Revisiting CroPA: A Reproducibility Study and Enhancements for Cross-Prompt Adversarial Transferability in Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we conduct a comprehensive reproducibility study of An Image is Worth 1000 Lies: Adversarial Transferability Across Prompts on Vision-Language Models. Beyond replicating the original Cross-Prompt Attack (CroPA) method, we identify key limitations and propose enhancements to improve its effectiveness. Our key contributions include: (1) Two novel initialization strategies that significantly improve Attack Success Rate (ASR) and transferability (2) a refined loss function that manipulates the vision encoder's attention mechanisms to improve generalization and (3) a broader evaluation by benchmarking CroPA against multiple robust attack baselines. We evaluate our approach across a range of prevalent VLMs, including Flamingo, BLIP-2, and InstructBLIP, validate the original results while demonstrating consistent improvements. Our work reinforces the importance of studying adversarial vulnerabilities in VLMs and provides a more robust and versatile framework for generating transferable adversarial examples, with significant implications for understanding and improving the security of VLMs in real-world applications.

1 Introduction

The advent of large Vision-Language Models (VLMs) has significantly transformed the field of computer vision by enabling a wide range of tasks, including image classification, captioning, and visual question answering. This versatility has fostered deeper exploration into visual-linguistic interactions. However, recent studies Zhao et al. (2023); Qi et al. (2023); Zhang et al. (2022); Carlini et al. (2024) have demonstrated that VLMs remain highly vulnerable to adversarial attacks. These attacks involve subtle perturbations to input images, leading VLMs to produce incorrect or even harmful outputs. Furthermore, the inclusion of textual modalities introduces additional attack vectors, expanding the range of threats beyond those faced by traditional vision models.

Several studies have investigated the adversarial robustness of VLMs. For example, Zhao et al. Zhao et al. (2023) conducted a comprehensive analysis of the adversarial robustness of VLMs such as BLIP and BLIP-2, exploring both query-based and transfer-based adversarial attack methods in black-box settings. Additionally, Schlarmann et al. Schlarmann & Hein (2023) examined targeted and untargeted adversarial attacks in white-box settings. While these works primarily focused on adversarial image attacks, subsequent research has also explored adversarial perturbations in textual inputs. Qi et al. Qi et al. (2023) demonstrated that adversarial images could manipulate VLMs into executing harmful instructions, while Tu et al. Tu et al. (2023) systematically evaluated both visual and textual adversarial attacks.

Traditionally, the generalization of adversarial examples in VLMs has been classified into two primary categories:

- **Cross-Model Transferability**: The ability of adversarial examples to maintain their adversarial nature across different VLM architectures, commonly referred to as transferability.

- **Cross-Image Transferability**: The ability of adversarial perturbations to generate adversarial examples that generalize across multiple images, often known as Universal Adversarial Perturbations (UAPs).

Luo et al. (2024) introduced the novel concept of **Cross-Prompt Transferability**, which describes the ability of adversarial images to remain effective across varying textual prompts. Unlike prior work that treated visual and textual adversarial perturbations independently, Luo et al. proposed the **Cross-Prompt Attack** (CroPA), which employs learnable prompts to ensure adversarial images retain their effectiveness regardless of textual input. Their work demonstrated CroPA's efficacy across multiple vision-language tasks, including image classification, captioning, and visual question answering.

Our work aims to address the following goals:

- **[Reproducibility Study] Reproducing the results from the original paper:** Through our experiments we were able to reproduce and verify the main claim of the paper by showing that CroPA achieves cross-prompt transferability across various target texts.
- **[Extended Work] Better Initialization:** We propose two new initialization strategies that substantially increase the Attack Success Rate (ASR) as well as Transferability.
- **[Extended Work] Loss Function:** We propose a novel loss function building on the idea that specific components within the vision encoder's attention mechanism control and determine the level of interaction between patches, manipulating the value vectors of the vision encoder in a targeted manner leads to greater generalization as well as ASR.
- **[Extended Work] Additional Baselines:** We expanded the scope of our research and experiments by comparing CroPA against multiple robust Attack Methods to provide a stronger baseline for comparison.

