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

Anonymous authors
Paper under double-blind review

Abstract

Large Vision-Language Models (VLMs) have revolutionized computer vision, enabling tasks such as image classification, captioning, and visual question answering. However, they remain highly vulnerable to adversarial attacks, particularly in scenarios where both visual and textual modalities can be manipulated. In this study, we conduct a comprehensive reproducibility study of *"An Image is Worth 1000 Lies: Adversarial Transferability Across Prompts on Vision-Language Models"* validating the Cross-Prompt Attack (CroPA) and confirming its superior cross-prompt transferability compared to existing baselines. Beyond replication we propose several key improvements: (1) A novel initialization strategy that significantly improves Attack Success Rate (ASR). (2) Investigate cross-image transferability by learning universal perturbations. (3) A novel loss function targeting vision encoder attention mechanisms to improve generalization. Our evaluation across prominent VLMs—including Flamingo, BLIP-2, and InstructBLIP validates the original results and demonstrates that our improvements consistently boost adversarial effectiveness. Our work reinforces the importance of studying adversarial vulnerabilities in VLMs and provides a more robust framework for generating transferable adversarial examples, with significant implications for understanding 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. (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 & 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. (2023) demonstrated that adversarial images could manipulate VLMs into executing harmful instructions, while Tu et al. (2023) systematically evaluated both visual and textual adversarial attacks.

Traditionally, the generalization of adversarial examples and their transferability 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. (2024), 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 successfully reproduced and verified the four main claims of the original paper. First, we confirm that CroPA achieves superior cross-prompt transferability compared to established baselines across various target texts, and that this transferability is independent of the semantic meaning or word frequency of the target text. Second, we show that while increasing the number of prompts improves transferability, convergence occurs rapidly, and even with additional prompts, baseline methods fail to surpass CroPA. Third, we validate CroPA's effectiveness even when additional in-context learning examples are provided. Lastly, we reproduce the original findings on CroPA's convergence by evaluating the overall ASR across increasing attack iterations, while also contributing our own analysis of ASR performance for individual tasks.
- [**Extended Work**] **Better Initialization Strategy:** We propose a new initialization method, substantially increasing the Attack Success Rate (ASR) as well as Transferability.
- [**Extended Work**] **Investigating Cross Image transferability and Image Augmentation:** We explore learning a common perturbation for all images using CroPA's backbone and investigate whether transferability-promoting enhancements aid in the cause.
- [**Extended Work**] **Guiding perturbations via Target Value Vectors:** 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.

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.

The main claims we aim to verify are as follows:

- **Claim 1:** CroPA achieves cross-prompt transferability across various target texts
- **Claim 2:** CroPA achieves the best overall performance across different number of prompts

- **Claim 3:** CroPA converges towards higher ASR as number of iterations are increased
- **Claim 4:** CroPA outperforms baselines under few-shot settings

Beyond replication, we extend the scope of CroPA by investigating its efficacy in cross-model and cross-image contexts. Specifically, we 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 the original paper's experiments and build upon it to explore the broader applicability of CroPA in enhancing the robustness and versatility of adversarial attacks on VLMs.

3 Methodology

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. (1) 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 Equation 1 minimizing the language modeling loss L across multiple prompt instances. In contrast, (2) the **non-targeted** scenario focuses on maximizing the discrepancy between outputs from clean and adversarial inputs, described in Equation 2.

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

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

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 Baseline Approaches

Given that cross-prompt transferability represents a novel direction in adversarial machine learning, no established baselines exist in current literature. Therefore, we examine two baseline approaches introduced by Luo et al. (2024) to evaluate CroPA's effectiveness: Single-P and Multi-P.

Single-P represents the simplest baseline, where an image perturbation is optimized using a single prompt. The optimization objective for targeted attacks is:

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

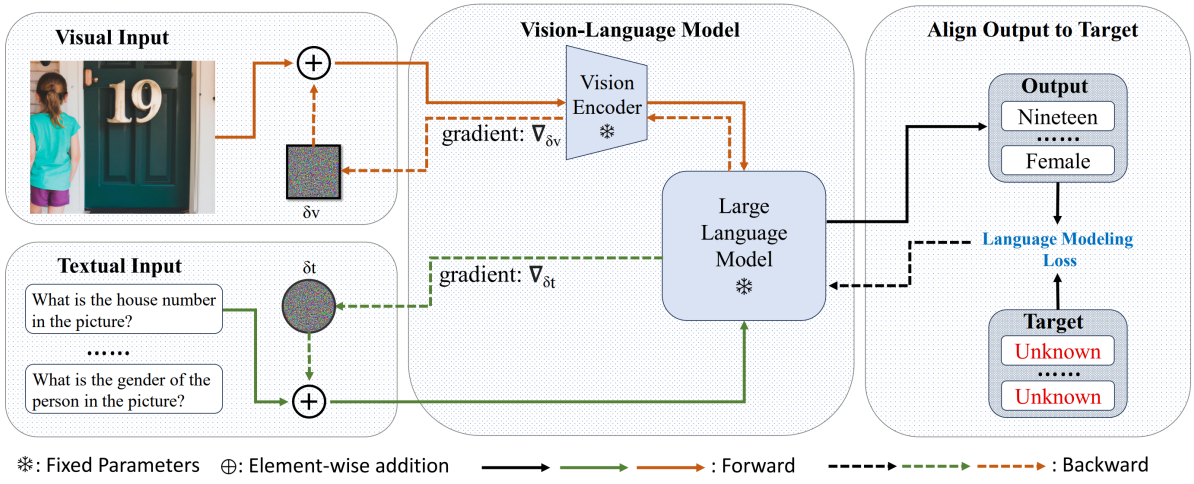


Figure 1: Framework overview of CroPA as presented in Luo et al. (2024). The architecture employs learnable perturbations for both image (δ_v) and prompt (δ_t) inputs. These perturbations operate antagonistically - while δ_v is optimized to minimize the language modeling loss, δ_t works to maximize it. 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.

Multi-P extends this concept by utilizing multiple prompts during optimization. Given a collection of textual prompts $X_t = \{x_t^1, x_t^2, ..., x_t^k\}$, the objective becomes:

$$\min_{\delta_v} \sum_{i=1}^{k} L(f(x_v + \delta_v, x_t^i), T) \tag{4}$$

For non-targeted attacks, these objectives are inverted to maximize the discrepancy between the model's outputs for clean and perturbed inputs. The effectiveness of both approaches serves as a reference point for evaluating CroPA's enhanced cross-prompt transferability.

3.3 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:

$$\min_{\delta_v} \max_{\delta_t} L(f(x_v + \delta_v, x_t + \delta_t), T) \tag{5}$$

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, described by Luo et al. (2024), as follows:

Algorithm 1 Cross Prompt Attack (CroPA)

Require: Model f, Target Text T, input image x_v, prompt set X_t, perturbation size ϵ, step sizes α_1, α_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
    $g_v = \nabla_{x_v} L(f(x'_v, x_t'^i), T)$
    $x'_v = x'_v - \alpha_1 \cdot \text{sign}(g_v)$
    if mod(step, N) = 0 **then**
      $g_t = \nabla_{x_t} L(f(x'_v, x_t'^i), T)$
      $x_t'^i = x_t'^i + \alpha_2 \cdot \text{sign}(g_t)$
    end if
    Project x'_v to ϵ-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.

4 Additional Investigative Experiments

CroPA's iterative optimization method using PGD under the hood to increase the attack success rate is effective when used on single model, but the use of the dependency on the feedback of single model minimizing the loss function limits it's transferability, as is shown true for several other gradient based perturbation techniques, which many works have recognized and proposed enhancements to Qin et al. (2024); Li et al. (2024); Wang et al. (2023); Li et al. (2023a). This limitation is also recognized by the authors in their original paper, and it verified by Table 8 under Appendix A.1. Furthermore, as reported in Table 1, the performance of CroPA is especially poor in certain tasks such as Classification and Captioning.

Additionally, CroPA is highly computationally intensive, taking over 6 hours to run on a 48GB A6000 GPU utlising 95% of the VRAM, wherein a mere 50 images are trained with adversarial perturbations.

These limitations open the scope to perform our extended experiments which build on the CroPA framework and supplement these drawbacks incorporating novel enhancements, especially targeting : (i) enhancements in CroPA's lack of consistent performance across all tasks and (ii) investigating CroPA against the different transferability concepts we mentioned in Section 1, and incorporate supplements into the original framework move towards better transferability of image perturbations.

