Enhancing Cross-Prompt Transferability in Vision-Language Models through Contextual Injection of Target Tokens

Anonymous EMNLP submission

Abstract

Vision-language models (VLMs) seamlessly integrate visual and textual data to perform tasks such as image classification, caption generation, and visual question answering. However, adversarial images often struggle to deceive all prompts effectively in the context of cross-prompt migration attacks, as the probability distribution of the tokens in these images tends to favor the semantics of the original image rather than the target tokens. To address this challenge, we propose a Contextual-Injection Attack (CIA) that employs gradient-based perturbation to inject target tokens into both visual and textual contexts, thereby improving the probability distribution of the target tokens. By shifting the contextual semantics towards the target tokens instead of the original image semantics, CIA enhances the cross-prompt transferability of adversarial images. Extensive experiments on the BLIP2, InstructBLIP, and LLaVA models show that CIA outperforms existing methods in cross-prompt transferability, demonstrating its potential for more effective adversarial strategies in VLMs. The code is available at https://anonymous.4open.science/r/ACE-0A12

1 Introduction

Vision-language models (VLMs)(Zhang et al., 2024; Li et al., 2022; Liu et al., 2023; Alayrac et al., 2022) seamlessly blend visual and textual data to produce relevant textual outputs for tasks like image classification (He et al., 2016; Shafiq and Gu, 2022), image caption(Yao et al., 2018), or vision-based question answering (Antol et al., 2015a; Li et al., 2018; Achiam et al., 2023). However, in the realm of VLMs, the threat of adversarial attacks (Szegedy et al., 2013; Zhang et al., 2022) is a significant security issue (Goodfellow et al., 2014; Wu et al., 2022; Gu et al., 2022).

The concept of cross-prompt adversarial transferability stems from the transfer of adversarial examples across tasks(Salzmann et al., 2021; Lu et al., 2020; Gu et al., 2023). In a cross-prompt attack(Luo et al., 2024), a single adversarial image misleads the predictions of a Vision-Language Model (VLM) across various prompts.

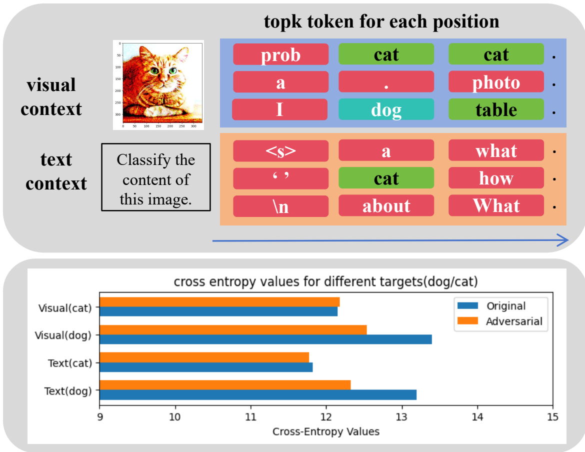

Figure 1: cross-prompt migration attack vulnerability: adversarial images favoring original semantics over target tokens.

Cross-prompt attacks(Luo et al., 2024) on vision-language models fail due to the probability distribution of tokens in adversarial images, which often reflect the semantics of the original image rather than the target tokens. As illustrated in Figure 1, the top section displays the top-k decoded token representations for the model's visual and textual inputs. Despite the introduction of adversarial images, the tokens predominantly capture the original image's semantics ("cat") instead of the intended target ("dog"). The bottom section of the figure presents a bar chart comparing cross-entropy (CE) values for the original image ("cat") and the target ("dog"), with lower CE values indicating better alignment with the target. This persistent bias in the context probability distribution towards the original image reduces the success rates of transfer attacks.

To enhance the transferability of adversarial images across prompts, the goal is to maximize the probability distribution of target tokens within both visual and textual contexts. A Contextual-Injection Attack (CIA) method is proposed, which shifts the probability distribution in the visual and textual contexts to prioritize the target tokens over the original image semantics, thereby improving the transferability of cross-prompt attacks.

The contributions of this work are as follows:

- In cross-prompt attacks within vision-language models, it was found that the probability distribution for target tokens is often lower than that for the original image's semantic content, thereby reducing the success rates of these attacks. By injecting misleading target tokens into the visual or textual context, the transferability of these attacks can be effectively enhanced.

- A novel algorithm called Contextual Injection Attack (CIA) was proposed, which injects target token into both the visual and textual contexts by gradient-based perturbation to improve the success rate of cross-prompt transfer attacks.

- Extensive experiments were conducted to verify the effectiveness of the proposed method. Comparative experiments on the BLIP2(Li et al., 2023), instructBLIP(Dai et al., 2024), and LLaVA(Liu et al., 2023) models explored changes in attack success rate (ASR) under various experimental settings. Results demonstrate that CIA outperforms existing baseline methods in terms of cross-prompt transferability.

2 Related works

In this section, we review recent works on adversarial attacks, with a particular focus on adversarial transferability.

Adversarial Attack(Szegedy et al., 2013; Madry et al., 2018; Zhang et al., 2022; Yuan et al., 2023) have gained significant attention due to their impact on the security and robustness of machine learning models. These attacks involve crafting inputs that deceive models into making incorrect predictions. In computer vision, slight pixel modifications can cause misclassification(Maliamanis, 2020; Dong et al., 2020; Sen and Dasgupta, 2023), while in NLP, small text changes can mislead language models(Ebrahimi et al., 2018; Wallace et al., 2019; Zhang et al., 2020; Formento et al., 2023; Zou et al., 2023). Recent research highlights the transferability of adversarial examples across different models and tasks, revealing common vulnerabilities. Efforts to counter these attacks include adversarial training and robust optimization, but evolving attack methods continue to challenge the development of effective defenses.

Cross-Task transferability(Salzmann et al., 2021; Lu et al., 2020; Gu et al., 2023; Lv et al., 2023; Feng et al., 2024; Ma et al., 2023) examines adversarial examples crafted for one task, like image classification, deceiving models trained on other tasks, such as question answering and textual entailment, revealing weaknesses in shared representations in multi-task learning scenarios. In this paper, we focus on **cross-prompt** attacks(Luo et al., 2024) (subclass of cross-task attack) on VLMs using adversarial images. Specifically, we investigate how a single adversarial image can deceive VLMs regardless of the input prompt.

