setting	Avg. Best Baseline (%)	LAMP (%)	Δ (pp)
Per Target Model			
Mantis-CLIP	51.5	71.9	+20.4
Mantis-SIGLIP	51.6	71.9	+20.3
Mantis-Idefics2	49.2	72.4	+23.2
VILA-1.5	56.1	76.2	+20.1
LLaVA-v1.6	58.5	78.9	+20.4
Qwen-VL-Chat	64.4	79.9	+15.5
Qwen-2.5	62.5	79.4	+16.9
Overall	56.3	75.8	+19.5
Per Dataset			
Mantis Eval	59.4	77.7	+18.4
NLVR2	39.4	59.7	+20.3
BLINK	66.9	85.7	+18.8
Q-Bench	52.5	76.0	+23.4
MVBench	63.1	80.0	+16.9

Table 1: Average ASR (%) and absolute improvement of **LAMP** over the strongest prior attack. The first block averages across datasets for each target model; the second block averages across target models for each dataset. Except for the shaded row, all are zero-shot cross-model evaluation. We include complete results in Appendix D. Our model outperforms all baselines significantly in all settings.

and Bai 2024; Fang et al. 2025; Lu et al. 2023). ASR is calculated as the percentage of adversarial examples that successfully deceive the model by generating incorrect outputs. The higher the ASR, the better the attack performance.

Experimental Results

LAMP outperforms by a significant margin across multi-image benchmarks and models. LAMP is compared with other baselines e.g. CPGC-UAP (Fang et al. 2025), UAP-VLP (Zhang, Huang, and Bai 2024), Doubly-UAP (Kim, Kim, and Kim 2024), Jailbreak-MLLM (Schaeffer et al. 2025). Here, the last two baselines are multimodal baselines, and the first two are encoder-decoder baselines. Additionally, we also compared our method with other transferability-based methods like (Liu et al. 2024a; Zhao et al. 2023). These methods are designed to learn universal adversarial perturbations and can be directly adapted to our problem. Other methods (Wu et al. 2020; Wei et al. 2022; Xie et al. 2019; Wang et al. 2024) either fully rely on the output from classifiers or combine feature perturbation with classification loss (Huang et al. 2019; Inkawhich et al. 2020b,a). Note that we learned UAPs based on the Mantis-CLIP pre-trained model, and the learned UAPs are applied across target MLLMs. LAMP achieves 19.5% average ASR gain across all models and datasets, as shown in Tab. 1. The specific per-model, per-dataset results are shown in Appendix D. Here, the optimal number of perturbations is $|\delta| = 2$. **LAMP outperforms single image and multi-image VQA tasks.** In Tab. 4, LAMP outperforms baselines by a significant margin on both single-image tasks such as LLaVA Bench and MM-Vet, as well as multi-image VQA tasks like Mantis Eval. We also present additional selection free VQA and image captioning task result in Appendix H

Defense	Method	ASR
Qin et al. (2021)	LAMP	70.23%
Li et al. (2022a)	LAMP	69.21%
Qin et al. (2021)	Liu et al. (2024a)	56.33%
Li et al. (2022a)	Liu et al. (2024a)	20.21%

Table 2: ASR against blackbox defense strategies on Mantis Eval dataset and Mantis-CLIP model. Detailed defense results are provided in Appendix G.

Ablation over loss components. We have evaluated the different combination of the loss function Eq. 9 in Tab. 3. If we skip $\mathcal{L}_{adv}^{ctg}$ and $\mathcal{L}_{adv}^{ias}$, the performance of LAMP drops.

Robustness to defense strategy. Following (Liu et al. 2024a), we evaluate the robustness of our attack method against defense mechanisms designed for different threat models. PatchCleanser (Xiang and Mittal 2022) is a certifiable defense against adversarial patch attacks that uses a double masking strategy to certify predictions. As our method does not depend on visible patches, PatchCleanser does not apply to imperceptible attacks like ours. Instead, we evaluate against query-based defenses, which specifically detect malicious queries in black-box settings. Tab. 2 shows that our attack remains robust in the presence defenses.

Complexity analysis and hyperparameter sensitivity. The attack complexity and hyperparameters ablations are shown in Appendix I and J.

