QAVA: Query-Agnostic Visual Attack to Large Vision-Language Models

Yudong Zhang[1,2], Ruobing Xie[2,✉], Jiansheng Chen[3,✉], Xingwu Sun[2,4], Zhanhui Kang[2], Yu Wang[1,5,✉]

[1]Department of Electronic Engineering, Tsinghua University,
[2]Machine Learning Platform Department, Tencent,
[3]School of Computer and Communication Engineering, University of Science and Technology Beijing, [4]Faculty of Science and Technology, University of Macau,
[5]State Key Laboratory of Space Network and Communications, Tsinghua University

zhangyd16@mails.tsinghua.edu.cn, xrbsnowing@163.com, jschen@ustb.edu.cn, sunxingwu01@gmail.com, kegokang@tencent.com, yu-wang@mail.tsinghua.edu.cn. (✉: Corresponding authors)

Abstract

In typical multimodal tasks, such as Visual Question Answering (VQA), adversarial attacks targeting a specific image and question can lead large vision-language models (LVLMs) to provide incorrect answers. However, it is common for a single image to be associated with multiple questions, and LVLMs may still answer other questions correctly even for an adversarial image attacked by a specific question. To address this, we introduce the query-agnostic visual attack (QAVA), which aims to create robust adversarial examples that generate incorrect responses to unspecified and unknown questions. Compared to traditional adversarial attacks focused on specific images and questions, QAVA significantly enhances the effectiveness and efficiency of attacks on images when the question is unknown, achieving performance comparable to attacks on known target questions. Our research broadens the scope of visual adversarial attacks on LVLMs in practical settings, uncovering previously overlooked vulnerabilities, particularly in the context of visual adversarial threats. The code is available at https://github.com/btzyd/qava.

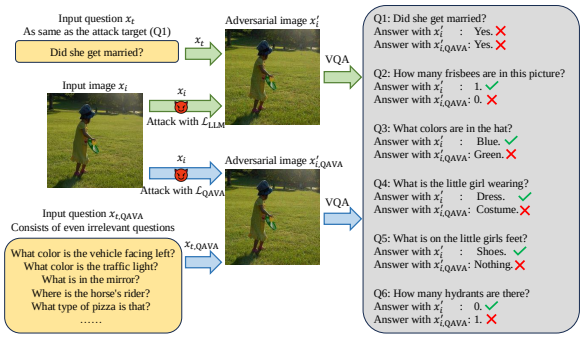


Figure 1: Traditional adversarial attacks involve inputting an image x_i and a specified target question $x_{t,target}$ into LVLMs, with adversarial images generated through gradient-based methods. This approach typically results in incorrect answers for x_i and $x_{t,target}$ (*i.e.*, Q1). However, for other questions $x_{t,other} \in \{x_{t,other} \in \mathcal{T} | x_{t,other} \neq x_{t,target}\}$ within the question set $\mathcal{T}$ that are not the same as the $x_{t,other}$, it remains possible for LVLMs to provide correct answers (*i.e.*, Q2-Q6). Our QAVA samples a set of questions $x_{\text{t,QAVA}}$ and performs attacks on these questions, even if they are unrelated to the original image x_i. QAVA generates adversarial images that are likely to produce incorrect responses when faced with unknown target questions.

1 Introduction

With the expansion in model parameters and training datasets, large vision-language models (LVLMs) have gained significant popularity, demonstrating exceptional performance across various tasks, including image classification, image captioning, semantic segmentation, and visual question answering (VQA) (Liu et al., 2023a; Alayrac et al., 2022; Wang et al., 2023). However, training LVLMs from the ground up is resource-intensive. As a result, the prevailing approach involves fine-tuning pre-trained visual encoders and large language models (LLMs) while training a vision-language alignment module. This process adapts visual tokens to the input space of the LLMs after they pass through the alignment module, enabling the LLMs to effectively process visual tokens. A well-known and efficient visual-language alignment module is Q-former (Li et al., 2023), which is employed by many popular LVLMs to bridge the visual encoder and the LLM (Dai et al., 2023; Zhu et al., 2023).

Despite their robust capabilities, LVLMs remain vulnerable to adversarial attacks. Attack-Bard (Dong et al., 2023) employs surrogate models to manipulate images, causing LVLMs to err on the image captioning task using the fixed prompt, "Describe this image". Similarly, VLAttack (Yin et al., 2023) targets both visual and textual modalities to disrupt the output of LVLMs for a specific question and image. However, these methods focus

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10205–10218

exclusively on attacks against *one image and one question at a time*. Consequently, the adversarial examples generated by these attacks may not be effective when confronted with different questions.

To develop more potent attacks, our objective is to manipulate images so that they yield incorrect answers to an *unknown set of target questions*, noted as **query-agnostic visual attack (QAVA)**. These query-agnostic adversarial examples have the potential to cause significant disruption to the model. For instance, during the inference phase, LVLMs may consistently provide incorrect answers to any question regarding the manipulated image. Moreover, employing these adversarial examples during the training or supervised fine-tuning phase could be even more detrimental, particularly when used for data poisoning. We enhance the attack by scrutinizing both the attack's location and the selection of questions used.

In terms of attack positioning, traditional attack methods typically employ the end-to-end loss function of the entire LVLM as the attack objective function. However, for adversarial attacks, it is crucial to identify and target the most vulnerable component of the LVLM. Given the extensive number of parameters in large language models, we posit that end-to-end attacks on LVLMs may not be as effective as targeting the inputs to the LLMs. Consequently, we focus on **attacking the output of the visual-language alignment module**. The visual-language alignment module's output encompasses multimodal interactions, and we can disrupt these critical multimodal interactions by targeting the alignment module, potentially leading to more effective attack outcomes.

Regarding the questions employed in attacks, we observed that when targeting the visual-language alignment module, effective attack performance can be achieved even with the use of *randomized questions that are unrelated to the image*. Furthermore, the attack is enhanced when a larger number of random, irrelevant questions are utilized, which verifies our QAVA's flexibility and effectiveness.

In conclusion, QAVA diverges from traditional attacks in two key aspects: (1) We focus on attacking the output of the visual-language alignment module within LVLMs, which is verified to be more vulnerable to query-agnostic attacks. (2) We utilize multiple randomized, image-independent questions in our attacks, ensuring that the adversarial examples are maximally incorrect when confronted with unknown potential inputs.

Our main contributions are summarized as follows: (1) We introduce a query-agnostic attack method, QAVA, which enhances the practicality of adversarial attacks on images within LVLMs. (2) We identify the vulnerability of visual-language alignment modules in LVLMs to adversarial attacks and leverage this vulnerability to execute query-agnostic attacks. (3) Extensive experiments demonstrate the efficacy of our QAVA approach in both white-box and black-box attack scenarios. Additionally, our QAVA method exhibits inter-task transferability, such as transferring from the VQA task to the image captioning task. This serves as an important alert regarding the security of LVLMs.

