Say What I Want! Prompt-Agnostic Adversarial Attacks on Large Vision Language Models

Tingchao Fu, Renyang Liu, Ziyao Liu, Peiyuan Si, Fanxiao Li, Jinhong Zhang, Wei Zhou, *Member, IEEE*

***Abstract*—Large Vision Language Models (LVLMs) exhibit vulnerability to *cross-prompt attacks*: a prompt-agnostic adversarial example can easily deceive existing LVLMs. Unlike previous task-specific adversarial attacks on visual or text models, cross-prompt attacks are more challenging as they need to formulate valid adversarial images across different prompts for LVLMs. However, current cross-prompt adversarial attacks often suffer from *gradient instability*—a phenomenon where significant semantic differences in training prompts lead to unstable gradient updates during adversarial example generation. Such gradient instability causes adversarial examples to overfit training prompts, resulting in limited generalization to unseen prompts. To address these issues, we propose a novel method called *Prompt-Agnostic Attack (PAA)*, which can effectively alleviate the instability of gradient updates caused by varying prompts during the optimization process. Specifically, PAA divides training prompts into distinct mini-prompt batches and optimizes the adversarial perturbation in an inner-outer loop way. Within each inner loop, gradients from various mini-prompt batches are accumulated to serve as the gradient for the next outer loop optimization iteration. This approach enables PAA to achieve stable perturbation updates across semantically diverse training prompts, generating adversarial examples with stronger generalizations to new prompts. Extensive experiments conducted on numerous test prompts and various LVLMs demonstrate that PAA outperforms existing methods. Code and data are available at https://github.com/TingchaoFu/PAA.**

***Index Terms*—Cross-Prompt Attack, Adversarial Example, Adversarial Attack, Visual Question Answering, Large Vision Language Model, Model Vulnerability.**

I. Introduction

LARGE Vision Language Models (LVLMs) have demonstrated outstanding performance across various tasks, such as Visual Question Answering [1], Image Caption [2] and Information Governance [3]–[5]. These models allow users to input images and interact via natural language to obtain specific outputs [6]. LVLMs are more robust than previous visual or textual models due to the huge number of parameters and complex structure. However, LVLMs are still vulnerable to adversarial examples [7], [8], which can be crafted by adding imperceptible perturbations to the benign image to mislead LVLMs. Particularly in the case of cross-prompt attacks [9], an adversarial example can be combined with different clean prompts to deceive LVLMs.

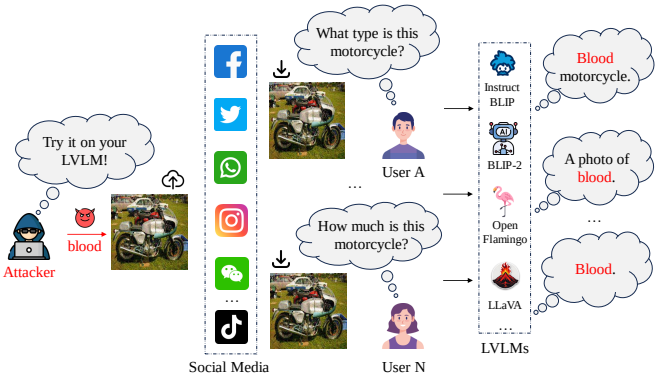


Fig. 1. Attackers add imperceptible noise to benign images, causing LVLMs to respond with malicious outputs predefined by the attacker across multiple unseen prompts.

Different from previous task-specific adversarial attacks in visual or text models [12], [13], cross-prompt attacks are characterized by the attacker's lack of knowledge regarding the specific task or prompt the victims will employ [9]. Adversarial examples generated by cross-prompt attacks must be capable of deceiving LVLMs across multiple different prompts. There are various methods for tricking users into using such adversarial examples. As shown in Fig. 1, an attacker might disseminate the adversarial example on social media, enticing users to test it by saying, "Try this image on your LVLMs!" The user then inputs the adversarial example along with their prompt, which leads the LVLM to generate the output desired by the attacker. This cross-prompt adversarial example poses a significant threat to LVLMs, enabling attackers to manipulate users' LVLMs to output arbitrary malicious commands or jailbreak, threatening the user's privacy and security [14]. Consequently, the transition from task-specific attack to cross-prompt attack inherently expands the attack surfaces while escalating the burden of defenses. In order to highlight these risks and support research into strengthening the security of LVLMs, it is vital to design effective cross-prompt attacks to explore the vulnerability of LVLMs.

This work was supported by the Yunnan Research Project (Grant Nos. 202503AG380006, 202401AT070474, 202501AU070059, and 202403AP140021), the 17th Graduate Student Research and Innovation Program of Yunnan University (Grant No. KC-252512297), the National Natural Science Foundation of China (Grant Nos. 62562061, 62502422, and 62462067), and the Yunnan Provincial Department of Education Science Research Project (Grant Nos. 2025J0006, 2024J0010, and 2025J0007). *(Corresponding authors: Renyang Liu, Wei Zhou).*

Tingchao Fu, Fanxiao Li and Jinhong Zhang are with the School of Information Science and Engineering, Yunnan University, Kunming 650500, China (e-mail: futingchao@stu.ynu.edu.cn, lifanxiao@stu.ynu.edu.cn).

Renyang Liu is with the Institute of Data Science, National University of Singapore, Singapore 117602 (e-mail: ryliu@nus.edu.sg).

Ziyao Liu is with the Digital Trust Centre, Nanyang Technological University, Singapore 639798 (e-mail: liuziyao@ntu.edu.sg).

Peiyuan Si is with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798 (e-mail: peiyuan001@e.ntu.edu.sg).

Wei Zhou is with the School of Engineering, Yunnan University, Kunming 650500, China (e-mail: zwei@ynu.edu.cn).

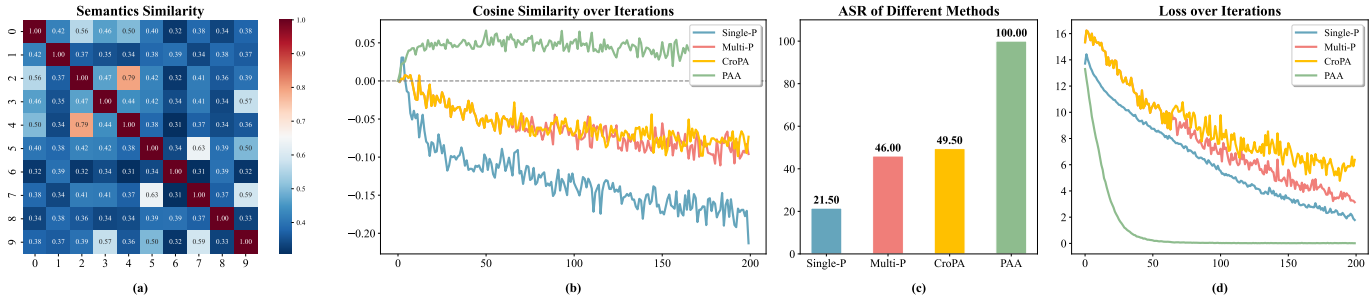


Fig. 2. Gradient stability in cross-prompt attack, we evaluated key metrics for adversarial example generation on InstructBLIP [10] using the Single-P, Multi-P, CroPA [9] and PAA. (a) shows the semantic similarity between training prompts, we sampled 10 prompts from training prompts, input these prompts into a BERT [11], and computed the cosine similarity between them to quantify the semantic differences. (b) shows the cosine similarity across these methods during generation reveals that, unlike PAA, the existing methods exhibit cosine similarities below zero, indicating gradient angles exceeding 90° and causing unstable perturbation updates. (c) demonstrates that higher cosine similarity between gradients correlates with increased attack performance. (d) shows PAA's stability in generating cross-prompt adversarial examples, achieving more effective attacks in fewer iterations.

The generalization of cross-prompt attacks on unseen prompts can be improved by increasing the number and diversity of training prompts [9]. In this work, we are trying to answer the following questions: ***Can the limited perturbation optimization space of adversarial examples effectively accommodate the increasing semantic differences introduced by a larger and more diverse set of training prompts? Moreover, does the semantic differences brought by diverse prompts enhance or hinder the cross-prompt generalization performance on unseen prompts?***

To answer the above questions, we first evaluate the generalization ability of adversarial examples generated by existing methods when trained with prompts exhibiting large semantic differences. We also investigate the stability of perturbation updates during the optimization process. Our findings reveal a key limitation in existing approaches: ***gradient instability***—a phenomenon where considerable semantic differences between various prompts lead to unstable gradient updates during the iterative generation of adversarial examples. The gradient instability problem will cause the adversarial examples to overfit the training prompts, limiting the generalization of adversarial examples on unseen prompts. As demonstrated in Fig. 2 (a), we sampled from the training prompts, encoded them using BERT [11], and found their average semantic similarity at a low level. Additionally, we calculated the cosine similarity between the gradients at each iteration to illustrate the gradient instability scenario. As shown in Fig. 2 (b), previous attack methods have experienced unstable gradient updates due to negative cosine similarity between gradients. In Fig. 2 (c) and (d), this gradient instability results in a slower reduction in loss, thereby limiting the generalization of adversarial examples on unseen prompts. An intuitive solution to this problem is to use a large batch for computation; however, large batches tend to a sharp minima [15], [16], and the large sensitivity of train loss at sharp minima negatively impacts the ability of train examples to generalize on unseen data. Additionally, large batch computations on LVLMs incur significant resource overhead due to the huge number of parameters in LVLMs.