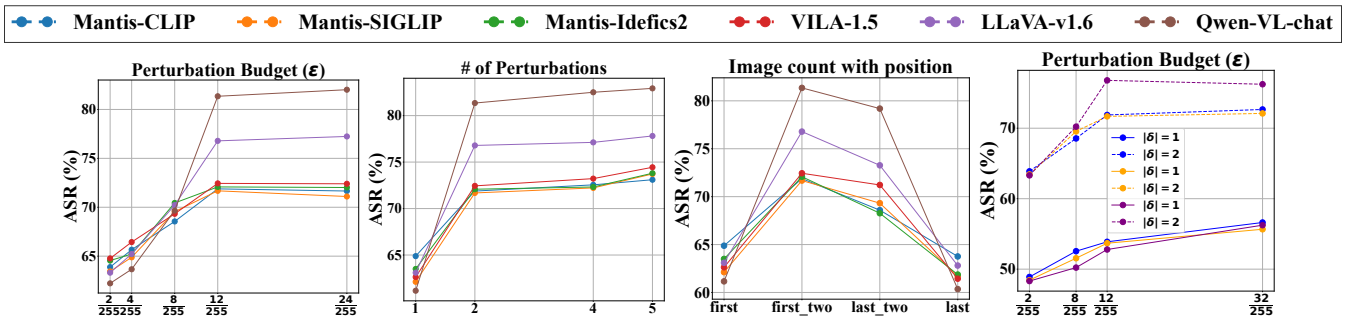
Loss					Datasets		
$\mathcal{L}_{adv}^{lm}$	$\mathcal{L}_{\text{adv}}^{dec}$	$\mathcal{L}_{\text{adv}}^{h}$	$\mathcal{L}_{adv}^{ctg}$	$\mathcal{L}_{adv}^{ias}$	Mantis Eval	NLVR2	BLINK
✓	✓	✓	✓	✓	73.43	52.73	84.86
✓	✓	✓		✓	70.32	49.33	82.64
	✓	✓	✓		67.12	48.56	78.90
✓		✓	✓		67.89	48.10	77.87
✓	✓		✓		67.32	48.34	78.90
✓	✓	✓			68.66	44.35	74.34

Table 3: ASR(%) on three benchmark tasks with different combination of loss and comparison with LAMP.

Ablations

Perturbation budget vs ASR. We compare the performance of ASR for different perturbation budget in Fig. 3a. Here, the method with **"contagious"** attack. We observe that ASR improves significantly with the increasing perturbation budget. We experiment with $\epsilon = 12/255$ for imperceptibility, and increasing this value does not significantly improve ASR, but it compromises imperceptibility.

# of perturbations vs ASR. We compare the performance of ASR for different number of perturbations for Mantis-CLIP and Mantis-Eval datasets with **"contagious"** attack Fig. 3b. We observe that ASR improves significantly when the number of universal perturbations goes from 1 to 2, but after that, it does not improve significantly. We argue that, the "conta-



(a) Perturbation budget vs ASR (b) # of perturbation vs ASR (c) Image count vs ASR (d) # Perturbation vs. budget

Figure 3: Impact of different hyperparameters in ASR.

Target Model	Method	Mantis Eval	MM-Vet	LLaVA Bench
Mantis-CLIP	CPGC-UAP	49.23	50.32	45.60
	Jailbreak-MLLM	44.45	48.21	35.67
	Doubly-UAP	46.45	50.21	37.67
	LAMP	70.32	73.45	68.31
VILA - 1.5	CPGC-UAP	49.23	48.27	43.76
	Jailbreak-MLLM	15.56	17.24	18.65
	Doubly-UAP	45.45	49.13	38.74
	LAMP	71.32	72.54	67.13
MiniGPT4	CPGC-UAP	45.23	44.75	42.64
	Jailbreak-MLLM	13.56	12.24	13.65
	Doubly-UAP	43.34	47.67	36.46
	LAMP	69.23	68.54	66.13

Table 4: Performance comparison on benchmark datasets for single and multi-image VQA tasks.

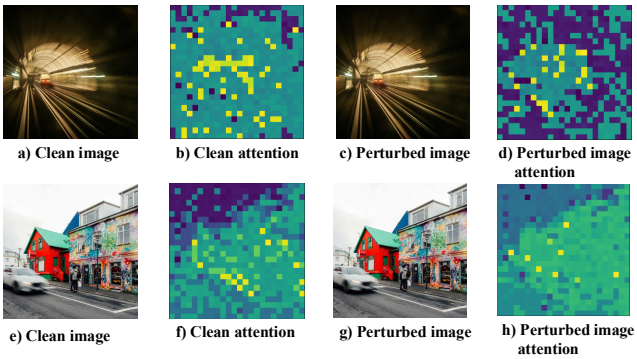
gious" attack impacts the clean images in attention spaces that help us to gain the similar ASR even number of perturbations more than 2.