2 Related Work

Large vision-language models. LVLMs are typically composed of a pre-trained LLM, a visual encoder, and a projector that aligns visual and textual modalities. Recent popular LVLMs include InstructBLIP (Dai et al., 2023) and MiniGPT-4 (Zhu et al., 2023). Both models utilize EVA-CLIP (Sun et al., 2023) as the visual encoder and employ the Q-Former (Li et al., 2023) for aligning textual and visual modalities. For the LLM component, models such as Vicuna (Chiang et al.) and FlanT5 (Chung et al., 2022) are viable options. These LVLMs have demonstrated outstanding performance across various multimodal tasks, including image classification and VQA (Antol et al., 2015), among others.

Adversarial Attacks. By introducing small perturbations to the inputs of neural networks, adversarial attacks (Szegedy et al., 2014; Nguyen et al., 2015) can cause models to produce incorrect outputs. These attacks can be categorized into white-box and black-box (or gray-box) attacks (Papernot et al., 2016). In white-box attacks, the adversary has full access to the model's parameters. Conversely, in black-box or gray-box attacks, the adversary has limited information, such as the ability to make a certain number of queries to the model or knowledge of some of the model's parameters. Furthermore, adversarial attacks can be classified as either targeted or untargeted. Untargeted attacks aim to generate incorrect outputs, while targeted attacks strive to manipulate the output to meet the adversary's specific expectations. Initial research on adversarial attacks concentrated mainly on the visual modality, given its high-dimensional and continuous input space (Moosavi-Dezfooli et al., 2016; Goodfellow et al., 2015; Carlini and Wagner,

2017). More recent studies have extended the attacks to discrete textual modalities (Alzantot et al., 2018; Jia and Liang, 2017; Wallace et al., 2019). Additionally, some research has focused on targeting the fusion of visual and textual modalities (Zhang et al., 2022; Lu et al., 2023).

LVLMs and Adversarial Attacks. With the increasing popularity of LVLMs, numerous recent studies have focused on adversarial attacks against these models. Recent research has demonstrated the feasibility of generating adversarial examples to jailbreak LVLMs (Shayegani et al., 2023b). This includes attacking images using gradient-based approaches (Carlini et al., 2023), targeting texts through prompt engineering (Liu et al., 2023c), and embedding malicious instructions into images as text, with the aim of having the model execute these commands via optical character recognition (OCR) (Shayegani et al., 2023a). While these studies primarily address the security concerns surrounding LVLMs, our research is specifically focused on the safety and integrity of images within these models. Some studies (Luo et al., 2024) have also concentrated on query-agnostic adversarial attacks, in which LVLMs are prompted to respond with answers such as "none" or "don't know" to various inquiries. In contrast, our study specifically examines scenarios in which LVLMs are induced to provide incorrect answers.

3 Method

3.1 Preliminary

We provide a concise overview of the adversarial attack pipeline. This study specifically investigates gradient-based white-box adversarial attacks. Let the LVLM be represented as $y = f(f_i(x_i), x_t)$, where x_i and x_t denote the input image and text, respectively, with $f_i(\cdot)$ serving as a visual encoder, and y as the textual output generated by the LVLM. Given the input image x_i and text x_t, our objective is to identify an adversarial image x_i' such that $y' = f(f_i(x_i'), x_t)$ is semantically distant from y. This is subject to the condition that the difference between x_i and x_i' remains within the constraints ϵ, *i.e.*, $|x_i' - x_i|_p \leq \epsilon$.

FGSM (Goodfellow et al., 2015) generates the adversarial example x_i' by updating the original input x_i using a single gradient computation. In contrast, PGD (Madry et al., 2018) executes multiple iterative gradient updates, projecting x_i' after each update to ensure adherence to the perturbation constraints defined by ϵ_∞. The C&W attack (Carlini and Wagner, 2017) employs Eq. (1) as its optimization objective, which is iterated multiple times. The first component of Eq. (1) seeks to modify the model output to diverge significantly from the original output by employing a specific loss function $\mathcal{L}$. Meanwhile, the second component ensures that the adversarial example x' remains sufficiently close to the original input x. The constant c serves as a hyperparameter, balancing the divergence in model outputs against the l_2 distance between x' and x.

$$\mathcal{L}_{CW}(x, x', \dots) = \mathcal{L}(x, x', \dots) - c \times ||x - x'||_2^2 \quad (1)$$

The objective of QAVA is to manipulate images such that they yield incorrect responses to unknown target questions. Consequently, our task involves employing a specific method to adversarially attack a given image. We explore the selection of surrogate questions for these attacks in Sec. 3.2 and detail the associated loss functions in Sec. 3.3.

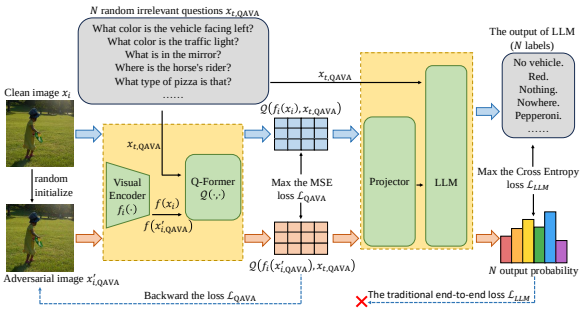


Figure 2: The framework of QAVA is structured as follows: Initially, we generate N randomly sampled questions, denoted as $x_{\text{t,QAVA}}$, which are not pertinent to the input image x_i. Subsequently, we introduce random perturbations to x_i to create the initial variant, $x'_{\text{i,QAVA}}$. Both x_i and $x_{\text{t,QAVA}}$ are then input into the LVLM, and the LVLM's response serves as a label. Despite the fact that the question $x_{\text{t,QAVA}}$ is unrelated to the image x_i, the LVLM still provides a response. Following this, we input $x'_{\text{i,QAVA}}$ and $x_{\text{t,QAVA}}$ into the LVLM to calculate the MSE loss based on the Q-former output features. Adversarial attacks are executed using techniques such as PGD or C&W by employing the loss functions, denoted as $\mathcal{L}_{\text{QAVA}}$. The traditional end-to-end attack loss function, $\mathcal{L}_{\text{LLM}}$, is also shown.

3.2 Strategies for Sampling Questions

We outline the four question sampling strategies utilized in our QAVA attack as follows.

White-box targeting questions (WTQ). WTQ em-

ploys a predefined set of target questions that are used to evaluate the adversarial example generated by attack. In contrast, our QAVA method does not have access to any information regarding the target questions at the time of the attack.

Visual question generation (VQG). VQG (Mostafazadeh et al., 2016) is capable of generating questions for input images, including questions with specific anticipated answers (*e.g.*, "yes" or "green"). In this strategy, we input the images into LVLMs and generate N questions using VQG prompts (*e.g.*, "Taking the image into account, generate N questions.").

Random sample questions (RSQ_N). RSQ_N randomly samples N questions from the validation set of VQA v2, which comprises 214,354 questions.