Furthermore, we conduct in-depth analyses to elucidate the mechanisms behind our improvements, offering insights into the nature of adversarial vulnerabilities in VLMs. Our work not only reinforces the importance of studying these vulnerabilities but also provides a more robust and versatile framework for generating transferable adversarial examples.

2 Scope of reproducibility

This study aims to examine and validate the results demonstrated by Luo et al. (2024). Our primary objective is to confirm that CroPA significantly enhances the transferability of adversarial examples across various prompts by meticulously reproducing their experimental procedures.

Beyond replication, we intend to extend the scope of CroPA by investigating its efficacy in cross-model and cross-image contexts. Specifically, we will assess whether adversarial images generated through CroPA can consistently deceive diverse Vision-Language Models (VLMs), regardless of the input prompt or specific model parameters. Addressing these points will enable us to faithfully reproduce your experiments and build upon your work to explore the broader applicability of CroPA in enhancing the robustness and versatility of adversarial attacks on VLMs.

3 Methodology

Our reproduction efforts were based on the code provided in the authors' public repository. While the overall implementation was well-documented, we encountered several challenges that required modifications for successful reproduction. Notably, the code for BLIP-2 and InstructBLIP models triggered multiple runtime errors, necessitating significant debugging and adjustments to achieve functional execution.

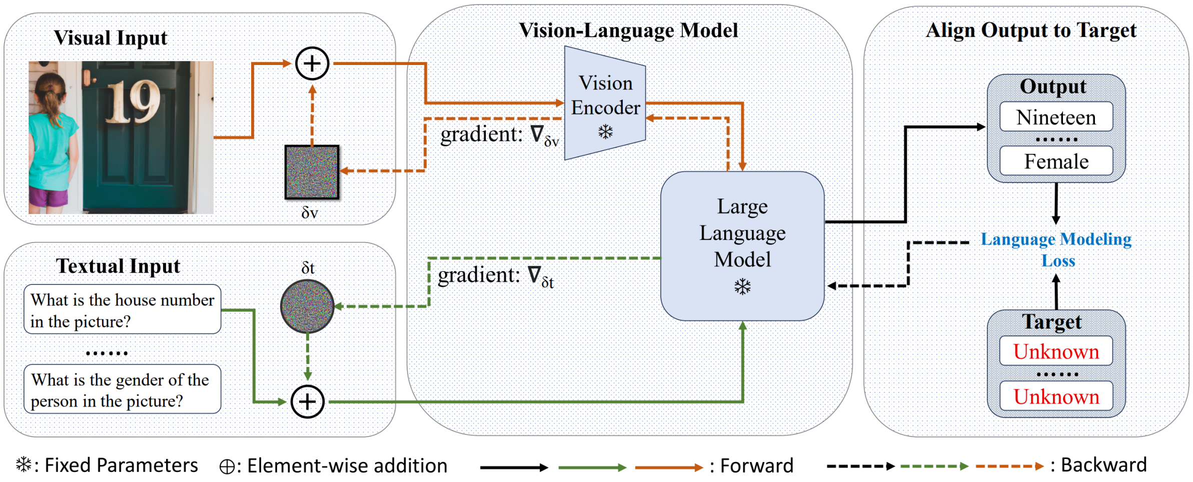

Figure 1: Framework overview of CroPA as presented in Luo et al. (2024). The architecture employs learnable perturbations for both image (δ_v) and prompt (δ_t) inputs. These perturbations operate antagonistically - while δ_v is optimized to minimize the language modeling loss, δ_t works to maximize it. The model's forward pass (solid arrows) processes the perturbed inputs through the vision encoder and language model, with backpropagation (dashed arrows) updating the perturbations at configurable frequencies. The model parameters (marked with *) remain fixed during the attack. $\oplus$ denotes element-wise addition.

