Adversarial Attacks on VQA-NLE: Exposing and Alleviating Inconsistencies in Visual Question Answering Explanations

Yahsin Yeh, Yilun Wu, Bokai Ruan, Honghan Shuai
National Yang Ming Chiao Tung University

Abstract

Natural language explanations in visual question answering (VQA-NLE) aim to make black-box models more transparent by elucidating their decision-making processes. However, we find that existing VQA-NLE systems can produce inconsistent explanations and reach conclusions without genuinely understanding the underlying context, exposing weaknesses in either their inference pipeline or explanation-generation mechanism. To highlight these vulnerabilities, we not only leverage an existing adversarial strategy to perturb questions but also propose a novel strategy that minimally alters images to induce contradictory or spurious outputs. We further introduce a mitigation method that leverages external knowledge to alleviate these inconsistencies, thereby bolstering model robustness. Extensive evaluations on two standard benchmarks and two widely used VQA-NLE models underscore the effectiveness of our attacks and the potential of knowledge-based defenses, ultimately revealing pressing security and reliability concerns in current VQA-NLE systems.

1 Introduction

Visual Question Answering with Natural Language Explanations (VQA-NLE) has recently become an active area of research (Sammani et al., 2022; Suo et al., 2023; Lai et al., 2024), as it promises both accurate answers and human-readable justifications. By augmenting conventional VQA pipelines with textual rationales, it can offer deeper transparency and facilitate trust in black-box models (Kayser et al., 2021). Moreover, generating explanations has been shown to reinforce question-answering performance itself, surpassing models trained solely on image-question pairs (Kayser et al., 2021). Despite this potential, critical questions remain about the quality and consistency of the explanations produced, prompting further investigation into how these models truly reason about visual and linguistic inputs.

Specifically, while VQA-NLE models can produce explanations for their decisions, we observe that they can yield contradictory or inconsistent outputs even when the input scenario remains essentially the same. For instance, consider an image that depicts a woman wearing goggles and skiing downhill. If a person asks, "Why is the woman wearing goggles?" the model answers, "to protect eyes because the woman is wearing goggles to protect eyes," along with an explanation mentioning the goggles. However, rephrasing the prompt slightly to "Why is the woman using goggles?" causes the system to respond, "to photograph because the woman is using a camera," thereby contradicting its previous statement. Such inconsistent responses raise doubts about whether these models truly ground their reasoning in the image-question pair or instead rely on superficial cues, thereby questioning the extent to which VQA-NLE models genuinely "understand" their inputs when generating explanations.

To explore whether these contradictory outputs reflect genuine weaknesses in VQA-NLE systems, we leverage an existing attack and also propose a new adversarial sample generation framework designed to uncover vulnerabilities across both textual and visual modalities. In particular, we systematically rephrase questions (while preserving semantic equivalence) or selectively remove objects from images—even those seemingly irrelevant to the query. These controlled yet minimal edits often cause the model to produce inconsistent explanations, revealing the model's reliance on shallow patterns rather than robust visual-textual reasoning.[1]

In addition to exposing such vulnerabilities, we present an alleviation strategy based on integrating

[1]Although we primarily evaluate text-based and image-based manipulations separately, in principle they can also be combined to further stress-test a model's consistency.

external knowledge into the question. Concretely, for each query, we use a language model to generate short, relevant knowledge statements (e.g., clarifying synonyms or describing contextual details). Appending these statements to the input helps the VQA-NLE model anchor its reasoning in genuine semantic understanding rather than superficial cues. Experimental results demonstrate that this knowledge-driven approach can reduce contradictory explanations, offering a practical pathway toward more reliable and transparent VQA-NLE systems.

In summary, our contributions are as follows:

- To the best of our knowledge, this is the first adversarial framework specifically aimed at revealing potential vulnerabilities in VQA-NLE models, offering a systematic way to probe their security and consistency.
- Our attacks indeed degrade the semantic consistency of the VQA-NLE models on standard benchmarks (VQA-X and A-OKVQA), highlighting the models' reliance on brittle cues.
- We propose a knowledge-based mitigation strategy that reduces inconsistencies introduced by adversarial textual, improving robustness in VQA-NLE.

