VQAttack: Transferable Adversarial Attacks on Visual Question Answering via Pre-trained Models

Ziyi Yin[1], **Muchao Ye**[1], **Tianrong Zhang**[1], **Jiaqi Wang**[1], **Han Liu**[2]
Jinghui Chen[1], **Ting Wang**[3], **Fenglong Ma**[1*]
[1]The Pennsylvania State University
[2]Dalian University of Technology
[3]Stony Brook University
{ziyiyin, muchao, tbz5156, jcz5917, fenglong}@psu.edu
liu.han.dut@gmail.com, twang@cs.stonybrook.edu

Abstract

Visual Question Answering (VQA) is a fundamental task in computer vision and natural language process fields. Although the "pre-training & finetuning" learning paradigm significantly improves the VQA performance, the adversarial robustness of such a learning paradigm has not been explored. In this paper, we delve into a new problem: using a pre-trained multimodal source model to create adversarial image-text pairs and then transferring them to attack the target VQA models. Correspondingly, we propose a novel VQATTACK model, which can iteratively generate both image and text perturbations with the designed modules: the large language model (LLM)-enhanced image attack and the cross-modal joint attack module. At each iteration, the LLM-enhanced image attack module first optimizes the latent representation-based loss to generate feature-level image perturbations. Then it incorporates an LLM to further enhance the image perturbations by optimizing the designed masked answer anti-recovery loss. The cross-modal joint attack module will be triggered at a specific iteration, which updates the image and text perturbations sequentially. Notably, the text perturbation updates are based on both the learned gradients in the word embedding space and word synonym-based substitution. Experimental results on two VQA datasets with five validated models demonstrate the effectiveness of the proposed VQATTACK in the transferable attack setting, compared with state-of-the-art baselines. This work reveals a significant blind spot in the "pre-training & fine-tuning" paradigm on VQA tasks. Source codes will be released.

Introduction

Visual Question Answering (VQA) is dedicated to extracting essential information from images to formulate responses to textual queries. While this application has proven to be highly versatile across various domains, including recommendation systems (Yu, Shen, and Jin 2019), medicine (Zhan et al. 2020), and robotics (Kenfack et al. 2020), the exploration of VQA system robustness remains a challenging endeavor. Current research primarily revolves around investigating the robustness of *end-to-end trained VQA models* through the development of effective attack methodologies, exemplified by Fool-VQA (Xu et al. 2018) and TrojVQA (Walmer et al. 2022). However, models trained end-to-end often exhibit inferior performance compared to the prevalent "pre-training & fine-tuning" paradigm. Within this paradigm, models are initially pre-trained on extensive collections of image-text pairs from the public domain, facilitating the acquisition of inter-modal relationships. Subsequently, the models undergo fine-tuning using specific VQA datasets to enhance their performance on downstream tasks. This instructional framework has yielded commendable predictive accuracy (Bao et al. 2022; Kim, Son, and Kim 2021; Li et al. 2021b). Nevertheless, the aspect of **adversarial robustness** within the context of the VQA task, as governed by this paradigm, remains insufficiently explored.

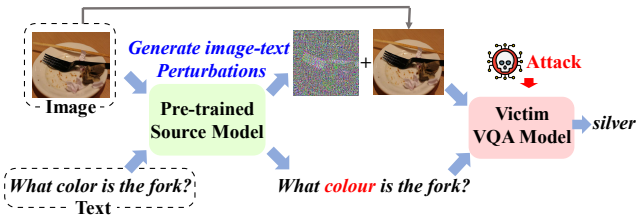

Figure 1: An example of Transferable adversarial attacks on VQA via pre-trained models.

This attack scenario presents notable complexities, which arise from the following two fundamental aspects:

- **C1 – Transferability across models**. The challenge here involves the transferability of adversarial attacks across distinct models. An example is shown in Figure 1. Pre-trained source models and victim target VQA models are usually trained for dissimilar tasks and trained on separate datasets. Furthermore, their structural disparities may result from variations introduced during fine-tuning. While the concept of transferability has been widely validated in the context of image models (Madry et al. 2018; Xie et al. 2019), such property within the domain of pre-trained models has yet to be comprehensively explored.
- **C2 – Joint attacks across different modalities**. Our task is centered around a multi-modal problem, necessitating the introduction of perturbations to both images and textual questions to achieve improved performance. Al-

*Corresponding author.

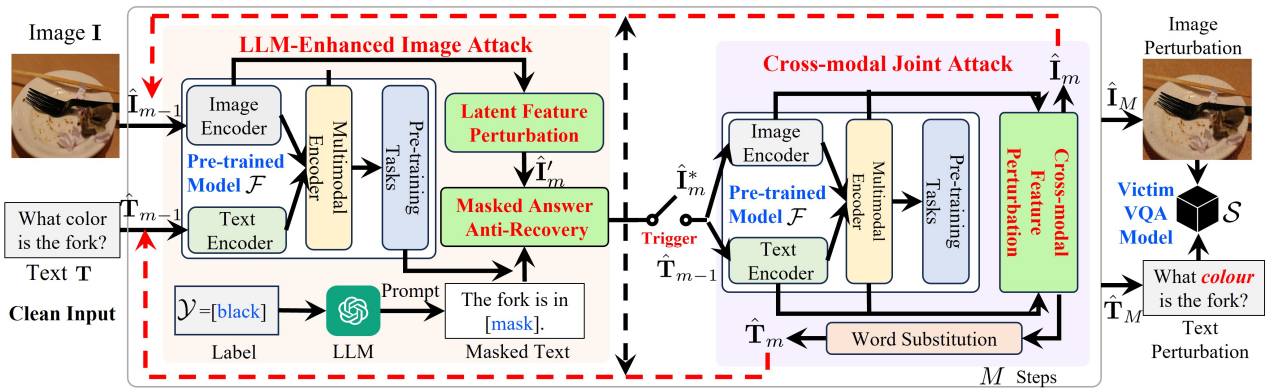

Figure 2: Overview of the proposed VQATTACK.

though previous methodologies have effectively devised attack strategies for each individual modality (Li et al. 2020; Madry et al. 2018), the intricate challenge lies in the simultaneous optimization of perturbations on images with continuous values and textual content characterized by discrete tokens. This joint attack task continues to pose a significant hurdle that requires innovative solutions.

To address these challenges, we propose a novel method named VQATTACK to explore the adversarial transferability between pre-trained source and victim target VQA models. As shown in Figure 2, the proposed VQATTACK generates image and text perturbations solely based on the pre-trained source model $\mathcal{F}$ with a novel multi-step attacking framework. After initializing the input image-text pair $(\mathbf{I}, \mathbf{T})$, VQATTACK will iteratively generate both image and text perturbations at each iteration m via two key modules: large language model (LLM)-enhanced image attack and cross-modal joint attack.

In the **LLM-enhanced image attack** module, VQATTACK first follows existing work (Naseer et al. 2020; Zhang, Yi, and Sang 2022) to minimize the similarity of latent features between the clean and perturbed input and then uses the clipping technique to obtain the image perturbation $\hat{\mathbf{I}}'_m$. To further enhance the transferability of attacks, VQATTACK introduces a new masked answer anti-recovery loss with the help of ChatGPT (OpenAI 2023), which differs from the existing latent feature-level attack by involving the correct answer label $\mathcal{Y}$ during the perturbation generation. The LLM-enhanced image attack module will be executed at each iteration, and the output from this module is denoted as $\hat{\mathbf{I}}^*_m$.