3.1 Problem Formulation

The vulnerability of neural networks to adversarial attacks has been well-documented since the seminal work of Goodfellow et al. (2014). Building on this foundation, we examine the authors' novel formulation that extends these concepts to cross-prompt scenarios in Vision-Language Models (VLMs). Their work introduces a critical perspective on how adversarial perturbations can maintain effectiveness across varying textual inputs Luo et al. (2024).

The authors develop their formulation around a VLM function f that processes both visual and textual inputs, denoted as x_v and x_t respectively.To ensure real-world applicability, the authors constrain the adversarial perturbation δ_v within human-imperceptible bounds, enforcing $\|\delta_v\|_p \leq \epsilon_v$. This constraint mirrors established practices in adversarial machine learning while adapting them to the multi-modal context of VLMs Carlini et al. (2024).

The authors establish two distinct attack scenarios that we reproduced in our study:

The targeted attack scenario aims to manipulate the VLM into generating a specific predetermined text T, regardless of the input prompt. This objective manifests mathematically as minimizing the language modeling loss L across multiple prompt instances:

$$\min_{\delta_v} \sum_{i=1}^{k} L(f(x_v + \delta_v, x_t^i), T) \tag{1}$$

In contrast, the non-targeted scenario focuses on maximizing the discrepancy between outputs from clean and adversarial inputs:

$$\max_{\delta_v} \sum_{i=1}^{k} L(f(x_v + \delta_v, x_t^i), f(x_v, x_t^i)) \tag{2}$$

The effectiveness of these attacks is quantified through the Attack Success Rate (ASR). For targeted attacks, success requires generating the exact target text, while non-targeted attacks succeed by producing any output that differs from the clean image's prediction.

3.2 Cross Prompt Attack

The Cross-Prompt Attack (CroPA) method introduced by Luo et al. (2024) employs learnable prompts during optimization to enhance cross-prompt transferability. The key innovation lies in using prompt perturbations that compete with image perturbations during the optimization process rather than collaborating to deceive the model.

The algorithm optimizes both visual perturbation δ_v and prompt perturbation δ_t but with opposing objectives. While δ_v aims to minimize the language modeling loss for generating the target text, δ_t maximizes this loss. This adversarial relationship between the perturbations forces δ_v to develop stronger transferability across different prompts.

The optimization process can be formally expressed as a min-max problem:

$$\min_{\delta_v} \max_{\delta_t} L(f(x_v + \delta_v, x_t + \delta_t), T) \tag{3}$$

where f represents the VLM, T is the target text for targeted attacks, and L denotes the language modeling loss.

The implementation follows an iterative approach using Projected Gradient Descent (PGD) Madry et al. (2017). The visual perturbation updates use gradient descent to minimize the loss, while prompt perturbation updates employ gradient ascent to maximize it. The update frequency can be controlled via a parameter N, where image perturbation updates occur N times for each prompt perturbation update. The framework can be visualized formally in Figure 1.

We reproduce the algorithm as follows:

Algorithm 1 Cross Prompt Attack (CroPA)

```
Require: Model $f$, Target Text $T$, input image $x_v$, prompt set $X_t$, perturbation size $\epsilon$, step sizes $\alpha_1$, $\alpha_2$,
  iterations $K$, update interval $N$
Ensure: Adversarial example $x'_v$
  Initialize $x'_v = x_v$
  for step = 1 to $K$ do
    Sample prompt $x_t^i$ from $X_t$
    if $x_t^i$ not initialized then
      Initialize $x_t'^i = x_t^i$
    end if
    $g_v = \nabla_{x_v} L(f(x'_v, x_t^i), T)$
    $x'_v = x'_v - \alpha_1 \cdot \mathrm{sign}(g_v)$
    if mod(step, $N$) = 0 then
      $g_t = \nabla_{x_t} L(f(x'_v, x_t^i), T)$
      $x_t'^i = x_t'^i + \alpha_2 \cdot \mathrm{sign}(g_t)$
    end if
    Project $x'_v$ to $\epsilon$-ball around $x_v$
  end for
  return $x'_v$
```

During evaluation, only the optimized image perturbation is applied, while prompt perturbations are discarded. This ensures that the attack's effectiveness stems from the image perturbation's inherent transferability rather than prompt modifications.