Random Sample Questions by Types (RSQ^t). RSQ^t ensures a balanced representation of each question type in the final set of sampled questions. This approach involves categorizing questions by type, such as "What is on the", "What animal is", "What color is", among others.

3.3 Design of the Loss Function $\mathcal{L}_{\text{QAVA}}$

For the LVLM f, the forward process $f(f_i(x_i), x_t)$ with a label generates a native loss function, $\mathcal{L}_{\text{LLM}}$, which is typically employed to optimize the adversarial image. Our aim is to identify more effective loss functions to enhance attack performance.

Revisiting the forward process of LVLMs, the Q-former, introduced by BLIP-2 and utilized in LVLMs such as InstructBLIP and MiniGPT-4, is instrumental in aligning visual and textual modalities within the feature space. For instance, in InstructBLIP, the image x_i is encoded into a vector of shape $[257, 1408]$, denoted as $f_i(x_i)$, by the image encoder f_i. The Q-former $\mathcal{Q}(\cdot,\cdot)$ then extracts a feature of shape $[32, 768]$, represented as $q = \mathcal{Q}(f_i(x_i), x_t)$, guided by the text x_t. This feature q is subsequently upscaled to $[32, 4096]$ and input into the LLM along with the text x_t, as illustrated in Fig. 2.

The feature vectors output by the Q-former can be utilized as supervised signals to optimize the adversarial image. Specifically, for a given question x_t and clean image x_i, we first input them into the Q-former to obtain $q = \mathcal{Q}(f_i(x_i), x_t)$. Next, we input the question x_t and the perturbed image x_i' into the Q-former to derive $q' = \mathcal{Q}(f_i(x_i'), x_t)$. We then optimize the MSE loss functions, *i.e.*

$$\begin{aligned} \max \quad & \mathcal{L}_{\text{QAVA}}(q, q') = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(q_{i,j} - q'_{i,j}\right)^2 \\ \text{subject to} \quad & |x_i - x_i'|_\infty \leq \epsilon_\infty \end{aligned} \tag{2}$$

The loss function $\mathcal{L}_{\text{QAVA}}$ is calculated for all white-box questions (*e.g.*, WTQ) or surrogate questions (*e.g.*, RSQ). After summing these individual computations, we obtain the overall loss function $\mathcal{L}_{\text{QAVA}}$, as detailed in Algorithm 1 in Appendix A.

4 Experiments

4.1 Experiment Settings

Datasets. We utilize the validation set of VQA v2 as the foundational dataset, comprising a total of 40,504 images and 214,354 questions. The distribution of questions across images in VQA v2 is uneven; for instance, over 18,000 images have only three questions, whereas merely 50 images have 50 or more questions. We define the dataset VQA v2 m+n as a subset of VQA v2, including m images, each associated with n questions, resulting in a total of $m \times n$ questions. To construct the dataset VQA v2 m+n, we randomly sample m images from those with at least n corresponding questions, followed by randomly selecting n questions for each image. Our experiments were performed on the VQA v2 32+50 data subset. We adhere to the official evaluation procedure designated for the VQA v2 dataset. It is crucial to acknowledge that the ground truth answers for each question in VQA v2 are not singular; each validation question possesses ten ground truth answers, which are utilized to compute the VQA scores.

Models. In our experiments, we employ BLIP-2 (Li et al., 2023) and InstructBLIP (Dai et al., 2023). Both LVLMs utilize CLIP as the visual encoder and incorporate a set of learnable queries along with a Q-former trained on a frozen visual encoder and LLM. Additionally, we assess the transferability of our QAVA approach on LLaVA (Liu et al., 2023a) and MiniGPT-4 (Zhu et al., 2023).

Adversarial Attacks. We employ two standard adversarial attack methods: PGD-l_∞ and CW-l_2. For the PGD attack, unless otherwise specified, we typically set the number of attack steps n to 20, the attack step size α to 2 (*i.e.*, $2/255$), and the maximum perturbation magnitude ϵ_∞ to 8. For the CW attack, we generally configure the number of attack steps n to 50, the attack step size α to

0.01 (*i.e.*, approximately $2.55/255$), and set the confidence level to 0. The choice of the constant c in the CW attack is contingent upon the loss function employed. Specifically, since $\mathcal{L}_{\text{LLM}}$ is approximately 20 times larger than $\mathcal{L}_{\text{QAVA}}$, we set the constant $c = 0.1$ when using $\mathcal{L}_{\text{LLM}}$ as the loss function. Conversely, when utilizing $\mathcal{L}_{\text{QAVA}}$, we set the constant $c = 0.005$.

4.2 Main Results of QAVA

$\mathcal{L}_{\text{LLM}}$(RSQ$_{\text{N}}$) can effectively attack images. We assess the baseline VQA performance of InstructBLIP with FlanT5$_{\text{XL}}$ and InstructBLIP with Vicuna-7B on the VQA v2 32+50 dataset. Subsequently, we apply PGD and CW attacks using WTQ to evaluate the maximum potential of attack performance. Following this, we randomly select N questions from the entire set of VQA v2 questions (214,354 questions) to form RSQ$_{\text{N}}$ and use these for the attack. The results, presented in Tab. 1, demonstrate that both PGD and CW can effectively reduce VQA scores when white-box questions are used. However, when using randomly sampled irrelevant questions, PGD and CW do not reach the same level of performance as with WTQ, displaying a difference of approximately 10 to 15 points.

Attack method	Question strategy	InstructBLIP(Vicuna-7B) Overall	Other	Number	Yes/No
✘	✘	78.00	66.98	69.68	94.29
PGD	WTQ$_{50}$	42.41	17.56	50.59	71.13
	RSQ$_1$	62.81(±1.34)	45.20	60.13	85.68
	RSQ$_5$	58.46(±0.66)	39.72	56.37	82.58
	RSQ$_{10}$	57.61(±1.50)	37.71	55.24	83.28
	RSQ$_{15}$	55.08(±1.85)	34.39	53.22	81.57
	RSQ$_{20}$	55.42(±0.72)	34.58	55.42	81.55
	RSQ$_{25}$	54.53(±1.40)	33.65	54.26	80.79
CW	WTQ$_{50}$	40.66	16.93	43.05	69.68
	RSQ$_1$	61.30(±1.73)	44.65	58.45	83.03
	RSQ$_5$	58.58(±0.63)	40.74	56.20	81.65
	RSQ$_{10}$	57.65(±1.33)	39.24	53.22	82.05
	RSQ$_{15}$	57.05(±1.99)	38.01	54.28	81.74
	RSQ$_{20}$	57.13(±0.95)	37.42	55.79	82.25
	RSQ$_{25}$	55.68(±1.42)	35.64	56.03	80.69

Table 1: $\mathcal{L}_{\text{LLM}}$(RSQ$_{\text{N}}$) can effectively attack images, but its performance still lags behind that of WTQ.