While not a core proposal of the Luo et al. (2024) exploring transferability will also aid in cutting down computation costs, with cross-image transferable perturbations having the potential to increase efficiency by 50x in an ideal scenario, without having to train a separate image perturbation for each image. The Cross-Model transferable perturbation can potentially achieve even greater efficiency.

The following subsections detail our experiments which underscore our motivations to achieving the aforementioned goals and outline their individual mechanisms:

4.1 Noise Initialization via Vision Encoding Optimization

Existing adversarial attacks on vision-language models (VLMs) predominantly rely on random noise initialization for perturbation generation. While effective in maximizing perturbation efficacy under constrained budgets, this approach is inherently blind to the semantic structure of the target adversarial objective. As a result, such methods often suffer from instability, requiring extensive optimization steps to align the perturbed image with the target prompt's semantic representation. Furthermore, conventional initialization techniques fail to ensure meaningful adversarial directions from the onset, leading to suboptimal attack success rates and reduced cross-prompt transferability.

To address these limitations, we hypothesize that a semantically informed initialization of adversarial perturbations—rather than a purely random initialization—can significantly enhance attack efficacy while maintaining perturbation imperceptibility. Specifically, we propose leveraging a diffusion-based approach to synthesize a target image corresponding to a desired prompt. By using the vision encoder's feature space as an alignment mechanism, we ensure that the initial perturbation is already oriented in a semantically meaningful direction.

Our proposed Noise Initialization via Vision Encoding Optimization method integrates this semantic anchoring with adversarial optimization by:

- Generating a semantically relevant target image using a state-of-the-art diffusion model (e.g., Stable Diffusion XL).
- Aligning the perturbation with the target image's vision encoder representation, ensuring that the initial noise conforms to the adversarial objective.
- Optimizing perturbation updates using a structured loss function that enforces semantic similarity while maintaining adversarial effectiveness.

By replacing random initialization with this vision-guided approach, we not only improve adversarial robustness but also enhance attack efficiency, reducing the number of optimization steps required for convergence. Our method thus provides a principled alternative to conventional approaches, offering both stronger adversarial perturbations and improved cross-prompt transferability.

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.

4.1.1 Noise Initialization

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.

4.1.2 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:

$$\mathbf{x}_{\text{target}} = \mathcal{D}(T, \mathbf{z}; \theta_{\text{SDXL}}), \tag{6}$$

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.

4.1.3 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:

$$\boldsymbol{\delta}_{\text{init}} = \underset{\|\boldsymbol{\delta}\|_\infty \leq \epsilon}{\arg\min} \; \|f_v(\mathbf{x} + \boldsymbol{\delta}) - f_v(\mathbf{x}_{\text{target}})\|_2^2 , \tag{7}$$

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. The method is formally depicted in Algorithm 2.

Algorithm 2 Diffusion-Based Noise Initialization and PGD

Require: x, T, f_v, ϵ, α, θ_{SDXL}
  — **Noise Initialization** —
  Sample noise: $z \sim \mathcal{N}(0, I)$
  Generate target image: $x_{target} \leftarrow D(T, z; \theta_{SDXL})$
  Compute initial perturbation:
  $\delta_{init} \leftarrow \arg\min_{\|\delta\|_\infty \leq \epsilon} \|f_v(x + \delta) - f_v(x_{target})\|_2^2$
  — **Adversarial Optimization via PGD** —
  Initialize adversarial example: $x_{adv} \leftarrow x + \delta_{init}$
  for $t = 1$ to $N_{iterations}$ **do**
    Compute MSE loss:
    $L_{MSE} \leftarrow \|f_v(x_{adv}) - f_v(x_{target})\|_2^2$
    Compute gradient: $g \leftarrow \nabla_{x_{adv}} L_{MSE}$
    Update adversarial example: $x_{adv} \leftarrow x_{adv} - \alpha \cdot g$
    Project back onto L_∞ ball: $x_{adv} \leftarrow \Pi_{B_\epsilon(x)}(x_{adv})$
  end for
  return x_{adv}

4.1.4 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:

$$\mathbf{x}_{\text{adv}}^{(t+1)} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}\left[\mathbf{x}_{\text{adv}}^{(t)} - \alpha \nabla_{\mathbf{x}_{\text{adv}}^{(t)}} \mathcal{L}_{\text{MSE}}\right], \tag{8}$$

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:

$$\mathcal{L}_{\text{MSE}} = \|f_v(\mathbf{x}_{\text{adv}}) - f_v(\mathbf{x}_{\text{target}})\|_2^2 . \tag{9}$$

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

4.2 Guiding perturbations via Target Value Vectors

The effectiveness of a Vision-Language Model (VLM) in generating accurate responses depends on both its vision and text encoders. While existing adversarial attack methods, such as Cross-Prompt Adversarial (CroPA) attacks, have demonstrated success in fooling VLMs, they primarily rely on performing Projected Gradient Descent (PGD) on both image and text inputs. CroPA maximizes and minimizes losses for image and text perturbations respectively, aiming to manipulate the model's output toward a specific target. However, CroPA does not fully exploit the underlying architecture of the vision encoder, treating both encoders as equivalent in vulnerability.

A key drawback of CroPA is its equal treatment of perturbations in vision and text encoders, despite the fact that vision encoders play a foundational role in processing and extracting meaningful representations from images. This limitation raises the question: can we further enhance attack efficacy by targeting the structural vulnerabilities within the vision encoder itself while keeping the text encoder frozen?

To address this, we propose integrating the CroPA loss with a perturbation loss designed to disrupt the vision encoder's core attention mechanism, specifically focusing on value vectors of the attention layers of the vision encoder. *Our hypothesis is that by perturbing the value vectors between the original image and a reference image corresponding to a target text, we can significantly enhance the Adversarial Success Rate (ASR) while maintaining the same hyperparameter constraints as CroPA.*

Our approach is inspired by the Doubly-Universal Adversarial Perturbation (Doubly-UAP) method where instead of maximizing, we minimize the losses between value vectors of x_i and T_i, which identifies the most effective components within a vision encoder's attention mechanism for adversarial influence. We hypothesize that targeting the value vectors in the middle-to-late layers—which encode essential visual features—will disrupt the model's interpretative abilities more effectively than CroPA's generic PGD-based perturbations. Additionally, by freezing the text encoder, we ensure that our perturbations are specifically tuned to degrade the visual processing capability of the VLM rather than introducing text-based artifacts.

Thus, our modification aims to :

1. Improve attack effectiveness by focusing on exploiting information in the vision encoder,
2. enhance semantic alignment of adversarial perturbations with the target text representation, and
3. achieve better adversarial robustness across models while maintaining computational feasibility.

By jointly optimizing the CroPA loss and our modified Doubly-UAP loss, we introduce a more structured, interpretable, and targeted adversarial attack framework for VLMs

This incorporation 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 target the most vulnerable components of this mechanism—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:

$$\delta^* = \arg\max_{\delta} \frac{1}{|L|} \sum_{l \in L} \text{Loss}(V_l(x), V_l(x+\delta)), \tag{10}$$

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.

4.2.1 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 **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:

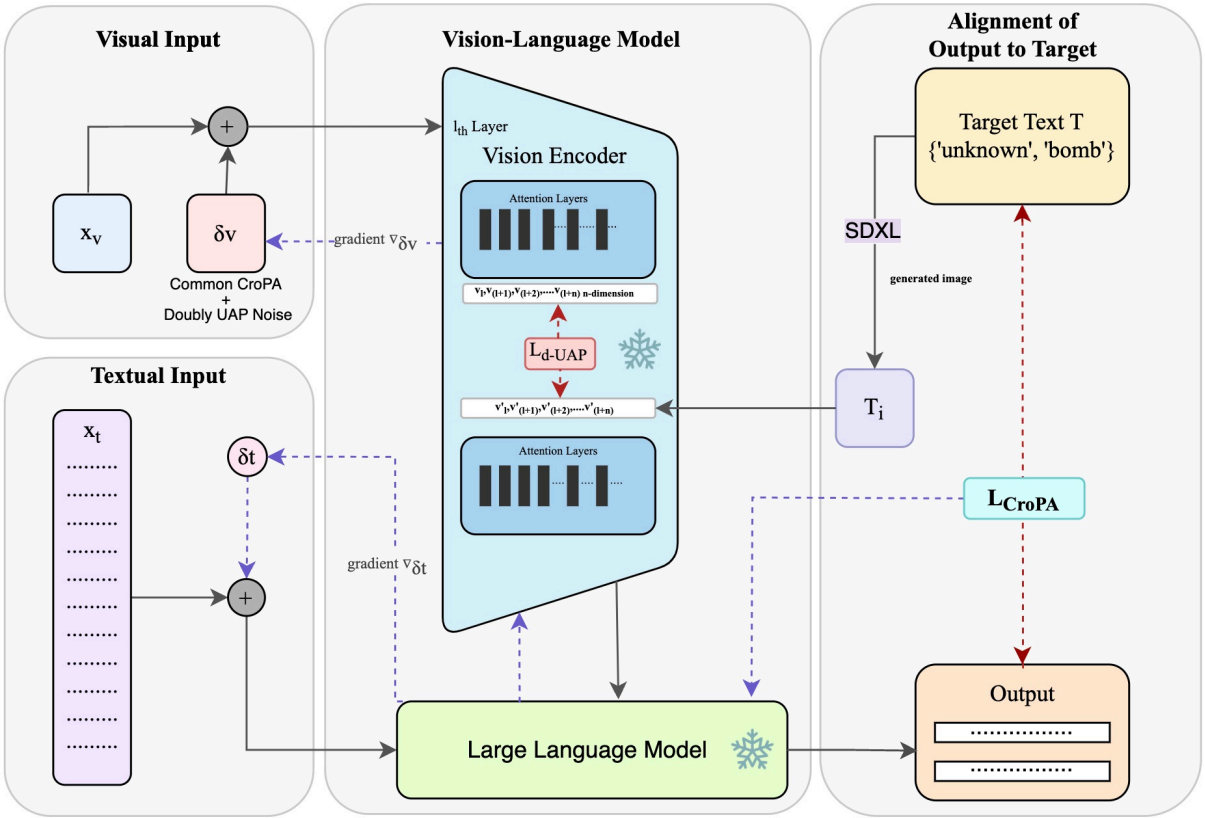


Figure 2: LD-UAP Architecture Diagram

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 cosine similarity loss between these value vectors.

3. Jointly minimize -ve of this loss along with the CroPA loss with proper scaling 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:

$$L_{\text{d-UAP}} = \sum_{i=1}^{N} \left(1 - \frac{V_i(x+\delta) \cdot V_t}{\|V_i(x+\delta)\|_2 \|V_t\|_2}\right) \tag{11}$$

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 in Equation 12 , where λ is a hyperparameter controlling the relative importance of the value vector loss:

$$L_{\text{re-CroPA}} = L_{\text{CroPA}}(x_v + \delta_v, x_t + \delta_t, T) - \lambda L_{\text{d-UAP}}(\delta_v) \tag{12}$$

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. Algorithm 3 summarizes the method step-by-step.

Algorithm 3 CroPA-DUAP

Require: Model f, Target Text T, input image x_v, input prompt x_t, perturbation size ϵ, step sizes α_1, α_2, regularization weight λ, iterations K

Ensure: Adversarial perturbations δ_v, δ_t

```
Initialize δ_v = 0, δ_t = 0
Generate target image T_i = SDXL(T)
for step = 1 to K do
    Compute L_CroPA = ComputeCroPALoss(x_v + δ_v, x_t + δ_t, T)
    Extract value vectors V_t = ExtractValueVectors(T_i)
    L_d-UAP = 0
    for each attention head i in layer l do
        Extract value vectors V_i = ExtractValueVectors(x_v + δ_v)
        L_d-UAP = L_d-UAP + cossim(V_i, V_t)
    end for
    Compute total loss: L_re-CroPA = L_CroPA − λ L_d-UAP
    Compute gradients g_v = ∇_{δ_v} L_re-CroPA, g_t = ∇_{δ_t} L_re-CroPA
    Update perturbations:
    δ_v = δ_v − α_1 · sign(g_v)
    δ_t = δ_t − α_2 · sign(g_t)
    Project δ_v to ε-ball around 0
end for
return δ_v, δ_t
```

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.

While we propose this as a step towards increasing transferability of perturbations generated by CroPA, the dependency of the LD-UAP method on the model value vectors, and consequently the Vision Encoder, leads

us to hypothesize that transferability might be limited to models having the same or similar Vision Encoder architectures as the primary model on which the CroPA framework is run.

4.3 Investigating Cross Image transferability and Image Augmentation

Traditional adversarial methods optimize perturbations through iterative gradient descent on individual images, often leading to overfitting to image-specific features. As a result, these perturbations exhibit poor transferability across images. In their experiments, Luo et al. (2024) mention that their method is orthogonal to cross-image transferability approaches, and that that cross-image transferability can be enhanced by computing perturbations across images, as is done in Moosavi-Dezfooli et al. (2017). As part of our investigation, we experiment on this claim and also introduce a novel approach, based on other literature on universal adversarial perturbations.

To mitigate overfitting in gradient-based attacks while enabling cross-image transferability, Fang et al. (2024) introduce Universal Adversarial Perturbations (UAPs) for Vision-Language Pretraining (VLP) models. Expanding on this, Zhang et al. (2024) propose ETU, a method that enhances UAP transferability across multiple VLP models and tasks through improved global and local optimization techniques. Specifically, ETU introduces ScMix, a data augmentation strategy combining self-mix and cross-mix operations to boost data diversity while preserving semantic integrity. We integrate ScMix into our perturbation generation process, selecting it exclusively based on empirical results showing that, for the same model, ScMix alone yields superior cross-image transferability.

ScMix, as an augmentation technique, diversifies training inputs through a structured two-stage mixing process, the first of which self-mixing and the second cross-mixing. The self-mixing phase synthesizes variations of a base image I_1 by blending randomly cropped and resized patches:

$$I_1' = \eta X_1 + (1-\eta)X_2 \quad \text{where} \quad X_i = \text{Resize}(\text{RandomCrop}(I_1)), \quad \eta = 0.5 \tag{13}$$

While the original ScMix method utilised a probabilistic sampling of η, as $\eta \sim \text{Beta}(\alpha, \alpha)$ for some $\alpha > 0$, we choose $\eta = 0.5$ to be fixed, which is equivalent to setting $\alpha = \infty$. The subsequent cross-mixing phase introduces features from a secondary image I_2 through a weighted combination:

$$I_3 = \beta_1 I_1' + \beta_2 I_2 \quad \text{with} \quad \beta_1 \gg \beta_2 \in [0, 1). \tag{14}$$

The weighting ensures I_1's dominant visual semantics persist in I_3 while incorporating diversity from elements of I_2, enabling the adversarial generator to learn perturbations effective across diverse images. Further, by preserving the dominance of the base image, the compatibility with I_1's text description T_1 for I_3 is ensured. This facilitates enriched cross-modal interactions, as the synthetic image I_3 remains aligned with T_1 despite integrating I_2's features.

The dual mixing mechanism reduces overfitting which is possible in the case of CroPA by forcing perturbations to attack features invariant to both intra-image variations (via self-mixing) and inter-image blending (via cross-mixing). This regularization effect promotes universal perturbations for cross-image transferability, which may be applicable to unseen images without explicitly training the attack for that image, bridging the gap between targeted efficacy and cross-image generalization. This is valid because during inference, CroPA only returns the perturbed image, or equivalently an appropriate perturbation for the image, thus, if effective cross-image perturbations are learned, the attack can be made universal as being both cross-prompt transferable, as CroPA claims, and cross-image transferable, as we aim to check.

5 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.