Due to the discrete nature of text data and the limited number of informative words in each text input, attacking the text at every iteration might not be necessary or beneficial for perturbation generation. Consequently, the **cross-modal joint attack** module will be triggered when m satisfies a specific condition. During this stage, VQATTACK first updates the perturbation of the image (i.e., $\hat{\mathbf{I}}^*_m$) via the cross-modal feature perturbation and the clipping technique. It then updates the text perturbation $\hat{\mathbf{T}}_m$ using the learned gradients and word synonym-based substitution in the word embedding space.

VQATTACK will return the output after iterating for M steps as the final adversarial image-text pair, i.e., $(\hat{\mathbf{I}}_M, \hat{\mathbf{T}}_M)$, which will be used to attack the victim VQA model $\mathcal{S}$. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first study on the adversarial robustness of the VQA task under the "pre-training & fine-tuning" paradigm. It does not only discuss the robustness of this paradigm but, more importantly, probes the potential security concern under a realistic scenario.
- We propose VQATTACK, which is a novel method to generate adversarial image-text pairs on the pre-trained vision language models. It consists of two novel modules, utilizes an LLM to generate masked text, and enables the iterative joint attack between image and text modalities.
- Five pre-trained models and two VQA datasets are involved in our experiment. Experimental results verify the effectiveness of the proposed VQATTACK under the transferable attack setting.

The Proposed VQATTACK

Problem Formulation

We use $\mathcal{F}$ to denote the publicly available pre-trained VL source model and $\mathcal{S}$ to represent the victim VQA target model. The goal of the transferable VQA attack is to generate an adversarial image-text pair $(\hat{\mathbf{I}}, \hat{\mathbf{T}})$ on the pre-trained source model $\mathcal{F}$ using the clean input $(\mathbf{I}, \mathbf{T})$, which will make the target victim model $\mathcal{S}$ have a wrong prediction, i.e., $S(\hat{\mathbf{I}}, \hat{\mathbf{T}}) \notin \mathcal{Y}$, where $\mathcal{Y}$ is the set of correct answers.

However, the victim model $\mathcal{S}$ in our setting is a black-box, arbitrary, and unknown model, and the only model that we can access is the pre-trained source model $\mathcal{F}$. Let $\mathcal{G}$ denote the proposed transferable attack strategy VQATTACK. We use the following function to generate an adversarial image-

Algorithm 1: The proposed VQATTACK

Input: A pre-trained source model $\mathcal{F}$, a clean image-text pair $(\mathbf{I}, \mathbf{T})$ and the ground-truth label $\mathcal{Y}$, step-size ϵ, prompt $\mathcal{P}$, LLM;

Input: Perturbation budget σ_i on image, σ_s on text, and the number of total iterations M.

1: **Initialization** $\hat{\mathbf{I}}_0 = \mathbf{I} + \delta$, $\delta \in \mathcal{U}(0, 1)$, ; $\hat{\mathbf{T}}_0 = \mathbf{T}$, and use BERT model to generate candidate token set $\mathcal{C}$.
2: **for** $m = 1$ to M **do**
3: *// LLM-enhanced Image Attack*
4: // Perturbation Generation with Latent features
5: Calculate $\nabla_i \mathcal{L}_f^m$ via Eq. (3) using $(\hat{\mathbf{I}}_{m-1}, \hat{\mathbf{T}}_{m-1})$;
6: $\hat{\mathbf{I}}'_m = \text{clip}_{\sigma_i}(\hat{\mathbf{I}}_{m-1} + \epsilon \text{sign}(\nabla_i \mathcal{L}_T))$;
7: // LLM-based Perturbation Enhancement
8: Masked text generation with LLM using $\hat{\mathbf{T}}_{m-1}$, label $\mathcal{Y}$, and prompt $\mathcal{P}$;
9: Calculate gradiants $\nabla_i \mathcal{L}_a^m$ via Eq. (4);
10: $\hat{\mathbf{I}}^*_m = \text{clip}_{\sigma_i}(\hat{\mathbf{I}}'_m + \epsilon \text{sign}(\nabla_i \mathcal{L}_a^m))$;
11: *// Cross-modal Joint Attack*
12: **if** $m \mod \lfloor \frac{M}{|\mathcal{W}|+1} \rfloor = 0$ **then**
13: // Image Perturbation Update
14: Calculate $\nabla_i \mathcal{L}_c^m$ via Eq. (5) using $(\hat{\mathbf{I}}^*_m\ \hat{\mathbf{T}}_{m-1})$;
15: $\hat{\mathbf{I}}_m = \text{clip}_{\sigma_i}(\hat{\mathbf{I}}^*_m + \epsilon \text{sign}(\nabla_i \mathcal{L}_c^m))$;
16: // Text Perturbation Update
17: Latent word embedding estimation via Eq. (6);
18: Obtain the synonym ranks $\mathcal{R}(\mathcal{C})$ according to Eq. (7);
19: Conduct synonym substitution to obtain $\hat{\mathbf{T}}_m$;
20: **else** $\hat{\mathbf{I}}_m = \hat{\mathbf{I}}^*_m$, $\hat{\mathbf{T}}_m = \hat{\mathbf{T}}_{m-1}$;
21: **end if**
22: **end for**
23: **return** $(\hat{\mathbf{I}}_M, \hat{\mathbf{T}}_M)$

text pair $(\hat{\mathbf{I}}, \hat{\mathbf{T}})$:

$$(\hat{\mathbf{I}}, \hat{\mathbf{T}}) = \mathcal{G}(\mathcal{F}, (\mathbf{I}, \mathbf{T}), M, \sigma_i, \sigma_s), \tag{1}$$

where $\mathcal{G}$ is an iterative attacking function, and M is the number of iterations. σ_i and σ_s are two hyperparameters to control the quality of adversarial images and text, which are defined as follows:

$$\begin{aligned} &\text{Image:} \quad \|\hat{\mathbf{I}} - \mathbf{I}\|_\infty < \sigma_i, \\ &\text{Text:} \quad Cos(U(\hat{\mathbf{T}}), U(\mathbf{T})) > \sigma_s. \end{aligned} \tag{2}$$

For an adversarial image $\hat{\mathbf{I}}$, we add pixel-level perturbations under the L_∞-norm distance. The distance threshold is set to σ_i. For an adversarial sentence $\hat{\mathbf{T}}$, we replace words with their synonyms and enforce a semantic similarity constraint σ_s, which is implemented through the cosine similarity $Cos(\cdot, \cdot)$ between the sentence embeddings $U(\hat{\mathbf{T}})$ and $U(\mathbf{T})$. Here, $U(\cdot)$ represents the universal sentence encoder (Cer et al. 2018), which has been widely adopted in text attack methods (Jin et al. 2020; Li et al. 2021a, 2019).

Overview

As shown in Figure 2, the proposed VQATTACK $\mathcal{G}$ first initializes the input pair $(\mathbf{I}, \mathbf{T})$ as $(\hat{\mathbf{I}}_0, \hat{\mathbf{T}}_0)$, and then updates $(\hat{\mathbf{I}}_m, \hat{\mathbf{T}}_m)$ at each iteration m through the proposed large language model (LLM)-enhanced image attack and cross-modal joint attack until the maximum iteration M. The final output $(\hat{\mathbf{I}}_M, \hat{\mathbf{T}}_M)$ is then used to attack the victim model $\mathcal{S}$. Algorithm 1 shows the algorithm flow of the proposed VQATTACK. Next, we provide the details of our model design step by step.