Furthermore, to address the gradient instability caused by semantic differences across prompts, we propose the **Prompt-Agnostic Attack (PAA)**. Specifically, we construct an inner-outer loop to generate adversarial examples iteratively. The available prompts in the generated adversarial examples are divided into different mini-prompt batches. In the inner loop, we iteratively select multiple mini-prompt batches with benign images to calculate the perturbation. The gradients from these mini-prompt batches are accumulated in the outer loop as a single update step. This approach alleviates gradient instability caused by different prompts within the inner loop, and focusing more on stable perturbation updates in the outer loop enables mini-prompt batches to converge to flat minima with stronger generalization compared to large batches [15]. We evaluate the performance of the proposed PAA against different LVLMs and tasks on benchmark datasets. The experimental results demonstrate that the PAA crafts adversarial examples with remarkable attack performance compared to existing methods. Our contributions are summarized as follows:

- We extensively investigate current cross-prompt attacks that suffer from gradient instability problems: significant semantic differences between prompts lead to unstable gradient updates during the iterative generation of adversarial examples.
- We propose the Prompt-Agnostic Attack (PAA), which stabilizes the gradient updates during adversarial example generation through an inner-outer loop mechanism, thereby avoiding overfitting to the training prompts and enhancing generalization to unseen prompts.
- We validated the effectiveness of PAA on a broader set of test prompts. Experimental results show that PAA exhibit a remarkable attack performance compared with existing cross-prompt attacks.

II. Related Work

A. Large Vision Language Model

In the context of the success achieved by Large Language Models (LLMs), the past year has witnessed the emergence of LVLMs. These models extend the capabilities of LLMs to process and analyze both visual and textual information concurrently. Li *et al.* [17] propose BLIP-2. a generic and efficient pretraining strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen LLMs. A Q-Former was introduced to

align images and text in BLIP-2. BLIP-2 achieves remarkable performance on various vision-language tasks despite having significantly fewer trainable parameters. Dai *et al.* [10] explore the vision-language instruction tuning on LVLMs and propose InstructBLIP, which introduces an instruction-aware Query Transformer to extract informative features tailored to the given instruction. Awadalla *et al.* [18] provide an open-source replication of DeepMind's Flamingo models, they augment the layers of pretrained, frozen language models so that they cross attend to the outputs of a frozen vision encoder while predicting the next token. Wan *et al.* [19] propose GRID, a framework that reformulates temporal visual generation tasks as grid layouts, enabling efficient processing of visual sequences using existing image generation models. GRID significantly improves computational efficiency while maintaining strong performance across diverse tasks. Peng *et al.* propose CIA [20], which introduces an instance encoder using cross-modal self-attention to generate instance-specific features, leading to improved performance across various benchmark datasets. Liu *et al.* [21] present the first attempt to use language-only GPT-4 [6] to generate multimodal language-image instruction-following data and use these generated data to train LVLMs called LLaVA, an end-to-end trained large multimodal model that connects a vision encoder and an LLM for general purpose visual and language understanding. Despite demonstrating superior performance across various domains, these LVLMs still exhibit vulnerabilities when faced with adversarial examples [22]–[24].

B. Adversarial Attack

Research on adversarial attacks has surged since Szegedy *et al.* [7] revealed the vulnerability of deep neural networks to adversarial examples. Most previous vision or text models have been designed for specific tasks, such as image classification, object detection, semantic segmentation, and sentiment classification. In these cases, the attacker only needs to define the relevant target loss for the specific task to generate adversarial examples [25]–[32]. For example, Fast Gradient Sign Method (FGSM) [33] performs a one-step update in the direction of the sign of the gradient to generate the adversarial examples. Projected Gradient Descent (PGD) [34] utilizes random start and updates adversarial images in small steps iteratively. Dong *et al.* [35] integrating the momentum term into the iterative process for the attack to achieve a higher attack success rate. Lin *et al.* [36] regarded the process of generating adversarial examples to model training and proposed the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM). NI-FGSM adapts Nesterov accelerated gradient into the iterative attacks so as to effectively look ahead. Based on the scale-invariant property of the deep learning models, SIM optimizes the adversarial examples over the scale copies of the input.

Wang *et al.* proposed Block Shuffle and Rotation (BSR) [37], which enhances transferability by disrupting the attention heatmaps through random block shuffling and rotation. Zhu *et al.* proposed the Gradient Relevance Attack (GRA) to enhance the transferability of adversarial examples by utilizing gradient relevance frameworks and a decay indicator to counter fluctuations in adversarial perturbations [38]. Ren *et al.* introduced the Efficient Polar Coordinates Attack (EPCA) [39], which enhances query efficiency in decision-based attacks by addressing the update contradiction and reallocating queries, and further improves performance with an Adaptive Activation Strategy to prevent local optima.

In addition, there has been some exploration into adversarial attacks on multimodal models. Zhang *et al.* investigated the vulnerability of Vision-Language Pre-training (VLP) models by combining PGD [34] and BERT-Attack [40] into Co-Attack, marking the first exploration of VLP model robustness [41]. Lu *et al.* examined the correspondence between image and text embeddings and proposed the SGA for VLP models [42]. Wang *et al.* leveraged contrastive learning to improve the transferability of adversarial samples in VLP models [43]. In terms of adversarial robustness improvement, Wang *et al.* [44] introduce MAVA to enhance the robustness of VLP models against multimodal adversarial attacks. By incorporating cross-modal supervision, MAVA outperforms previous fine-tuning methods in defending against both image and multimodal adversarial threats. However, VLP models are often only used as visual encoders in LVLMs, and merely investigating the adversarial robustness of VLP models is insufficient to reveal the vulnerabilities of LVLMs. Wang *et al.* [45] propose a novel multimodal universal jailbreak attack that combines adversarial suffixes and images to exploit vulnerabilities in LVLMs. They enhance the effectiveness of bypassing MLLM safety measures by leveraging the intricate interactions between image and text modalities.

These attack methods demonstrate superior results on specific tasks, however, when applied to other tasks and models, the adversarial examples generated by these methods yield unsatisfactory results compared to the original tasks.

C. Cross-Prompt Attack

Wan *et al.* first explore prompt-agnostic attacks on text-to-image generation models and introduce the Prompt-Agnostic Adversarial Perturbation (PAP) method, which generates robust adversarial perturbations for customized text-to-image diffusion models without relying on specific prompts [46]. PAP models the prompt distribution using Laplace approximation and maximizes the disturbance expectation through Monte Carlo sampling. Compared to previous task-specific attacks in visual or text models, there has been limited effort in cross-prompt attacks on LVLMs. Different from earlier visual models, LVLMs have a larger number of parameters and a more complex model structure, which also contributes to their enhanced robustness [47]–[49]. Luo *et al.* [9] were the first to explore cross-prompt attacks on LVLMs, revealing their vulnerability to such adversarial examples. They proposed three methods: Single-P, Multi-P, and CroPA. Single-P uses a single prompt as input for generating adversarial examples, while Multi-P employs multiple prompts. CroPA adopts a min-max approach by maximizing the target loss for the prompt embedding while minimizing it for the adversarial examples. Luo *et al.* demonstrated that using multiple prompt inputs

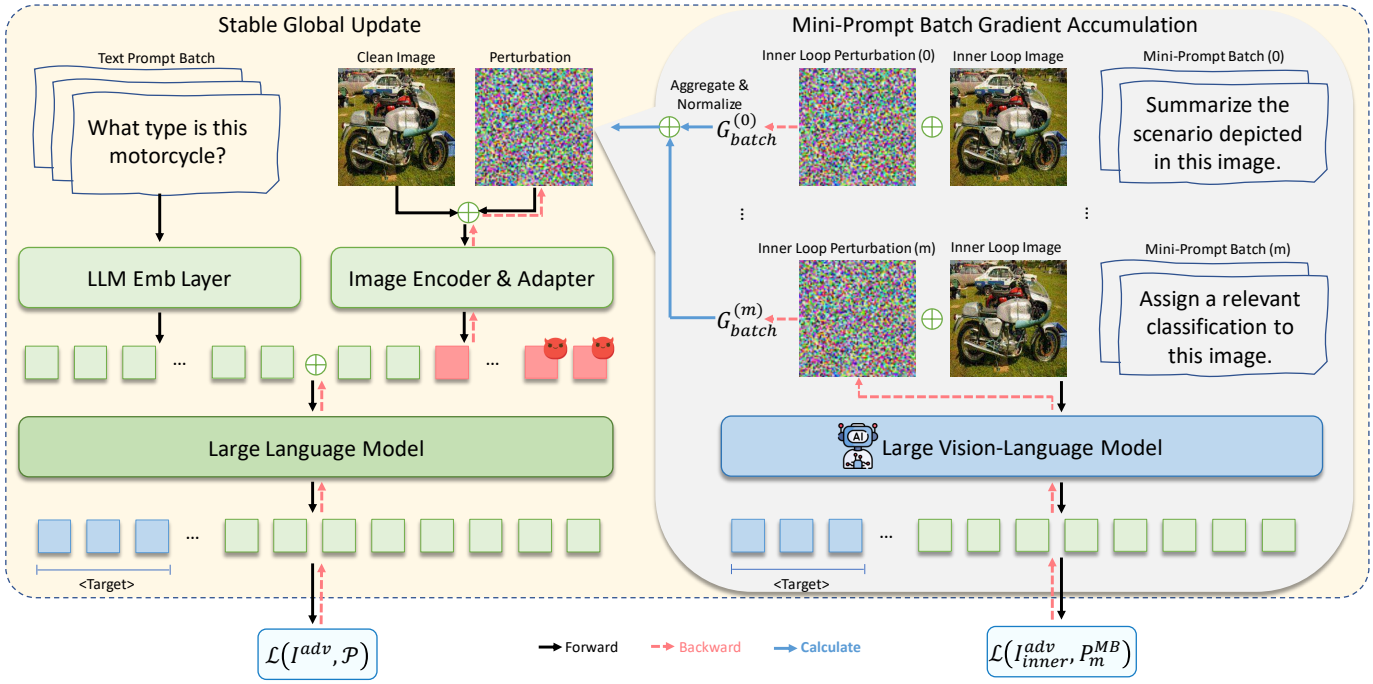


Fig. 3. The overall framework of PAA consists of an inner loop and an outer loop. In the inner loop, different mini-prompt batches and the input image are fed into the LVLM to compute perturbations and gradients. The gradient instability caused by semantic differences is mitigated within this loop. In the outer loop, which performs stable global updates, PAA accumulates the inner-loop gradients to compute perturbations, enabling stable optimization under large semantic variations and generating adversarial examples with stronger generalization to unseen prompts.

during the generation of adversarial examples can generate more aggressive adversarial examples. However, semantic differences between prompts result in gradient instability, causing the perturbation update to become unstable and more likely to converge to local optima, thereby limiting the attack performance of the adversarial examples. Therefore, our study focuses on addressing gradient instability during the optimization of adversarial examples, enabling improved generalization on unseen prompts.