2 Related Work

2.1 VQA Explanations

Natural language explanations (NLEs) for question-answering (QA) have garnered increasing attention, driven by findings that explicit rationales can bolster a model's reasoning capabilities and interpretability (Camburu et al., 2018; Park et al., 2018; Narang et al., 2020; Wei et al., 2022a). In the context of visual question answering (VQA), these so-called VQA-NLE methods provide human-readable justifications alongside predicted answers (Park et al., 2018; Kayser et al., 2021). Approaches broadly split into post hoc (predict first, then explain) (Park et al., 2018; Wu and Mooney, 2019; Kayser et al., 2021) and self-rationalizing methods, where answer prediction and explanation generation occur jointly (Sammani et al., 2022; Suo et al., 2023). Recent work has introduced contrastive objectives to further align explanations with visual evidence (Lai et al., 2024). Despite improvements in accuracy, concerns remain regarding how faithfully these explanations reflect genuine model reasoning, as opposed to exploiting spurious patterns (Ray et al., 2019; Agarwal et al., 2020). Few studies have probed how small alterations in text or images can undermine explanation consistency, leaving open questions about the robustness of VQA-NLE outputs.

2.2 VQA Robustness

Meanwhile, VQA robustness research has largely focused on ensuring answer consistency under various input perturbations. For instance, (Ray et al., 2019; Shah et al., 2019) investigate how changes in question phrasing affect predictions, whereas (Agarwal et al., 2020) examine the impact of altering semantic elements in images. Augmentation techniques have also been proposed to mitigate inconsistent or brittle answers (Agarwal et al., 2020; Chen et al., 2020). However, most such work overlooks natural language explanations, and the few efforts targeting NLE consistency (Camburu et al., 2020; Jang et al., 2023) address primarily text-based variations. In contrast, we adopt a broader perspective on VQA-NLE robustness by implementing and introducing adversarial attack frameworks that target both linguistic and visual inputs. We then propose a knowledge-based defense to bolster model reliability against these minimal yet strategically chosen perturbations. By assessing both answer correctness and explanation consistency, our work expands robustness research into interpretability-focused VQA systems.

3 Method

While robustness in linguistic variations and image semantics has been respectively studied in the fields of language modeling and VQA, it remains an underexplored area for VQA-NLE models. To this end, we structure our attack method into two approaches, each targeting a different aspect: **text-based** attack and **image-based** attack.

3.1 Text-based Attack

For the text-based approach, existing adversarial attack methods for discrete data, such as BERT-Attack (Li et al., 2020) and R&R (Xu et al., 2022), are well-developed for generating adversarial text samples. Here, we directly employ BERT-Attack to generate text perturbations with synonym-based word substitution, aiming to fool models while maintaining grammatical and semantic coherence. By doing so, we expect to expose the model's reliance on superficial linguistic cues rather than gen-

uine contextual understanding. Specifically, the attack highlights cases where the model associates certain words or phrases with specific answers and explanations, revealing a dependency on dataset biases rather than robust reasoning.

To implement this attack, we leverage a masked language model (MLM) as part of the candidate ranking mechanism. The core objective function aims to rank substitution sequences based on contextual fluency. Specifically, given a set of token-level substitution candidates generated via Byte Pair Encoding (BPE), the method constructs full-length sequences by exhaustively combining candidates across token positions. Each sequence is then evaluated by computing the token-wise cross-entropy loss between the MLM's predicted distribution and the actual substitute token IDs. These losses are aggregated and exponentiated to estimate each sequence's perplexity, which serves as a proxy for fluency. By ranking sequences in ascending order of perplexity, the model promotes those that are both semantically coherent and grammatically well-formed.

Building upon this ranking mechanism, we then apply a filtering step to ensure semantic consistency between the adversarial and original inputs. Concretely, given the original question Q, we first use BERT-Attack to generate an adversarial candidates set containing n samples, denoted by $\{Q'_i\}_{i=1}^{n}$. Afterwards, we compute the semantic similarity between Q'_i and the original question Q using a universal sentence encoder U_s (Cer et al., 2018). If the cosine similarity between Q'_i and Q, denoted by γ_i, is lower than the predefined threshold σ_s, i.e., $\gamma_i = \cos(U_s(Q'_i), U_s(Q)) < \sigma_s$, then the candidate Q'_i is rejected. Otherwise, Q'_i is considered a valid adversarial example for the given image I. Following the previous work (Garg and Ramakrishnan, 2020), we set $\sigma_s = 0.8$.

3.2 Image-based Attack

To evaluate the robustness of VQA-NLE models against semantic changes in images, we propose a pipeline for editing images corresponding to each question. We hypothesize that any change influencing the model's predictions can reveal its weak contextual understanding. For instance, in Figure 4 (left), for the question about the type of the event, removing the table should not influence the model's prediction. To guarantee that the question's context remains unchanged after the image modification, we must ensure that objects relevant to the question's context are preserved.