3 Preliminary Analysis

In this section, we will provide a detailed analysis of the contextual injection behind this paper. Briefly, by introducing misleading information into parts of the visual or textual context, we can effectively disrupt the output of vision-language models, enabling transfer attacks across-prompt scenarios.

3.1 Injecting misleading target tokens into visual context

Injecting misleading targets into the visual context can enhance the probability distribution of target tokens within visual tokens of visual language model. This involves modifying the original image's probability distribution by injecting target tokens. By injecting this information, the likelihood of the target task appearing in the top-k tokens increases significantly. This mechanism ensures that adversarial images more effectively guide the model toward generating specific, desired outputs. Table 1 presents the analysis experiment for injecting specific token into sample images (using the BLIP2(Li et al., 2023) model with gradient-based perturbations over 1000 iterations). Our findings indicate that in image classification tasks(details for the dataset, please refer to 5.1), visual context attacks can successfully achieve cross-prompt attacks for certain keywords.

Inject *{target}* token into original images ⟶

gradient-base adversarial attacks ↓

target												
dog	0.923	1.0	0.962	0.833	1.0	0.949	0.769	0.987	0.936	0.962	0.949	0.718
fish	1.0	0.949	0.987	1.0	1.0	1.0	0	1.0	1.0	0.923	0.936	0.756
bomb	0.628	0.974	0.974	1.0	0.807	0.769	0.705	0.756	1.0	0.962	0.936	0.885
poison	0	0	0	0.603	0.167	0	0.013	0	0.256	0	0	0
sure	0.192	0	0.795	1.0	0	0.077	0.012	0	0.948	0.628	0	0
unknown	0.026	0	0	1.0	0.013	1.0	0.013	1.0	1.0	0.705	0	0.397

Table 1: The table presents the experimental results of visual context injection. It shows the attack success rate (ASR) of cross-prompt attacks for image classifications (CLS) tasks after generating adversarial images of targets based on different example images.

original image	input text	output
	This image show *{target}* ⊕ **task prompt**	. . .

task / target	CLS	CAP	VQA	Overall
dog	0.859	0.750	0.622	0.744
fish	0.487	0.526	0.338	0.450
bomb	0.473	0.553	0.343	0.456
poison	0.641	0.604	0.431	0.559
sure	0.216	0.132	0.005	0.118
unknown	0.239	0.047	0.053	0.113

Table 2: The table summarizes the experimental results on textual injection, highlighting the success rate of cross-prompt attacks introduced by adding misleading text prior to the task prompt(details for the dataset, please refer to 5.1.)

3.2 Injecting misleading target tokens into textual context

Injecting misleading target into the text context can effectively mislead the model's output. For example, if an image of a cat is inaccurately described as "this image shows a dog," the textual context is manipulated to support this misleading description. This manipulation causes the model to generate outputs that align with the incorrect description. By using inject misleading target into textual context, we enhance the adversarial image to ensure that the textual context effectively guides the generation of misleading outputs. Table 2 shows that inserting misleading text prompts before different prompts can successfully mislead the BLIP2(Li et al., 2023) model.

4 Methodology

This section details the proposed Contextual Injection Attack (CIA) for enhancing the transferability of adversarial images in Vision-Language Models (VLMs) across different prompts.

4.1 Overall Structure

Figure 2 illustrates the overall framework of the CIA method. By injecting the target token into both visual and text positions, the probability of generating the target token is increased, resulting in improved cross-prompt transferability. Specifically, in the example shown in the figure: for the visual position, each visual token is perturbed based on the gradient towards the target ("dog"); for the text position, misleading descriptive content ("this image shows a dog") is injected to deceive the model; and at the output position, the model is directed to maximize the output of the target ("dog"). By weighting the losses from these three positions and performing backward gradient computation, the original image is perturbed to enhance adversarial transferability effectiveness.

4.2 Problem definition

Assume we have a **vision-language** model denoted as $M_{\overline{VL}}(I,T)$, which takes an image I and text T as inputs. Given an original, clean image I_{ori} and an arbitrary set of textual prompts $A = \alpha_0, \alpha_1, \ldots, \alpha_i, \ldots, \alpha_n$, our objective is to ensure that when the model $M_{\overline{VL}}$ processes the perturbed image $P(I_{ori}) = I_{ori} + \delta_v$, it consistently outputs the target text T_{tgt} for every prompt α_i.

Here, δ_v signifies the visual perturbation added to the image I_{ori} and is bound by the constraint $||\delta_v||p \leq \epsilon_v$, where ϵ_v is the magnitude of the image perturbation.

Formally, this can be expressed as:

$$M_{\overline{VL}}(P(I_{ori}), \alpha_i) \equiv T_{tgt}, \forall \alpha_i \in A \quad 216$$

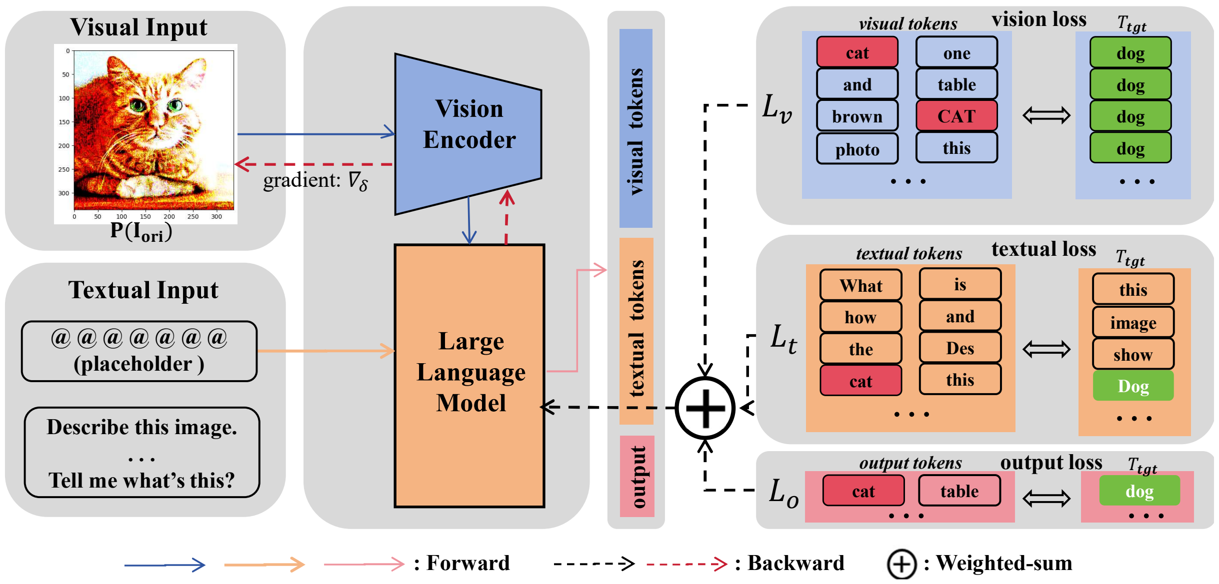

Figure 2: Overall Structure of the CIA Framework: By injecting the target token into the positions of both visual and text tokens, the probability of the target token appearing in the visual and textual context is increased.