III. Preliminary

A. Large Vision Language Model

An LVLM model typically consists of a visual encoder, an adapter for modality alignment, and a LLM for generation. Let $f_\theta(\cdot)$ represent a LVLM model with input $X = [I, P]$, where I is the image and P are the prompts. The visual encoder is used to initially extract image embeddings, which are then aligned with text embeddings through an adapter. The aligned image embeddings are concatenated with the text embeddings and fed into the LLM to generate textual outputs. The output of the autoregressive LVLM is formulated as follows:

$$p(X_{i+1:i+H}|X_{0:i}) = \prod_{j=1}^{H} p(X_{i+j}|X_{0:i+j-1}), \tag{1}$$

where p is the probability distribution of the next H tokens given the input $X_{0:i}$. An autoregressive LLM predicts the subsequent H tokens based on the joint probability distribution of the preceding context $X_{0:i}$, which consists of both the image I and the prompt $\mathcal{P}$.

B. Adversarial Example Generation

The objective of the adversarial attack is to identify a carefully crafted and imperceptible perturbation δ, which added to a benign example X^{clean} to form an adversarial example $X^{adv} = X^{clean} + \delta$, can cause the target model $f_\theta(\cdot)$ with parameters θ to make an incorrect prediction. The perturbation δ is optimized by feeding the adversarial example into the target model $f_\theta(\cdot)$ to obtain a malicious output $f_\theta(X^{adv})$, and then updating the perturbation via backpropagation to push the output away from the ground truth $\mathcal{Y}$. This process can be formulated as follows:

$$\arg\max_{X^{adv}} \mathcal{L}(f_\theta(X^{adv}), \mathcal{Y}), s.t. \left\|X^{adv} - X^{clean}\right\|_\infty \leq \epsilon, \tag{2}$$

where loss function $\mathcal{L}$ measures the distance between the model's malicious out $f_\theta(X^{adv})$ and the ground truth $\mathcal{Y}$, $\|\cdot\|_\infty$ denotes the constraint L_∞ norm which is used to constrain the perturbation δ to ensure its imperceptibility. The above process is known as the untargeted attack, which lacks a predefined target and thus leads to uncontrollable malicious outputs from the model. Instead, a target $\mathcal{Y}^{target}$ can be specified during the optimization of the adversarial example, guiding the adversarial input X^{adv} to produce a specific output $\mathcal{Y}^{target}$ when fed into the model $f_\theta(\cdot)$. This type of attack is known as the targeted attack and can be formulated as follows:

$$\arg\min_{X^{adv}} \mathcal{L}(f_\theta(X^{adv}), \mathcal{Y}^{target}), s.t. \left\|X^{adv} - X^{clean}\right\|_\infty \leq \epsilon. \tag{3}$$

The main difference between untargeted and targeted attacks lies in whether a specific target output is defined for the model. While untargeted attacks optimize the perturbation to push the adversarial output $f_\theta(X^{adv})$ away from the ground

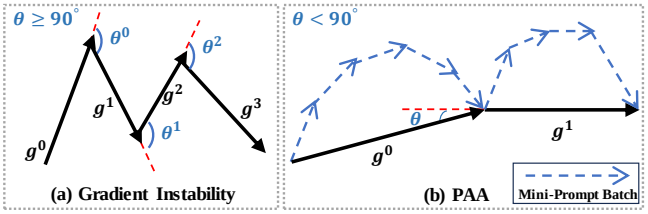


Fig. 4. (a) illustrates gradient instability in cross-prompt attacks, which cause instability in perturbation updates. (b) shows that PAA achieves stable perturbation updates by effectively accumulating gradients.

truth, targeted attacks are more challenging, as they require the perturbation to be optimized toward producing a predefined target output from the model. The cross-prompt attack proposed falls under the category of targeted attacks. It is particularly challenging, as it requires the adversarial example to consistently induce the model to output a predefined target $\mathcal{Y}^{target}$ under different unseen prompts.

C. Problem Definition

The goal of a cross-prompt adversarial attack is to input an $X^{adv} = [I^{adv}, \mathcal{P}]$ consisting of an adversarial example I^{adv} and N clean prompts list $\mathcal{P} = [P_0, P_1, \cdots, P_N]$, manipulate the model to generate the target output $\mathcal{Y}^{target}$:

$$f_\theta(I^{adv}, P_0) = \cdots = f_\theta(I^{adv}, P_N) = p(\mathcal{Y}^{target}|I^{adv}, \mathcal{P}). \tag{4}$$

To ensure that LVLM outputs the intended target, the negative log-likelihood probability of the LVLM generating the target must be minimized, where the loss function is:

$$\mathcal{L}(I^{adv}, \mathcal{P}) = -\log p(\mathcal{Y}^{target}|I^{adv}, \mathcal{P}), \tag{5}$$

therefore, the objective of generating cross-prompt adversarial examples I^{adv} can be formulated as follows:

$$\arg\min_{I^{adv}} \mathcal{L}(I^{adv}, \mathcal{P}), \text{s.t.} \left\|I^{adv} - I\right\|_\infty \leq \epsilon, \tag{6}$$

where $\|\cdot\|_\infty$ denotes the constraint L_∞ norm such that the adversarial perturbations are constrained to be within ϵ.

IV. Methodology

A. Gradient Instability in Cross-Prompt Attack

In this section, we provide a detailed analysis of the gradient instability caused by significant semantic differences. The generation of adversarial examples can be viewed as a model training process [36]. Let the fixed model parameters be θ, and let the training prompt set be:

$$\mathcal{P}_{train} = \{P_0 \cdots P_N\}, \tag{7}$$

with perturbation δ constrained by $\|\delta\|_\infty$ the perturbation is treated as a "parameter" to be optimized, while prompts serve as training data. We optimize δ by minimizing the empirical loss:

$$J_{\text{train}}(\delta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(f_\theta(I^{clean} \oplus \delta, P_i), y_i), \tag{8}$$

where $\mathcal{L}(\delta)$ is the task loss (e.g. negative log-likelihood) and $\oplus$ denotes adding δ to the benign image. Similar to model training, unstable gradient updates during this process may cause the perturbation to fall into local optima. Concretely, using projected gradient descent (PGD) [34]:

$$\delta^{t+1} = \prod_{i=1}^{N}(\delta^t - \alpha\nabla_\delta\mathcal{L}(\delta^t)), \tag{9}$$

the variance of the per-prompt gradient:

$$\text{Var}_i[\nabla_\delta\mathcal{L}(f_\theta(I^{clean} \oplus \delta, P_i), Y^{target})] \gg 0, \tag{10}$$

captures the gradient instability introduced by semantic diversity across prompts.

Achieving strong zero-shot capability in model training typically requires exposure to a diverse set of samples. Likewise, to enhance the generalization of adversarial examples to unseen prompts, it is necessary to optimize δ over as many prompts as possible. However, compared to the model's parameters, the perturbation has a significantly smaller parameter space:

$$d_\delta \ll d_\theta, \tag{11}$$

so the limited-capacity δ struggles to adapt when N (and hence the semantic diversity) grows, and thus tends to overfit the training prompts. As a result, the risk on an unseen prompt distribution $\mathcal{P}_{test}$:

$$J_{\text{test}}(\delta) = \mathbb{E}_{P\sim\mathcal{P}_{\text{test}}}\left[\mathcal{L}(f_\theta(I^{clean} \oplus \delta, P), y)\right], \tag{12}$$

obeys a standard generalization bound of the form:

$$J_{\text{test}}(\delta) \leq J_{\text{train}}(\delta) + \mathcal{O}\left(\sqrt{\frac{\mathcal{C}(d_\delta)}{N}}\right), \tag{13}$$

where $\mathcal{C}(d_\delta)$ measures the complexity of the perturbation space. This bound is a standard result from statistical learning theory, which connects the test error to the training error, where the complexity term $\mathcal{C}(d_\delta)$ depends on the complexity of the perturbation space δ (e.g. Rademacher complexity [50]) and N is the number of training samples. As shown in Fig. 4 (a), the angle between the gradient at each step and the previous step exceeds $90°$, leading to unstable gradient directions. This instability hinders the optimization of adversarial examples in the correct direction.