Initialization

As shown in Algorithm 1 line 1, for the input image $\mathbf{I}$, we follow the Projected Gradient Decent (PGD) (Madry et al. 2018) method to initialize $\hat{\mathbf{I}}$ by adding noise δ sampled from the Gaussian distribution $\mathcal{U}$, i.e., $\hat{\mathbf{I}}_0 = \mathbf{I} + \delta$, where $\delta \in \mathcal{U}(0, 1)$. For the text modality, we directly use the original input as the initialization, i.e., $\hat{\mathbf{T}}_0 = \mathbf{T}$.

Intuitively, the initialized pair $(\hat{\mathbf{I}}_0, \hat{\mathbf{T}}_0)$ can serve as the initial input for the cross-modal joint attack module, where iterative updates are performed on $(\hat{\mathbf{I}}_m, \hat{\mathbf{T}}_m)$ at each iteration. However, it is worth noting that this seemingly straightforward approach may not yield adversarial examples of high quality for effectively attacking the targeted model $\mathcal{S}$.

One aspect to consider is the intrinsic disparity between the numerical pixel representation of the input image $\mathbf{I}$ and the sequence-based nature of the input text $\mathbf{T}$. Frequent perturbations to the discrete $\mathbf{T}$ can often result in significant gradient fluctuations, which could subsequently adversely impact the perturbation of the numerical $\mathbf{I}$. As such, strictly coupling the updates of these two modalities throughout the entire attack process may not be the most optimal strategy. Besides, the input text $\mathbf{T}$ is typically characterized by a relatively short average length[1], containing only a limited number of informative words. This leads us to recognize that attacking the text at every iteration might not be necessary or beneficial.

It is due to these considerations that we put forth a novel module, namely the LLM-enhanced image attack. This module is designed to first learn an effective image perturbation independently, subsequently followed by a collaborative update of both image and text perturbations iteratively.

LLM-enhanced Image Attack

Perturbation Generation with Latent Features Several approaches have been proposed to generate the image perturbations using pre-trained models, such as Co-Attack (Zhang, Yi, and Sang 2022) and BadEncoder (Jia, Liu, and Gong 2022). The goal of these approaches is to minimize the similarity between the latent features learned by the pre-trained model $\mathcal{F}$ using the clean $\mathbf{I}$ and the perturbed $\hat{\mathbf{I}}_{m-1}$ at each iteration m, respectively.

Most multimodal VL pre-trained models such as ViLT (Kim, Son, and Kim 2021) and VLMO (Bao et al. 2022) usually consist of three encoders to learn latent features, including an image encoder, a text encoder, and a multimodal encoder. To generate the perturbation of $\hat{\mathbf{I}}_m$, we first

[1]According to our investigation on the VQAv2 validation set, each sentence is only composed of an average of 6.21 words.

follow existing work to update the image perturbation by minimizing the following loss function:

$$\mathcal{L}_f^m = \underbrace{\sum_{i=1}^{L_p}\sum_{j=1}^{D_p} Cos(\mathbf{f}_{i,j}^p,\ \hat{\mathbf{f}}_{i,j}^p)}_{\text{image encoder}} + \underbrace{\sum_{i=1}^{L_q}\sum_{j=1}^{D_q} Cos(\mathbf{f}_{i,j}^q,\ \hat{\mathbf{f}}_{i,j}^q)}_{\text{multimodal encoder}}, \tag{3}$$

where L_p and L_q denote the number of layers in the image encoder and multimodal encoder, respectively. D_p and D_q represent the number of input tokens of the image encoder and multimodal encoder. For the image encoder, the input tokens are image patches; and the multimodal encoder takes the representations from both image patches and text words as the input tokens. $\mathbf{f}_{i,j}^p$ and $\mathbf{f}_{i,j}^q$ are the output feature representation vectors of the j-th token in the i-th layer with the clean input pair $(\mathbf{I}, \mathbf{T})$. $\hat{\mathbf{f}}_{i,j}^p$ and $\hat{\mathbf{f}}_{i,j}^q$ denote the output feature representation vectors of the j-th neuron in the i-th layer with the perturbed input pair $(\hat{\mathbf{I}}_{m-1}, \hat{\mathbf{T}}_{m-1})$.

Let $\hat{\mathbf{I}}'_m$ denote the output by optimizing Eq. (3) with the clipping technique, which is further used to generate an enhanced image perturbation in the following section. This step is shown in Algorithm 1 lines 4-6.

LLM-based Perturbation Enhancement In the context of transferable attacks, it is common for the pre-trained source model $\mathcal{F}$ to exhibit notable dissimilarities when compared to the victim target model $\mathcal{S}$. Consequently, relying solely on perturbing the latent representations using Eq. (3) may prove insufficient in ensuring the creation of high-quality adversarial samples capable of effectively attacking $\mathcal{S}$. To tackle this challenge, we present a solution that leverages the capabilities of Large Language Models (LLMs), such as ChatGPT (OpenAI 2023), and the corresponding answers $\mathcal{Y}$ to bolster the process of perturbation generation.

• **Masked Text Generation with LLM.** In a given visual-question pairing, multiple correct answers can exist, represented as $\mathcal{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N]$, where N corresponds to the count of correct answers. The primary objective of the transferable attack is to create adversarial instances in such a manner that the output of $\mathcal{S}(\hat{\mathbf{I}}, \hat{\mathbf{T}})$ does not belong to the set $\mathcal{Y}$. To maximize the effectiveness of this transferable attack, a straightforward approach could involve compelling the pre-trained model $\mathcal{F}$ to produce incorrect predictions at each iteration. More specifically, this would entail ensuring that $\mathcal{F}(\hat{\mathbf{I}}'_m, \hat{\mathbf{T}}_{m-1}) \notin \mathcal{Y}$.

However, it is important to note that this approach is impractical for the current state of pre-trained models $\mathcal{F}$, as they are not explicitly designed for predicting VQA answers during their pre-training phase. Fortunately, a viable alternative stems from the fact that many of these models incorporate the masked language modeling (MLM) task as part of their pre-training. In this context, we can transform the answer prediction task into a masked answer recovery task using the MLM framework.

Towards this end, we need to combine the perturbed question $\hat{\mathbf{T}}_{m-1}$ and each correct answer $\mathbf{y}_i \in \mathcal{Y}$ with a predefined prompt $\mathcal{P}$ using LLMs. Let $\hat{\mathbf{Z}}_{m,i} = \text{LLM}(\hat{\mathbf{T}}_{m-1}, \mathbf{y}_i, \mathcal{P})$ denote the combined sentence for the i-th correct answer. The next step is to mask the answer $\mathbf{y}_i$ from the generated sentence $\hat{\mathbf{Z}}_{m,i}$. Note that each answer $\mathbf{y}_i$ may contain multiple words. Let $\mathcal{M}_i$ denote the set of masked indices, and we can use $\hat{\mathbf{Z}}_{m,i\setminus\mathcal{M}_i}$ to represent the masked sentence.