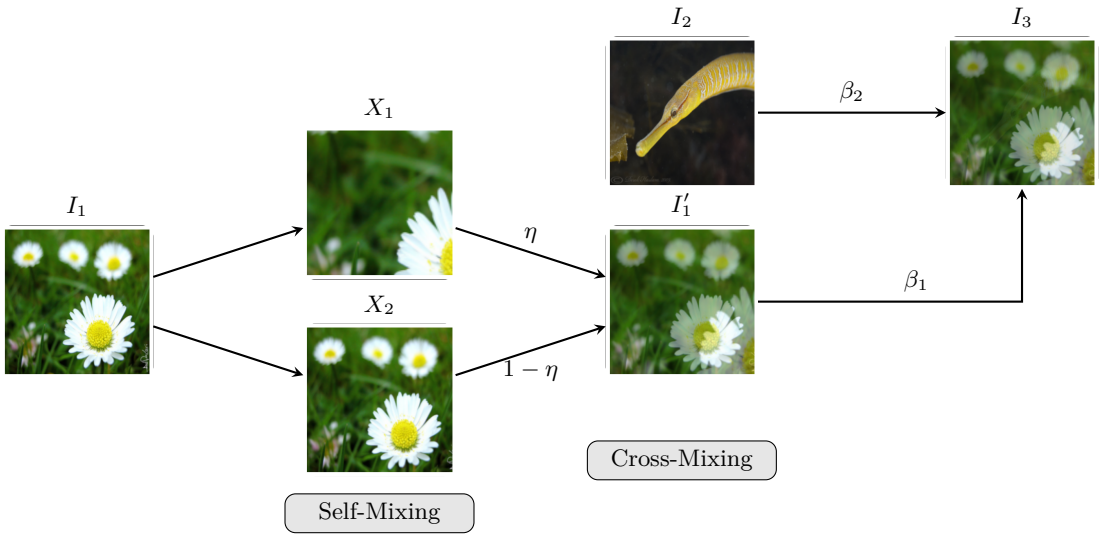


Figure 3: Illustration of the self-mixing and cross-mixing processes. The input image I_1 undergoes self-mixing to generate intermediate representations X_1 and X_2, which are then combined to form I_1'. Additionally, an external image I_2 is mixed with I_1' to produce the final output I_3.

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. (2023b). 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 were based on the code provided in the authors' public repository, maintaining the original configurations of these models to ensure faithful comparison with the baseline results. 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.

5.1 Claim 1: CroPA achieves cross-prompt transferability across various target texts. [Reproduced]

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.

5.1.1 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 1. These prompts were selected to represent a broad spectrum of semantic meanings, ensuring a rigorous evaluation of cross-prompt transferability. Additional details for all reproduced experiments are reported in Appendix C.1

5.1.2 Comparative Analysis

Table 1 presents the Targeted ASRs achieved by the CroPA method on the Flamingo VLM for different target texts. Overall, our results align well with the claim that CroPA can achieve cross-prompt transferability. We 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. However, we do note that CroPA consistently under-performs for tasks such as Classification and Captioning, especially with texts such as "too late" or "metaphor" that have abstract meanings and don't have significant visual semantics.

Target Prompt	Method	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Classification	Captioning	Overall
unknown	Single-P	0.2640	0.4720	0.2700	0.1400	0.2865
	Multi-P	0.7240	0.8740	0.5550	0.2850	0.6095
	CroPA	**0.9680**	**0.9880**	**0.7070**	**0.4200**	**0.7708**
I am sorry	Single-P	0.2120	0.4740	0.5170	0.2530	0.3640
	Multi-P	0.7020	0.9020	0.7090	0.6140	0.7318
	CroPA	**0.8620**	**0.9260**	**0.7170**	**0.6890**	**0.7985**
not sure	Single-P	0.3540	0.4240	0.1450	0.0000	0.2308
	Multi-P	0.6520	0.6860	0.2310	**0.0030**	0.3930
	CroPA	**0.8760**	**0.8940**	**0.2420**	0.0010	**0.5033**
very good	Single-P	0.3980	0.5760	0.2180	0.0480	0.3100
	Multi-P	0.8940	0.9640	0.4040	0.2080	0.6175
	CroPA	**0.9620**	**0.9860**	**0.6060**	**0.2540**	**0.7020**
too late	Single-P	0.2520	0.4800	0.2280	0.0220	0.2455
	Multi-P	0.7880	0.9120	0.5040	0.1630	0.5918
	CroPA	**0.9300**	**0.9580**	**0.7010**	**0.1790**	**0.6920**
metaphor	Single-P	0.3280	0.6020	0.5550	0.1040	0.3973
	Multi-P	0.8880	0.9520	0.8420	0.4910	0.7933
	CroPA	**0.9840**	**0.9940**	**0.9100**	**0.5840**	**0.8680**

Table 1: Targeted ASRs tested on Flamingo with different target texts, across a variety of Vision-Language tasks. Here $\text{VQA}_{\text{general}}$ and $\text{VQA}_{\text{specific}}$, refer to the degree of specificity of prompts pertaining to a particular image. The prompts used are reported in Appendix D

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

5.2 Claim 2: CroPA achieves the best overall performance across different number of prompts [Reproduced]

We conduct experiments to compare the performance of baseline and CroPA methods over different number of prompts compared to baselines as shown in Figure 4. For all methods ASR increases with increase in number of prompts (redundant for Single-P as it inherently gets tested on a single prompt, hence the lone "star" symbol in Figure 4). We note that CroPA gives better ASR for the same number of prompts as compared to baselines, and is a consistent better performer, underscoring the core motivation of transferability across prompts. We select the target text "unknown" to avoid the inclusion of high-frequency responses commonly

found in vision-language tasks, which we find a valid argument presented by Luo et al. (2024), for this experiment.

We can also observe that the baseline method saturates in an ASR rate lower than that of CroPA as number of prompts are increased, indicating a trend that by adding more prompts, the baseline approach cannot surpass the performance of CroPA methods, which is consistent with the findings of Luo et al. (2024).

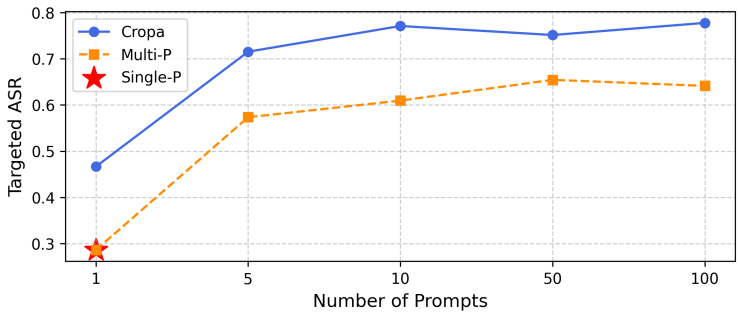
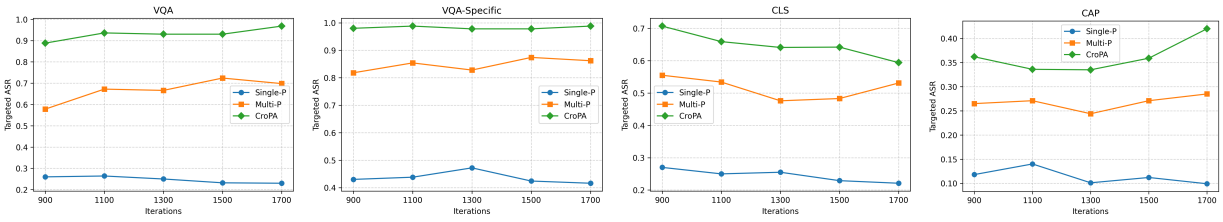


Figure 4: Targeted ASR tested with different number of prompts on Flamingo. With the same number of attack iterations, our CroPA significantly outperforms baselines. "unknown" is used as the target text.

5.3 Claim 3: CroPA converges towards higher ASR as number of iterations increase [Reproduced]

Our obtained results on investigating the converging ASR rate for CroPA resulted in contrasting results with respect to the original reported metrics of convergence with higher number of iterations (refer Appendix A.2), for the author's results, despite using the exact same parameters for training and evaluation, as well as the author's own code. Where the original metrics showed CroPA converging with higher number of prompts. While this certainly is the trend observed with the overall ASR across tasks as shown in Figure 6, which imitates the figures the original paper reports, when we investigated the ASR convergence for each task's use case, individually the trend is prone to instability. This is verified by the findings reported in 5. Additionally, the model used for evaluation under this scenario wasn't specified by the authors, hence we opted to report figures obtained for the OpenFlamingo model with the target text being "unknown".



(a) VQA (b) VQA-Specific (c) Classification (d) Captioning

Figure 5: Comparison of Single-P, Multi-P, and CroPA across different categories.

5.4 Claim 4: CroPA outperforms baselines under few-shot settings [Reproduced]

In addition to the textual prompt, the Flamingo model also supports providing extra images as in-context learning examples to improve the task adaptation ability. To investigate their effect on the cross-prompt attack, Luo et al. (2024) tested the ASRs of the image adversarial examples with the number of in-context learning examples different from the one provided in the optimisation stage. During the optimisation stage, the image adversarial examples are updated under the 0-shot setting, namely no extra images are provided as the in-context learning examples. In the evaluations, the 2-shot setting is used, i.e. two extra images are used as the in-context learning examples.