B. Prompt-Agnostic Attack (PAA)

1) Overview: The core idea of Prompt-Agnostic Attack is to stabilize gradient updates across semantically diverse prompts by organizing training prompts into multiple mini-prompt batches and using a two-level optimization loop. As shown in Fig. 3, at each outer iteration, PAA first partitions the full prompt set into M mini-prompt batches. Within each mini-batch (inner loop), an inner perturbation is iteratively refined against only those prompts, producing a batch-specific gradient that has been normalized to prevent any single prompt from dominating. The detailed design of this inner-loop mechanism is described in Sec. IV-B2. Once all M mini-batches have been processed, their normalized gradients are summed (outer loop)

Algorithm 1 Prompt-Agnostic Attack

Input: benign image I^{clean}, prompt set $\mathcal{P}_{train}$, LVLM model $f_\theta(\cdot)$, perturbation budget ε, target text $\mathcal{Y}^{target}$, outer iterations T, number of mini-batches M, inner steps per batch K, step size α:

Output: adversarial example I^{adv}

Initialize $I^{\text{adv}} \leftarrow I$
for $t = 1$ to T **do**
  $g_{\text{outer}} \leftarrow 0$
  for $m = 1$ to M **do**
    Sample mini-batch $P_m^{MB} \subset \mathcal{P}_{train}$
    $I_{\text{inner}}^{\text{adv}} \leftarrow I^{\text{adv}}$
    $G_{\text{batch}} \leftarrow 0$
    for $k = 1$ to K **do**
      Compute normalized inner gradient:
      $g_{\text{inner}}^{(m,k)} = \frac{\left|\nabla_{\delta_{\text{inner}}} \mathcal{L}\left(f_\theta(I^{\text{clean}} \oplus \delta_{\text{inner}}^{(m,k-1)}, P), \mathcal{Y}^{\text{target}}\right)\right|}{\left\|\nabla_{\delta_{\text{inner}}} \mathcal{L}\left(f_\theta(I^{\text{clean}} \oplus \delta_{\text{inner}}^{(m,k-1)}, P), \mathcal{Y}^{\text{target}}\right)\right\|}$
      Update inner adversarial example:
      $\delta_{\text{inner}}^{(m,k)} = \Pi_{\|\delta\| \le \varepsilon}\left(\delta_{inner}^{(m,k-1)} - \alpha \cdot sign(\frac{g_{\text{inner}}^{(m,k)}}{\| g_{\text{inner}}^{(m,k)} \|})\right)$
      Accumulate batch gradient:
      $G_{\text{batch}} \leftarrow G_{\text{batch}} + g_{\text{inner}}^{(m,k)}$
    end for
    $g_{\text{outer}} = \frac{1}{M} \sum_{m=1}^{M} G_{\text{batch}}^{(m)}$.
  end for
  Update outer adversarial example:
  $\delta \leftarrow \Pi_{\|\delta\| \le \varepsilon}\Big(\delta - \alpha \cdot sign(g_{\text{outer}})\Big)$
  $I^{\text{adv}} \leftarrow I^{\text{clean}} + \delta$
end for

to form a single global update direction. By updating the adversarial perturbation using this accumulated gradient, instead of averaging across all prompts at once, PAA mitigates the gradient instability caused by large semantic gaps and steers the perturbation toward a "flatter" optimum that transfers more reliably to unseen prompts. We elaborate on this outer-loop aggregation and its role in mitigating cross-prompt gradient conflict in Sec. IV-B3. In essence, the inner loops ensure each small group of prompts contributes a balanced gradient, while the outer update aggregates these contributions into a stable step that maximizes cross-prompt attack success.

2) Inner-Loop: Mini-Prompt Batch Gradient Accumulation: In order to address the gradient instability in the previous cross-prompt attack, PAA first divides the training prompt set $\mathcal{P}_{\text{train}}$ into random mini-batches P_m^{MB} of size B. For each inner iteration $k = 1, \ldots, K$ on batch m, PAA computes a normalized gradient with a regularization term to prevent domination by "easy" prompts:

$$g_{\text{inner}}^{(m,k)} = \frac{\left|\nabla_{\delta_{\text{inner}}} \mathcal{L}\big(f_\theta(I^{\text{clean}} \oplus \delta_{\text{inner}}^{(m,k-1)}, P), \mathcal{Y}^{\text{target}}\big)\right|}{\left\|\nabla_{\delta_{\text{inner}}} \mathcal{L}\big(f_\theta(I^{\text{clean}} \oplus \delta_{\text{inner}}^{(m,k-1)}, P), \mathcal{Y}^{\text{target}}\big)\right\|}, \tag{14}$$

$$\delta_{\text{inner}}^{(m,k)} = \Pi_{\|\delta\| \le \varepsilon}\Big(\delta_{inner}^{(m,k-1)} - \alpha \cdot sign(\frac{g_{\text{inner}}^{(m,k)}}{\| g_{\text{inner}}^{(m,k)} \|})\Big), \tag{15}$$

where α is the inner-loop step size, $P \in P_m^{\text{MB}}$, and Π is the projection onto the ℓ_∞ ball. After K inner steps, PAA accumulates:

$$G_{\text{batch}}^{(m)} = \sum_{k=1}^{K} g_{\text{inner}}^{(m,k)}. \tag{16}$$

3) Outer-Loop: Stable Global Update: Across M mini-prompt batches, the full accumulated gradient is:

$$g_{\text{outer}} = \frac{1}{M} \sum_{m=1}^{M} G_{\text{batch}}^{(m)}. \tag{17}$$

We then perform a single step update:

$$\delta \leftarrow \Pi_{\|\delta\| \le \varepsilon}\Big(\delta - \alpha \cdot sign(g_{\text{outer}})\Big), \tag{18}$$

By normalizing g_{outer}, we ensure balanced progress even if some batches produce larger magnitudes. As shown in Fig. 4 (b), PAA mitigates this by introducing an inner-loop gradient accumulation over mini-prompt batches, followed by a more stable outer-loop update. The details of our proposed method are outlined in Alg. 1.

C. Generalization Analysis of Prompt-Agnostic Attack

To further understand the generalization capability of our PAA framework, we analyze how mini-prompt batch optimization contributes to lower test loss on unseen prompts. Recall the generalization bound from Eq. 13.

In the cross-prompt setting, N is relatively small due to the high cost of backpropagation through large-scale LVLMs, and the perturbation dimension d_δ is significantly lower than the model parameters d_θ. This creates a risk of overfitting to seen prompts. To mitigate this, we employ a variance-reduced optimization by aggregating gradients over diverse mini-prompt batches.

Under standard smoothness and bounded-variance assumptions (variance bounded by σ^2), we can derive the convergence behavior of PAA as:

$$\mathbb{E}[\|\nabla J_{\text{train}}(\delta)\|^2] \le \mathcal{O}\left(\frac{1}{T} + \frac{\sigma^2}{MK}\right), \tag{19}$$

where T is the number of outer iterations, M is the number of mini-batches, and K is the number of inner steps. This shows that increasing the number of inner gradient samples MK reduces the stochastic variance, stabilizing the outer updates and avoiding convergence to sharp minima.

More importantly, the accumulation over semantically diverse prompts (i.e., batches with low inter-prompt cosine similarity) helps align gradients across the prompt space. Let θ_m be the angle between gradient directions from two mini-prompt batches, then the following condition promotes gradient alignment:

$$\cos(\theta_m) = \frac{(g_{outer}^i, g_{outer}^j)}{\|g_{outer}^i\| \cdot \|g_{outer}^j\|} > 0, \tag{20}$$

which empirically correlates with attack transferability, as shown in Fig. 4 (b). Therefore, PAA can be interpreted as implicitly minimizing a prompt-alignment regularizer:

TABLE I
ATTACK PERFORMANCE ON INSTRUCTBLIP (INB), BLIP-2 (B2), OPENFLAMINGO (OF) AND LLAVA (LA). COMPARISON WITH EXISTING CROSS-PROMPT ATTACKS ON DIFFERENT LVLMS. A HIGHER ASR INDICATES BETTER CROSS-PROMPT ATTACK PERFORMANCE.