$\mathcal{L}_{\text{QAVA}}$(RSQ$_{\text{N}}$) can further improve attack performance than $\mathcal{L}_{\text{LLM}}$(RSQ$_{\text{N}}$). $\mathcal{L}_{\text{LLM}}$(RSQ$_{\text{N}}$) approach typically employs the end-to-end loss function $\mathcal{L}_{\text{LLM}}$ of the LLMs. In contrast, we propose targeting the visual-language alignment module using $\mathcal{L}_{\text{QAVA}}$, as defined in Sec. 3.3. We compared the effectiveness of $\mathcal{L}_{\text{LLM}}$ and $\mathcal{L}_{\text{QAVA}}$, with results presented in Tab. 2. For WTQ, there is no significant difference in performance between $\mathcal{L}_{\text{QAVA}}$ and $\mathcal{L}_{\text{LLM}}$. However, for RSQ, ***using $\mathcal{L}_{\text{QAVA}}$ results in a significantly better attack performance than $\mathcal{L}_{\text{LLM}}$***. This highlights the effectiveness of our approach $\mathcal{L}_{\text{QAVA}}$(RSQ$_{\text{N}}$), indicating its potential to enhance adversarial attack capabilities.

Attack method	Loss $\mathcal{L}$	Question strategy	InstructBLIP(Vicuna-7B) Overall	Other	Number	Yes/No
✘	✘	✘	78.00	66.98	69.68	94.29
PGD	$\mathcal{L}_{\text{LLM}}$	WTQ$_{50}$	42.41	17.56	50.59	71.13
		RSQ$_{25}$	54.53(±1.40)	33.65	54.26	80.79
	$\mathcal{L}_{\text{QAVA}}$	WTQ$_{50}$	44.41	21.12	43.96	73.75
		RSQ$_1$	45.70(±1.06)	21.70	46.29	75.60
		RSQ$_5$	43.59(±1.16)	20.76	42.00	72.69
		RSQ$_{10}$	44.85(±1.07)	22.11	43.25	73.83
		RSQ$_{25}$	44.07(±0.83)	20.15	46.60	73.32
CW	$\mathcal{L}_{\text{LLM}}$	WTQ$_{50}$	40.66	16.93	43.05	69.68
		RSQ$_{25}$	55.68(±1.42)	35.64	56.03	80.69
	$\mathcal{L}_{\text{QAVA}}$	WTQ$_{50}$	41.82	17.63	39.68	72.79
		RSQ$_1$	43.16(±1.03)	19.39	42.24	73.25
		RSQ$_5$	42.18(±1.15)	18.37	41.48	72.24
		RSQ$_{10}$	41.85(±0.83)	17.38	41.03	72.77
		RSQ$_{25}$	40.98(±1.71)	16.45	40.48	71.88

Table 2: $\mathcal{L}_{\text{QAVA}}$ is better than $\mathcal{L}_{\text{LLM}}$ in use of RSQ$_{\text{N}}$.

Generalizability of QAVA to other LVLMs. Previously, our experiments focused solely on the InstructBLIP Vicuna-7B. To assess the broader applicability of QAVA, we extended our evaluation to include additional LVLMs, with results summarized in Tab. 3. The findings demonstrate that QAVA consistently delivers effective attack performance across a wider spectrum of LVLMs, showcasing its robustness and adaptability in diverse LVLMs.

Model	Clean	$\mathcal{L}_{\text{LLM}}$(RSQ$_{25}$)	$\mathcal{L}_{\text{QAVA}}$(RSQ$_{10}$)
BLIP-2 opt-2.7B	45.64	34.56	19.15
BLIP-2 FlanT5$_{\text{XL}}$	62.05	29.39	32.28
BLIP-2 opt-6.7B	48.26	15.25	19.49
BLIP-2 FlanT5$_{\text{XXL}}$	62.10	29.11	26.93
InstructBLIP FlanT5$_{\text{XL}}$	74.54	49.78	34.31
InstructBLIP Vicuna-7B	78.00	54.53	44.85
InstructBLIP FlanT5$_{\text{XXL}}$	73.02	48.57	34.32
InstructBLIP Vicuna-13B	67.87	53.19	42.90

Table 3: Results of QAVA on various LVLMs.

Generalizability of QAVA to other datasets. To further assess the applicability of QAVA, we conducted evaluations using the VizWiz test-dev dataset (Gurari et al., 2018, 2019), which consists of 8,000 image-question pairs, each image paired with a single question. We explored two attack scenarios: $\mathcal{L}_{\text{LLM}}$(WTQ$_1$), which represents a traditional end-to-end adversarial attack targeting the specific question, and $\mathcal{L}_{\text{QAVA}}$(RSQ$_{10}$), which illustrates the QAVA attack using 10 randomly selected

questions from VQA v2 without prior knowledge of the target questions. As shown in Tab. 4, QAVA reliably performs effective adversarial attacks using these randomized questions against unknown target questions on VizWiz. These results underscore the versatility and robustness of QAVA across different datasets and question distributions.

Attack method	VizWiz test-dev evaluation				
	overall	yes/no	number	other	unanswerable
Clean	33.08	81.74	28.10	39.14	10.99
$\mathcal{L}_{\text{LLM}}(\text{WTQ}_1)$	10.48	63.71	8.25	8.44	6.93
$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	10.69	59.04	7.78	9.54	5.85

Table 4: Results of QAVA on VizWiz test-dev dataset.

4.3 Ablation Study of Question Sampling

No need for image-related surrogate questions. VQG for images requires LVLMs to process the image and generate numerous tokens, which can be resource-intensive and time-consuming. While this approach ensures that the questions used for the attack are closely related to the image, the results presented in Tab. 5 indicate that this does not lead to a significant improvement in attack performance. Consequently, in practical applications, it is unnecessary to employ VQG to create the set of surrogate questions, as the benefits in terms of attack efficacy do not justify the additional computational cost.

Attack method	Question strategy	VQA v2 scores			
		Overall	Other	Number	Yes/No
PGD	WTQ_{50}	44.41	21.12	43.96	73.75
	RSQ_{10}	44.85(± 1.07)	22.11	43.25	73.83
	$\text{RSQ}^{\text{t}}_{10}$	43.85(± 1.29)	20.32	44.54	73.13
	VQG_{10}	43.26(± 0.67)	20.83	39.44	72.53
CW	WTQ_{50}	41.82	17.63	39.68	72.79
	RSQ_{10}	41.85(± 0.83)	17.38	41.03	72.77
	$\text{RSQ}^{\text{t}}_{10}$	42.05(± 1.22)	17.79	40.22	73.01
	VQG_{10}	39.38(± 1.07)	16.00	38.10	69.08

Table 5: The ablation study of the question sampling strategy as outlined in Sec. 3.2. In all experiments, the loss function $\mathcal{L}_{\text{QAVA}}$ is utilized on InstructBLIP Vicuna-7B. RSQ_{10} involves the random sampling of 10 questions from the entire list of available questions. In contrast, $\text{RSQ}^{\text{t}}_{10}$ first randomly selects 10 question types from the 67 types identified in VQA v2 and subsequently samples one question from each selected type, ensuring that no more than one question per type is sampled. The approach VQG_{10} employs MiniGPT-4 to generate 10 questions specifically for the images.