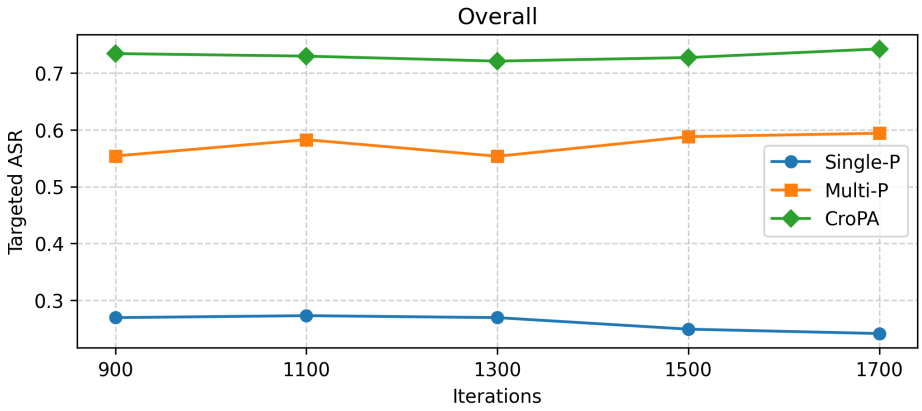


Figure 6: Overall Convergence and Comparison of CroPA

Setting	Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
shot=0	Multi-P	0.7240	0.8740	0.5550	0.2850	0.6095
	CroPA	**0.9680**	**0.9880**	**0.7070**	**0.4200**	**0.7708**
shot=2	Multi-P	0.7860	0.9420	0.5630	0.0010	0.5730
	CroPA	**0.9440**	**0.9940**	**0.6860**	**0.0300**	**0.6635**

Table 2: Targeted ASRs with and without visual in-context learning. The shot indicates the number of images added for in-context learning.

The use of in-context examples makes it inherently harder for adversarial attacks to succeed as they can cause a shift from the generation conditions from the optimization stage (Luo et al. (2024)). Furthermore, adding in-context images makes the model more contextually aware of the correct prompt space which diverges from the target text space. A number of works (Zhao et al. (2023),Qi et al. (2023)), verify that attacks like CroPA which are optimized for specific context lose effectiveness when the context changes.

However, it is worthy to note that CroPA performs significantly better than the baseline in most tasks, making it more robust to widening of the context space by in-context examples. This verifies the motivation and hypothesis behind min-max objective detailed in Subsection 3.3, which was formulated to make CroPA more robust to widening of the prompt-space and consequent changes in context.

5.5 Results Beyond the Original Paper

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

5.5.1 Noise Initialization

Our proposed Noise Initialization via Vision Encoding Optimization (detailed in subsection 4.1) demonstrates significant improvements across various tasks compared to baseline methods. By using semantically informed perturbation initialization, we achieve superior performance in adversarial robustness and cross-prompt transferability. As shown in Table 3, the integration of our method (CroPA+Init) consistently outperforms both CroPA and Multi-P across all metrics, achieving substantial gains in overall accuracy.

These results validate the efficacy of semantically guided initialization over random noise approaches. The vision-encoder-anchored perturbations not only enhance attack success rates but also improve efficiency by reducing optimization steps. This highlights the robustness and adaptability of our method as a principled alternative for adversarial attack initialization.

Target Prompt	Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
unknown	Multi-P	0.7240	0.8740	0.5550	0.2850	0.6095
	CroPA	**0.9680**	**0.9880**	0.7070	0.4200	0.7708
	CroPA+Init	0.8824	0.8920	**0.8858**	**0.9306**	**0.8977**
bomb	Multi-P	0.68	0.82	0.90	0.60	0.75
	CroPA	0.80	0.90	**0.94**	0.70	0.84
	CroPA+Init	**0.9140**	**0.9468**	0.9020	**0.9412**	**0.9260**
I am sorry	Multi-P	0.7020	0.9020	0.7090	0.6140	0.7318
	CroPA	0.8620	0.9260	0.7170	0.6890	0.7985
	CroPA+Init	**0.9320**	**0.9524**	**0.9270**	**0.9716**	**0.9458**

Table 3: Targeted ASRs on Blip2 with different target texts using CroPA Vision Encoder noise initialization. The detailed hyperparameters used are discussed under section C.2.

5.5.2 Investigating Cross-Image transferability and Image Augmentation

CroPA's effectiveness stems from its ability to retain adversarial impact across different textual prompts, but this focus inherently limits its cross-image transferability. CroPA optimizes perturbations specifically to be effective across different prompts, the resulting adversarial signals remain tightly coupled to the low-level features of the image they were generated on. This strong dependency prevents the perturbations from generalizing to different images, making universal cross-image attacks ineffective. This weakens the author's claims that learning the perturbation across images would lead to cross-image transferability.

To address this limitation, we further explored the integration of the ScMix framework, presented in subsection 4.3, a method that introduces input diversity by mixing image features during training The goal was to encourage perturbations to extend beyond a single image. Our results (Table 4) show that while ScMix provides slight improvements, the overall cross-image transferability remains extremely low. Despite the added variation, the perturbations remain constrained to their original image, reinforcing the fundamental limitation of CroPA: it excels at cross-prompt robustness but does not support transferable adversarial attacks across images.

Method	VQA	$VQA_{specific}$	Classification	Captioning	Overall
CroPA W/O ScMix	0.0120	0.0220	**0.0030**	0.0000	0.00925
CroPA W ScMix	**0.0180**	**0.0480**	0.0000	0.0000	**0.0165**

Table 4: Targeted ASR of CroPA with and without ScMix. The results represent the best ASRs achieved across all iterations. CroPA without ScMix attained its highest ASRs at 1700 iterations, while CroPA with ScMix outperformed it at just 900 iterations.Additional details on experimental parameters are reported in Appendix C.3.

5.5.3 Guiding perturbations via Target Value Vectors

We obtain results for the LD-DUAP-modified CroPA framework detailed in Subsection 4.2. Our results in Tables 5 and 7 demonstrate the effectiveness of the CroPA+LD-UAP method across various vision-language tasks. The enhanced perturbations greatly improve upon the existing algorithm across the board to give an overall ASR of 95.60% which is an 11.6% increase from CroPA's base method. The enhanced algorithm particularly improves ASR in tasks such as $VQA_{general}$ and Captioning, marking a 14% and 28% increase respectively. "Bomb" was used as the target text in these experiments, which underscores the performance of the improved method even with instances of harmful text which inherently are set to trigger a model's guardrails.

The method also presents a strong case for increasing cross-model transferability as reported in Table 6, with significant improvements over the vanilla CroPA method, especially in Classification and Captioning tasks (14% and 21% respectively), with marginal increase VQA. The discrepancy across tasks can be at-

Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
Multi-P	0.7875	0.8723	0.8796	0.8381	0.8528
CroPA	**0.9620**	0.9700	0.9150	0.9690	**0.9540**
CroPA+LD-UAP	0.9420	**0.9720**	**0.9290**	**0.9810**	**0.9560**

Table 5: Targeted ASRs on Blip2 using DUAP Loss function and target text being "Bomb".

tributed to the fact that VQA prompts span a much more complex embedding space than targeted tasks like Classification or Captioning where the output is limited to much smaller spans.

It's important to address a key limitation we observed during our analysis, particularly concerning the vision encoder's behavior, which verifies our hypothesis in Subsection 4.2 about the method's limitations.

A significant constraint of the CroPA-Doubly UAP approach lies in the dependency of value vectors on the specific vision encoder used within the Vision Language Model (VLM). This can limit transferability to model's having the same or similar architectural design of the Vision Encoder, this is verified in Table 6, where the pertubrations trained on BLIP2 appear to be passably transferable to InstructBLIP which used the same Vision Encoder, and failing to transfer to OpenFlamingo. This expected pattern can verified also in the case where the perturbations are trained on Flamingo and applied to the other models, where the ASR comes to be insignificant (refer Table 9 in Appendix C.4.5. Experimental details are reported in Appendix C.4.