3.3 Models Used

In reproducing the work of Luo et al. (2024), we evaluated three state-of-the-art Vision-Language Models (VLMs): Flamingo, BLIP-2, and InstructBLIP. For Flamingo, we utilized the open-source OpenFlamingo-9B implementation Awadalla et al. (2023), which provides comparable performance to the original model while being publicly accessible.

BLIP-2 introduces a two-stage approach that first extracts visual features using a frozen CLIP image encoder, then processes these features through a Querying Transformer Li et al. (2023). This architecture enables efficient adaptation to diverse vision-language tasks. The model employs OPT-2.7b as its language model component, facilitating flexible text generation capabilities.

InstructBLIP builds upon BLIP-2's architecture while incorporating instruction tuning Dai et al. (2023). A key distinction is its use of the Vicuna-7b language model, which enhances the model's ability to follow task-specific instructions. This modification enables more precise control over the model's outputs through carefully crafted prompts.

Each model offers distinct advantages in handling vision-language tasks. Flamingo excels at few-shot learning through visual examples, BLIP-2 demonstrates strong zero-shot generalization capabilities, and InstructBLIP shows improved performance on instruction-guided tasks. Our reproduction efforts maintained the original configurations of these models to ensure faithful comparison with the baseline results.

3.4 Datasets

Following the original work, our evaluation utilized images from the MS-COCO validation dataset Lin et al. (2014). This dataset provides a diverse collection of natural images suitable for testing cross-prompt transferability across various visual scenarios.

For the textual component, we employed two categories of Visual Question Answering (VQA) prompts. The first category, $\text{VQA}_{\text{general}}$, consists of general questions applicable to any image, focusing on common visual attributes and objects. The second category, $\text{VQA}_{\text{specific}}$, derives from the VQA-v2 dataset Goyal et al. (2017) and contains questions specifically tailored to individual image content.

This combination of a standard vision dataset with both general and specific VQA prompts enables comprehensive evaluation of cross-prompt transferability across different types of queries and visual contexts. The prompts were designed to test both broad visual understanding and specific detail recognition capabilities of the models.

3.5 Experimental Setup and Code

The experimental setup followed specific parameters for attack configuration and evaluation.For the attack implementation, we maintained consistency with the original setup by utilizing the same seeds. By default, the experiments were conducted as targeted attacks, with "unknown" chosen as the target text to avoid high-frequency responses typical in vision-language tasks. The perturbation size was fixed at 16/255, and all adversarial examples were optimized and tested under zero-shot settings.

For multi-prompt experiments, both Multi-P and CroPA implementations used ten prompts. We maintained three evaluation runs for each experiment, averaging the Attack Success Rate (ASR) scores to ensure reliable results. The prompts spanned multiple task types including general visual questions, image-specific queries, classification tasks, and image captioning, with varying lengths and semantic structures.

For model implementations, we used the public OpenFlamingo-9B Awadalla et al. (2023) as our Flamingo variant, along with BLIP-2 and InstructBLIP models. Our reproduction maintained these core experimental parameters to ensure comparable results with the original work.

3.6 Computational Requirements

The computational demands of reproducing the CroPA experiments were substantial, reflecting the resource-intensive nature of modern Vision-Language Models. Our primary experiments were conducted on a PyTorch Lightning platform using an L40S GPU with 48GB VRAM and a 4-core CPU with 16GB RAM, matching the original paper's minimum requirement of 45GB VRAM for stable execution.

Working within the constraints of the free-tier platform credits posed significant challenges. Our experiments were limited by a pooled allocation of 120 credits shared across four accounts. This necessitated careful resource management, particularly given the computational intensity of large-scale VLMs. To overcome these limitations, we implemented several memory optimization strategies to enable partial execution on local machines with 16GB VRAM, though this required significant code modifications.

The total computational cost of our reproduction study amounted to approximately 140 GPU hours and 90 CPU hours. This includes time spent on model training, attack generation, and evaluation across multiple experimental configurations. The substantial computational requirements underscore the importance of efficient resource allocation in modern machine learning reproducibility studies.