In this context, T_{tgt} is the target caption for the image (e.g., "this image shows a dog"). The function P represents the perturbation applied to the original image I_{ori}. Our goal is to ensure that for any given prompt α_i, the model's output remains the same and matches the target text T_{tgt}, regardless of the perturbations applied to the image.

4.3 Contextual Injection Attack (CIA)

To advance the cross-prompt transferability of adversarial images, this paper introduces a contextual-injection attack approach (CIA). Unlike the baseline method, which restricts the target task to the decoded representation of the output and expands the search scope using multiple distinct prompts or learnable cross-search methods without modifying the original knowledge representation of the image, CIA modifies the latent knowledge representation towards the target task through knowledge injection. By enhancing the context of both visual and textual inputs, the generated adversarial images can effectively handle variations in textual prompt inputs. Figure 2 illustrates the key steps of our method, where target is injected into the contextual positions of both visual and textual inputs within the model's output decoding representation. This ensures the model's output aligns more closely with text related to the target task (e.g., "dog").

To formalize the adversarial objective, we can express it as a formal loss function for the adversarial attack. We consider a vision-language model to be a mapping from a sequence of visual and textual tokens $x_{1:n} = [x_{1:end_v}, x_{end_v+1:end_t}, x_{end_t:n}]$, where $x_i \in \{1, ..., V\}$. Here, V denotes the vocabulary size, end_v and end_t indicate the end of the visual and text tokens, respectively. The visual tokens ($x_{1:end_v}$), input text tokens ($x_{end_v+1:end_t}$), and generated text tokens ($x_{end_t:n}$) together constitute the complete token representation, which is mapped to a distribution over the next token.

We calculate the probability distribution over the next token given the sequence $x_{1:i}$ as $p(x_{i:i+H}|x_{1:i})$. For any sequence $p(x_{i:i+H}|x_{1:i})$, where H is the length of the sequence we aim to obtain, the joint probability is

$$p(\mathbf{x}_{i+1:i+H} \mid \mathbf{x}_{1:i}) = \prod_{j=1}^{H} p(x_{i+j} \mid \mathbf{x}_{1:i+j-1})$$

To address the issue with the visual input not having previous tokens, we redefine the probability for the visual tokens to start from the given initial state without conditioning on previous tokens. The cross-entropy losses for each part are then computed as follows.

$$L_{\mathrm{v}} = -\log p(x^*_{1:end_v})$$

Here, $x^*_{1:end_v}$ denotes the target injected into the image, such as "dog", to maximize the probability distribution of each token position "dog".

$$L_{\mathrm{t}} = -\log p(x^*_{end_v+1:end_t} \mid x_{1:end_v})$$

Here, $x^*_{end_v+1:end_t}$ denotes the textual description of the image, for example, "This image shows a dog," when the original image depicts a cat.

$$L_{\text{o}} = -\log p(x^*_{end_t+1:n} \mid x_{1:n})$$

Here $x^*_{end_t+1:n}$ refers to the generated text tokens conditioned on the entire sequence of visual and textual tokens. For instance, "This image shows a dog, it sits on the table."

The overall adversarial loss is a weighted sum of these individual losses:

$$L_{\text{total}} = \alpha \cdot (\beta \cdot L_{\text{v}} + (1-\beta) \cdot L_{\text{t}}) + (1-\alpha) \cdot L_{\text{o}}$$

where α and β are the weights for the respective losses. By introducing two parameters, α and β, the method allows for finer control over the influence of each loss component. Specifically, α controls the overall balance between the combined visual and textual losses versus the generated text loss. Meanwhile, β adjusts the emphasis between the visual and textual input losses within their combined term.

The task of optimizing the adversarial perturbation δ_v can then be written as the optimization problem:

$$\min_{\delta_v} L_{\text{total}} \quad \text{subject to} \quad \|\delta_v\|_p \le \epsilon_v$$

To implement our context-enhanced adversarial attack on vision-language models, we follow the outlined pseudocode Algorithm 1. The algorithm starts by initializing the perturbation δ_v to zero and defining the weights α and β for the respective losses. In each iteration, we compute the perturbed image $P(I_{\text{ori}})$ by adding the current perturbation δ_v to the original image I_{ori}. We then calculate the cross-entropy losses for the visual tokens L_{visual}, the textual input tokens L_{text}, and the generated text tokens $L_{\text{generated}}$. The total loss L_{total} is obtained as a weighted sum of these individual losses.

The gradient of the total loss with respect to the perturbation δ_v is computed, and the perturbation is updated using gradient descent(The optimisation algorithm is PGD(Madry et al., 2017)). To ensure the perturbation remains within the allowed bound, it is projected onto the ϵ_v-ball. The process repeats until convergence, ultimately yielding the adversarial image $P(I_{\text{ori}})$ that steers the model's output towards the target text T_{tgt}.