• **Masked Answer Anti-Recovery.** To achieve the transferable attack, we will prevent the model from recovering the correct answer tokens for each masked text $\hat{\mathbf{Z}}_{m,i\setminus\mathcal{M}_i}$, by minimizing the following anti-recovery loss:

$$\mathcal{L}_a^m = \sum_{i=1}^{N}\sum_{j\in\mathcal{M}_i} \log(p_c(z_{m,i,j}|\hat{\mathbf{Z}}_{m,i\setminus\mathcal{M}_i}, \hat{\mathbf{I}}'_m)), \tag{4}$$

where $z_{m,i,j}$ is the j-th token in $\hat{\mathbf{Z}}_{m,i}$, and p_c is the conditional probability score generated from the MLM head the pre-trained model $\mathcal{F}$ that is composed of a fully-connected layer and a softmax layer. After optimizing Eq. (4), we clip the learned image perturbation again, and the output is denoted as $\hat{\mathbf{I}}^*_m$. This step is shown in Algorithm 1 lines 7-10.

Cross-modal Joint Attack

Due to the differences of input image $\hat{\mathbf{I}}^*_m$ and text $\hat{\mathbf{T}}_{m-1}$, we cannot use a unified approach to update their perturbations. For the numerical image, we can still use gradients and the clipping technique to update the perturbation, but for the discrete text, we propose to use the word substitution technique to replace words in the text with the help of continuous word embeddings.

Joint Attack Trigger As discussed before, updating the text perturbation at each iteration is unnecessary. We design a heuristic function to determine when to trigger the joint attack by taking the number of informative words in the text (denoted as $|\mathcal{W}|$) and the maximum iterations M into consideration. When $m \mod \lfloor \frac{M}{|\mathcal{W}|+1} \rfloor = 0$, then VQATTACK triggers the joint attack. Here, the "+1" operation is to prevent attacking $\hat{\mathbf{T}}_{m-1}$ only in the last iteration step. The trigger is shown in Algorithm 1 line 12. Otherwise, VQATTACK will output $(\hat{\mathbf{I}}^*_m, \hat{\mathbf{T}}_{m-1})$ as $(\hat{\mathbf{I}}_m, \hat{\mathbf{T}}_m)$ for the m-the iteration (Algorithm 1 line 20).

Next, we introduce how to identify informative words and extract their synonyms. Given a clean text $\mathbf{T}$, we first tokenize it and filter out all stop words using the Natural Language Toolkit (NLTK)[2], which results in a set $\{t_i | i \in \mathcal{W}\}$, where $\mathcal{W}$ represents the indices of the unfiltered tokens. For each token t_i, we follow BERT-Attack (Li et al. 2020) and employ the BERT model (Devlin et al. 2019) to predict the top-K candidate words that share similar contexts, which results in a set of candidate words $\{c_{i,1}, \cdots, c_{i,K}\}$. We then obtain the candidate set for all tokens $i \in \mathcal{W}$, and obtain a set $\mathcal{C} = \{c_{i,j} | 1 \le j \le K\}_{i\in\mathcal{W}}$. The motivation for using BERT is that it can better capture the context of a word, compared to other methods like Glove (Pennington, Socher, and Manning 2014) and Word2Vec (Mikolov et al. 2013).

[2]https://www.nltk.org/

These candidates can retain more accurate syntactic and semantic information, making them more likely to satisfy semantic constraints. Note that this step can be done during "Initialization", which is fixed during word substitutions.

Cross-modal Perturbation Generation After triggering the cross-modal attack, VQATTACK will update the gradients with regard to both perturbed image and text via minimizing the following latent feature-level loss function:

$$\mathcal{L}_c^m = \mathcal{L}_f^m + \underbrace{\sum_{i=1}^{L_t}\sum_{j=1}^{D_t} Cos(\mathbf{f}_{i,j}^t,\ \hat{\mathbf{f}}_{i,j}^t)}_{\text{text encoder}}, \tag{5}$$

where $\mathcal{L}_f^m$ is the loss function from image and multimodal encoders with Eq. (3) using $(\hat{\mathbf{I}}_m^*, \hat{\mathbf{T}}_{m-1})$ and their corresponding token representations as the inputs, respectively. The second loss term is used to measure the feature similarity from the text encoder. L_t denotes the mumbler of layers in the text encoder, and D_t represents the number of input word tokens. $\mathbf{f}_{i,j}^t$ and $\hat{\mathbf{f}}_{i,j}^t$ denote the output feature representation vectors of from the clean input $(\mathbf{I}, \mathbf{T})$ and the perturbed input pair $(\hat{\mathbf{I}}_m^*, \hat{\mathbf{T}}_{m-1})$, respectively.

• **Image Perturbation Update.** Since the image perturbations are numerical values, we can directly calculate the gradients using Eq. (5) and then apply the clipping technique to generate the output $\hat{\mathbf{I}}_m$ for the m-th iteration, as shown in Algorithm 1 lines 13-15.

• **Text Perturbation Update.** Due to the discrete nature of text words, we need to unitize the learned gradients with Eq. (5) in the latent word embedding space to generate text perturbations motivated by (Ye et al. 2022b). Toward this end, we propose to use word substitution attacks to generate text perturbations.

Latent Word Embedding Estimation. The word substitution attack aims to replace the original, informative words in text $\hat{\mathbf{T}}_{m-1}$ with their synonyms, i.e., the words in set $\mathcal{C}$. To this end, we need to estimate the word representations after the attack first using the original informative word embeddings $\mathbf{E}(t_i)$ $(i \in \mathcal{W})$ and its gradient $\nabla\mathcal{L}_c^m(t_i)$ learned by Eq. (5) as follows:

$$\mathbf{E}(\hat{t}_i) = \mathbf{E}(t_i) + \nabla\mathcal{L}_c^m(t_i). \tag{6}$$

Synonym Ranking. The goal of synonym substitution is to find a synonym of t_i from $\{c_{i,1}, \cdots, c_{i,K}\}$ to replace the original informative word t_i and make the embedding of the synonym close to $\mathbf{E}(\hat{t}_i)$. Since there may be several informative words in $\mathcal{W}$, we need to decide the order of replacement. Intuitively, the larger similarity between $\mathbf{E}(\hat{t}_i)$ and the embedding of a synonym $c_{i,j}$, the higher chance of $c_{i,j}$ being a perturbation. To this end, we replace the original word with each synonym $c_{i,j}$ to generate each synonym's context-aware word embedding $\mathbf{E}(c_{i,j})$. We then calculate the pairwise cosine similarity between the estimated latent representation and the synonym context-aware word embedding as follows:

$$\gamma_{i,j} = Cos(\mathbf{E}(\hat{t}_i), \mathbf{E}(c_{i,j})). \tag{7}$$

According to the similarity score values, we rank all the synonyms in $\mathcal{C}$ in descending order, denoted as $\mathcal{R}(\mathcal{C})$.

Synonym Substitution. We replace the original word in $\hat{\mathbf{T}}_{m-1}$ with its synonym that has the largest similarity in $\mathcal{R}(\mathcal{C})$. Let $\hat{\mathbf{T}}'_{m-1}$ denote the new text sample. Then we check whether the new sample $\hat{\mathbf{T}}'_{m-1}$ satisfies the constraint listed in Eq. (2). If $Cos(U(\hat{\mathbf{T}}'_{m-1}), U(\mathbf{T})) > \sigma_s$, then we keep the replacement in $\hat{\mathbf{T}}_{m-1}$, remove all the other synonyms of this word in $\mathcal{R}(\mathcal{C})$, and move to the next informative word. If $Cos(U(\hat{\mathbf{T}}'_{m-1}), U(\mathbf{T})) \leq \sigma_s$, we do not conduct the replacement and use the synonym with the second largest value in $\mathcal{R}(\mathcal{C})$.