Randomly sampling questions is simple and efficient. The findings in Tab. 5 show that RSQ^{t} does not significantly enhance attack performance. Consequently, a straightforward approach of randomly sampling questions RSQ_{N} is sufficient and effective. This method not only simplifies the process but also reduces computational overhead, all while maintaining robust attack performance.

4.4 Abalation Study of the Imperceptibility of Images Generated by QAVA

Although we employed a smaller attack strength (*e.g.*, $\epsilon_\infty = 8$ in PGD), adversarial examples may still be detectable upon careful examination, as shown in Fig. 3a. Previous studies have developed techniques to enhance image imperceptibility in classical attack methods, such as SSAH (Luo et al., 2022). By integrating QAVA with SSAH, we can produce adversarial images that are both imperceptible and query-agnostic, as demonstrated in Fig. 3b. The VQA scores for QAVA and QAVA+SSAH are demonstrated in Tab. 6. There exists a trade-off between the imperceptibility of adversarial examples and the effectiveness of adversarial attacks. Nevertheless, QAVA+SSAH successfully generates adversarial examples that are both imperceptible and exhibit a significant attack impact. Additional adversarial images are presented in Fig. 4. Furthermore, we investigate the effect of varying attack strengths on the efficacy of our approach, as detailed in Sec. 5.

(a) The QAVA image. (b) The QAVA+SSAH image.

Figure 3: Visualization of image imperceptibility.

Attack method	VQA v2 scores			
	Overall	Other	Number	Yes/No
QAVA	44.85	22.11	43.25	73.83
QAVA+SSAH	50.67	28.64	50.37	78.37

Table 6: The combination of QAVA and SSAH generates imperceptible adversarial examples.

5 Ablation Study of QAVA Attack Strength

Our experiments concentrate on two main attack methods: PGD and C&W, as described in Sec. 4.1. To evaluate the balance between attack efficacy and image imperceptibility, we investigated different levels of attack strength. As demonstrated in Tab. 7, our QAVA approach consistently generates effective adversarial examples across various attack strengths. Even at lower attack intensities, there is a notable reduction in VQA scores, while the noise introduced remains nearly imperceptible.

Attack method	ϵ_∞ of PGD c of C&W	VQA v2 scores Overall	Other	Number	Yes/No
PGD	4/255	48.93	27.57	46.42	76.46
	8/255	44.85	22.11	43.25	73.83
	16/255	38.86	16.02	38.29	67.67
C&W	0.05	51.72	32.11	48.13	77.37
	0.005	41.85	17.38	41.03	72.77
	0.0005	36.44	11.41	38.45	67.22

Table 7: The alternative attack strength of $\mathcal{L}_{\text{QAVA}}$ with RSQ_{10}. The other settings are the same as Sec. 4.1.

5.1 The Efficiency of QAVA Attacks

Table 2 illustrates that our QAVA $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$ attains performance levels comparable to traditional end-to-end adversarial attacks method $\mathcal{L}_{\text{LLM}}(\text{WTQ}_{50})$ that directly target 50 specific questions. To further underscore the efficiency of our QAVA approach, we evaluate three scenarios: traditional end-to-end attacks targeting 1 question and 50 questions, and QAVA targeting 10 randomly selected questions. In comparison to the traditional end-to-end approach, our QAVA method offers significant improvements in both time and memory efficiency. As indicated in Table 8, QAVA achieves approximately 80% savings in GPU memory usage because it targets only the output of the vulnerable visual language alignment module, bypassing the resource-intensive LLM. Regarding time efficiency, while the traditional approach is rapid when attacking a single question, it proves ineffective for a series of target questions. Conversely, attacking all 50 questions using the traditional method is effective but requires prior knowledge of all target questions and is time-intensive. In contrast, our QAVA approach, which employs 10 random questions without prior knowledge of specific targets, achieves attack results comparable to those of the traditional method across all target questions, while requiring only 5% of the time consumed by the traditional end-to-end adversarial attack.

Attack method	Time (s)	GPU Mem (GB)	VQA scores
Clean	✘	✘	78.00
$\mathcal{L}_{\text{LLM}}(\text{WTQ}_1)$	208	32.62	57.07
$\mathcal{L}_{\text{LLM}}(\text{WTQ}_{50})$	7788	32.62	42.41
$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	387	6.30	44.85

Table 8: The efficiency of QAVA over traditional end-to-end adversarial attacks. Red cells denote worse performance, while green cells indicate better performance.

5.2 Experiments on QAVA's Transferability

Surrogate model	LLaVA VQA scores $m = 0$	$m = 0.9$
Clean image	78.32	
InstructBLIP $\text{FlanT5}_{\text{XL}}$	65.83	64.99
InstructBLIP Vicuna-7B	68.76	64.69
InstructBLIP $\text{FlanT5}_{\text{XXL}}$	66.97	66.82
InstructBLIP Vicuna-13B	67.17	63.86

Table 9: Results of transfer attacking LLaVA using DI+MI with $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$, where m denotes the momentum. We use the LLaVA-v1.5-7b model with CLIP-ViT-L-336px, which differs significantly from InstructBLIP. Despite this, QAVA still has good transferability.

Transferability of QAVA between InstructBLIP and LLaVA on VQA tasks. We investigated the transferability of the QAVA attack on the VQA tasks between InstructBLIP and LLaVA, with the results presented in Tab. 9. To enhance the transferability of the adversarial examples, we incorporated the momentum attack (Dong et al., 2018) and diverse input methods (Xie et al., 2019).

Transferability of QAVA between BLIP-2 and InstructBLIP on VQA tasks. We examined the transferability of QAVA across the LVLMs utilized in Tab. 3. Adversarial examples were generated on each LVLM independently, and their VQA scores were evaluated on the other models, with the results presented in Tab. 10. The experiments were conducted using the VQA v2 32+50 dataset with $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$. Each row represents the surrogate model used to generate the adversarial perturbation, while each column represents the target model used to test the VQA scores. The diagonal cells indicate white-box model settings, whereas the non-diagonal cells represent black-box model settings, demonstrating transferability. The results indicate that our QAVA exhibits strong transferability.