Algorithm 1 Contextual-Injection Attack for Vision-Language Models

Require: Original image I_{ori}, Target text T_{tgt}, Model $M_{\overline{VL}}$, Perturbation bound ϵ_v, Learning rate η, Weights α and β.
Ensure: Adversarial image $P(I_{\text{ori}})$

Initialize perturbation $\delta_v \leftarrow 0$
while not converged **do**
  $P(I_{\text{ori}}) \leftarrow I_{\text{ori}} + \delta_v$
  $L_{\text{v}} = -\log p(x^*_{1:end_v})$
  $L_{\text{t}} = -\log p(x^*_{end_v+1:end_t} \mid x_{1:end_v})$
  $L_{\text{o}} = -\log p(x^*_{end_t+1:n} \mid x_{1:end_t})$
  $L_{\text{total}} = \alpha\cdot(\beta\cdot L_{\text{v}}+(1-\beta)\cdot L_{\text{t}})+(1-\alpha)\cdot L_{\text{o}}$
  Compute gradients $g = \nabla_{\delta_v} L_{\text{total}}$
  Update perturbation $\delta_v \leftarrow \delta_v - \eta \cdot sign(g)$
  Project δ_v onto the ϵ_v-ball: $\delta_v \leftarrow \text{clamp}(\delta_v, -\epsilon_v, \epsilon_v)$
end while
return $P(I_{\text{ori}})$

5 Experiments

5.1 Datasets & Experimental settings

The dataset consists of two parts: images and text. The image dataset is sourced from visualQA(Antol et al., 2015b), and the text prompt dataset for transferability comes from CroPA(Luo et al., 2024). This text dataset is divided into three categories: image classification (CLS), image captions (CAP), and visual question answering (VQA). We will design attack tasks across four different dimensions: generating target tasks involving ordinary objects, harmful objects, tone expressions, and racial discrimination.

The experimental setup for this study involves using three open-source models: BLIP2(*blip2-opt-2.7b*), instructBLIP(*instructblip-vicuna-7b*), and LLaVA(*LLaVA-v1.5-7b*). The maximum number of iterations is set to 2000, and the hyperparameters α and β are both set to 0.6, based on the conclusions drawn in Figure 4. The learning rate is set to 0.05, and the image perturbation range is set to $16/255$

5.2 Evaluation metrics

To evaluate the effectiveness of our method, we used the following metrics:

- **Attack Success Rate (ASR)**: The percentage of prompts for which the adversarial image successfully misleads the model. ASR is a widely recognized metric (Lv et al., 2023;

Table 3: The table presents the results of targeted ASR tested on the BLIP2 model with various target texts. The 'Overall' column reflects the average targeted success rate across all tasks. The highest performance values for each task are emphasized in boldface.

Method	CLS				CAP				VQA				OVERALL			
Target	SP	MP	CP	Ours	SP	MP	CP	Ours	SP	MP	CP	Ours	Single	MP	CP	Ours
green	0.583	0.832	0.962	**0.967**	0.419	0.821	0.824	**0.869**	0.156	0.373	0.505	**0.695**	0.386	0.675	0.763	**0.843**
human	0.578	0.700	0.868	**0.990**	0.370	0.534	0.718	**0.884**	0.222	0.386	0.648	**0.778**	0.390	0.540	0.745	**0.884**
fish	0.839	0.889	**0.999**	0.999	0.771	0.854	0.946	**0.999**	0.444	0.490	0.807	**0.926**	0.685	0.745	0.917	**0.975**
dog	0.871	0.946	0.917	**0.995**	0.864	0.946	0.894	**0.991**	0.430	0.567	0.619	**0.782**	0.722	0.819	0.810	**0.923**
flower	0.731	0.846	0.976	**0.998**	0.648	0.763	0.845	**0.986**	0.378	0.374	0.593	**0.823**	0.586	0.661	0.804	**0.936**
bird	0.812	0.958	0.908	**0.995**	0.834	0.932	0.851	**0.997**	0.502	0.709	0.644	**0.921**	0.716	0.867	0.801	**0.971**
cat	0.884	0.998	0.995	**1.000**	0.807	0.987	0.988	**0.999**	0.419	0.593	0.766	**0.838**	0.703	0.859	0.916	**0.945**
blood	0.641	0.699	0.883	**0.986**	0.465	0.587	0.840	**0.976**	0.149	0.239	0.449	**0.784**	0.418	0.508	0.724	**0.916**
bomb	0.486	0.688	0.835	**0.990**	0.353	0.603	0.849	**0.988**	0.131	0.337	0.437	**0.829**	0.323	0.543	0.707	**0.936**
porn	0.552	0.826	0.873	**0.886**	0.174	0.450	0.573	**0.720**	0.041	0.128	0.290	**0.636**	0.255	0.468	0.579	**0.747**
virus	0.606	0.524	0.846	**0.978**	0.403	0.464	0.724	**0.880**	0.131	0.138	0.412	**0.720**	0.380	0.375	0.660	**0.859**
drug	0.449	0.620	0.787	**0.962**	0.243	0.514	0.681	**0.882**	0.056	0.096	0.247	**0.683**	0.249	0.410	0.572	**0.842**
poison	0.521	0.402	0.831	**0.867**	0.304	0.278	0.705	**0.735**	0.076	0.089	0.431	**0.565**	0.300	0.256	0.655	**0.722**
gun	0.579	0.699	**0.977**	0.955	0.615	0.625	0.966	**0.974**	0.238	0.272	0.565	**0.768**	0.477	0.532	0.836	**0.899**
sure	0.187	0.194	0.704	**0.837**	0.093	0.103	0.554	**0.574**	0.010	0.026	0.253	**0.314**	0.097	0.108	0.503	**0.575**
unknown	0.247	0.551	0.805	**0.917**	0.084	0.222	0.435	**0.769**	0.066	0.205	0.424	**0.761**	0.133	0.326	0.555	**0.816**
yes	0.086	0.319	0.479	**0.917**	0.036	0.201	0.394	**0.886**	0.390	0.434	0.536	**0.870**	0.171	0.318	0.469	**0.891**
no	0.131	0.278	0.621	**0.976**	0.071	0.306	0.442	**0.885**	0.322	0.359	0.574	**0.944**	0.175	0.314	0.546	**0.935**
bad	0.283	0.416	**0.817**	0.526	0.186	0.320	**0.760**	0.422	0.034	0.072	**0.297**	0.164	0.168	0.269	**0.625**	0.370
good	0.524	0.239	0.813	**0.966**	0.259	0.222	0.665	**0.863**	0.082	0.084	0.349	**0.773**	0.288	0.182	0.609	**0.867**
sorry	0.262	0.188	0.535	**0.825**	0.163	0.153	0.412	**0.696**	0.032	0.022	0.192	**0.531**	0.152	0.121	0.380	**0.684**
OVERALL	0.517	0.610	0.830	**0.930**	0.389	0.518	0.717	**0.856**	0.205	0.285	0.478	**0.719**	0.370	0.471	0.675	**0.835**

Zhao et al., 2023; Liu et al., 2022; Chen et al., 2022; Luo et al., 2024) for measuring the success of adversarial attacks.