We will repeat this procedure until all informative words are replaced or all synonyms in $\mathcal{R}(\mathcal{C})$ are checked. The output from this step is the perturbed text $\hat{\mathbf{T}}_m$ as shown in Algorithm 1 lines 16-19.

After executing all the above steps for M iterations, we generate the final perturbed image-text pair $(\hat{\mathbf{I}}_M, \hat{\mathbf{T}}_M)$, which will be fed into different unknown victim models to conduct the transferable adversarial attack.

Experiments

Experimental Setup

Datasets & **Models** We evaluate the proposed VQATTACK on the VQAv2 (Antol et al. 2015) and TextVQA (Singh et al. 2019) datasets. We randomly select **6,000** and **1,000 correctly predicted samples** from the VQAv2 and TextVQA validation datasets, respectively. Because an image-question pair may have multiple candidate answers provided by crowd workers, we define a correct prediction only if the predicted result is the same as the label with the highest VQA score[3]. Each selected sample is correctly classified by all target models. We also development experiments on **five models**, including ViLT (Kim, Son, and Kim 2021), TCL (Yang et al. 2022), ALBEF (Li et al. 2021b), VLMO-Base (VLMO-B) (Bao et al. 2022), and VLMO-Large (VLMO-L) (Bao et al. 2022). Note that VLMO-B and VLMO-L share the same structure but have different model sizes. These models are first pre-trained on public image-text pairs and then fine-tuned on VQA datasets.

Baselines We comprehensively compare VQATTACK with text, image, and multi-modal adversarial attack methods. Specifically, we first adopt BERT-Attack (B&A) (Li et al. 2020) and Rewrite-Rollback (R&R) (Xu et al. 2022) as **text-attack** baselines. For **image attack** methods, we adopt DR (Lu et al. 2020), SSP (Naseer et al. 2020), and FDA (Ganeshan, S., and Radhakrishnan 2019) as baselines. These methods generate adversarial images by only perturbing intermediate features and can thus be directly utilized in our problem. VQATTACK is also compared with the **multimodal attack** approach Co-Attack (CoA) (Zhang, Yi, and Sang 2022). To the best of our knowledge, it is the only scheme that attempts to simultaneously add image and text

[3]VQA score calculates the percentage of the predicted answer that appears in 10 reference ground truth answers. More details can be found via https://visualqa.org/evaluation.html

Source Model	Target Model	VQAv2							TextVQA						
		Text Only		Image Only			Multi-modality		Text Only		Image Only			Multi-modality	
		B&A	R&R	DR	FDA	SSP	CoA	VQATTACK	B&A	R&R	DR	FDA	SSP	CoA	VQATTACK
ViLT	ALBEF	10.28	5.20	8.78	9.84	24.90	16.70	**30.36**	13.00	5.80	8.20	9.40	17.00	15.40	**22.20**
	TCL	11.86	6.08	8.74	9.62	22.54	17.84	**27.96**	12.20	4.80	7.10	8.20	13.60	14.90	**19.80**
	VLMO-B	6.34	1.82	5.08	5.70	21.48	13.64	**25.72**	7.30	3.20	7.40	5.70	13.90	12.90	**19.50**
	VLMO-L	5.02	2.18	5.58	5.72	13.08	10.64	**25.98**	6.60	0.30	2.80	2.40	7.20	7.60	**8.40**
TCL	ViLT	6.68	2.52	5.74	5.78	11.04	11.22	**21.80**	9.30	2.60	4.60	5.20	7.10	10.80	**16.30**
	ALBEF	5.58	2.92	11.10	12.52	38.26	33.24	**58.42**	10.80	8.70	9.10	10.50	31.80	26.10	**46.80**
	VLMO-B	7.52	3.84	15.82	9.00	23.88	18.32	**47.48**	7.82	2.54	6.50	7.60	16.70	15.50	**34.00**
	VLMO-L	5.64	2.22	8.04	6.14	15.26	12.64	**30.46**	2.40	5.96	3.80	4.70	9.50	10.00	**18.60**
ALBEF	ViLT	6.72	2.42	6.90	7.02	11.42	11.36	**21.60**	8.70	2.60	4.60	5.80	8.20	11.70	**15.60**
	TCL	6.96	1.80	12.64	11.78	35.46	27.24	**61.32**	9.90	2.90	9.60	8.80	13.10	20.50	**43.70**
	VLMO-B	5.68	2.04	8.14	9.04	21.48	16.16	**42.32**	8.50	3.30	7.70	8.10	15.20	14.50	**28.30**
	VLMO-L	5.02	2.18	5.58	5.72	21.56	10.64	**25.98**	5.70	2.20	4.10	4.50	8.20	7.40	**16.20**
VLMO-B	ViLT	7.72	2.04	4.36	5.34	10.20	10.90	**18.70**	10.90	0.80	3.20	3.40	7.80	11.70	**15.20**
	TCL	12.20	6.26	10.98	13.64	20.24	21.52	**43.62**	13.50	4.50	8.20	9.30	14.30	18.00	**28.30**
	ALBEF	10.74	6.30	11.22	14.52	22.66	22.46	**48.06**	13.50	6.10	9.50	12.70	16.80	19.60	**32.60**
	VLMO-L	5.98	3.96	4.58	5.48	10.66	12.52	**30.82**	6.70	0.60	2.70	4.20	6.80	9.60	**17.40**
VLMO-L	ViLT	7.50	1.62	7.48	3.52	7.94	8.78	**13.08**	10.30	1.30	3.00	2.90	5.80	9.20	**13.10**
	TCL	12.20	6.14	12.10	10.92	21.18	15.48	**32.96**	12.90	4.40	6.90	6.80	15.60	13.70	**21.70**
	ALBEF	10.84	5.98	24.84	10.90	24.50	15.14	**37.48**	13.00	6.40	9.30	9.40	17.00	12.30	**26.80**
	VLMO-B	8.22	1.86	20.96	7.58	19.60	12.70	**33.78**	8.70	1.90	6.00	4.50	14.20	11.60	**25.20**

Table 1: Comparison between VQATTACK and baselines on different models using the VQAv2 and TextVQA datasets (%).

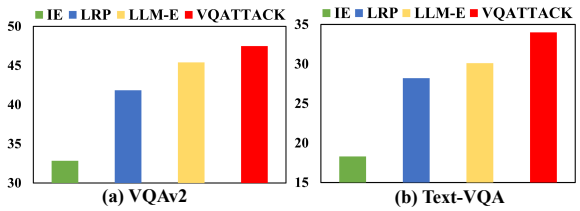

Figure 3: Ablation study results on the source model TCL and the victim model VLMO-B.

perturbations. We adopt the *Attack Success Rate* (**ASR**) as the evaluation metric, which measures the ratio of samples whose predicted labels are not in the correct answers.