4 Results

As stated in Section 3, a core objective of our research is to reproduce the cross-prompt transferability results presented by Luo et al. (2024) for the CroPA attack. This section details our efforts to replicate those findings and provides a comparative analysis of our reproduced results against the original paper.

4.1 Claim : CroPA achieves cross-prompt transferability across various target texts. [Verified]

To validate the central claim of Luo et al. (2024), which posits that CroPA can achieve cross-prompt transferability, we meticulously replicated their experimental setup. This involved training adversarial examples using the CroPA method and evaluating their transferability across a diverse set of target prompts on the Flamingo VLM. Our evaluation focused on measuring the Targeted Attack Success Rate (ASR) across the following tasks: VQA (both general and specific), image classification, and image captioning.

4.1.1 Experimental Details

Following the experimental protocol outlined in Luo et al. (2024), we generated adversarial examples using CroPA and assessed their effectiveness against a range of target prompts distinct from those used during training. The specific target prompts used in our evaluation are listed in Table 4.1.2. These prompts were selected to represent a broad spectrum of semantic meanings, ensuring a rigorous evaluation of cross-prompt transferability.

4.1.2 Comparative Analysis

Table 4.1.2 presents the Targeted ASRs achieved by the CroPA method on the Flamingo VLM for different target texts. Overall, our results strongly support the claim that CroPA can achieve cross-prompt transferability. We consistently observed non-zero ASRs across all evaluated tasks, indicating that adversarial examples generated using CroPA were effective in misleading the VLM, even when presented with diverse target prompts.

Our experiments successfully reproduce the core finding that CroPA exhibits robust cross-prompt transferability, thereby validating the adversarial vulnerability of VLMs to such attacks.

4.2 Results Beyond the Original Paper

The following subsections detail our extended experiments building upon CroPA's framework and incorporating novel enhancements.

Target Prompt	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Classification	Captioning	Overall
unknown	0.9680	0.9880	0.5940	0.4200	0.7425
I am sorry	0.8620	0.9180	0.6630	0.6400	0.7708
not sure	0.8540	0.8940	0.2420	0.0010	0.4978
very good	0.9620	0.9860	0.6020	0.2020	0.6880
too late	0.9300	0.9580	0.7010	0.1530	0.6855
metaphor	0.9840	0.9940	0.9100	0.5840	0.8680

Table 1: Targeted ASRs tested on Flamingo with different target texts using CroPA.

4.3 ScMix

For the purpose of universal adversarial perturbations resulting in cross-image transferability, we adapt a component of the ETU method proposed in Zhang et al. (2024). In this work, the ETU attack is inspired by adding two variations to a normal VLM adversarial attack, viz. a local utility reinforcement and an augmentation by the name of ScMix. In particular, we only adopt the ScMix augmentation, which increases input diversity using cross and self-mixing strategies between the input images. This is useful in learning adversarial perturbations which are universal across images. The ScMix strategy for data augmentation is described below:

Given two images, say I_1 and I_2, the self-mixing aspect involves extracting two random crops (patches), say x_1 and x_2 from I_1 and resizing them to the size of the original image, X_1 and X_2, respectively. Then, a weighted summation of these gives us the self-mixed image from I_1, $I_1' = \eta X_1 + (1-\eta)X_2$ where η is a random variable such that $\eta \sim \text{Beta}(\alpha, \alpha)$ for some $\alpha > 0$.