Category	Target	Method	VQA-Agnostic				VQA-Specific				Classification				Caption				Overall			
			INB	B2	OF	LA	INB	B2	OF	LA	INB	B2	OF	LA	INB	B2	OF	LA	INB	B2	OF	LA
Object	dog	Single-P	29.00	31.00	2.00	16.00	19.00	58.00	4.00	12.00	66.00	94.00	21.00	17.00	86.00	95.00	43.00	25.00	50.00	69.50	17.50	17.50
		Multi-P	61.00	54.00	2.00	21.00	45.00	70.00	5.00	16.00	96.00	**100.00**	37.00	25.00	99.00	**100.00**	55.00	29.00	75.25	81.00	24.75	22.75
		CroPA	62.00	54.00	4.00	18.00	45.00	70.00	7.00	11.00	96.00	99.00	44.00	22.00	99.00	99.00	63.00	28.00	75.50	80.50	29.50	19.75
		PAA	**99.00**	**100.00**	**94.00**	**29.00**	**99.00**	**100.00**	**95.00**	**27.00**	**100.00**	**100.00**	**97.00**	**34.00**	**100.00**	**100.00**	**97.00**	**36.00**	**99.50**	**100.00**	**95.75**	**31.50**
	cat	Single-P	24.00	49.00	1.00	20.00	15.00	77.00	3.00	12.00	63.00	97.00	24.00	45.00	76.00	98.00	27.00	40.00	44.50	80.25	13.75	29.25
		Multi-P	55.00	80.00	3.00	19.00	37.00	87.00	4.00	10.00	98.00	**100.00**	35.00	45.00	99.00	**100.00**	40.00	35.00	72.25	91.75	20.50	27.25
		CroPA	54.00	81.00	5.00	20.00	38.00	88.00	6.00	14.00	97.00	**100.00**	47.00	45.00	97.00	**100.00**	55.00	35.00	71.50	92.25	28.25	28.50
		PAA	**99.00**	**100.00**	**96.00**	**27.00**	**97.00**	**100.00**	**93.00**	**19.00**	**100.00**	**100.00**	**94.00**	**53.00**	**100.00**	**100.00**	**93.00**	**41.00**	**99.00**	**100.00**	**94.00**	**35.00**
Jailbreak	blood	Single-P	13.00	35.00	0.00	6.00	9.00	59.00	1.00	7.00	45.00	87.00	13.00	8.00	53.00	87.00	8.00	11.00	30.00	67.00	5.50	8.00
		Multi-P	56.00	79.00	1.00	7.00	27.00	88.00	0.00	7.00	87.00	99.00	21.00	8.00	92.00	**100.00**	14.00	8.00	65.50	91.50	9.00	7.50
		CroPA	59.00	76.00	2.00	9.00	34.00	91.00	3.00	8.00	92.00	99.00	26.00	12.00	97.00	**100.00**	24.00	11.00	70.50	91.50	13.75	10.00
		PAA	**99.00**	**100.00**	**99.00**	**21.00**	**94.00**	**100.00**	**99.00**	**22.00**	**100.00**	**100.00**	**99.00**	**34.00**	**100.00**	**100.00**	**99.00**	**31.00**	**98.25**	**100.00**	**99.00**	**27.00**
	drug	Single-P	17.00	25.00	0.00	2.00	11.00	52.00	0.00	3.00	54.00	87.00	1.00	3.00	53.00	88.00	11.00	3.00	33.75	63.00	3.00	2.75
		Multi-P	58.00	51.00	1.00	2.00	44.00	67.00	1.00	1.00	95.00	95.00	19.00	4.00	96.00	96.00	16.00	2.00	73.25	77.25	9.25	2.25
		CroPA	57.00	55.00	1.00	1.00	39.00	67.00	1.00	1.00	95.00	94.00	23.00	4.00	95.00	94.00	21.00	1.00	71.50	77.50	11.50	1.75
		PAA	**99.00**	**100.00**	**93.00**	**8.00**	**97.00**	**100.00**	**93.00**	**9.00**	**100.00**	**100.00**	**94.00**	**9.00**	**100.00**	**100.00**	**86.00**	**8.00**	**99.00**	**100.00**	**91.50**	**8.50**
Emotion	hello	Single-P	8.00	26.00	1.00	8.00	10.00	37.00	2.00	7.00	32.00	61.00	5.00	8.00	35.00	60.00	15.00	10.00	21.25	46.00	5.75	8.25
		Multi-P	39.00	48.00	0.00	4.00	21.00	55.00	2.00	4.00	75.00	81.00	13.00	8.00	84.00	87.00	23.00	5.00	54.75	67.75	9.50	5.25
		CroPA	34.00	52.00	1.00	6.00	18.00	59.00	3.00	6.00	70.00	85.00	11.00	9.00	78.00	90.00	23.00	8.00	50.00	71.50	9.50	7.25
		PAA	**100.00**	**100.00**	**96.00**	**26.00**	**99.00**	**100.00**	**96.00**	**26.00**	**100.00**	**100.00**	**95.00**	**39.00**	**100.00**	**100.00**	**95.00**	**37.00**	**99.75**	**100.00**	**95.50**	**32.00**
	unknown	Single-P	4.00	11.00	2.00	0.00	8.00	22.00	5.00	0.00	14.00	37.00	29.00	1.00	10.00	16.00	6.00	0.00	9.00	21.50	10.50	0.25
		Multi-P	36.00	34.00	1.00	12.00	21.00	40.00	1.00	12.00	68.00	64.00	7.00	19.00	59.00	46.00	2.00	15.00	46.00	46.00	2.75	14.50
		CroPA	36.00	39.00	3.00	17.00	26.00	44.00	5.00	19.00	64.00	64.00	32.00	21.00	54.00	51.00	10.00	17.00	45.00	49.50	12.50	18.50
		PAA	**100.00**	**100.00**	**82.00**	**21.00**	**98.00**	**100.00**	**90.00**	**22.00**	**100.00**	**100.00**	**79.00**	**27.00**	**100.00**	**100.00**	**84.00**	**23.00**	**99.50**	**100.00**	**83.75**	**23.25**

$$\mathcal{R}_{\text{prompt}} = \sum_{i<j} \left(1 - \cos(\theta_{i,j})\right). \tag{21}$$

This regularization encourages updates that benefit a broader prompt distribution $\mathcal{P}_{\text{test}}$, leading to better generalization on unseen prompts.

V. EXPERIMENTS

A. Experimental Setting

*a) **Datasets**:* We utilize image text pairs from the VQAv2 [1] and VizWiz [51] datasets, incorporating prompts from the previous work by Luo *et al.* [9]. The datasets are organized into training and testing prompts. Each image in the training set is paired with $N = 65$ prompts, encompassing visual question answering, image classification, and image captioning tasks. The visual question answering task includes both VQA-Specific and VQA-Agnostic prompts, with the latter assessing whether the prompt is relevant to the image content. In contrast to previous approaches, our evaluation of PAA's generalization is conducted using a larger set of test prompts, comprising a total of 285 test prompts across the aforementioned tasks. Additionally, to enhance the training and testing prompts for VizWiz, we leverage GPT-5 to generate image prompt, which are subsequently subjected to manual quality checks.

*b) **Models**:* We evaluate attack performance on four widely used LVLMs, covering diverse architectures and training paradigms: open-flamingo-9B (OpenFlamingo, OF) [18], blip2-opt-2.7B (BLIP-2, B2) [17], instructblip-vicuna-7B (InstructBLIP, INB) [10], and llava-1.5-7B-hf (LLaVA, LA) [52].

*c) **Baselines**:* We compared PAA with existing cross-prompt attacks on LVLMs, including Single-P, Multi-P and CroPA [9]. The Single-P and Multi-P methods introduce single or multiple textual prompts, respectively, during the optimization of adversarial examples. CroPA incorporates multiple prompts as well, but further distinguishes itself by performing gradient-based updates on the textual prompts during the optimization process.

*d) **Metrics**:* In the experiments, the evaluation metric used is Attack Success Rate (ASR), which is the ratio of successfully manipulating the LVLMs to generate target text among all generated cross-prompt adversarial examples. A higher ASR indicates better cross-prompt transferability. In all experiments, we strictly enforced constraints on the perturbation magnitude and image validity. Furthermore, to align with real-world scenarios, adversarial images were quantized via local storage before being used to attack evaluation.

*e) **Training prompts**:* The training prompts are distributed across four tasks. Among the 65 training prompts, there are 5 prompts per image for VQA-Specific, Classification, and Caption tasks, while VQA-Agnostic has 50 prompts per image. For example, the prompt "*Classify the content of this image.*" template for each model is as follows:

- BLIP-2: "*Question: Classify the content of this image. Answer: *"
- InstructBLIP: "*Classify the content of this image.*"
- OpenFlamingo: "*<image> Question: What is on the bike? Short answer: basket <|endofchunk|> <image> Question: Is the water churning? Short answer: no <|endofchunk|> <image> Question: Classify the content of this image. Short answer:*"

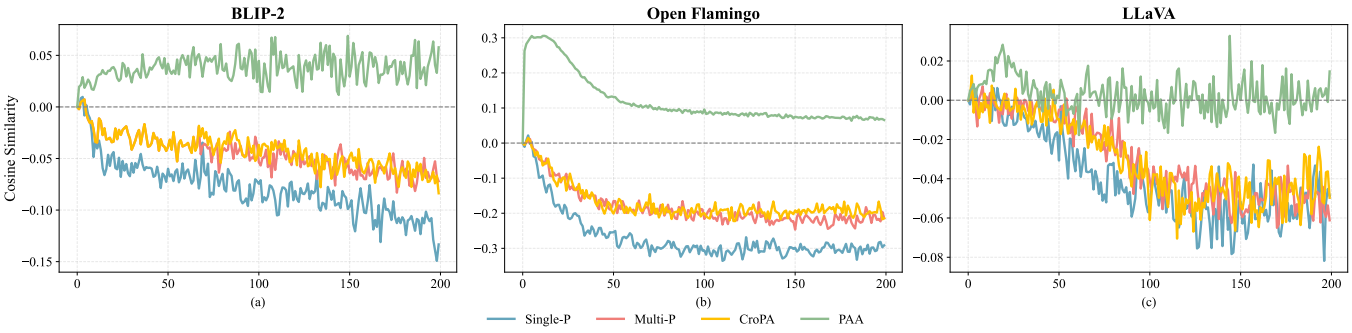


Fig. 5. The cosine similarity between gradients during adversarial examples generation on BLIP-2, OpenFlamingo and LLaVA. Due to the enhanced robustness of the safety-aligned LLaVA compared to other models, optimizing adversarial examples against it is inherently more challenging. Nevertheless, it can be observed that PAA exhibits more stable gradient updates during the adversarial optimization process compared to Single-P, Multi-P, and CroPA.

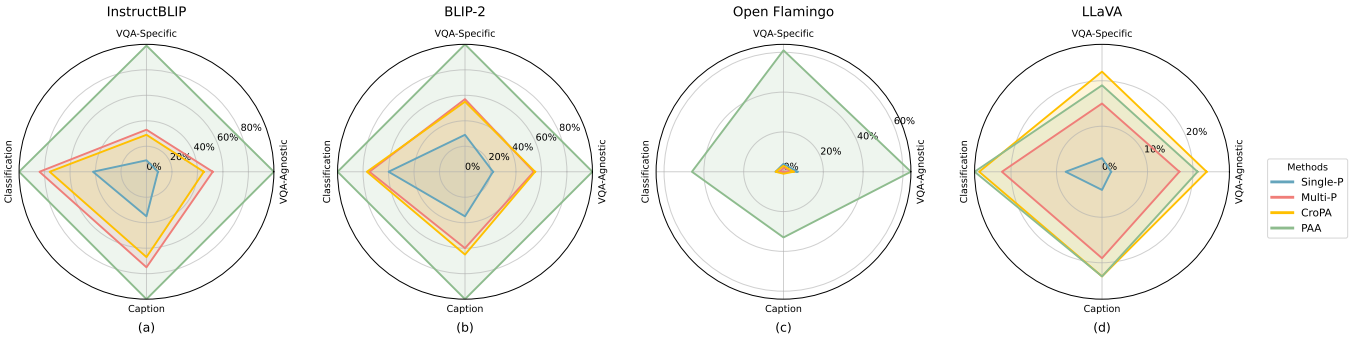


Fig. 6. Attack performance for target "unknown" on VizWiz Datasets. Comparison with existing cross-prompt attacks on different LVLMs. A higher ASR indicates better cross-prompt attack performance.