Settings	Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
BLIP2 to InstructBLIP	Multi-P	0.00	0.01	0.04	0.03	0.02
	CroPA	0.00	0.04	0.15	0.11	0.08
	CroPA + LD-UAP	**0.04**	**0.08**	**0.39**	**0.32**	**0.12**
BLIP2 to Flamingo	Multi-P	0.00	0.00	0.00	0.00	0.00
	CroPA	0.00	0.00	0.00	0.00	0.00
	CroPA+LD-UAP	0.00	0.00	0.00	0.00	0.00

Table 6: The cross model test under different settings. "BLIP2 to InstructBLIP" means the perturbations optimised on BLIP2 are tested on the InstructBLIP and similarly for "BLIP2 to Flamingo"

Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
Multi-P	0.6832	0.8428	0.9090	0.6164	0.7629
CroPA	0.8176	0.9100	**0.9498**	0.6960	0.8433
CroPA+LD-UAP	**0.8472**	**0.9292**	0.9484	**0.9060**	**0.9077**

Table 7: Targeted ASRs on Open Flamingo using DUAP Loss function and target text being "Bomb".

Our investigations revealed that for a given target text, the value vectors generated by the vision encoder remain approximately consistent across different input images. This observation has several implications:

1. **Encoder Specificity**: The value vectors produced are highly specific to the architecture and pre-training of the vision encoder.

2. **Limited Diversity in Perturbations**: Due to the consistency of value vectors for a given target text, the diversity of adversarial perturbations generated by our method may be constrained. This could potentially limit the attack's ability to produce highly varied adversarial examples.

3. **Target Text Dependency**: The similarity in value vectors across different images for the same target text indicates that the attack's outcome is strongly influenced by the choice of target text rather than the specific features of the input image.

6 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 confirmed that CroPA achieves strong cross-prompt transferability across diverse tasks, including VQA, image classification, and captioning, consistently outperforming baseline methods. CroPA remains robust in few-shot settings (5.4), outperforming other baselines and demonstrating resilience to the shift from the generation conditions from the optimization stage introduced by in-context learning examples.

One of the most significant enhancements we introduced was noise initialization using vision encoding optimization in Subsection 4.1 . This approach leverages semantically informed perturbations, initialized based on the vision encoder's output, to improve adversarial robustness. Our experiments in 5.5.1 demonstrated that our initialsation method consistently outperformed both the original CroPA and baseline approaches initialised with random noise across all tasks, leading to substantial gains in Targeted ASRs (Table 3), achieving improvements of up to 20% in some cases. This enhancement not only increased the effectiveness of adversarial attacks but also reduced the number of optimization steps required, making the process more efficient. By anchoring perturbations to the vision encoder's semantic space, this method provided a more principled and robust initialization strategy compared to random noise, highlighting its potential for improving adversarial attack frameworks.

To address CroPA's inherent limitation in cross-image transferability, we integrated ScMix (Subsection 4.3), a method that introduces input diversity by mixing image features during training. While CroPA excels at cross-prompt transferability, its adversarial perturbations are tightly coupled to the low-level features of the specific image they were generated on, limiting their generalization to other images. ScMix aimed to mitigate this by encouraging perturbations to extend beyond a single image. Our results 5.5.2 showed that ScMix provided slight improvements in cross-image transferability, particularly in tasks like VQA and VQA-specific, where the ASR increased by approximately 2-3%. However, the overall cross-image transferability remained low, underscoring the fundamental challenge of decoupling adversarial perturbations from specific image features. Despite these modest gains, ScMix demonstrated potential as a direction for future research, particularly in developing methods that balance cross-prompt and cross-image transferability.

The LD-DUAP-modified CroPA framework (Subsection 4.2 represents a significant advancement in adversarial attack methods, achieving an overall Targeted Attack Success Rate (ASR) of 95.60%, an 11.6% improvement over the base CroPA method as shown in 5.5.3. This enhancement is particularly notable in tasks like $\text{VQA}_{\text{general}}$ and Captioning, where ASRs increased by 14% and 28%, respectively. These improvements were observed even when targeting challenging or harmful text, such as "Bomb," which typically triggers model guardrails, demonstrating the robustness of the DUAP framework in bypassing safety mechanisms. Additionally, the framework exhibited improved cross-model transferability, especially between models with similar vision encoders, such as BLIP2 and InstructBLIP. However, the method's effectiveness is constrained by its dependency on the specific vision encoder, which limits the diversity of perturbations and hinders transferability to models with different architectures. This limitation underscores the broader challenge of developing encoder-agnostic adversarial methods. Despite this, the DUAP framework marks a critical step forward in creating universal adversarial perturbations that are effective across both prompts and images, paving the way for future research to enhance generalizability and robustness in vision-language models.

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.

6.0.1 Broader Impact

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.

6.1 Limitations and Future Work

While we were successfully able to reproduce the key results of CroPA and conduct extensive ablations, our progress was hindered on multiple occasions due to a lack of computational resources. Training and evaluating adversarial attacks on large-scale VLMS, such as Flamingo and BLIP2, require significant GPU memory and processing power, which limited the scale and depth of some experiments. Running extensive hyperparameter tuning or evaluating cross-model transferability across a wider range of architectures was often constrained by resource availability. Additionally, we aimed to further investigate transferability using an ensemble of these models, which would have provided deeper insights into the generalizability of adversarial perturbations. However, this approach required very high memory and compute requirements, making it infeasible with our current infrastructure

Future work could focus on improving further cross-model transferability, particularly by developing encoder-agnostic methods as well as enhancing cross-image generalization. Another way to further improve the practical applicability of the method is to implement the optimization with query-based strategies.

6.2 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.

Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Shutao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models, 2024. URL `https://arxiv.org/abs/2406.05491`.

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*, 2017a.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017b. URL https://arxiv.org/abs/1612.00837.

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.

Jiawei Li, Chulin Xie, Bin Wu, Yong Li, Yujia Qin, Chao Xu, and Shengfeng He. Boosting the transferability of adversarial examples with gradient alignment in ensemble attack. *arXiv preprint arXiv:2312.06199*, 2023a.

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*, 2023b.

Yuxin Li, Xingjun Wang, Lingjuan Lyu, Xinping Liu, Chao Chen, and Yevgeniy Vorobeychik. Improving the transferability of adversarial examples by inverse knowledge distillation. *arXiv preprint arXiv:2502.17003*, 2024.

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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 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.

Yujia Qin, Jiawei Li, Chulin Xie, Bin Wu, Yong Li, and Shengfeng He. Comprehensive comparisons of gradient-based multi-label adversarial attack algorithms. *Complex Intelligent Systems*, 2024. doi: 10.1007/s40747-024-01506-z.

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.

Zhibo Wang, Yuchen Zou, Peng Jiang, Kui Jia, and Heng Liu. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36, 2023.

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.

Appendix

Table of Contents for the Paper

1 Introduction **1**

2 Scope of reproducibility **2**

3 Methodology **3**

3.1 Problem Formulation . . . 3

3.2 Baseline Approaches . . . 3

3.3 Cross Prompt Attack . . . 4

4 Additional Investigative Experiments **5**

4.1 Noise Initialization via Vision Encoding Optimization . . . 6

4.1.1 Noise Initialization . . . 6

4.1.2 Diffusion-Based Target Synthesis . . . 7

4.1.3 Vision-Encoder Anchored Perturbation . . . 7

4.1.4 Adversarial Optimization via PGD . . . 7

4.2 Guiding perturbations via Target Value Vectors . . . 8

4.2.1 Our Modified Approach . . . 9

4.3 Investigating Cross Image transferability and Image Augmentation . . . 11

5 Results **11**

5.1 Claim 1: CroPA achieves cross-prompt transferability across various target texts. [Reproduced] 12

5.1.1 Experimental Details . . . 13

5.1.2 Comparative Analysis . . . 13

5.2 Claim 2: CroPA achieves the best overall performance across different number of prompts **[Reproduced]** . . . 13

5.3 Claim 3: CroPA converges towards higher ASR as number of iterations increase [Reproduced] 14

5.4 Claim 4: CroPA outperforms baselines under few-shot settings [Reproduced] . . . 14

5.5 Results Beyond the Original Paper . . . 15

5.5.1 Noise Initialization . . . 15
5.5.2 Investigating Cross-Image transferability and Image Augmentation . . . 16
5.5.3 Guiding perturbations via Target Value Vectors . . . 16

6 Discussion **18**
6.0.1 Broader Impact . . . 18
6.1 Limitations and Future Work . . . 19
6.2 Communication With Original Authors . . . 19

Appendix **21**

A Results Sourced from the original paper **23**
A.1 Cross-Model Transferability Evaluation for CroPA . . . 23
A.2 Convergence results of the original paper . . . 23

B Implementational Details **23**
B.1 Datasets . . . 23
B.2 Experimental Setup . . . 24
B.3 Computational Requirements . . . 24