Thereafter, cross-mixing is applied by adding I_2 to I_1' in a weighted manner, where I_1' is given a higher weightage to preserve the visual semantics from I_1, to obtain the augmented image as $I_3 = \beta_1 I_1' + \beta_2 I_2$ where $\beta_1 > \beta_2$ and $\beta_1, \beta_2 \in [0, 1)$. Formulaically, for $X_1 = \text{Resize}(\text{RandomCrop}(I_1))$ and $X_2 = \text{Resize}(\text{RandomCrop}(I_1))$,

$$I_3 = \beta_1(\eta X_1 + (1-\eta)X_2) + \beta_2 I_2 \tag{4}$$

where $\eta \sim \text{Beta}(\alpha, \alpha)$ for $\alpha > 0$ and $\beta_1 > \beta_2$ for $\beta_1, \beta_2 \in [0, 1)$

Here, the choice α, β_1 and β_2 acts as a hyperparemeter, for which we have chosen the values $\alpha = 4$, $\beta_1 = 0.9$ and $\beta_2 = 0.1$, which are close to the values given in Zhang et al. (2024).

4.4 Noise Initialization via Vision Encoding Optimization

Recent advances in adversarial attacks on vision-language models (VLMs) have underscored vulnerabilities arising from cross-prompt transferability. In this work, we propose a novel strategy for adversarial perturbation initialization by leveraging diffusion-based semantic anchoring. Instead of employing a conventional random noise initialization, our method synthesizes a target image corresponding to a desired prompt using a state-of-the-art diffusion model, such as Stable Diffusion XL (SDXL). The generated image serves as a semantic anchor for aligning the adversarial example via a basic mean squared error (MSE) loss calculated between the outputs of the VLM's vision encoder. This approach provides a more effective initialization, ensuring that the adversarial perturbations are semantically informed from the start.

4.4.1 Diffusion-Based Target Synthesis

Given a target prompt T, we first generate an image $\mathbf{x}_{\text{target}}$ that embodies the semantic attributes described by T. This is accomplished via a diffusion model, specifically Stable Diffusion XL (SDXL). The generation process can be formalized as follows:

$$\mathbf{x}_{\text{target}} = \mathcal{D}(T, \mathbf{z}; \theta_{\text{SDXL}}), \tag{5}$$

where $\mathcal{D}$ denotes the diffusion process, $\mathbf{z}$ is sampled from a Gaussian distribution, and θ_{SDXL} represents the pre-trained weights of the diffusion model. The resulting image $\mathbf{x}_{\text{target}}$ effectively captures the semantic essence of the prompt T, thus providing an informative basis for initializing adversarial perturbations.

4.4.2 Vision-Encoder Anchored Perturbation

Let $f_v(\cdot)$ denote the vision encoder component of the VLM. Our objective is to craft an adversarial perturbation $\boldsymbol{\delta}$ such that the perturbed image $\mathbf{x}+\boldsymbol{\delta}$ mimics the semantic representation of the target image in the vision encoder's output space. To achieve this, we derive the initial perturbation by minimizing the following objective:

$$\boldsymbol{\delta}_{\text{init}} = \underset{\|\boldsymbol{\delta}\|_\infty \le \epsilon}{\arg\min} \; \|f_v(\mathbf{x}+\boldsymbol{\delta}) - f_v(\mathbf{x}_{\text{target}})\|_2^2 \,, \tag{6}$$

where ϵ is the maximum allowable perturbation (ensuring imperceptibility under an ℓ_∞ constraint). This initialization ensures that the adversarial example starts within a semantically meaningful neighborhood of the target prompt's representation.

4.4.3 Adversarial Optimization via PGD

After initializing the perturbation, we refine the adversarial example using projected gradient descent (PGD). For the t^{th} iteration, the update rule is given by:

$$\mathbf{x}_{\text{adv}}^{(t+1)} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}\Big[\mathbf{x}_{\text{adv}}^{(t)} - \alpha \nabla_{\mathbf{x}_{\text{adv}}^{(t)}} \mathcal{L}_{\text{MSE}}\Big], \tag{7}$$

where α is the step size, and $\Pi_{\mathcal{B}_\epsilon(\mathbf{x})}$ projects the updated input back onto the admissible ℓ_∞ ball around the original image $\mathbf{x}$. The loss function used during optimization is a simple mean squared error (MSE) between the vision encoder outputs:

$$\mathcal{L}_{\text{MSE}} = \|f_v(\mathbf{x}_{\text{adv}}) - f_v(\mathbf{x}_{\text{target}})\|_2^2 \,. \tag{8}$$

This loss ensures that each update incrementally aligns the adversarial example with the target's semantic embedding.