- LLaVA: "*USER: <image> \n Classify the content of this image. ASSISTANT:*"

f) ***Parameters****:* We adopt PGD [34] to craft adversarial examples. Unless otherwise specified, we set the maximum perturbation budget to $\epsilon = 16/255$, the outer iteration number to $T = 200$, the inner-loop iterations to $K = 2 \cdot |P|/|P^{MB}|$, and the step size to $\alpha = 1/255$. Our target set spans multiple categories, including object targets (e.g., "cat" and "dog"), emotion-related targets (e.g., "hello" and "unknown"), and jailbreak-oriented targets (e.g., "blood" and "drug").

g) ***Experimental Environment****:* All experiments were conducted on a server equipped with an Intel Xeon (3rd Gen) 5318Y CPU, an NVIDIA Tesla A100 GPU (40 GB), and 256 GB of RAM. Our implementation is based on PyTorch 2.3.1. The same hardware and software configuration was used for all runs.

B. Performance on VQAv2

Tab. I presents the attack performance of our method on the InstructBLIP, BLIP-2 and OpenFlamingo. Due to the limited number of training prompts and the significant semantic differences between prompts, CroPA demonstrates unsatisfactory attack performance under this setup. Even for the "drug" "cat" and "unknown" targets, the difference in performance between CroPA and Multi-P on InstructBLIP is not substantial. In contrast, the adversarial examples generated by PAA achieved superior performance across different targets on InstructBLIP, BLIP-2, and OpenFlamingo, attaining an ASR close to 100%.

TABLE II
ATTACK PERFORMANCE WITH AND WITHOUT DEFENSE ON THE VQAV2 AND VIZWIZ DATASETS. THE TABLE REPORTS THE AVERAGE ASR OF ADVERSARIAL EXAMPLES ACROSS FOUR TASKS.

Dataset	Setting	Method	INB	B2	OF	LA
VQAv2	Without Defense	Single-P	9.00	21.50	10.50	0.25
		Multi-P	46.00	46.00	2.75	14.50
		CroPA	45.00	49.50	12.50	18.50
		PAA	**99.50**	**100.00**	**83.75**	**23.25**
	With Defense	Single-P	2.50	11.25	5.50	0.00
		Multi-P	24.50	27.50	1.25	8.00
		CroPA	22.50	28.25	6.50	10.25
		PAA	**54.50**	**58.25**	**40.50**	**13.25**
VizWiz	Without Defense	Single-P	23.75	36.50	4.50	4.25
		Multi-P	61.00	61.50	2.25	18.25
		CroPA	54.25	63.00	3.25	**23.75**
		PAA	**99.75**	**100.00**	**51.00**	22.75
	With Defense	Single-P	13.75	19.75	4.75	3.00
		Multi-P	35.75	34.00	3.25	8.25
		CroPA	29.00	35.00	3.75	**14.25**
		PAA	**55.50**	**65.25**	**27.00**	10.50

Furthermore, we evaluate the attack performance on the safety-aligned LLaVA. Due to the safety alignment of its language backbone, LLaVA demonstrates greater robustness than BLIP-2, InstructBLIP, and OpenFlamingo, thereby limiting the ASR of adversarial examples. Nevertheless, PAA achieves a higher ASR on LLaVA compared to the baseline attacks. As shown in Fig. 5, gradient stability on LLaVA is slightly lower compared to BLIP-2, InstructBLIP, and OpenFlamingo.

To further analyze the experiment, we encoded all training

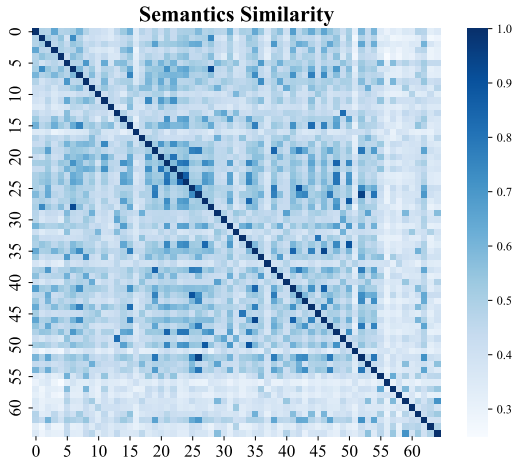


Fig. 7. The semantic similarity among all training prompts. The training set includes prompts from four task types: VQA-Agnostic, VQA-Specific, Classification, and Caption. The average similarity score is 0.46.

prompts using BERT [11] and computed their pairwise semantic similarity to quantify the semantic discrepancy among them. As shown in Fig. 7, most prompt similarities fall within the range of 0.3 to 0.4. Notably, prompts indexed from 0 to 55 exhibit higher semantic similarity compared to those indexed from 55 to 65. This is attributed to the fact that prompts 55–65 correspond to captioning and classification tasks, which are generally longer and more complex than the VQA prompts in the 0–55 range, thereby resulting in lower semantic similarity. Nevertheless, the overall mean similarity is 0.46, indicating a relatively low level of semantic overlap and thus confirming the existence of semantic diversity among the prompts.

Notably, under the three distinct attack target settings, the baseline attack reveals that object targets are the most vulnerable to exploitation, while emotional targets prove to be the most challenging. This increased difficulty stems from the inherent complexity of attacking an image with an unrelated emotional target, which does not naturally align with the image's content. In contrast, PAA demonstrates consistently superior performance across all target types. By accounting for semantic differences and optimizing over a small set of carefully selected training prompts, PAA is able to generate adversarial examples with high ASR across a broad range of test prompts. It is worth acknowledging that the inner-outer loop mechanism in PAA incurs higher computational overhead compared to baseline methods. However, given the substantial improvements in cross-prompt generalization and the typically offline nature of adversarial generation, this cost-performance trade-off is justified.

C. Performance on VizWiz

We report the performance of PAA compared to other methods on the VizWiz dataset in Fig. 6, with results consistent with previously observed conclusions. PAA demonstrates superior cross-prompt transferability across most models and tasks, slightly outperforming CroPA on the securely aligned LLaVA. In contrast, InstructBLIP and BLIP-2 are more susceptible to cross-prompt attacks compared to OpenFlamingo and LLaVA, with CroPA performing similarly to Multi-P on these two models. Due to the in-context learning mechanism of OpenFlamingo, prior methods exhibit lower cross-prompt transferability when evaluated on a larger test dataset.

D. Adversarial Robustness to Defense Strategies

To further assess the robustness of the adversarial examples, we applied random rotation pre-processing to the adversarial examples and then evaluated their performance across different models and tasks. As shown in Tab. II, we report the performance of the adversarial examples on four models, evaluated in VQA-Agnostic, VQA-Specific, Classification, and Captioning tasks. It can be observed that the adversarial examples generated by PAA, after defense, exhibit robustness. However, on the VizWiz dataset, it is worth noting that on the securely aligned LLaVA model CroPA slightly outperforms PAA. This exception is due to the combination of LLaVA's robust safety alignment and the significant domain shift in the VizWiz dataset. These factors result in PAA's strategy of accumulating gradients to find a generalized direction faces greater resistance. In contrast, CroPA's Min-Max optimization strategy proves slightly more effective in penetrating the defense in this specific robust setting. Additionally, the performance of the with defense OpenFlamingo model under Single-P, Multi-P, and CroPA attacks outperforms without defended configuration. This improvement is due to the relatively low attack performance of these models on the OpenFlamingo when trained on a limited dataset, with certain examples experiencing successful but accidental attacks after defense. Although PAA's ASR also decreases, its performance drop is significantly smaller than that of the baseline methods. Most importantly, under the random rotation defense, PAA maintains a substantial performance margin over most baselines.

E. Loss Landscape Visualization

To intuitively compare the geometric properties of the local minima found by different attack methods, we employed a loss landscape visualization technique. We visualize the local geometry adversarial example by projecting the high dimensional loss function onto a two dimensional plane. Specifically, the visualization is centered at the final solution δ^* obtained from optimization. We begin by generating two random Gaussian vectors of the same shape as δ^*, which are then processed using the Gram Schmidt procedure to yield two orthonormal direction vectors, d_1 and d_2, that define the axes of a 2D plane passing through δ^*. Any perturbation $\delta(\alpha, \beta)$ on this plane is parameterized as $\delta^* + \alpha \cdot d_1 + \beta \cdot d_2$, where the scalar coordinates α and β control the displacement along the d_1 and d_2 directions, respectively. We define a 30×30 uniform grid over the range $[-1.0, 1.0]$ for both α and β. For each coordinate pair (α, β) on the grid, we compute the corresponding perturbation, add it to the clean image, and then calculate its average loss over the entire set of training prompts. Finally, the collected loss values are rendered as a 2D contour plot, where the density of the contour lines provides a direct visual representation of the surface's flatness: a wide

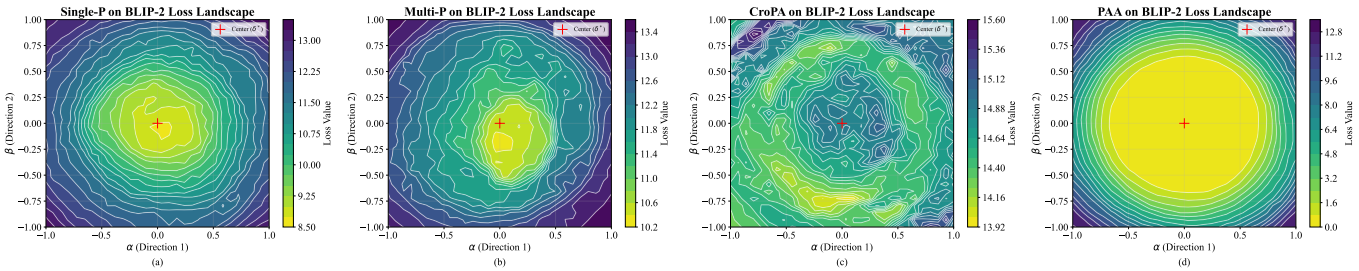


Fig. 8. Loss Landscape Visualization of Adversarial Examples on BLIP-2. We visualize the loss landscape of Single-P, Multi-P, CroPA, and PAA on BLIP-2. Lighter colors represent lower loss values.