Perturbation position vs ASR. We compare the performance of ASR with different number of perturbations at different position of interleaved image-text inputs in Fig. 3c. When the front two images are perturbed the ASR is the best, the ASR decreases a little as the last two images are perturbed, and least ASR for first and last image.

Comparing perturbation count vs. budget In Fig. 3d, when $|\delta| = 2$, we achieve a very high ASR compared to when $|\delta| = 1$ for all perturbation budget. We can also infer that if $|\delta| = 2$, we can maintain a very ASR in low budget maintaining the imperceptibility. **Quantification of imperceptibility.** To evaluate the imperceptibility of the adversarial perturbations, we adopt perceptual similarity metrics, including LPIPS, as detailed in Appendix K. Since lower LPIPS values correspond to more imperceptibility, our method (0.021) demonstrates significantly improved stealth compared to the best-performing baseline (0.068). **Interaction between losses.** We present interaction between losses and analysis of "contagious" losses in Appendix L.

Qualitative Analysis

We visualize the effect of our proposed method through attention maps in Fig. 4. For clean images, the downstream model pays attention to object of the images (distribution of the yellow dots). However, when the imperceptible perturbations are added to the images, the model starts to pay attention to a different location of the images. We have also shown position- invariant attack in Appendix E.



How many running white compact cars are there in all images? (A) One (B) Two (C) Three.
Answer with the option's letter from the given choices directly. GT: A, Output: B

Figure 4: Attention maps of clean and perturbed images. Here, two input images and one question. The incorrect answer is in red. Yellow indicates high attention.

Conclusion

In this paper, we investigate to learn UAPs that are capable of transferring across different target multi-image MLLMs models, datasets and downstream tasks. We propose a novel UAP learning method LAMP that incorporates different constraints exploiting the self attention module of the LLM backbone. We propose a novel "contagious" constraint that enables an attacker to learn perturbation by infecting the clean tokens through self attention. We also propose an index-attention suppression objective so that the attack remains position-invariant. We test the proposed methods across different target MLLMs, downstream tasks, and promising results demonstrate the superiority of the proposed method.

Acknowledgments

This work was supported in part by a Google Research Scholar award and Virginia Commonwealth Cyber Initiative Award #469112. We acknowledge Advanced Research Computing (ARC) at Virginia Tech for providing the computational resources and technical support that contributed to the results reported in this paper. The authors would also like to thank the reviewers for their constructive feedback.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; and Almeida. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4971–4980.

Antol, S.; Agrawal, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Berinde, V.; and Pacurar, M. 2013. The role of the Pompeiu-Hausdorff metric in fixed point theory. *Creative Mathematics and Informatics*, 22(2): 143–150.

Broomfield, J.; Ingebretsen, G.; Iranmanesh, R.; Pieri, S.; Kosak-Hine, E.; Gibbs, T.; Rabbany, R.; and Pelrine, K. 2024. Decompose, Recompose, and Conquer: Multi-modal LLMs are Vulnerable to Compositional Adversarial Attacks in Multi-Image Queries. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*.

Fang, H.; Kong, J.; Yu, W.; Chen, B.; Li, J.; Wu, H.; Xia, S.-T.; and Xu, K. 2025. One Perturbation is Enough: On Generating Universal Adversarial Perturbations against Vision-Language Pre-training Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: learning new dimensions of human visual similarity using synthetic data. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 50742–50768.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, 148–166. Springer.

Hao, S.; Hooi, B.; Liu, J.; Chang, K.-W.; Huang, Z.; and Cai, Y. 2025. Exploring Visual Vulnerabilities via Multi-Loss Adversarial Search for Jailbreaking Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19890–19899.

Hayes, J.; and Danezis, G. 2018. Learning universal adversarial perturbations with generative models. In *2018 IEEE Security and Privacy Workshops (SPW)*, 43–49. IEEE.

Huang, Q.; Katsman, I.; He, H.; Gu, Z.; Belongie, S.; and Lim, S.-N. 2019. Enhancing adversarial example transferability with an intermediate level attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4733–4742.

Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; and Lu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Inkawhich, N.; Liang, K.; Wang, B.; Inkawhich, M.; Carin, L.; and Chen, Y. 2020a. Perturbing across the feature hierarchy to improve standard and strict blackbox attack transferability. *Advances in Neural Information Processing Systems*, 33: 20791–20801.

Inkawhich, N.; Liang, K. J.; Carin, L.; and Chen, Y. 2020b. Transferable Perturbations of Deep Feature Distributions. In *International Conference on Learning Representations*.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. MANTIS: Interleaved Multi-Image Instruction Tuning. *Transactions on Machine Learning Research*.

Khrulkov, V.; and Oseledets, I. 2018. Art of singular vectors and universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8562–8570.