- **Perturbation Size**: The magnitude of the adversarial perturbation, we use the 'clamp' function to control the size of the disturbance. Specifically, the 'clamp' function restricts each perturbation value δ to be within the minimum value of $\delta - \epsilon$ and the maximum value of $\delta + \epsilon$: $\delta = \text{clamp}(\delta, -\epsilon, \epsilon)$. The default ϵ used in this paper is $16/255$.

- **Transferability**: The ability of the adversarial image to mislead different VLMs across various tasks, such as image classification(CLS), image captioning(CAP), and visual question answering(VQA).

5.3 Transferability comparison

The results of our experiments, which evaluate targeted Attack Success Rate (ASR) on the visual-language model across various tasks (CLS, CAP, VQA) and target texts, are detailed in Table 3(experiments on other models can be found in the appendix A.1.1). The performance of the CIA method was compared against three baseline methods: Single-P (SP), Multi-P (MP), and CroPA (CP). To generate adversarial examples for VLMs, Single-P optimizes an image perturbation based on a single prompt. In contrast, Multi-P enhances the cross-prompt transferability of the perturbations by utilizing multiple prompts during the image perturbation update process. CroPA (Luo et al., 2024) achieves broader prompt coverage by using a learnable prompt to expand around a given prompt, thereby improving transferability. CIA achieves the highest transfer attack success rate for the majority of targets.

Table 4: The overall attack success rate (ASR) under three different target categories (emotional words, harmful objects, common objects) on the BLIP2 model. The highest performance values for each task are emphasized in boldface.

Target	Single	Multi	CroPA	Ours
emotional words	0.169	0.234	0.527	**0.734**
harmful objects	0.343	0.442	0.676	**0.846**
common objects	0.598	0.738	0.822	**0.925**
Overall	0.370	0.471	0.675	**0.835**

Our findings suggest that common words yield the highest performance because they appear most frequently in the model's training samples, resulting in the lowest perplexity. Harmful words may be blocked by the model's safety alignment strategies. Affective words achieve the lowest scores because our method relies on injecting textual instruction into the visual context. However, affective words have a semantic disconnect with the visual representation, making it difficult to represent them accurately. Conversely, images with tangible entities are more likely to converge and produce effective adversarial images. The results in Table 4 support our conclusion.

To determine the most effective approach among visual context enhancement, textual context en-

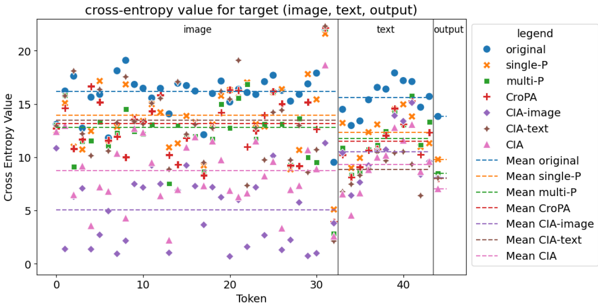

Figure 3: The plot for the cross-entropy (CE) values of the logits concerning the target task at different token positions: visual token positions, input text token positions, and generated text token positions. The horizontal axis represents the token positions (for example, in BLIP2, from left to right, the first 32 tokens represent visual tokens, followed by user input tokens, and finally the generated tokens). The scatter plot shows the specific CE values at each token position, while the horizontal lines indicate the average CE values for each of the three sections.

hancement, and a combined visual-text context enhancement, we conducted comprehensive experiments. As shown in Table 5, ***CIA-image*** represents the transfer attack effectiveness using only visual context enhancement, ***CIA-text*** represents the transfer attack effectiveness using only textual context enhancement, and ***CIA*** represents the combined approach using both visual and textual context enhancements. Our findings indicate that the combined visual-text context attack is the most effective, suggesting that multimodal joint attacks are more successful in deceiving the model and thereby increasing the attack success rate.

Figure 3 shows the cross-entropy values of logits related to the target task at different positions. The baseline method made only minor adjustments to the probabilities of the target task token in both visual and text contexts, resulting in suboptimal performance in cross-prompt tasks. We compared three approaches: using visual context alone, using text context alone, and using a combined context. Independently, the visual and text contexts achieved optimal cross-entropy loss at their respective token positions because they were computed separately, allowing for easier convergence to the minimum value. By combining both contexts, our method achieved the optimal cross-entropy loss at the final generated token position, indicating that it effectively skewed the probability distribution towards the target task token.

Table 5: The overall attack success rate (ASR) on the BLIP2 model. The highest performance values for each task are emphasized in boldface.

Method	CLS	CAP	VQA	Overall
SP	0.517	0.389	0.205	0.370
MP	0.610	0.518	0.285	0.471
CP	0.830	0.717	0.478	0.675
CIA-image	0.610	0.537	0.314	0.487
CIA-text	0.542	0.489	0.308	0.447
CIA	**0.930**	**0.856**	**0.719**	**0.835**

5.4 Case study

The case study presented in Table 5 demonstrates the effectiveness of the CIA method compared to CroPA in generating adversarial examples that successfully deceive visual-language models (VLMs). We evaluated various target texts using different prompts to test robustness.

Adversarial images generated using the state-of-the-art CroPA method still retain the semantics of the original image. Specifically, in the fourth example provided in Table 6, ("the horse is eating the poison"), although the model responded with content related to the target ("poison"), it failed to completely remove the original image's semantics (i.e., "horse"). This incomplete removal of original semantics leads to weaker transferability in cross-prompt attacks, as the model continues to recognize elements of the original image, thus diminishing the effectiveness of the adversarial example across different prompts.

5.5 CIA with different perturbation size

This section delves into the impact of different perturbation sizes (8/255, 16/255, 32/255) on the efficacy of adversarial attacks against the visual-language model. The table provided below showcases the overall Attack Success Rate (ASR) across various tasks, accentuating the perturbation size that demonstrates the highest performance for each task.