4.4.4 Experimental Considerations

Preliminary experiments performed on models such as BLIP-2 and InstructBLIP suggest that initializing adversarial perturbations with semantically informed noise significantly enhances the attack's efficacy over traditional random initializations while only increasing the computation time for a single image by 20-25 seconds on a single GPU. The diffusion-based approach not only improves alignment in the vision encoder's feature space but also preserves the visual fidelity of the resulting adversarial examples while conforming to strict perturbation budgets.

4.5 Doubly- Universal Adversarial Perturbation

4.5.1 Doubly-UAP (Value Vector)

This incorporaion is adapted from a component of the Doubly-UAP method proposed in Kim et al. (2024) which introduced a UAP by identifying which specific components within the vision encoder's attention mechanism most effectively influence the performance of the VLM. We choose to focus on the two components with the most fundamental roles in the attention mechanism:

1. **Attention Weights**: These control how much each patch should focus on other patches, determining the level of interaction or relevance between patches. We hypothesize that by targeting the attention weights, we can effectively interfere with the encoder's ability to establish these relationships.

2. **Value Vectors**: These hold the actual information within each patch. We expect that perturbing the value vectors will disrupt the essential information content within patches, further impairing the model's interpretative abilities.

Additionally, since the attention mechanism spans multiple layers, we explore whether their impact on LLM output varies across layers viz. Early, Middle and Late. We target the vision encoders within VLM, as they are crucial for visual interpretation. Specifically, we focus on the attention mechanism within the vision encoder, the core process responsible for interpreting visual features. We aim to disrupt this mechanism by targeting its most vulnerable components—the value vectors at the middle-to-late layers—based on prior analysis.

Formally, in the standard Doubly-UAP attack, the perturbation δ^* is obtained as:

$$\delta^* = \arg\max_{\delta} \frac{1}{|L|} \sum_{l \in L} \text{Loss}(V_l(x), V_l(x+\delta)), \tag{9}$$

where $V_l(x)$ represents the value vectors associated with the l-th layer with input image x, and $\text{Loss}(\cdot)$ is the loss function applied to the target vectors.

4.5.2 Our Modified Approach

We adapt this approach by introducing a modified loss function that incorporates a target value vector derived from a reference image corresponding to a desired target text $\mathbf{T}$. Instead of solely maximizing the deviation of value vectors from their original representation, we enforce alignment between the vision encoder's output and a predefined target representation. Our method consists of the following steps:

1. Extract the value vectors from the l-th layer of the vision encoder for both the input image and the target text's associated image.

2. Compute the Language Model Loss or cosine similarity loss between these value vectors.

3. Jointly minimize this loss along with the CroPA loss to ensure adversarial robustness and target alignment.

Formally, let $V_i(x+\delta)$ denote the perturbed value vectors of the i-th attention head for the input image x, and let V_t represent the value vectors of the target text's reference image. We define our value vector loss as:

$$L_{\text{d-UAP}} = \sum_{i=1}^{N} \left\| V_i(x+\delta) - V_t \right\|_2^2, \tag{10}$$

where N is the number of attention heads. This loss encourages the perturbed image's value vectors to closely align with the target text's representation.

We integrate this loss with the CroPA loss to formulate our final objective function:

$$L_{\text{re-CroPA}} = L_{\text{CroPA}}(x_v + \delta_v, x_t + \delta_t, T) + \lambda L_{\text{d-UAP}}(\delta_v) \tag{11}$$

where λ is a hyperparameter controlling the relative importance of the value vector loss.

By jointly optimizing both losses, our approach not only preserves the adversarial nature of the perturbation but also enforces semantic alignment with the target text. Specifically, during optimization, the gradients from both loss components are combined to update the perturbation δ, ensuring that the generated adversarial example exhibits both cross-prompt transferability and guided semantic influence.