Kim, H.-S.; Kim, M.; and Kim, C. 2024. Doubly-Universal Adversarial Perturbations: Deceiving Vision-Language Models Across Both Images and Text with a Single Perturbation. *arXiv preprint arXiv:2412.08108*.

Kim, T.; and Ghosh, J. 2019. On single source robustness in deep fusion models. *Advances in Neural Information Processing Systems*, 32.

Laurençon, H.; Tronchon, L.; Cord, M.; and Sanh, V. 2024. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37: 87874–87907.

Li, C.; Gao, S.; Deng, C.; Xie, D.; and Liu, W. 2019. Cross-modal learning with adversarial samples. *Advances in neural information processing systems*, 32.

Li, H.; Shan, S.; Wenger, E.; Zhang, J.; Zheng, H.; and Zhao, B. Y. 2022a. Blacklight: Scalable defense for neural networks against Query-Based Black-Box attacks. In *31st USENIX Security Symposium (USENIX Security 22)*, 2117–2134.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; and Luo. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.

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 *European conference on computer vision*, 740–755. Springer.

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 *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.

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*.

Lu, D.; Wang, Z.; Wang, T.; Guan, W.; Gao, H.; and Zheng, F. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 102–111.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1765–1773.

Mopuri, K. R.; Ganeshan, A.; and Babu, R. V. 2018. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10): 2452–2465.

Qin, Z.; Fan, Y.; Zha, H.; and Wu, B. 2021. Random noise defense against query-based black-box attacks. *Advances in Neural Information Processing Systems*, 34: 7650–7663.

Radford, A.; Kim, J. W.; Hallacy, C.; and Ramesh. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Schaeffer, R.; Valentin, D.; Bailey, L.; Chua, J.; Eyzaguirre, C.; Durante, Z.; Benton, J.; Miranda, B.; and Sleight. 2025. Failures to Find Transferable Image Jailbreaks Between Vision-Language Models. In *International Conference on Learning Representations*. ICLR.

Shah, M.; Chen, X.; Rohrbach, M.; and Parikh, D. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6649–6658.

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6418–6428.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; and Soricut. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2153–2162.

Wang, J.; Chen, Z.; Jiang, K.; Yang, D.; Hong, L.; Guo, P.; Guo, H.; and Zhang, W. 2024. Boosting the transferability of adversarial attacks with global momentum initialization. *Expert Systems with Applications*, 255: 124757.

Wang, Y.; Hu, W.; Dong, Y.; Liu, J.; Zhang, H.; and Hong, R. 2025. Align is not Enough: Multimodal Universal Jailbreak Attack against Multimodal Large Language Models. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wei, Z.; Chen, J.; Goldblum, M.; Wu, Z.; Goldstein, T.; and Jiang, Y.-G. 2022. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2668–2676.

Wu, D.; Wang, Y.; Xia, S.-T.; Bailey, J.; and Ma, X. 2020. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *International Conference on Learning Representations*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; and Wang. 2024. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-level Vision. In *The Twelfth International Conference on Learning Representations*.

Xiang, S.; and Mittal, P. 2022. PatchCleanser: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX security symposium (USENIX Security 22)*, 2065–2082.

Xiao, J.; Shang, X.; Yao, A.; and Chua, T.-S. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9777–9786.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Xu, X.; Chen, X.; Liu, C.; Rohrbach, A.; Darrell, T.; and Song, D. 2018. Fooling vision and language models despite localization and attention mechanism. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4951–4961.

Xue, J.; Zheng, M.; Hua, T.; Shen, Y.; Liu, Y.; Bölöni, L.; and Lou, Q. 2024. Trojllm: A black-box trojan prompt attack on large language models. *Advances in Neural Information Processing Systems*, 36.

Yang, F.; Huang, Y.; Wang, K.; Shi, L.; Pu, G.; Liu, Y.; and Wang, H. 2024. Efficient and Effective Universal Adversarial Attack against Vision-Language Pre-training Models. *arXiv preprint arXiv:2410.11639*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, 57730–57754.

Zhang, J.; Yi, Q.; and Sang, J. 2022. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5005–5013.

Zhang, P.-F.; Huang, Z.; and Bai, G. 2024. Universal Adversarial Perturbations for Vision-Language Pre-trained Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 862–871.

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

Zhou, S.; Li, M.; Zhang, H.; Zhang, Y.; and Jin, H. 2023. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6311–6320.

Zhou, Y.; Wu, J.; Wang, H.; and He, J. 2022. Adversarial robustness through bias variance decomposition: A new perspective for federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2753–2762.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.