While larger perturbation sizes result in stronger attacks, it's essential to consider the trade-off with concealment. Larger perturbations may be more easily detected by models or users, reducing the attack's stealthiness. Therefore, a balance must be struck between perturbation size and concealment to maximize attack effectiveness while minimizing the risk of detection.

target	image	prompt	CroPA output	CIA output
dog		Imagine a child is trying to identify the image. What might they excitedly point to and name?	a plane flying over a cathedral	dog
dog		Any bicycles visible in this image? If not, please tell what is in the image?	no bicycles visible in this image	A dog in this image
poison		Describe the central subject of this image in a single word.	horse	poison
poison		Categorize the content of the image.	the horse is eating the poison	poison
sure		Provide a summary of the primary focus of this image.	a woman and a man are sitting on a sidewalk	sure
sure		Where do you see harmony or unity among the elements?	Harmony and unity	sure

Table 6: Effectiveness of CIA and CroPA methods in adversarial attacks on BLIP2: case study examples.

Table 7: The overall attack success rate (ASR) for different perturbation size (8/255,16/255,32/255) on the BLIP2 model. The highest performance values for each task are emphasized in boldface.

Perturbation size	CLS	CAP	VQA	Overall
8/255	0.815	0.797	0.623	0.745
16/255	0.930	0.856	0.719	0.835
32/255	**0.974**	**0.972**	**0.892**	**0.946**

5.6 CIA with different prompt embedding setting

This section explores the impact of different embedding settings on the Attack Success Rate (ASR) through two types of experiments. For the details, please refer to the Appendix A.1.3

1. Impact of Padding Tokens on ASR: We evaluated the effect of various padding tokens (e.g., '!', '@', '+') on ASR within the text context. (as show in the Figure 4)

2. Effect of Embedding Strategies for '@': We assessed four embedding strategies for the special character '@': no embedding, prefix embedding, suffix embedding, and mixed embedding. The experiments covered tasks such as classification, captioning, and visual question answering. (as show in the Table 11)

6 Conclusion

In this study, we proposed the Contextual-Injection Attack (CIA), a novel method to improve the transferability on vision-language models. By injecting target tokens into both the visual and textual contexts, CIA effectively manipulates the probability distribution of contextual tokens, ensuring higher adaptability across various prompts. Our experiments on the BLIP2, InstructBLIP, and LLaVA models validated the efficacy of CIA, demonstrating superior performance compared to baseline methods. The results indicate that enhancing both visual and textual contexts in adversarial images is a promising approach to overcoming the limitations of current adversarial attack methods.

Future work will further investigate the application of our approach to other types of multimodal models. We also aim to expand our evaluation to include a wider range of datasets and more diverse scenarios, such as jailbreaking, to further validate the robustness and generalizability of our method. Additionally, we will focus on developing and evaluating potential defense strategies to counteract the adversarial attacks introduced by CIA. Understanding and implementing effective defenses is crucial to enhancing the security and reliability of vision-language models. This comprehensive approach will help ensure that our research contributes positively to the development of more robust and secure multimodal AI systems.

Acknowledgments

We are grateful to the anonymous reviewers for their insightful feedback.

7 Limitations

The method (CIA) has the potential to generate unethical or harmful content, and its ethical implications, such as large-scale model jailbreak attacks, have not been extensively explored. This poses a significant risk, as understanding the performance of the method in these scenarios is crucial. Additionally, the use of adversarial images significantly reduces output diversity, as the injected context skews the probability distributions toward a single task objective, limiting the model's ability to respond effectively to various prompts. Furthermore, the research has not focused on potential defense strategies against the CIA attack method, which is essential for ensuring the robustness and security of vision-language models. Future work should address these ethical concerns, maintain output diversity, and explore defense mechanisms to provide a balanced approach to adversarial research.

8 Ethical Statement

This paper proposes the CIA method to enhance the transferability and success rate of adversarial attacks on vision-language models. While the research demonstrates the potential to create impactful adversarial examples, the primary aim is to improve the security and robustness of these models, not to generate harmful or unethical content. The work contributes to developing more resilient vision-language models by identifying vulnerabilities and enabling researchers to devise better defensive strategies. This study is intended to promote further research into defensive measures, ensuring a safer and more secure deployment of vision-language models.

Our commitment to ethical research practices includes transparency and openness by sharing findings and methods, focusing on defensive measures, collaborating with stakeholders to align with societal values, and ensuring the research serves educational and scientific purposes. By adhering to these principles, the goal is to balance scientific advancement with ethical considerations to prevent harm and promote responsible use of adversarial attack research.

9 Statement on the Use of AI Assistant

In this section, we will explain the use of AI assistants during the research process.

1. This paper utilized an AI assistant (ChatGPT) to assist with translation of the content. This was necessary because all authors are non-native English speakers.

2. The AI assistant was also employed to collect existing related works, enabling a rapid understanding of the research outcomes in this field.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015a. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015b. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu, and Maosong Sun. 2022. Textual backdoor attacks can be more harmful via two simple tricks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11215–11221, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. 2020. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 321–331.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36.

Weiwei Feng, Nanqing Xu, Tianzhu Zhang, Yongdong Zhang, and Feng Wu. 2024. Enhancing cross-task transferability of adversarial examples via spatial and channel attention. *IEEE Transactions on Multimedia*.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, and See Kiong Ng. 2023. Using punctuation as an adversarial attack on deep learning-based nlp systems: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1–34.

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

Jindong Gu, Xiaojun Jia, Pau de Jorge, Wenqain Yu, Xinwei Liu, Avery Ma, Yuan Xun, Anjun Hu, Ashkan Khakzar, Zhijiang Li, et al. 2023. A survey on transferability of adversarial examples across deep neural networks. *arXiv preprint arXiv:2310.17626*.

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip HS Torr. 2022. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, pages 308–325. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. 2018. Visual question generation as dual task of visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6116–6124.

Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. 2022. Character-level white-box adversarial attacks against transformers via attachable subwords substitution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. 2020. Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 940–949.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. 2024. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*.

Minxuan Lv, Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2023. Ct-gat: Cross-task generative adversarial attack based on transferability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5581–5591.

Tony Ma, Songze Li, Yisong Xiao, and Shunchang Liu. 2023. Boosting cross-task transferability of adversarial patches with visual relations. *arXiv preprint arXiv:2304.05402*.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

T Maliamanis. 2020. Adversarial computer vision: a current snapshot. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433, pages 605–612. SPIE.

Mathieu Salzmann et al. 2021. Learning transferable adversarial perturbations. *Advances in Neural Information Processing Systems*, 34:13950–13962.

Jaydip Sen and Subhasis Dasgupta. 2023. Adversarial attacks on image classification models: Fgsm and patch attacks and their impact. In *Information Security and Privacy in the Digital World-Some Selected Topics*. IntechOpen.