Specifically, to generate an adversarial image without altering its overall meaning, we first identify any objects referenced in the question and answer, then limit our edits to regions unrelated to those objects. As the commonly-used VQA-NLE datasets, e.g., VQA-X (Park et al., 2018) and A-OKVQA (Schwenk et al., 2022) datasets, contain images sourced from MS-COCO (Lin et al., 2014), we consider the 80 predefined object classes and ground truth bounding boxes. To remove objects from the image, we utilize a diffusion-based inpainting approach (Zhuang et al., 2024), ensuring that the edited image remains semantically coherent. Our approach for maintaining contextual consistency in image modifications comprises two steps: (1) vocabulary mapping and (2) object removal.

Vocabulary Mapping To determine whether an object can be removed, we first map all object references in the question, answer, and explanation to the 80 COCO categories. These categories are often referred to using various synonyms or subset terms in the QA and explanation space. For example, van, taxi, trunk, truck, and SUV all correspond to the category "car," while table and desk refer to the category "dining table." To prevent erroneous removals, we compile a comprehensive mapping of nouns, pronouns, and synonyms used in the QA and explanation vocabulary to the 80 COCO categories. Due to the space constraint, the complete list of the mapping table is available in the supplementary materials.

Object Removal Let S_{I} represent the set of objects in the images (identified via COCO bounding boxes), S_{QA} represent the set of objects in the QA pair, and S_{E} represent the set of objects in the explanation. We define the set of candidate objects for removal as

$$S_{\text{candidate}} := S_{\text{E}} \cap \{S_{\text{I}} \setminus S_{\text{QA}}\}. \quad (1)$$

We then select the most frequent object in $S_{\text{candidate}}$ as our target object, S_{target}. Our underlying assumption is that a robust model should continue to generate explanations that accurately reflect the modified image content and do not mislead.

3.3 Alleviation

Because synonyms or paraphrases often cause inconsistencies between a model's outputs and those

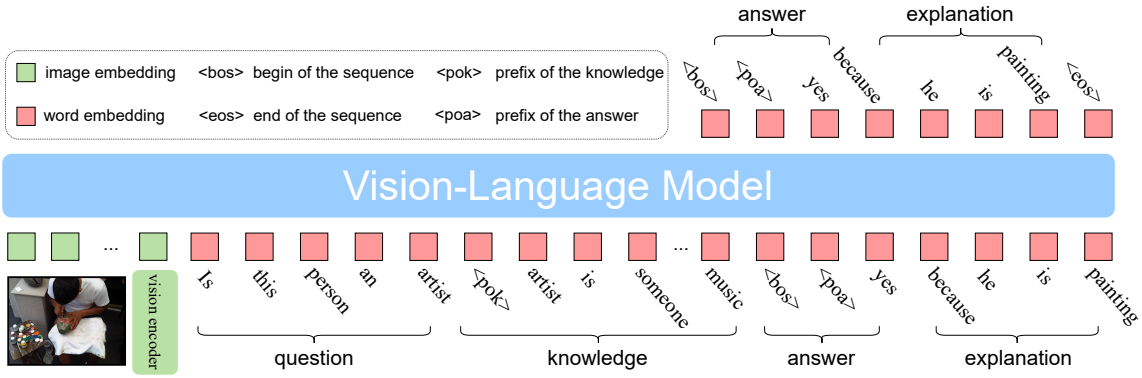

Figure 1: Architecture of our alleviation model. Given an image, its corresponding question, and external knowledge, we then append the "<bos>" token along with the predefined prefix for answer. The model then generates the answer and explanation in an auto-regressive manner. The external knowledge is retrieved from GPT-4o to provide additional context for the question. For the content prefix, "<pok>" represents the knowledge prefix ("based on the fact that") and "<poa>" denotes the answer prefix ("the answer is").

Prompt
Generate some knowledge about the concepts in the input. Generate up to 20 words. Examples: **Input**: Does the dog like Scooby Snacks? **Knowledge**: Scooby Snacks are treat rewards often enjoyed by dogs, famously associated with the character Scooby-Doo's motivation. **Input**: Does this appear to be a photo of multiple exposures of the black clad snowboarder? **Knowledge**: Multiple exposures overlay several captures to create a composite image showing motion or different poses. **Input**:{**question**} **Knowledge**:

Figure 2: Prompt template for knowledge generation.

of a benign model, a logical remedy is to ensure the model better captures synonym equivalences. To address this, we propose extending the model's input with external knowledge relevant to the question. The key insight is that incorporating question-specific knowledge can help the model interpret synonymous words more faithfully, leading to a more robust understanding of the query. Our approach to mitigating inconsistencies in VQA-NLE involves two main steps: (1) generating question-related knowledge, and (2) injecting that knowledge into the model's input.