This enhancement to the Doubly-UAP framework allows for more precise adversarial manipulation of VLMs, facilitating controlled and interpretable perturbations with applications in adversarial robustness and security analysis.

5 Discussion

Our study aimed to replicate key aspects and findings of Luo et al. (2024) on the Cross-Prompt Attack (CroPA) framework, we have substantiated the original claims regarding the enhancement of cross-prompt adversarial transferability in Vision-Language Models (VLMs). Our experiments corroborate that adversarial examples generated through CroPA maintain their deceptive efficacy across diverse textual prompts, thereby underscoring the robustness of this approach.

Despite these affirmations, the domain of cross-prompt adversarial transferability remains underexplored. The considerable lack of literature in this area highlights a gap in our understanding of prompt-induced vulnerabilities within VLMs. Addressing this gap is imperative, as it holds profound implications for the secure deployment of these models.

From an attacker standpoint, adversarial examples exhibiting high cross-prompt transferability pose significant threats. Such examples can manipulate VLMs to produce malicious or misleading outputs, even when prompted with harmless queries. This capability could be exploited to disseminate false information or to subvert systems dependent on VLMs for content generation and decision-making.

Conversely, from a defensive perspective, the application of imperceptible perturbations offers a novel mechanism to safeguard sensitive information. By embedding these perturbations into images, it is possible to induce VLMs to consistently output predetermined, non-sensitive text, thereby thwarting unauthorized attempts to extract confidential data from personal images. This technique serves as a proactive measure to enhance privacy and data security in an era where visual data is increasingly susceptible to exploitation.

Our study also introduces refinements to the CroPA framework, including improved initialization strategies and an enhanced loss function. These modifications have demonstrated a marked increase in both the Attack Success Rate (ASR) and the generalizability of adversarial examples across different models and images. Such advancements not only reinforce the efficacy of cross-prompt attacks but also pave the way for more resilient defenses against them.

In conclusion, while our reproducibility study affirms the foundational work of Luo et al. (2024), it also accentuates the necessity for deeper investigation into cross-prompt adversarial transferability. A comprehensive understanding of this phenomenon is crucial for developing robust VLMs capable of withstanding adversarial manipulations and for formulating effective countermeasures to protect user data and maintain the integrity of model outputs.

5.1 Limitations and Future Work

While our study successfully reproduces and validates the key findings of Luo et al. (2024) regarding the effectiveness of Cross-Prompt Attack (CroPA), certain aspects of our experiments remain incomplete due to significang computational constraints. Specifically, we were unable to conduct a comprehensive set of ablation studies to systematically analyze the impact of different components within our proposed modifications.

The high cost of generating and evaluating adversarial examples across multiple prompts and models also limited our ability to scale experiments. Despite these constraints, we are committed to completing our ablation studies and additional experiments before the end of the review period. Moving forward, we aim to explore more computationally efficient methods to reduce overhead. Furthermore, we plan to extend our evaluation to a broader set of Vision-Language Models beyond those originally tested as well as more robust attack methodologies for comparison.

5.2 Communication With Original Authors

No direct communication could be established with the original authors during the replication process. The issue raised on their GitHub repository regarding the BLIP-2 and InstructBLIP models remains unresolved at the time of writing.

References

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned?, 2024. URL `https://arxiv.org/abs/2306.15447`.

Wenliang Dai, Junnan Li, Dongxu Li, and Philip Torr. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Devi Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

Hee-Seon Kim, Minbeom Kim, and Changick Kim. Doubly-universal adversarial perturbations: Deceiving vision-language models across both images and text with a single perturbation, 2024. URL `https://arxiv.org/abs/2412.08108`.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models, 2024. URL `https://arxiv.org/abs/2403.09766`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. URL `https://arxiv.org/abs/2306.13213`.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models, 2023. URL `https://arxiv.org/abs/2308.10741`.

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms, 2023. URL `https://arxiv.org/abs/2311.16101`.

Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models, 2022. URL `https://arxiv.org/abs/2206.09391`.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. 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*, pp. 862–871, 2024.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models, 2023. URL `https://arxiv.org/abs/2305.16934`.