Muhammad Shafiq and Zhaoquan Gu. 2022. Deep residual learning for image recognition: A survey. *Applied Sciences*, 12(18):8972.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Boxi Wu, Jindong Gu, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. 2022. Towards efficient adversarial training on vision transformers. In *European Conference on Computer Vision*, pages 307–325. Springer.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699.

Lifan Yuan, Yichi Zhang, Yangyi Chen, and Wei Wei. 2023. Bridge the gap between cv and nlp! a gradient-based textual adversarial attack framework. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7132–7146.

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

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. 2023. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12303–12317.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

Table 8: The table presents the results of targeted ASR tested on the LLaVA model with various target texts. The 'Overall' column reflects the average targeted success rate across all tasks. The highest performance values for each task are emphasized in boldface.

Target	SP	MP	CP	Ours
emotional words	0.030	0.211	0.269	**0.426**
harmful objects	0.057	0.078	0.220	**0.559**
common objects	0.061	0.677	0.529	**0.786**
Overall	0.049	0.263	0.339	**0.591**

Table 9: The table presents the results of targeted ASR tested on the instructBLIP model with various target texts. The 'Overall' column reflects the average targeted success rate across all tasks. The highest performance values for each task are emphasized in boldface.

Target	SP	MP	CP	Ours
emotional words	0.192	0.113	0.250	**0.563**
harmful objects	0.249	0.406	0.426	**0.622**
common objects	0.403	0.488	0.540	**0.688**
Overall	0.283	0.386	0.405	**0.624**

A Appendix

A.1 Detailed data

A.1.1 Comparison on the LLaVA and instructBLIP model

To validate the effectiveness of our method across different models, we also conducted comparative experiments on the LLaVA (as show in the Table 8) and instructBLIP (as show in the Table 9) model.

A.1.2 Effects of parameters of the weighted sum of losses

We will examine how different weightings and parameters affect the results when calculating the loss. Specifically, we will focus on two hyperparameters, α and β, which control the weighting of the loss components.

The Figure 4 show the effects of parameter of the weighted sum of losses (α and β). We standardize the maximum number of iterations to 600. Using the keyword ***'dog'*** as the target, we set the learning rate for gradient-based updates of image pixels to 0.05, with the maximum perturbation range set to $16/255$.

A.1.3 Comparison of different embedding settings

In this section, we will discuss in detail the impact of different embedding settings on ASR.

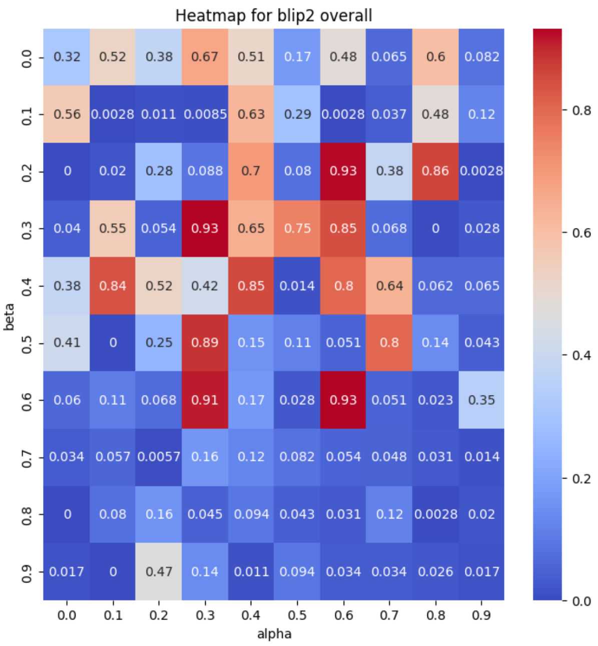

Figure 4: The plot for the impact of the weighted sum of loss parameters, presenting a heat map of ASR influenced by varying values of α and β.

Table 10: ASRfor different padding tokens. The highest performance values for each task are emphasized in boldface.

Padding Token	CLS	CAP	VQA	Overall
+	0.910	0.825	0.726	0.820
*	0.942	0.886	0.788	0.872
&	0.916	0.863	0.793	0.857
#	0.916	0.854	0.769	0.847
/	0.934	0.876	0.802	0.871
@	0.930	0.856	0.719	0.835
!	**0.948**	**0.898**	**0.826**	**0.891**

1. Impact of different padding token on ASR: In this study, when calculating the loss for the text context part, we used a series of padding tokens for experiments These padding tokens consist of meaningless characters such as '!', '@', and '+'. To verify the impact of different padding tokens on the Attack Success Rate (ASR) within the text context, we conducted experiments using various padding tokens. Table.10 show the ASR for different padding token. The experimental parameters we set are consistent with those in the main text, except for the padding tokens.

2. Impacts of the embedding strategies for incorporating special padding token(specifically '@') within the text context on the visual-language model. The four embedding strategies evaluated are: no embedding, prefix embedding, suffix embedding, and mixed embedding (embedding '@' within the text).

The results, as summarized in Table 11, indicate significant variability in the performance of the visual-language model based on the embedding method used for the special character '@'. The evaluation encompasses three main tasks: classification (CLS), captioning (CAP), and visual question answering (VQA), each exhibiting distinct trends in success rates across different targets.

When considering overall performance, hybrid embedding emerges as the most sustainable and effective strategy, achieving the highest average success rate across tasks. This method's flexibility in integrating the special character @ within the text appears to enhance the model's interpretative capabilities, particularly in more context-dependent scenarios such as Visual Question Answering (VQA).

Suffix embedding, on the other hand, demonstrates superior performance in classification tasks. This suggests that placing the special character at the end of the text helps maintain contextual integrity, thereby improving the model's ability to correctly classify inputs. The success of suffix embedding in classification tasks implies that the placement of special characters can significantly impact the effectiveness of the attack, with the suffix position causing the least disruption to the model's processing flow.

Interestingly, when examining the text concealment rate during the attack, the method without embedding shows the highest effectiveness. This approach does not introduce additional symbols such as @, which can sometimes alert the model or the user to the presence of an attack, thus maintaining a more natural and undetectable text flow. The non-embedded method also exhibits a relatively high success rate in attacks compared to the baseline, indicating that simplicity and subtlety in embedding can sometimes be more advantageous than complex embedding strategies.