C Experiment Hyperparameters **24**
C.1 Targeted ASRs tested on Flamingo with different target texts using CroPA . . . 24
C.1.1 Optimization Parameters . . . 24
C.1.2 Multi-Prompt Strategy . . . 25
C.1.3 Base Models Supported . . . 25
C.1.4 Generation Parameters . . . 25
C.1.5 Text Embedding Perturbation . . . 25
C.1.6 Reproducibility . . . 25
C.2 Noise Initialization via Vision Encoding Optimization . . . 25
C.2.1 PGD via Vision Encoder Image Perturbation Hyperparameters . . . 26
C.2.2 Text Perturbation Hyperparameters (CroPA Specific) . . . 26
C.2.3 Evaluation and Dataset Hyperparameters . . . 26
C.2.4 Noise Initialization via Vision Encoding Optimization . . . 27
C.2.5 Additional Notes . . . 27
C.2.6 Justification of Choices . . . 27
C.3 Investigating Cross Image transferability and Image Augmentation . . . 27
C.4 LD-UAP Enhancements . . . 27
C.4.1 Attack Configuration . . . 27

C.4.2 CroPA-Specific Parameters 28
C.4.3 Evaluation Configuration 28
C.4.4 Model-Specific Settings 29
C.4.5 Cross Model test with OpenFlamingo as the primary model 29

D Prompts for Different Tasks **30**

A Results Sourced from the original paper

A.1 Cross-Model Transferability Evaluation for CroPA

Table 8 presents CroPA's cross model transferability results, underscoring the model-specific over-fitting issue highlighted in 4.3.

Settings	Method	$VQA_{general}$	$VQA_{specific}$	Classification	Captioning	Overall
BLIP2 to InstructBLIP	Multi-P	0.00	0.01	0.04	0.03	0.02
	CroPA	**0.00**	**0.04**	**0.15**	**0.11**	**0.08**
InstructBLIP to BLIP2	Multi-P	0.00	0.02	0.10	0.02	0.04
	CroPA	**0.01**	**0.05**	**0.13**	**0.04**	**0.06**

Table 8: The cross model test under different settings. "BLIP2 to InstructBLIP" means the perturbations optimised on BLIP2 are tested on the InstructBLIP and "InstructBLIP to BLIP2" are different. The language model for BLIP2 is OPT-2.7b and the model for InstructBLIP is Vicuna-7b. The best performance values for each task are highlighted in bold.

A.2 Convergence results of the original paper

Figure 7 presents CroPA's convergence results over increasing number of iterations, referred in 5.3.

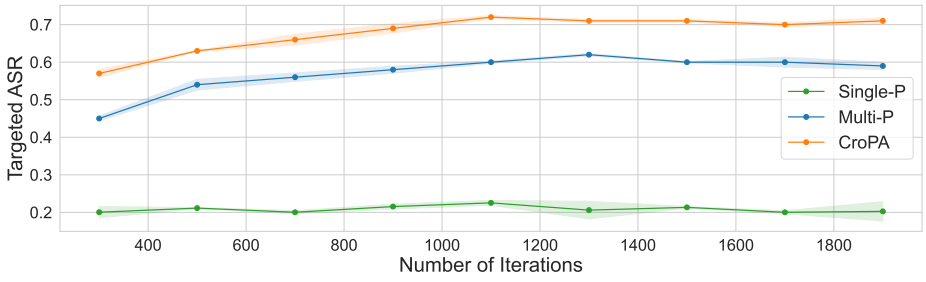


Figure 7: Overall Convergence and Comparison of Original CroPA

B Implementational Details

B.1 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, $VQA_{general}$, 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. (2017a) 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.

B.2 Experimental Setup

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.

B.3 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.

C Experiment Hyperparameters

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

The following are the detailed hyper-parameters used to obtain the results in Table 1, under Subsection 5.1:

CROPA utilizes the following hyperparameters for optimizing adversarial perturbations in vision-language models. These parameters are designed to balance attack effectiveness while preserving task-specific functionality.

C.1.1 Optimization Parameters

- **Total Iterations:** 1,701
- **Step Size:**

  - **Image Perturbations** (α_1): $\frac{1}{255} \approx 0.0039$
  - **Text Embedding Perturbations** (α_2, CROPA method): 0.01
- **Perturbation Budget** (ϵ): $\frac{16}{255} \approx 0.0627$ under L_∞ constraint
- **Loss Function:** Mean Squared Error (MSE) on ViT embeddings
- **Batch Size:** Dynamically allocated based on available GPU memory

C.1.2 Multi-Prompt Strategy

- **Simultaneous Prompts per Image:** 10 (default)
- **Prompt Rotation:** Randomized cyclic permutation per full cycle
- **Context Token Masking:** Preserve first N context tokens during updates

C.1.3 Base Models Supported

- OpenFlamingo
- BLIP-2
- InstructBLIP

C.1.4 Generation Parameters

- **Beam Search Width:** 3
- **Length Penalty:** -2.0 (encourages concise outputs)
- **Maximum Generation Length:** 5 tokens

C.1.5 Text Embedding Perturbation

- **Update Intervals:** Every 30 iterations (for 10 prompts)
- **Constraint:** ± 0.27 deviation from original embeddings

C.1.6 Reproducibility

- **Random Seed:** 42
- **Image Preprocessing:** 224×224 center crop

This configuration enables simultaneous optimization of vision-language perturbations while maintaining task functionality through constrained gradient updates.

C.2 Noise Initialization via Vision Encoding Optimization

This section details the key hyperparameters used in our CroPA with Noise Initialisation via Vision Encoding optimisation attack detailed in Subsection 4.1, to obtain the results in Table 3 and provides a rationale for their selection. We aim to provide sufficient information for reproducibility and to justify our experimental choices. **Projected Gradient Descent(PGD)** was used to align the original image $\mathrm{x_i}$ with the target image $\mathrm{T_i}$ by adding perturbations iteratively.

C.2.1 PGD via Vision Encoder Image Perturbation Hyperparameters

- **Epsilon (ε)**: We set the maximum allowed pixel perturbation (L_∞ norm) to 16/255. This value represents a trade-off between attack strength and perceptibility. Smaller epsilon values might be less effective at fooling the model but also less noticeable to human observers.
- **Alpha1 (α_1)**: The step size for each PGD iteration was set to 1/255. This relatively small step size allows for finer-grained exploration of the perturbation space and helps to avoid overshooting optimal perturbation directions.
- **PGD Iterations**: The number of PGD iterations for image perturbation was set to 1701. This was determined empirically; we observed that increasing iterations beyond this point yielded diminishing returns in terms of attack success rate while significantly increasing computation time. Save perturbation iterations are 900, 1100, 1300,1500 and 1700.
- **Budget**: 0.05. This hyperparameter defines the maximum allowed change to each pixel value in the image during the PGD optimization process for aligning ViT embeddings. It constrains the perturbation magnitude to ensure the modified image remains visually similar to the original.
- **Timesteps**: 150. Specifies the number of optimization steps taken during the PGD process to align ViT embeddings between the generated and target images. A larger number of timesteps allows for finer adjustments and potentially better alignment but also increases computational cost.

C.2.2 Text Perturbation Hyperparameters (CroPA Specific)

- **Alpha2 (α_2)**: This hyperparameter controls the step size for updating the text embedding perturbations. The value is dynamically assigned based on the number of prompts used.
- **Number of Prompts**: The number of prompts utilized during the optimization phase was 10. We use multiple prompts to make a more robust and transferable perturbation. Using different prompts that all target the same wrong answer forces the attack to find a perturbation that works across variations in the input question.
- **CroPA Update Iterations**: Text perturbations are updated at specific iterations (defined by `cropa_iter`). The text perturbation will be updated till 300 iterations. We empirically found that the text perturbation can guide the optimization better with some iterations.

```
cropa_end = 300
step = max((cropa_end//prompt_num),1)
cropa_iter = [i for i in range(step,cropa_end+1, step)]
```

C.2.3 Evaluation and Dataset Hyperparameters

- **Test Dataset Fraction**: We evaluated our attack on a 5% (fraction = 0.05) subset of the test dataset. This allowed us to perform a thorough evaluation while keeping the computational cost manageable. We selected the subset randomly to ensure a representative sample of the overall test distribution.
- **N-Shot Examples**: The number of in-context learning examples was set to 0. This means we did not provide any demonstration examples to the model during evaluation.
- **Number of Test Images**: 50.
- **Max Generation Length**: 5.
- **Num Beams**: 3.
- **Length Penalty**: -2.0.