Surrogate model	BLIP-2				InstructBLIP			
	opt-2.7B	FlanT5$_{XL}$	opt-6.7B	FlanT5$_{XXL}$	FlanT5$_{XL}$	Vicuna-7B	FlanT5$_{XXL}$	Vicuna-13B
Clean images	45.64	62.05	48.26	62.10	74.54	78.00	73.02	67.87
BLIP-2 opt-2.7B	19.15	27.63	19.48	28.09	42.37	44.93	41.28	45.84
BLIP-2 FlanT5$_{XL}$	27.26	32.28	24.68	35.88	45.64	48.1	46.29	48.58
BLIP-2 opt-6.7B	22.47	28.2	19.49	34.22	45.32	45.53	41.69	44.25
BLIP-2 FlanT5$_{XXL}$	20.22	29.07	25.07	26.93	44.81	47.35	44.39	46.91
InstructBLIP FlanT5$_{XL}$	22.99	28.18	23.52	24.29	34.31	45.34	43.35	47.98
InstructBLIP Vicuna-7B	20.15	22.85	19.35	27.1	38.53	44.85	38.83	43.61
InstructBLIP FlanT5$_{XXL}$	18.07	27.43	16.81	29.47	42.67	46.46	34.32	46.84
InstructBLIP Vicuna-13B	19.85	22.2	19.76	30.34	40.29	43.41	39.84	42.90

Table 10: Results of transferring attacks against BLIP-2 and InstructBLIP on VQA.

LLM of InstructBLIP	Attack method	Loss $\mathcal{L}$	Image Caption evaluation on InstructBLIP							
			CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
Vicuna-7B	✘	✘	160.5	82.5	68.4	54.6	42.9	61.4	31.4	25.4
	PGD	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	71.4	61.8	42.9	29.4	20.4	45.8	19.6	13.1
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	16.9	38.8	19.8	10.4	5.9	28.2	10.0	4.1
	CW	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	92.7	67.3	50.0	36.3	25.9	49.6	22.5	16.1
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	13.0	29.2	14.5	7.1	3.8	25.6	8.6	3.4

Table 11: Results of inter-task transferability (VQA $\rightarrow$ caption) of adversarial examples on InstructBLIP. The grey lines are the results of clean images. QAVA(RSQ$_{10}$) effectively improves the tasks transferability.

Target model	Surrogate model	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
MiniGPT-4	Clean image	89.7	65.9	49.9	35.8	25.0	52.9	29.4	24.6
	InstructBLIP FlanT5$_{XL}$	16.9	39.9	24.0	14.0	8.4	33.3	15.9	8.6
	InstructBLIP Vicuna-7B	12.9	37.8	21.7	12.3	7.4	31.8	14.8	7.4
	InstructBLIP FlanT5$_{XXL}$	18.0	39.6	24.0	14.5	9.0	33.7	15.9	8.6
	InstructBLIP Vicuna-13B	13.7	36.2	21.0	12.0	7.2	31.2	14.1	6.9
LLaVA	Clean image	116.9	73.1	56.8	42.0	30.3	56.6	29.9	24.4
	InstructBLIP FlanT5$_{XL}$	82.9	65.5	47.8	33.5	22.7	50.2	25.1	18.4
	InstructBLIP Vicuna-7B	82.3	64.6	47.4	33.5	23.2	49.7	25.0	18.1
	InstructBLIP FlanT5$_{XXL}$	86.5	65.6	48.1	33.9	23.5	50.5	25.5	18.8
	InstructBLIP Vicuna-13B	79.9	64.2	46.5	32.5	22.3	49.2	24.7	17.8

Table 12: Results of transferability of adversarial examples on both tasks (VQA $\rightarrow$ caption) and models (InstructBLIP $\rightarrow$ MiniGPT-4/LLaVA). The attack is all $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$.

Transferability of QAVA from VQA to caption task on InstructBLIP. We investigated the performance of images subjected to adversarial attacks on tasks beyond VQA. Specifically, we input adversarial images generated from the VQA v2 500+10 dataset into InstructBLIP to produce image captions using the prompt "A short image caption:". The results of these adversarial images on the image caption task are displayed in Tab. 11. Despite targeting unrelated random questions, the adversarial approach effectively reduces image caption performance. The efficacy of QAVA-generated adversarial images on the caption task underscores the effectiveness of RSQ and $\mathcal{L}_{\text{QAVA}}$. The transferability results for the captioning task on more LVLMs are shown in Tab. 16 in the Appendix.

Transferability of QAVA from InstructBLIP and VQA to LLaVA/MiniGPT-4 and caption task. We also explored the transferability of QAVA across tasks (from VQA to captioning) and models (from InstructBLIP to LLaVA/MiniGPT-4). The results presented in Tab. 12 demonstrate QAVA's strong transferability between tasks and LVLMs, even when the attacks are based on irrelevant random questions. Given that image captioning is a classical pre-training task for LVLMs, employing the adversarial examples generated by QAVA to disrupt the training process of LVLMs could potentially have a significant impact.

5.3 Generalizability of QAVA

Our primary experiments were conducted using the VQA v2 dataset. To further assess the gen-

eralizability of QAVA, we performed additional experiments on several other datasets, including ImageNet (Deng et al., 2009), OKVQA (Marino et al., 2019), NoCaps (Agrawal et al., 2019), and Flickr30k (Young et al., 2014). The results presented in Tab. 13 demonstrate the effectiveness of QAVA in executing successful attacks across a diverse array of datasets.

Dataset	Metric	Clean	QAVA
ImageNet	Accuracy	81.0	34.8
OKVQA	VQA score	56.9	23.85
NoCaps	CIDEr	120.2	15.2
Flickr30k	CIDEr	85.2	9.1

Table 13: The generalizability of QAVA on other datasets including classification, Q&A and captioning.

5.4 Discussions of QAVA

Potential Further Optimizations of QAVA. (1) Recent research on universal adversarial attacks on images has introduced Stochastic Gradient Aggregation (SGA) (Liu et al., 2023b), a technique that improves stability by calculating multiple gradients over a small batch of images and merging them into a single gradient. Inspired by SGA, we can further improve the attack performance of QAVA by sampling different small batches of stochastic questions at each step of the attack process, as illustrated in Tab. 14. (2) Further optimization could enhance the performance of QAVA. Specifically, the vulnerable vision-language alignment module consists of multiple layers, and our current loss function $\mathcal{L}_{QAVA}$ targets only the output of the last layer of the Q-former. We extended the loss function to incorporate outputs from all layers of the Q-former. The results presented in Tab. 15 demonstrate that this minor optimization of the loss function leads to an improvement in the performance of QAVA.

Attack method	VQA v2 scores			
	Overall	Other	Number	Yes/No
$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	44.85	22.11	43.25	73.83
$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$ + SGA	41.14	19.08	37.97	69.74

Table 14: The optimized version of QAVA, drawing inspiration from SGA, leads to more robust query-agnostic adversarial examples.

Potential Defense of QAVA. We discuss possible defenses against QAVA in Appendix B.

Method	VQA v2 scores
$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{25})$	44.07 ± 0.83
Multi-layer $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{25})$	42.54 ± 0.92

Table 15: The optimized version of QAVA, employing multi-layer loss function $\mathcal{L}_{\text{QAVA}}$.

6 Conclusion

In this paper, we introduce a robust adversarial attack method, QAVA, designed to generate adversarial examples that significantly mislead responses to unknown target questions for a given image. QAVA initially identifies the vulnerability within the LVLMs, specifically the vision-language alignment module. Subsequently, it executes adversarial attacks utilizing a broader range of randomly sampled, image-irrelevant questions. Extensive experiments demonstrate the effectiveness of QAVA in both white-box and black-box attack scenarios. Additionally, we verify the high transferability of QAVA across various LVLMs and different tasks. Our findings with QAVA serve as a critical alert regarding the security vulnerabilities of LVLMs.