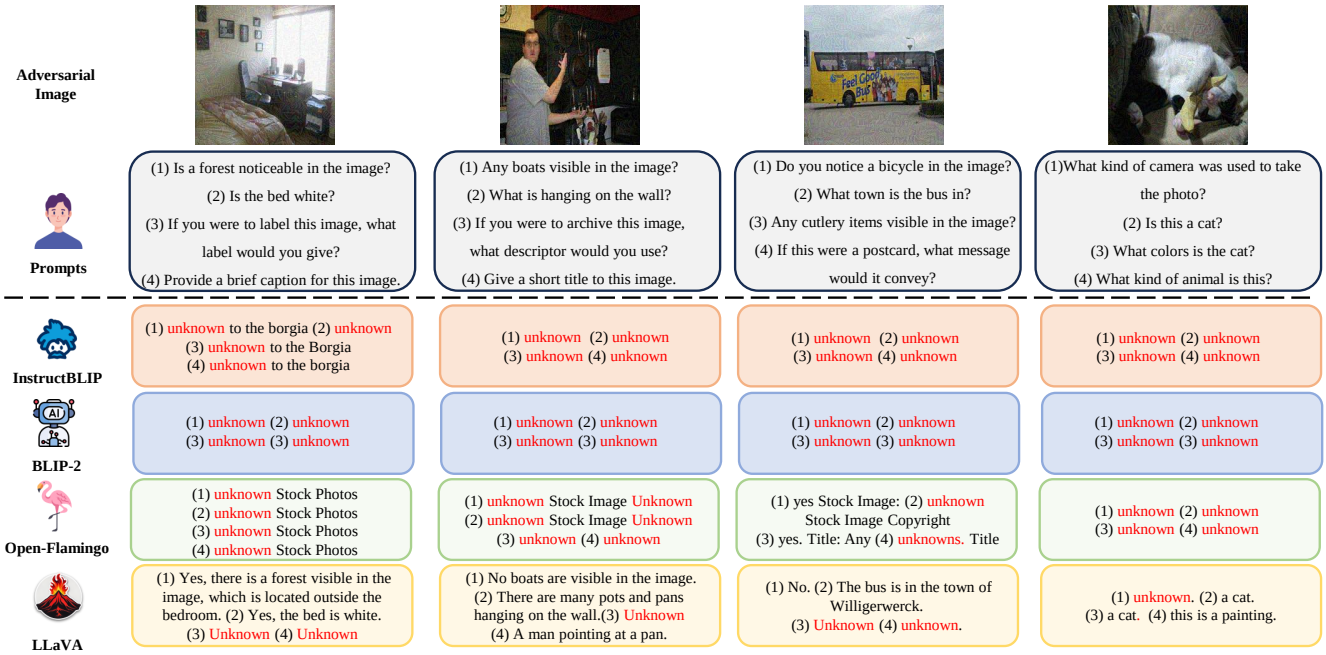


Fig. 9. Visualize cross-prompt adversarial examples generated by PAA across different LVLMs on VQAv2. We select the most challenging emotion target "unknown", and visualize the performance of adversarial examples generated by PAA across different LVLMs. As shown, PAA demonstrates better cross-prompt effectiveness on BLIP-2, InstructBLIP, and OpenFlamingo compared to the more robust LLaVA.

region with sparse contours corresponds to a flat minimum, whereas a narrow one indicates a sharp minimum.

As shown in Fig. 8, we visualize the loss landscape of adversarial examples on BLIP-2 across all training prompts. As analyzed previously, the minima for Single-P, Multi-P, and CroPA are significantly sharp, whereas PAA resides at a flatter minima, which facilitates stronger cross-text generalization.

F. Case Study

To provide a more in-depth analysis of PAA, we conducted a case study targeting the term "unknown" to evaluate the performance of adversarial examples generated by PAA across different LVLMs. As shown in Fig. 9, adversarial examples crafted by PAA successfully prompt the LVLMs to produce the desired output with high accuracy. Notably, the BLIP-2 model directly provides the correct targeted answer. In contrast, OpenFlamingo tends to frequently output "Stock Photos" following the term "unknown" with a high probability. We hypothesize that this behavior arises from OpenFlamingo's reliance on in-context learning during both training and inference. The presence of this in-context learning template seems to significantly influence OpenFlamingo's response, leading to the observed output.

G. Human Evaluation

We conducted a human evaluation with 5 well-educated volunteers. Each volunteer reviewed 50 PAA adversarial examples and formulated 3 questions per image, yielding 750 questions in total. Question content was unrestricted, participants could ask questions related or unrelated to the image.

In the human evaluation experiment, volunteers were free to ask any questions, making these questions entirely independent of training and test prompts. This setup better reflects PAA's generalization ability on unseen prompts. We sampled five images, each accompanied by three prompts, to visually demonstrate the results in Fig. 10. The average semantic similarity between the prompts provided by the volunteers was calculated to be 0.54. Additionally, we visualized the semantic similarity of a randomly sampled set of 50 prompts in Fig. 11.

Fig. 10. Visualize questions asked by humans. To evaluate the cross-prompt generalization capability of PAA, we instructed volunteers to freely ask questions about the images without any constraints on the content.

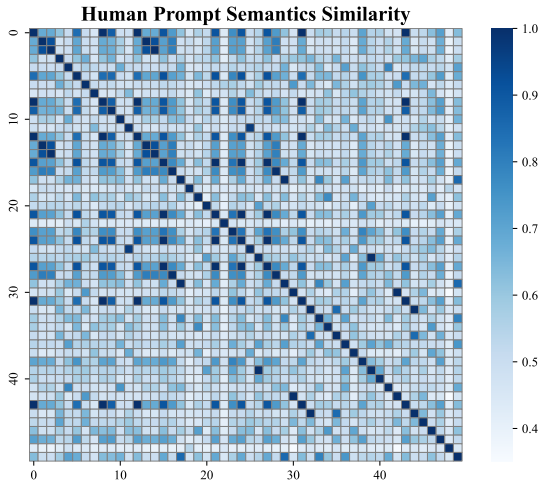


Fig. 11. To visualize the semantic similarity of the prompts provided by the volunteers, we randomly sampled 50 prompts and calculated their similarity.

TABLE III
HUMAN EVALUATION EXPERIMENT. WE TEST HUMAN-PROPOSED PROMPTS UNDER THE "UNKNOWN" TARGET SETTING.

Method	InstructBLIP	BLIP-2	OpenFlamingo	LLaVA
Single-P	5.60%	19.40%	2.60%	0.20%
Multi-P	29.60%	44.20%	0.00%	11.20%
CroPA	28.20%	47.40%	6.00%	18.00%
PAA	**97.80%**	**100.00%**	**91.20%**	**20.60%**

Tab. III shows the results, we observe that CroPA overfit to training prompts, leading to limited performance on unseen prompts. Moreover, since the volunteer-provided questions exhibit greater divergence from the training data, previous methods struggle to achieve targeted attacks on such inputs. In contrast, PAA demonstrates a higher ASR on unseen prompts, consistent with the findings of previous experiments.

H. Ablation Study

We conducted the ablation study to validate the effectiveness of PAA by examining the impact of mini-prompt batch accumulation and the number of inner loop iterations K on VQAv2 datasets.

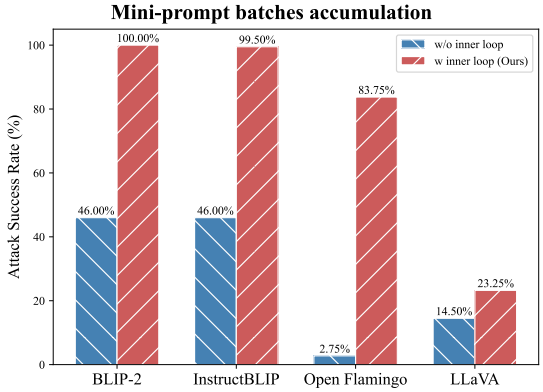


Fig. 12. Ablation study on mini-prompt batches accumulation. w/o inner loop represents the baseline attack without mini-prompt batches accumulation, w inner loop represents PAA with mini-prompt batches accumulation.

The impact of mini-prompt batches accumulation. We evaluated the performance differences between PAA with inner loops and the baseline attack without inner loops across four models, with "unknown" as the attack target. As shown in Fig. 12, our method achieves higher ASR compared to the baseline attack without inner loops across all four models, with especially significant improvements on InstructBLIP, BLIP-2, and OpenFlamingo. While LLaVA exhibits greater robustness than the other models, our method still outperforms the baseline attack without inner loops.

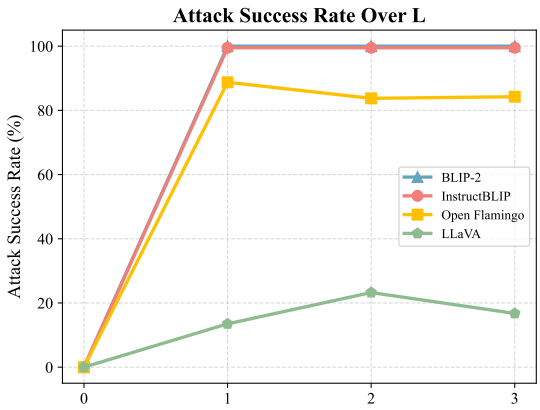


Fig. 13. Ablation study on inner loop iterations number K. The figure illustrates the impact of different inner loop iterations of PAA on ASR across the four models.

The inner loop iterations number K. We tested the ASR of PAA with varying numbers of inner loop iterations $K = L \cdot |P|/|P^{MB}|$. As shown in Fig. 13, PAA demonstrates an increase in ASR with the growth of L. Specifically, 100% ASR at $L = 2$ was achieved for InstructBLIP and BLIP-2. Although

ASR also improves with increasing L for other models, this comes with a significant time overhead.