Result Analysis

We alternatively select a pre-trained model as the source model to generate adversarial samples, which are then used to attack the remaining models treated as victims. Experimental results are listed in Table 1. We can observe that the proposed VQATTACK significantly outperforms all baselines on each dataset for the five transferable attack experiments. Specifically, VQATTACK achieves an average ASR of 22.49% using ViLT as the source model, 34.23% for TCL, 31.88% for ALBEF, 29.33% for VLMO-B and 25.51% for VLMO-L. The ASR value is comparatively lower when using ViLT as the source model because its model structure and pre-training strategies are greatly different from others. Also, the ASR value obtained by using VLMO-L as the source model is slightly lower than that of using VLMO-B as the source model. This observation demonstrates that the model owns larger parameters can present better adversarial robustness. Finally, all of these results have demonstrated the effectiveness of our proposed approach and also comprehensively reveal the huge threat of adversarial attacks in the "pre-training & fine-tuning" learning paradigm.

Ablation Study

This ablation study aims to validate the effectiveness of the two designed modules. Figure 3 shows the ablation study results using the adversarial samples generated by TCL to attack VLMO-B. "IE" means only using the latent presentations learned by the image encoder to generate adversarial samples in Eq. (3). "LRP" means the latent representation perturbation used in the LLM-enhanced image attack module, where we only use Eq. (3) to generate the adversarial samples. We can observe that using the multimodal encoder can make significant ASR improvements. "LLM-E" means using both Eqs. (3) and (4) to generate perturbations. Compared with "LRP", the performance can increase, which indicates the efficacy of introducing LLM to help generate masked text and the effectiveness of the designed masked answer anti-recovery loss in Eq. (4). The proposed VQATTACK achieves the best performance. The performance gap between LLM-E and VQATTACK demonstrates the effectiveness of the proposed cross-modal joint attack module.

Case Study

We conduct a case study on the VQAv2 dataset using the source model TCL, as shown in Figure 4. We can observe that the generated adversarial samples largely change the original correct prediction to a wrong answer. For instance, recognizing a kitchen as a bedroom (column 3). Furthermore, the generated adversarial samples still keep the nat-

Modality	Method	VQAv2					TextVQA				
		ViLT	ALBEF	TCL	VLMO-B	VLMO-L	ViLT	ALBEF	TCL	VLMO-B	VLMO-L
Text Only	B&A	15.16	8.24	8.96	10.16	11.72	20.20	11.50	14.90	13.00	10.10
	R&R	7.30	4.68	5.64	6.86	4.62	7.40	8.30	5.90	2.70	3.80
Image Only	DR	16.90	20.42	15.82	17.12	11.02	14.40	14.50	11.60	14.50	7.90
	FDA	20.08	17.72	16.74	22.16	9.92	13.90	12.80	11.70	19.50	7.50
	SSP	61.36	49.68	51.46	46.32	41.94	49.80	36.70	40.70	34.60	28.40
Multi-modality	Co-Attack	50.12	46.50	52.74	43.56	18.48	42.30	35.80	45.80	35.80	16.80
	VQATTACK	**79.00**	**75.16**	**76.46**	**75.04**	**61.60**	**65.00**	**61.90**	**65.70**	**66.20**	**48.70**

Table 2: Results of transferable attacks between $\mathcal{F}$ and $\mathcal{S}$ with the same pre-trained structures.

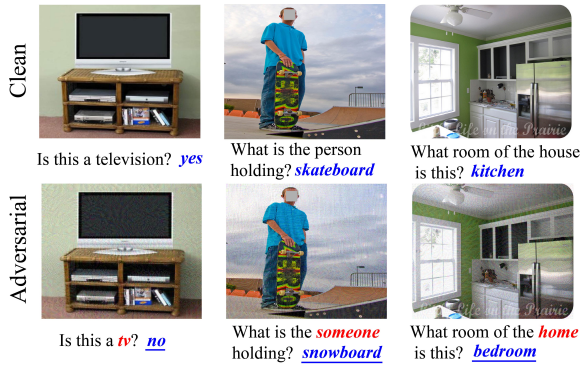

Figure 4: Qualitative results of VQATTACK on the VQAv2 dataset generated by the TCL model. The original answer and perturbed words are displayed in blue and red, respectively. The wrong prediction is shown with an underline.

ural appearance as the benign samples, which demonstrates a serious security threat in the present VQA systems.

Transferable Attacks with Shared Information

In this experiment, we use the pre-trained model $\mathcal{F}$ as the source model and its downstream VQA task as the target $\mathcal{S}$. $\mathcal{F}$ and $\mathcal{S}$ share most of the structures, and only the final prediction layers are different. Table 2 shows the experimental results. We can observe that the proposed VQATTACK still outperforms all the baselines on the two VQA datasets. Compared with the results listed in Table 1, we can observe that the performance of all approaches improves significantly under this setting. This experiment concludes that the shared information is sensitive, which may make the target models vulnerable.

Related Work

Robustness of VQA The robustness of VQA is moderately explored. Recently, Fool-VQA (Xu et al. 2018) explores the adversarial vulnerability of a VQA system by adding image noise constrained by l_∞ distance. TrojVQA (Walmer et al. 2022) performs a backdoor attack by injecting deliberate image patches and word tokens. These studies concentrate on the robustness of end-to-end trained VQA models and design algorithms based on the final predictions. Because the outputs of pre-trained and fine-tuned VQA models are different, they cannot be extended to our problem.

Adversarial Attacks Adversarial image attacks are initially explored in Fast Gradient Sign Method (Goodfellow, Shlens, and Szegedy 2015) and Projected Gradient Decent (Madry et al. 2018). An intriguing property of these adversarial images is their "transferability", which can be utilized to attack different image models with unknown parameters and structures. To enhance the transferability, the recently proposed methods exploit features from intermediate layers for adversarial attacks. They either combine the feature distortion loss with the classification cross-entropy term (Huang et al. 2019; Inkawhich et al. 2020a,b) or fully rely on the intermediate feature disruption (Ganeshan, S., and Radhakrishnan 2019; Naseer et al. 2020). Text attack methods are primarily divided into searching-based and gradient-based algorithms. Searching-based attacks include a set of heuristic ranking algorithms (Li et al. 2021a, 2020; Xu et al. 2022) with sub-optimal performance. Recently, gradient-based attacking approaches (Guo et al. 2021; Wang et al. 2022; Ye et al. 2022a,b) are proposed. Unlike image attacks, the gradient cannot be directly projected onto discrete text inputs. Accordingly, gradient change is instantiated either through distance matching on candidate word embeddings (Wang et al. 2022; Ye et al. 2022a,b), or by using Gumbel-softmax sampling (Jang, Gu, and Poole 2017) on a learnable distribution of all candidate words. For multi-modal attacks, the recently proposed Co-attack (Zhang, Yi, and Sang 2022) method firstly combines both image and text attacks, which utilizes word substitution to guide image adversarial attacks. It has demonstrated to some extent that perturbations across both modalities can be more effective than a single source. However, it does not take into account the dynamic connections between perturbations on different modalities, indicating potential space for significant improvements under more challenging scenarios.

Conclusion

In this paper, we investigate a novel transferable adversarial attack scenario, aiming to generate adversarial samples only using pre-trained models, which are used to attack different black-box victim models. Correspondingly, we propose a new model named VQATTACK, which can jointly update both image and text perturbations. Besides, we propose to incorporate the large language model to enhance the transferability of the source model. Experimental results on two VQA datasets with five models show the effectiveness of the proposed VQATTACK for the transferable attacks.

Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 1951729, 1953813, 2119331, 2212323, and 2238275.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433. IEEE Computer Society.

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *NeurIPS*.

Cer, D.; Yang, Y.; Kong, S.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In *EMNLP (Demonstration)*, 169–174. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*. OpenReview.net.

Ganeshan, A.; S., V. B.; and Radhakrishnan, V. B. 2019. FDA: Feature Disruptive Attack. In *ICCV*, 8068–8078. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Guo, C.; Sablayrolles, A.; Jégou, H.; and Kiela, D. 2021. Gradient-based Adversarial Attacks against Text Transformers. In *EMNLP (1)*, 5747–5757. Association for Computational Linguistics.

Huang, Q.; Katsman, I.; Gu, Z.; He, H.; Belongie, S. J.; and Lim, S. 2019. Enhancing Adversarial Example Transferability With an Intermediate Level Attack. In *ICCV*, 4732–4741. IEEE.

Inkawhich, N.; Liang, K. J.; Carin, L.; and Chen, Y. 2020a. Transferable Perturbations of Deep Feature Distributions. In *ICLR*. OpenReview.net.

Inkawhich, N.; Liang, K. J.; Wang, B.; Inkawhich, M.; Carin, L.; and Chen, Y. 2020b. Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability. In *NeurIPS*.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *ICLR (Poster)*. OpenReview.net.

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. BadEncoder: Backdoor Attacks to Pre-trained Encoders in Self-Supervised Learning. In *IEEE Symposium on Security and Privacy*, 2043–2059. IEEE.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In *AAAI*, 8018–8025. AAAI Press.

Kenfack, F. K.; Siddiky, F. A.; Balint-Benczedi, F.; and Beetz, M. 2020. RobotVQA - A Scene-Graph- and Deep-Learning-based Visual Question Answering System for Robot Manipulation. In *IROS*, 9667–9674. IEEE.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 5583–5594. PMLR.

Li, D.; Zhang, Y.; Peng, H.; Chen, L.; Brockett, C.; Sun, M.; and Dolan, B. 2021a. Contextualized Perturbation for Textual Adversarial Attack. In *NAACL-HLT*, 5053–5069. Association for Computational Linguistics.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *NDSS*. The Internet Society.

Li, J.; Selvaraju, R. R.; Gotmare, A.; Joty, S. R.; Xiong, C.; and Hoi, S. C. 2021b. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 9694–9705.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *EMNLP (1)*, 6193–6202. Association for Computational Linguistics.

Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Lu, Y.; Jia, Y.; Wang, J.; Li, B.; Chai, W.; Carin, L.; and Velipasalar, S. 2020. Enhancing Cross-Task Black-Box Transferability of Adversarial Examples With Dispersion Reduction. In *CVPR*, 937–946. Computer Vision Foundation / IEEE.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR (Poster)*. OpenReview.net.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR (Workshop Poster)*.

Naseer, M.; Khan, S. H.; Hayat, M.; Khan, F. S.; and Porikli, F. 2020. A Self-supervised Approach for Adversarial Robustness. In *CVPR*, 259–268. Computer Vision Foundation / IEEE.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *EMNLP*, 1532–1543. ACL.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In *CVPR*, 8317–8326. Computer Vision Foundation / IEEE.

Walmer, M.; Sikka, K.; Sur, I.; Shrivastava, A.; and Jha, S. 2022. Dual-Key Multimodal Backdoors for Visual Question Answering. In *CVPR*, 15354–15364. IEEE.

Wang, B.; Xu, C.; Liu, X.; Cheng, Y.; and Li, B. 2022. SemAttack: Natural Textual Attacks via Different Semantic Spaces. In *NAACL-HLT (Findings)*, 176–205. Association for Computational Linguistics.

Wu, D.; Wang, Y.; Xia, S.-T.; Bailey, J.; and Ma, X. 2020. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving Transferability of Adversarial Examples With Input Diversity. In *CVPR*, 2730–2739. Computer Vision Foundation / IEEE.

Xu, L.; Cuesta-Infante, A.; Berti-Équille, L.; and Veeramachaneni, K. 2022. R&R: Metric-guided Adversarial Sentence Generation. In *AACL/IJCNLP (Findings)*, 438–452. Association for Computational Linguistics.

Xu, X.; Chen, X.; Liu, C.; Rohrbach, A.; Darrell, T.; and Song, D. 2018. Fooling Vision and Language Models Despite Localization and Attention Mechanism. In *CVPR*, 4951–4961. Computer Vision Foundation / IEEE Computer Society.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-Language Pre-Training with Triple Contrastive Learning. In *CVPR*, 15650–15659. IEEE.

Ye, M.; Chen, J.; Miao, C.; Wang, T.; and Ma, F. 2022a. LeapAttack: Hard-Label Adversarial Attack on Text via Gradient-Based Optimization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2307–2315.

Ye, M.; Miao, C.; Wang, T.; and Ma, F. 2022b. TextHoaxer: budgeted hard-label adversarial attacks on text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3877–3884.

Yu, T.; Shen, Y.; and Jin, H. 2019. A Visual Dialog Augmented Interactive Recommender System. In *KDD*, 157–165. ACM.

Zhan, L.; Liu, B.; Fan, L.; Chen, J.; and Wu, X. 2020. Medical Visual Question Answering via Conditional Reasoning. In *ACM Multimedia*, 2345–2354. ACM.

Zhang, J.; Yi, Q.; and Sang, J. 2022. Towards Adversarial Attack on Vision-Language Pre-training Models. In *ACM Multimedia*, 5005–5013. ACM.

Zhou, W.; Hou, X.; Chen, Y.; Tang, M.; Huang, X.; Gan, X.; and Yang, Y. 2018. Transferable Adversarial Perturbations. In *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, 471–486. Springer.

Appendix

A. Details of LLM Utilization

In this section, we introduce how to use the LLM to combine the perturbed question $\hat{\mathbf{T}}_{m-1}$ and a correct answer $\mathbf{y}_i$ into a sentence. We use ChatGPTv4 (Liu et al. 2023) to accomplish this target because it has demonstrated stronger language reasoning capabilities compared to other LLMs. When utilizing ChatGPT-v4, we first design a prompt $\mathcal{P}$ as illustrated in Figure 5.

question: What color is the fork?

answer: Black

Prompt: Please paraphrase the "{***question***}" into a declarative sentence using the "{***answer***}". The output need fullfill following requirments:
1: The term {***answer***} must appear in the output.
2: Please only output the declarative sentence.

Figure 5: An example of Transferable adversarial attacks on VQA via pre-trained models.

As illustrated in Figure 1, the prompt consists of three parts. The first is the task description, in which we need to combine "{question}" and "{answer}" into a declarative sentence. To ensure the quality of the combined sentences, we add the following two constraints:

- *The term {answer} must appear in the output*. This constraint emphasizes that the output sentence must contain the answer, which is a prerequisite for masked text generation.
- *Please only output the declarative sentence*. This constraint aims to avoid redundant prompts and prefixes, such as "*Sure, here is the output sentence...*". Such a prefix is unnecessary and may even introduce interference in the masked answer anti-recovery step, thus affecting the attack performance.