7 Limitation

We summarize the limitations of our work as follows. We will try to do these in the future.

(1) Although extensive experiments have demonstrated the effectiveness of attacks utilizing irrelevant questions, we have not yet provided a plausible explanation for the impact that such irrelevant randomized questions can have on the attacks.

(2) Insufficient assessment of potential negative impacts of QAVA. As we analyzed, the adversarial examples obtained using QAVA are more aggressive and may have a larger negative impact if they are used for poisoning in the pre-training or supervised fine-tuning process of LVLMs. However, this aspect was not evaluated experimentally.

(3) We only evaluated the transferring attack of QAVA for image captioning tasks, not for broader visual-language tasks.

8 Acknowledgments

This work was supported by the National Natural Science Foundation of China (62376024, 62325405), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001) and Beijing National Research Center for Information Science and Technology (BNRist, BNR2024TD03001).

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

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.

Moustafa Alzantot, Yash Sharma Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

WL Chiang, Z Li, Z Lin, Y Sheng, Z Wu, H Zhang, L Zheng, S Zhuang, Y Zhuang, JE Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, mar. 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. 2023. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *stat*, 1050:20.

Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, and Weihong Deng. 2023b. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4435–4444.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023c. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. 2023. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111.

Cheng Luo, Qinliang Lin, Weicheng Xie, Bizhu Wu, Jinheng Xie, and Linlin Shen. 2022. Frequency-driven

imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15315–15324.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. 2024. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*.

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*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213*.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2023a. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023b. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*.

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)*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2023. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *arXiv preprint arXiv:2310.04655*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A The Algorithm of QAVA

Algorithm 1: The steps of QAVA(RSQ_N) using PGD to attack Q-former

Input: image x_i
Model: visual encoder f_i, Q-former Q
Data: the question set $\mathcal{T}$
Hyperparameter: questions number N, attack step α, perturbation limitation ϵ_∞, number of attack iterations n
Output: the adversarial image x_i'

for $k \leftarrow 1$ **to** n **do**
  $\mathcal{L}_{\text{QAVA}} = 0$;
  for $j \leftarrow 1$ **to** N **do**
    /* Randomly select a question from $\mathcal{T}$ */
    $x_t \sim \mathcal{T}$;
    $q = \mathcal{Q}(f_i(x_i), x_t)$;
    $q' = \mathcal{Q}(f_i(x_i'), x_t)$;
    $\mathcal{L}_{\text{QAVA}} = \mathcal{L}_{\text{QAVA}} + \frac{1}{N}\mathcal{L}_{\text{QAVA}}(q, q')$;
  end
  $x_i' = \text{Clip}_{\epsilon_\infty}(x_i' + \alpha \times \text{sign}(\nabla_{x_i'}\mathcal{L}_{\text{QAVA}}))$;
end
return x_i';

B The Discussion of Potential Defense Methods for QAVA

QAVA targets the output of the vulnerable visual-language alignment module within LVLMs. To mitigate such attacks, potential defense strategies include: (1) Implementing adversarial training of the visual-language alignment module. This approach is cost-effective, as it does not involve the computationally intensive LLM. (2) Developing mechanisms to suppress the module's output when faced with image-unrelated questions. For instance, earlier LVLMs (Qi et al., 2023) were susceptible to jailbreak attacks via adversarial images irrelevant to the input instructions. Conversely, modern LVLMs, such as GPT-4 and Gemini, would ignore input images unrelated to the instructions. However, these models might still be vulnerable to jailbreaks when adversarial images pertain to the instructions.

C Visualization results on the QAVA attack

In Sec. 4.4, we examine the imperceptibility of adversarial examples produced by the QAVA attack. The visual representations of these adversarial examples are provided in Fig. 4.

D Results of the extension to more LVLMs of Tab. 11

Table 11 presents the inter-task transferability of QAVA on InstructBLIP Vicuna-7B. Additionally, Tab. 16 provides further results for other versions of InstructBLIP, demonstrating QAVA's applicability across different LVLMs.

E More results on $\text{FlanT5}_{\text{XL}}$

For Tables 2 and 5 in the main paper, we also conducted the same experiments on $\text{FlanT5}_{\text{XL}}$. The results are shown in Tabs. 17 and 18.

LLM of InstructBLIP	Attack method	Loss $\mathcal{L}$	Image Caption evaluation on InstructBLIP							
			CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	SPICE
$\text{FlanT5}_{\text{XL}}$	✘	✘	154.6	81.9	67.1	52.0	39.0	59.7	30.2	24.5
	PGD	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	52.4	57.1	37.7	24.2	15.6	42.2	17.0	10.0
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	19.7	37.8	20.6	11.8	7.1	28.4	10.5	4.3
	CW	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	87.5	66.5	48.7	34.8	24.3	48.0	21.6	15.1
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	6.8	25.3	11.6	4.8	2.3	24.3	7.7	2.4
$\text{FlanT5}_{\text{XXL}}$	✘	✘	154.0	82.3	67.4	52.9	40.6	60.9	30.3	24.3
	PGD	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	43.1	54.6	34.8	22.0	14.4	40.6	16.1	8.8
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	23.2	41.4	23.2	12.3	6.7	30.3	11.3	5.6
	CW	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	88.8	58.6	42.8	30.1	21.3	47.7	20.7	15.2
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	11.6	23.8	11.6	5.7	3.0	25.2	7.9	3.1
Vicuna-13B	✘	✘	129.2	62.6	50.1	38.6	28.8	57.0	28.8	23.6
	PGD	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	60.6	49.0	34.1	22.9	15.2	42.7	19.1	12.7
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	12.7	25.4	12.6	6.0	3.2	25.9	8.9	3.6
	CW	$\mathcal{L}_{\text{LLM}}(\text{RSQ}_{10})$	72.5	52.0	36.7	25.4	17.4	44.6	20.9	14.8
		$\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$	7.2	25.0	11.7	5.4	2.9	22.4	7.8	3.0