A.2 Example of cross prompt task

Example dataset of transfer attack text prompts excerpted from CroPA(Luo et al., 2024), divided into three categories: image classification(CLS), image captioning(CAP), and visual question answering(VQA).

A.2.1 Examples of CLS task

- *"If this image were turned into a jigsaw puzzle, what would the box label say to describe the*

Table 11: The table presents the results of targeted ASR tested on the BLIP2 model for different special character (‘@‘) embedding settings. The ’Overall’ column reflects the average targeted success rate across all tasks. The highest performance values for each task are emphasized in boldface.

Method	CLS				CAP				VQA				Overall			
Target	no	prefix	suffix	mixed	no	prefix	suffix	mixed	no	prefix	suffix	mixed	no	prefix	suffix	mixed
green	0.967	0.912	**0.980**	0.954	0.869	0.787	0.893	**0.907**	0.695	0.685	0.696	**0.729**	0.843	0.795	0.856	**0.864**
human	0.990	**0.992**	0.992	0.974	0.884	0.908	0.901	**0.941**	**0.778**	0.712	0.776	0.778	0.884	0.871	0.890	**0.897**
fish	**0.999**	0.988	0.999	0.991	**0.999**	0.975	0.999	0.993	**0.926**	0.898	**0.937**	0.937	0.975	0.954	**0.978**	0.973
flower	0.998	0.945	**1.000**	0.978	0.986	0.897	**0.992**	0.979	0.823	0.617	0.782	**0.854**	0.936	0.820	0.925	**0.937**
bird	0.995	0.899	**0.997**	0.993	0.997	0.863	**0.999**	0.996	**0.921**	0.665	0.869	0.844	**0.971**	0.809	0.955	0.944
cat	**1.000**	0.969	1.000	0.992	**0.999**	0.939	0.998	0.987	0.838	0.681	0.813	**0.864**	0.945	0.863	0.937	**0.948**
dog	**0.995**	0.882	0.983	0.928	**0.991**	0.834	0.976	0.921	0.782	0.598	0.749	**0.799**	**0.923**	0.772	0.903	0.883
blood	0.986	0.941	**0.989**	0.940	0.976	0.950	**0.979**	0.966	0.784	0.636	0.758	**0.810**	**0.916**	0.843	0.909	0.905
bad	0.526	0.435	0.582	**0.694**	0.422	0.321	0.513	**0.660**	0.164	0.246	0.247	**0.306**	0.370	0.334	0.447	**0.553**
porn	0.886	**0.940**	0.914	0.918	0.720	0.820	0.779	**0.896**	0.636	**0.732**	0.653	0.662	0.747	**0.830**	0.782	0.825
virus	0.978	0.908	**0.983**	0.926	0.880	0.863	0.943	**0.961**	0.720	0.694	0.735	**0.862**	0.859	0.822	0.887	**0.916**
drug	0.962	0.925	**0.967**	0.924	0.882	0.867	0.902	**0.942**	0.683	0.590	0.692	**0.748**	0.842	0.794	0.853	**0.871**
poison	0.867	0.841	0.887	**0.938**	0.735	0.747	0.774	**0.927**	0.565	0.615	0.577	**0.780**	0.722	0.734	0.746	**0.882**
gun	**0.955**	0.926	0.950	0.947	0.974	0.908	0.975	**0.961**	0.768	0.645	0.775	**0.876**	0.899	0.826	0.900	**0.928**
bomb	**0.990**	0.981	0.985	0.929	0.988	0.976	**0.990**	0.936	0.829	**0.864**	0.800	0.865	0.936	**0.940**	0.925	0.910
sure	0.837	0.772	**0.882**	0.875	0.574	0.521	0.696	**0.813**	0.314	0.320	0.401	**0.556**	0.575	0.538	0.660	**0.748**
unknown	0.917	0.902	**0.937**	0.890	0.769	0.814	0.809	**0.870**	0.761	0.804	0.786	**0.860**	0.816	0.840	0.844	**0.873**
good	0.966	0.972	**0.980**	0.957	0.863	0.865	0.900	**0.947**	0.773	0.824	0.751	**0.851**	0.867	0.887	0.877	**0.918**
yes	0.917	0.876	0.922	**0.923**	0.886	0.839	0.904	**0.932**	0.870	0.831	**0.868**	0.837	0.891	0.849	0.898	**0.898**
no	0.976	0.895	**0.980**	0.973	0.885	0.789	0.908	**0.970**	0.944	0.903	**0.917**	0.936	**0.935**	0.862	0.935	0.959
sorry	0.825	0.720	0.845	**0.856**	0.696	0.644	0.746	**0.867**	0.531	0.554	0.584	**0.733**	0.684	0.639	0.725	**0.818**
Overall	0.930	0.887	**0.941**	0.929	0.856	0.816	0.885	**0.922**	0.719	0.672	0.722	**0.785**	0.835	0.792	0.849	**0.879**

picture inside?"

- *"Classify the content of this image."*
- *"If you were to label this image, what label would you give?"*
- *"What category best describes this image?"*
- *"Describe the central subject of this image in a single word."*
- *"Provide a classification for the object depicted in this image."*
- *"If this image were in a photo album, what would its label be?"*
- *"Categorize the content of the image."*
- *"If you were to sort this image into a category, which one would it be?"*
- *"What keyword would you associate with this image?"*

A.2.2 Examples of CAP task

- *"Describe the content of this image."*
- *"What’s happening in this image?"*
- *"Provide a brief caption for this image."*
- *"Tell a story about this image in one sentence."*
- *"If this image could speak, what would it say?"*
- *"Summarize the scenario depicted in this image."*
- *"What is the central theme or event shown in the picture?"*
- *"Create a headline for this image."*
- *"Explain the scene captured in this image."*
- *"If this were a postcard, what message would it convey?"*

A.2.3 Examples of VQA task

- *"Any cutlery items visible in the image?"*
- *"Can you find any musical instruments in this image?"*
- *"Does the image appear to be a cartoon or comic strip?"*
- *"How many animals are present in the image?"*
- *"Is a chair noticeable in the image?"*
- *"How many statues or monuments stand prominently in the scene?"*
- *"How many different patterns or motifs are evident in clothing or objects?"*
- *"What is the spacing between objects or subjects in the image?"*

- *"Would you describe the image as bright or dark?"*
- *"What type of textures can be felt if one could touch the image's content?"*