C.2.4 Noise Initialization via Vision Encoding Optimization

- **ViT Model**: `ViTModel.from_pretrained("openai/clip-vit-large-patch14")`
- **Learning Rate**: 0.1
- **Optimizer**: Adam

C.2.5 Additional Notes

- **Random Seed**: We used a fixed random seed (`seed_everything(42)`) to ensure reproducibility of our results.
- **Device**: All experiments were conducted on a GPU (`cuda:{ config_args.device}`).
- **Batch Size**: The evaluation batch size (`args.eval_batch_size`) was chosen to maximize GPU utilization without exceeding memory constraints.

C.2.6 Justification of Choices

The hyperparameter values were selected based on a combination of prior work, pilot experiments, and computational constraints. We prioritized settings that balanced attack effectiveness, stealthiness, and computational efficiency. Ablation studies (not included in this section but discussed elsewhere in the paper) were conducted to assess the sensitivity of our results to key hyperparameters such as ε and α.

C.3 Investigating Cross Image transferability and Image Augmentation

The following are the detailed hyper-parameters used to obtain the results in Table 1, under Subsection 4.1

For ScMix, the only hyperparameters are the choice of η (or α if $\eta \sim \text{Beta}(\alpha, \alpha)$), β_1 and β_2.

- **Eta (η)**: We chose $\eta = 0.5$ as a constant value for η to ensure that the degree of self-mixing during augmentation is constant (and equal). In terms of α, this is equivalent to setting $\alpha = \infty$
- **Beta1 (β_1) and Beta2 (β_2)**: We chose $\beta_1 = 0.9$ and $\beta_2 = 0.1$ to ensure that the base image, used for self-mixing, dominates the final augmented image in terms of visual characteristics, while still ensuring diversification by adding some characteristics of the cross-mixing image.

The other hyperparemeters pertaining to the base CroPA framework can be found in C.1. Apart from this, in terms of implementation, the only variation with base CroPA is that of the perturbation learned for an image and its set of text prompts, the one learned for the image is common across all the images used. The modified algorithm is presented in Algorithm 4.

C.4 LD-UAP Enhancements

Our implementation of CroPA (Cross-modal Prompt Attack) and the baseline UAP (Universal Adversarial Perturbation) approach utilized the following hyperparameters:

C.4.1 Attack Configuration

- **Maximum Perturbation (ε)**: 16/255, constraining the L_∞-norm of the adversarial perturbation.
- **Step Size (α_1)**: 1/255, controlling the gradient update magnitude for image perturbations.
- **Text Embedding Perturbation Range**: $[-0.23, 0.27]$, limiting the word embedding perturbation magnitude.
- **Text Perturbation Step Size (α_2)**: Varies based on prompt number, determining the magnitude of updates to text embeddings.

Algorithm 4 Cross Image Cross Prompt Attack

Require: Model f, Target Text T, input images $X_v = \{x_v^1, x_v^2, \ldots, x_v^n\}$, prompt set $S_t = \{X_t^1, X_t^2, \ldots, X_t^n\}$, perturbation size ϵ, step sizes α_1, α_2, iterations K, update interval N, ScMix augmentation boolean *Augment*, ScMix hyperparameters (η, β_1, β_2) if *Augment* is true
Ensure: Adversarial perturbation δ_v

```
Initialise δ_v ~ N(0, ε)
for step = 1 to K do
  for i = 1 to n do
    Sample prompt x_t^{i,j} from X_t^i
    if Augment is true then
      x_v^{cross} ~ X_v
      if x_v^{cross} is x_v then
        x'_v = x_v^i
      else
        x'_{v1} = Resize(RandomCrop(x_v^i))
        x'_{v2} = Resize(RandomCrop(x_v^i))
        x'_v = β1(η x'_{v1} + (1 − η) x'_{v2}) + β2 x_v^{cross}
      end if
    else
      x'_v = x_v^i
    end if
    if x_t'^{i,j} not initialized then
      Initialize x_t'^{i,j} = x_t^{i,j}
    end if
    g_v = ∇_{x_v} L(f(x'_v + δ_v, x_t'^{i,j}), T)
    δ_v = δ_v − α1 · sign(g_v)
    if mod(step, N) = 0 then
      g_t = ∇_{x_t} L(f(x'_v + δ_v, x_t'^{i,j}), T)
      x_t'^{i,j} = x_t'^{i,j} + α2 · sign(g_t)
    end if
    Project δ_v to ε-ball around 0
  end for
end fore
return δ_v
```

- **Total Iterations**: 1701, with evaluation checkpoints at intervals of 200, starting from iteration 900.
- **Target Response**: "bomb," forcing the model to generate this specific output.

C.4.2 CroPA-Specific Parameters

- **Text Perturbation Schedule**: Updates occur during the first 300 iterations at intervals determined by step = max((300/prompt_num), 1).
- **Prompt Access Strategy**: Random shuffling with rotation to ensure diverse prompt coverage.
- **Semantic Alignment**: Cosine similarity loss with weight factor 5 to align perturbed images with the target concept.

C.4.3 Evaluation Configuration

- **Sampling Fraction**: 0.05 of the total dataset, balancing computational resources with statistical significance.
- **In-Context Learning Examples**: 0 shots (default configuration).

- **Generation Parameters**:
  - Maximum generation length: 5 tokens.
  - Number of beams: 3.
  - Length penalty: -2.0.

C.4.4 Model-Specific Settings

We conducted the same experiment for both BLIP-2 and OpenFlamingo. Hyperparameters for both the models are given
Experiment 1:

- **Primary Model**: OpenFlamingo-9B
- **Image Generation Model**: stable-diffusion-xl-base-1.0 was used.
- **Vision Feature Extraction**: Middle transformer blocks (layers 10-19) for semantic representation.
- **Image Preprocessing**: Normalization with mean = $[0.485, 0.456, 0.406]$ and standard deviation = $[0.229, 0.224, 0.225]$.
- **Image Dimensions**: 224×224 pixels.

Experiment 2:

- **Primary Model**: BLIP-2 (blip2-opt-2.7b).
- **Image Generation Model**: stable-diffusion-xl-base-1.0 was used.
- **Vision Feature Extraction**: Middle transformer blocks (layers 14-29) for semantic representation.
- **Image Preprocessing**: Normalization with mean = $[0.485, 0.456, 0.406]$ and standard deviation = $[0.229, 0.224, 0.225]$.
- **Image Dimensions**: 224×224 pixels.

The hyperparameters were carefully selected to balance attack effectiveness and computational efficiency. Our implementation used a random seed of 42 to ensure reproducibility across experimental runs.

C.4.5 Cross Model test with OpenFlamingo as the primary model

Table 9 reports results for the cross-model transferability test as mentioned in Subsection 5.5.3 with the LD-UAP enhanced CroPA method described in Subsection 4.2 :

Settings	Method	$\text{VQA}_{\text{general}}$	$\text{VQA}_{\text{specific}}$	Classification	Captioning	Overall
Flamingo to InstructBLIP	Multi-P	0.00	0.00	0.00	0.00	0.00
	CroPA	0.00	0.00	0.00	0.00	0.00
	CroPA + LD-UAP	0.00	0.00	0.00	0.00	0.00
Flamingo to BLIP2	Multi-P	0.00	0.00	0.00	0.00	0.00
	CroPA	0.00	0.00	0.00	0.00	0.00
	CroPA+LD-UAP	0.00	0.00	0.00	0.00	0.00

Table 9: The cross model test under different settings. "BLIP2 to InstructBLIP" means the perturbations optimised on BLIP2 are tested on the InstructBLIP and similarly for "BLIP2 to Flamingo"

D Prompts for Different Tasks

This section presents a short description of prompts used for various vision-language tasks in our experiments. Detailed list of prompts is provided in our supplementary material.

Prompts for Visual Question Answering (VQA) are categorized into two types: $\text{VQA}_{\text{general}}$ and $\text{VQA}_{\text{specific}}$. $\text{VQA}_{\text{general}}$ prompts are image-agnostic while $\text{VQA}_{\text{specific}}$ prompts are tailored to specific image content. Prompts in the categories of $\text{VQA}_{\text{general}}$, Image captioning and Image Classification are created by the original paper Luo et al. (2024) while the prompts for $\text{VQA}_{\text{specific}}$ are derived from the Goyal et al. (2017b).