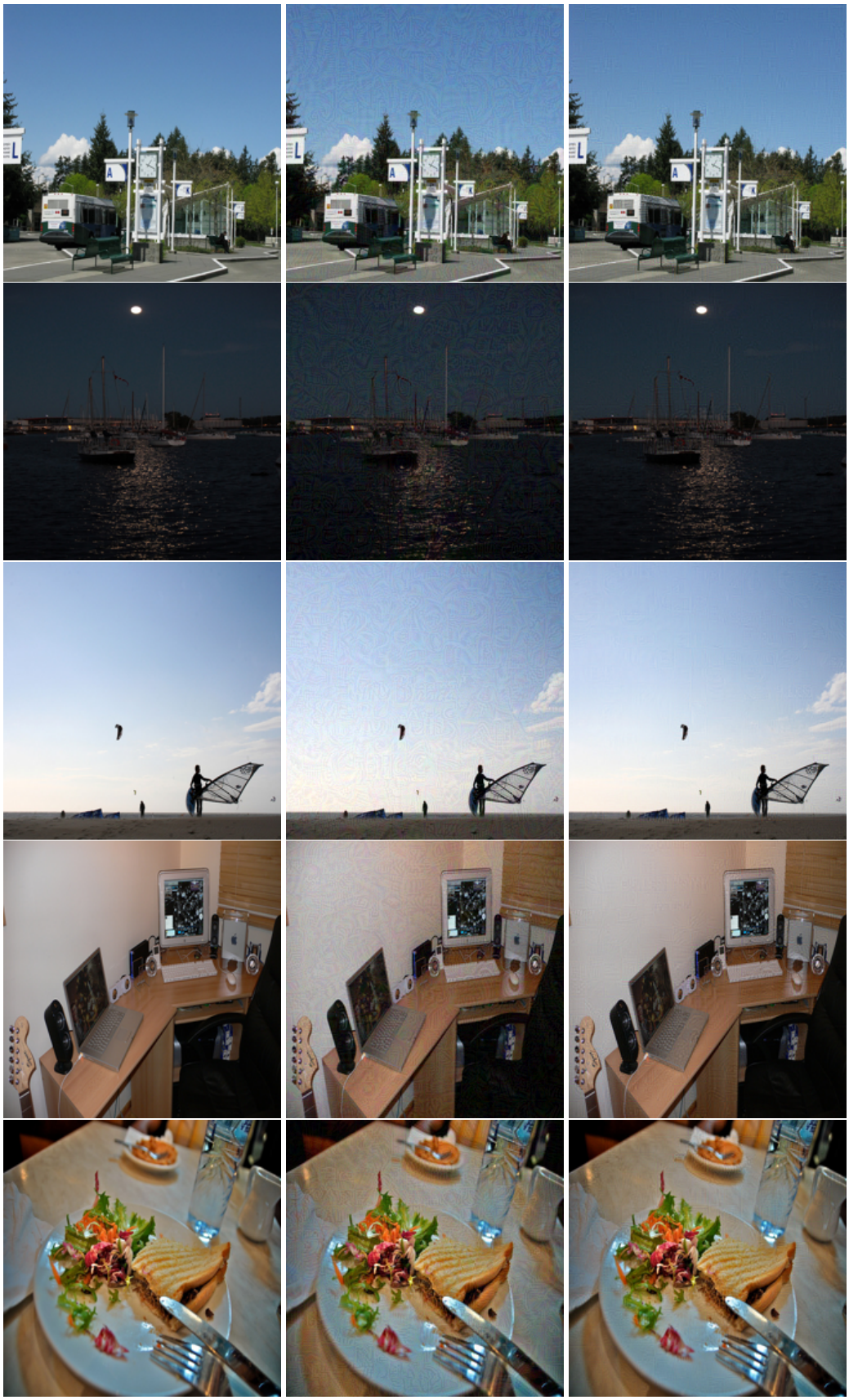
Table 16: Results of inter-task transferability (VQA $\rightarrow$ caption) of adversarial examples on InstructBLIP. The grey lines are the results of clean images. $\text{QAVA}(\text{RSQ}_{10})$ effectively improves the tasks transferability.

Attack method	Loss $\mathcal{L}$	Question strategy	InstructBLIP($\text{FlanT5}_{\text{XL}}$)			
			Overall	Other	Number	Yes/No
✘	✘	✘	74.54	63.07	66.52	91.31
PGD	$\mathcal{L}_{\text{LLM}}$	WTQ_{50}	40.91	18.77	31.02	71.63
		RSQ_{25}	49.78(±0.92)	26.98	46.60	79.33
	$\mathcal{L}_{\text{QAVA}}$	WTQ_{50}	36.11	17.71	34.33	59.70
		RSQ_{1}	41.54(±1.55)	22.94	41.50	64.87
		RSQ_{5}	36.30(±1.89)	19.76	32.81	58.07
		RSQ_{10}	34.31(±2.06)	18.10	31.61	55.42
		RSQ_{15}	35.22(±1.28)	17.76	33.13	57.72
		RSQ_{20}	35.50(±1.21)	18.14	34.50	57.56
		RSQ_{25}	35.89(±1.70)	18.47	35.19	57.92
CW	$\mathcal{L}_{\text{LLM}}$	WTQ_{50}	41.95	20.08	29.09	73.21
		RSQ_{25}	52.77(±0.90)	32.04	47.13	80.45
	$\mathcal{L}_{\text{QAVA}}$	WTQ_{50}	9.61	4.83	5.29	16.87
		RSQ_{1}	20.47(±4.21)	10.42	21.25	32.83
		RSQ_{5}	11.99(±1.99)	5.43	11.31	20.42
		RSQ_{10}	10.62(±1.55)	5.08	7.79	18.41
		RSQ_{15}	12.01(±1.42)	5.59	9.99	20.68
		RSQ_{20}	11.46(±1.92)	5.40	9.89	19.52
		RSQ_{25}	10.67(±2.29)	4.64	8.92	18.75

Table 17: $\mathcal{L}_{\text{QAVA}}$ is better than $\mathcal{L}_{\text{LLM}}$ in use of RSQ_{N}.

Attack method	Question strategy	InstructBLIP($\text{FlanT5}_{\text{XL}}$)			
		Overall	Other	Number	Yes/No
PGD	WTQ_{50}	36.11	17.71	34.33	59.70
	RSQ_{10}	34.31(±2.06)	18.10	31.61	55.42
	$\text{RSQ}^{\text{t}}_{10}$	34.54(±0.70)	18.23	29.67	56.43
	$\text{RSQ}^{\text{c}}_{10}$	34.28(±2.40)	18.36	31.93	54.94
	VQG_{10}	35.83(±2.48)	18.81	35.18	57.36
CW	WTQ_{50}	9.61	4.83	5.29	16.87
	RSQ_{10}	10.62(±1.55)	5.08	7.79	18.41
	$\text{RSQ}^{\text{t}}_{10}$	9.87(±1.01)	4.94	8.66	16.41
	$\text{RSQ}^{\text{c}}_{10}$	9.76(±2.89)	4.83	7.24	16.69
	VQG_{10}	9.04(±1.84)	4.61	7.58	15.04

Table 18: The ablation study of the question sampling strategy as outlined in Sec. 3.2.



(a) Clean image with VQA score 78.00.

(b) QAVA adversarial image with VQA score: 44.85.

(c) QAVA+SSAH adversarial image with VQA score: 50.67.

Figure 4: The clean images, the QAVA adversarial images, and the QAVA+SSAH adversarial images. All experiments are conducted using InstructBLIP Vicuna-7B with the attack $\mathcal{L}_{\text{QAVA}}(\text{RSQ}_{10})$.