VI. Conclusion

In this paper, we extensively examine current cross-prompt attacks, which encounter challenges due to gradient instability: significant semantic differences between prompts lead to unstable gradient updates during adversarial example generation. This instability limits the generalization of adversarial examples on unseen prompts. To address this issue, we propose PAA, which uses an inner-outer loop structure and divides the training prompts into mini-prompt batches. By accumulating gradients across these mini-prompt batches in the inner loop, PAA effectively mitigates gradient instability caused by prompt differences. Extensive experiments demonstrate that our method significantly improves the generalization of adversarial examples on unseen prompts across various settings.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[2] N. Rotstein, D. Bensaïd, S. Brody, R. Ganz, and R. Kimmel, "Fusecap: Leveraging large language models for enriched fused image captions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5689–5700.

[3] F. Li, J. Wu, C. He, and W. Zhou, "Cmie: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection," in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, 2025, pp. 9342–9354.

[4] J. Wu, F. Li, Z. Fu, M.-Y. Kan, and B. Hooi, "Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models," *arXiv preprint arXiv:2505.15489*, 2025.

[5] F. Li, J. Wu, T. Fu, Y. Dong, B. Song, and W. Zhou, "Drifting away from truth: Genai-driven news diversity challenges lvlm-based misinformation detection," *arXiv preprint arXiv:2508.12711*, 2025.

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

[7] C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[8] R. Liu, G. Li, T. Zhang, and S.-K. Ng, "Image can bring your memory back: A novel multi-modal guided attack against image generation model unlearning," in *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[9] H. Luo, J. Gu, F. Liu, and P. Torr, "An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

[10] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 49 250–49 267, 2023.

[11] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. Fu, J. Zhang, F. Li, P. Wei, X. Zeng, and W. Zhou, "Multimodal alignment augmentation transferable attack on vision-language pre-training models," *Pattern Recognition Letters*, vol. 191, pp. 131–137, 2025.

[13] R. Liu, K.-Y. Lam, W. Zhou, S. Wu, J. Zhao, D. Hu, and M. Gong, "Stba: Towards evaluating the robustness of dnns for query-limited black-box scenario," *IEEE Transactions on Multimedia*, 2025.

[14] X. Qi, K. Huang, A. Panda, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak large language models," *arXiv preprint arXiv:2306.13213*, 2023.

[15] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[16] Y. Zhao, H. Zhang, and X. Hu, "Penalizing gradient norm for efficiently improving generalization in deep learning," in *International conference on machine learning*. PMLR, 2022, pp. 26 982–26 992.

[17] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

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

[19] C. Wan, X. Luo, H. Luo, Z. Cai, Y. Song, Y. Zhao, Y. Bai, F. Wang, Y. He, and Y. Gong, "Grid: Omni visual generation," *arXiv preprint arXiv:2412.10718*, 2024.

[20] L. Peng, C. Wan, S. Wang, X. Song, Y. He, and Y. Gong, "Cia: Class-and instance-aware adaptation for vision-language models," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 2870–2879.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[22] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, P. W. W. Koh, D. Ippolito, F. Tramer, and L. Schmidt, "Are aligned neural networks adversarially aligned?" *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[23] Z. Wang, Z. Han, S. Chen, F. Xue, Z. Ding, X. Xiao, V. Tresp, P. Torr, and J. Gu, "Stop reasoning! when multimodal llms with chain-of-thought reasoning meets adversarial images," *arXiv preprint arXiv:2402.14899*, 2024.

[24] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[25] H. Fang, J. Zhang, Y. Qiu, J. Liu, K. Xu, C. Fang, and E.-C. Chang, "Tracing the origin of adversarial attack for forensic investigation and deterrence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4335–4344.

[26] C. Wen, X. Li, H. Huang, Y.-S. Liu, and Y. Fang, "3d shape contrastive representation learning with adversarial examples," *IEEE Transactions on Multimedia*, 2023.

[27] X. Wei and S. Zhao, "Boosting adversarial transferability with learnable patch-wise masks," *IEEE Transactions on Multimedia*, vol. 26, pp. 3778–3787, 2023.

[28] R. Ran, J. Wei, C. Zhang, G. Wang, Y. Yang, and H. T. Shen, "Adaptive multi-scale degradation-based attack for boosting the adversarial transferability," *IEEE Transactions on Multimedia*, 2024.

[29] J. Zhang, C. Dongdong, Q. Huang, J. Liao, W. Zhang, H. Feng, G. Hua, and N. Yu, "Poison ink: Robust and invisible backdoor attack," *IEEE Transactions on Image Processing*, vol. 31, pp. 5691–5705, 2022.

[30] S. Zheng, C. Zhang, and X. Hao, "Black-box targeted adversarial attack on segment anything (sam)," *IEEE Transactions on Multimedia*, 2024.

[31] M. Xue, K. Peng, X. Gong, Q. Zhang, Y. Chen, and R. Li, "Echo: Reverberation-based fast black-box adversarial attacks on intelligent audio systems," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–24, 2023.

[32] Y. Liu, X. He, M. Xiong, J. Fu, S. Deng, and B. Hooi, "Flipattack: Jailbreak llms via flipping," *arXiv preprint arXiv:2410.02832*, 2024.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[34] A. Madry, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[35] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[36] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

[37] K. Wang, X. He, W. Wang, and X. Wang, "Boosting adversarial transferability by block shuffle and rotation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 336–24 346.

[38] H. Zhu, Y. Ren, X. Sui, L. Yang, and W. Jiang, "Boosting adversarial transferability via gradient relevance attack," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4741–4750.

[39] Y. Ren, H. Zhu, C. Liu, and C. Li, "Efficient polar coordinates attack with adaptive activation strategy," *Expert Systems with Applications*, vol. 249, p. 123850, 2024.

[40] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," *arXiv preprint arXiv:2004.09984*, 2020.

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

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

[43] Y. Wang, W. Hu, Y. Dong, H. Zhang, H. Su, and R. Hong, "Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning," *IEEE Transactions on Multimedia*, 2025.

[44] Y. Wang, W. Hu, Q. Li, and R. Hong, "Boosting adversarial robustness of vision-language pre-training models against multimodal adversarial attacks," in *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

[45] Y. Wang, W. Hu, Y. Dong, J. Liu, H. Zhang, and R. Hong, "Align is not enough: Multimodal universal jailbreak attack against multimodal large language models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[46] C. Wan, Y. He, X. Song, and Y. Gong, "Prompt-agnostic adversarial perturbation for customized diffusion models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 136 576–136 619, 2024.

[47] Y. Dong, H. Chen, J. Chen, Z. Fang, X. Yang, Y. Zhang, Y. Tian, H. Su, and J. Zhu, "How robust is google's bard to adversarial image attacks?" *arXiv preprint arXiv:2309.11751*, 2023.

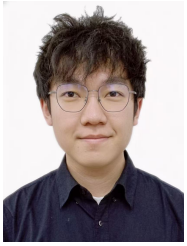
[48] C. Schlarmann and M. Hein, "On the adversarial robustness of multimodal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3677–3685.

[49] R. Schaeffer, D. Valentine, L. Bailey, J. Chua, C. Eyzaguirre, Z. Durante, J. Benton, B. Miranda, H. Sleight, J. Hughes *et al.*, "When do universal image jailbreaks transfer between vision-language models?" *arXiv preprint arXiv:2407.15211*, 2024.

[50] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[51] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3608–3617.

[52] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.



Tingchao Fu received his master's degree from Yunnan University in 2025. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Yunnan University, Kunming, China. His research focuses on trustworthy artificial intelligence, including adversarial examples, knowledge editing, and the safety of large language models. He is also interested in improving the robustness, reliability, and security of foundation models in real-world applications.

Renyang Liu is currently a Research Fellow with the Institute of Data Science, National University of Singapore, Singapore. He received the Ph.D. degree from Yunnan University, China, in 2024. From 2022 to 2023, he was a visiting student with the College of Computing and Data Science, Nanyang Technological University, Singapore. From 2023 to 2024, he was a Research Intern with the School of Cyber Security, Sun Yat-sen University, China. His research interests include AI security, data privacy, trustworthy generative AI, and AI for security.

Ziyao Liu received the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2023, the master's degree from Beijing Institute of Technology, China, in 2018, and the bachelor's degree from Zhengzhou University, China, in 2015. He is currently a Research Fellow with the National Centre for Research in Digital Trust, Singapore. His research interests include AI safety and privacy-enhancing technologies.

Peiyuan Si (Student Member, IEEE) received the bachelor's and master's degrees in communication engineering from Zhejiang University of Technology, Zhejiang, China, in 2018 and 2021, respectively. He received the Ph.D. degree from the College of Computing and Data Science, Nanyang Technological University, Singapore. His research interests include semantic communication, unmanned aerial vehicles, and reinforcement learning.

Fanxiao Li was born in 1998, received the B.Eng. degree in Computer Science from Fuzhou University in 2021 and the M.Eng. degree in Software Engineering from Yunnan University in 2024. He is currently a Ph.D. student at Yunnan University. He is also a visiting Ph.D. student at National University of Singapore (Sep 2025 - Sep 2026), supported by the China Scholarship Council (CSC). He is a student member of CCF and IEEE. His main research interests include trustworthy social intelligence and mis/disinformation governance.

Jinhong Zhang received the master's degree from Yunnan University in 2023. He is currently pursuing the Ph.D. degree with the School of Information Science and Engineering, Yunnan University. His research interests include artificial intelligence model security and multimedia data privacy protection, with particular focus on adversarial example attacks, model stealing attacks, image steganography, and visual privacy protection.

Wei Zhou (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences. He is currently a Full Professor with the School of Engineering, Yunnan University. He has hosted several National Natural Science Foundation projects. His current research interests focus on trustworthy artificial intelligence, including adversarial examples, mis/disinformation governance, privacy protection, and bioinformatics.