B. Details of VL Models

In this section, we illustrate the details of all five VL models evaluated in our paper, including ViLT (Kim, Son, and Kim 2021), ALBEF (Li et al. 2021b), TCL (Yang et al. 2022), VLMO-B and VLMO-L (Bao et al. 2022). These models consist of an image encoder, a text encoder, and a multimodal encoder.

- **ViLT**. We select the model ViLT because it employs a succinct model structure and significantly outperforms previous end-to-end trained VQA models on the VQAv2 dataset. The image encoder of ViLT is a linear projection layer. Given an input image $\mathbf{I}$, it first divides the image into patches with equal spatial resolution. Then each image patch is flattened into a vector and encoded by a linear transformation, which results in D_p image tokens. For the text $\mathbf{T}$, it is first tokenized and embedded by the commonly used byte-pair encoder. The word embeddings are then encoded through a text encoder, which is a word-vector linear projection layer. The latent token representations from image and text encoders are then concatenated with a learnable special token $\langle cls \rangle$ and fed into the multimodal encoder, which is a twelve-layer transformer encoder. The encoder attends tokens of different modalities through the self-attention mechanism. At the pre-training stage, the output features of text tokens are fed into the MLM head to predict the masked word tokens. When fine-tuning on VQA task, the output feature from the $\langle cls \rangle$ token is fed into a VQA prediction head to select the correct answer. The VQA prediction head is composed of a fully-connect layer and a softmax layer.
- **ALBEF.** The ALBEF model has a different structure from ViLT. Specifically, the image encoder is a twelve-layer visual transformer ViT-B/16 (Dosovitskiy et al. 2021). The text encoder is a six-layer transformer encoder. The structure of the multi-modal encoder is the same as a six-layer transformer decoder, where each layer contains a self-attention module, a cross-attention module and a feed-forward module. After obtaining the image and text token features through the image/text encoders, the multi-modal encoder first accepts the text token features as input and attends to them through the self-attention module. Then, the output will fuse with image token features through the cross-attention module. At the pre-training stage, the output features from the multi-modal encoder will be processed by the MLM head. For prediction on the VQA task, the multi-modal features are fed into a six-layer transformer decoder. The decoder also accepts a sequence initialized by the $\langle cls \rangle$ token as input and interacts with the multi-modal representations through cross-attention layers. As a result, the VQA answer is auto-regressively generated in an open-ended manner.
- **TCL.** TCL follows the same structures as ALBEF but has more different pre-training tasks. In addition to the traditional pre-training tasks like MLM, it introduces additional tasks through contrastive learning. These contrastive learning tasks are developed from the uni-modal and cross-modal levels based on additional image-text pairs. The experimental results indicate that TCL improves the quality of multimodal representations by pre-training with these extra tasks. After fine-tuning TCL on the VQA task, its performance also notably outperforms that of ALBEF.
- **VLMO-B.** VLMO-B adopts a transformer structure for each encoder. Both the image and text encoders are one-layer transformer encoders, and the multimodal encoder is a twelve-layer transformer encoder. VLMO-B also adopts a modular design for each encoder, replacing the original feed-forward (FFN) layer with a modality-aware FFN head. At the pre-training stage, each encoder has different parameters in the modality-aware FFN head, while the multi-head self-attention module is shared between the image and text encoders. For the VQA prediction, the output feature from the $\langle cls \rangle$ token is fed into a VQA prediction head to select the correct answer, which is the same as the ViLT model.

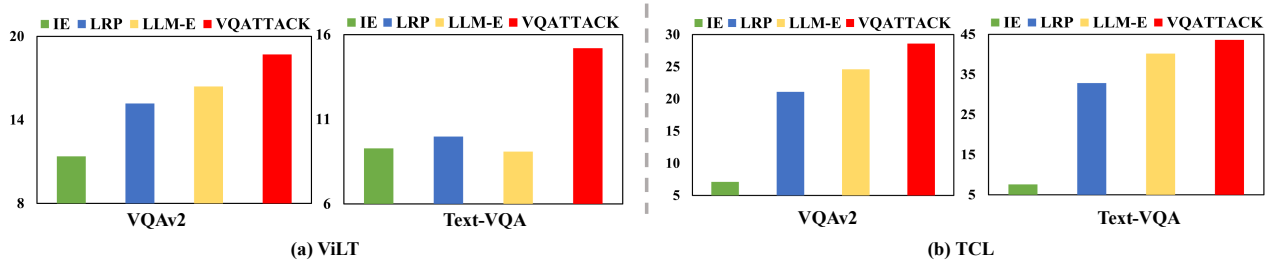

Figure 6: More examples of ablation study. We generate adversarial samples form the pre-trained VLMO-B model, and use the output to attack ViLT and TCL.

- **VLMO-L.** We also adopt a larger version of VLMO-B to evaluate the transferable attack performance under different model sizes. VLMO-L has the same structure as VLMO-B but with more layers and parameters. Specifically, the multi-modal encoder is a transformer encoder with 24 layers, and each latent representation owns a size of 1,024 dimensions, which is 768 in VLMO-B. As a result, the parameters of VLMO-L are three times larger than those of VLMO-B (562M vs 175M).

C. Implementation Details

In this section, we show the implementation details of our experiments. For the perturbation parameters on images, we follow the previous transferable image attack methods (Huang et al. 2019; Zhou et al. 2018) and set the L_∞-norm distance threshold σ_i to 16/255. Following the PGD, the maximum iteration steps M is set to 20 on VQATTACK and 40 on all image attack baselines, including SSP, FDA, DR, and the multi-modal attack Co-Attack (CoA). This is because, in most steps, VQATTACK needs to update gradients twice in one iteration step, while other baselines do it only once. Thus, we set a smaller M to our approach. Finally, all methods compared in experiments are optimized with the DIM (Xie et al. 2019), which is an image augmentation strategy and has been widely adopted in current transferable image attacks (Lu et al. 2020; Wu et al. 2020).

For the text modality, we set the semantic similarity constraint σ_s to 0.95, which follows the previous text-attack work (Li et al. 2020; Xu et al. 2022). Because the text-attack baselines B&A and R&R need to do queries to the target model, which is different from our transferable attack setting. Thus, when running B&A and R&R, we first generate adversarial texts by querying the VQA model fine-tuned on the source pre-trained one and then perform a transferable attack by sending the querying results to the victim model. Finally, all experiments are conducted on a single GTX A100 GPU.

D. More Ablation Study Analysis

As shown in Fig. 6, we also display more results of the ablation study. The adversarial examples are generated from the pre-trained ViLT source model, and they are transferred to attack the victim ViLT and TCL models. From the figure, we can observe that the performance is consistent with the analysis of the ablation study, except the ASR value of LLM-E is slightly lower than LRP on the ViLT model through the Text-VQA dataset. We attribute this to the huge differences in pre-training strategies on MLM tasks between VLMO-B and ViLT. Specifically, the ViLT model directly uses the MLM task to pre-trained the whole model from initialization. However, VLMO-B adopts a stage-wise pre-training strategy, which first pre-trains the image and text encoder on uni-modal tasks and then trains the multi-modal encoder on the MLM task with a good initialization of uni-modal representations. Finally, by combining the LLM-E image attack module with the cross-modal joint attack module, the performance of VQATTACK significantly surpasses that of individual components across all figures. This further demonstrates the effectiveness of our proposed method.