Knowledge Generation Drawing on (Liu et al., 2022), we employ a heuristic method to generate question-specific knowledge. Concretely, our system prompts OpenAI's GPT-4o API with a structured template that contains both clear instructions and illustrative examples, along with a placeholder to accommodate the new question. We then request GPT-4o to produce concise but relevant knowledge statements focusing on the core concepts of the query. Figure 2 outlines the complete prompt configuration, illustrating how these fixed demonstrations and guidelines steer the model toward generating succinct yet targeted knowledge.

Knowledge Injection Once the knowledge is obtained, we concatenate it with the question before feeding both into the model. Let the **image features** be

$$Z_I = \left(Z_1^I, Z_2^I, \ldots, Z_i^I, \ldots, Z_x^I\right), \tag{2}$$

where each $Z_i^I \in \mathbb{R}^d$ represents the i-th image patch, and x is the total number of image patches. For a **question** $Q = \left(q_1, q_2, \ldots, q_t\right)$ (length t) and a **knowledge statement** $K = \left(k_1, k_2, \ldots, k_n\right)$ (length n), we embed each word into a d-dimensional space via a pretrained image-caption model (DistilGPT2 (Sammani et al., 2022)). This yields the **question features**

$$Z_Q = \left(Z_1^Q, Z_2^Q, \ldots, Z_m^Q\right), \tag{3}$$

where m is the total number of question tokens after tokenization, and the **knowledge features**

$$Z_K = \left(Z_1^K, Z_2^K, \ldots, Z_r^K\right), \tag{4}$$

where r is the number of tokens in the knowledge statement. We then form the final **multimodal input** by concatenating all features:

$$Z = (\underbrace{Z_1^I, \ldots, Z_x^I}_{\text{image features}}, \underbrace{Z_1^Q, \ldots, Z_m^Q}_{\text{question features}}, \underbrace{Z_1^K, \ldots, Z_r^K}_{\text{knowledge features}}). \quad (5)$$

Finally, we fine-tune the VQA-NLE models on this extended input until their accuracy returns to its original level. The overall architecture is shown in Figure 1 where our alleviation model comprises a vision-language model and a vision encoder.

4 Experiment

This section offers a comprehensive assessment of our proposed attack and mitigation strategy. First, in Section 4.1, we detail our experimental setup. We then present quantitative findings, examining how our attacks influence textual variations (Section 4.2) and image manipulations (Section 4.3). Finally, in Section 4.4, we provide a case study that offers deeper, qualitative insights into model behavior.

4.1 Experimental Setup

4.1.1 Dataset

VQA-X (Park et al., 2018) This vision-and-language dataset extends the original Visual Question Answering (VQA) benchmark (Agrawal et al., 2015) by appending detailed, human-written explanations to each question–answer pair. In total, VQA-X encompasses 28,000 images and 33,000 Q&A pairs drawn from the COCO dataset (Lin et al., 2014). Of these, 29,000 pairs are allocated for training, while 1,400 are held out for validation. This additional explanatory text enables richer supervision, encouraging models to justify their answers rather than simply outputting them.

A-OKVQA (Schwenk et al., 2022) A-OKVQA is another vision–language dataset that leverages images from COCO (Lin et al., 2014), but its core feature is the inclusion of rationales. Each sample thus provides a question, an answer, and a rationale explaining the reasoning behind that answer. The dataset consists of 25,000 such triplets, split into 17,100 for training and 1,100 for validation. By including explicit rationales, A-OKVQA further challenges models to exhibit both accuracy and interpretability in their responses.

4.1.2 Evaluation Metrics

In line with prior research (Lai et al., 2024), the quality of generated explanations is measured using the following metrics: BLEU (from B1 to B4, corresponding to BLEU-1 through BLEU-4) (Papineni et al., 2002), METEOR (M) (Banerjee and Lavie, 2005), ROUGE-L (RL) (Lin, 2004), and BERT score (BS) (Zhang et al., 2020). As these datasets are for VQA tasks, we also provide accuracy to measure the correctness of the predicted answers. Additionally, we also follow (Maaz et al., 2024) to assess the VQA-NLE model for its response correctness, detail, and context comprehension.

4.1.3 Implementation Details

Victim Models. Since existing works (Sammani et al., 2022; Suo et al., 2023; Lai et al., 2024) primarily rely on DistilGPT2, we adopt DistilGPT2, pretrained on image-caption pairs, as our "victim" model for both attacks and evaluations on the VQA-X and A-OKVQA datasets, respectively (Suo et al., 2023; Lai et al., 2024).

Alleviation Model. To perform our knowledge-injection (alleviation) strategy, we adopt DistilGPT2 (Sammani et al., 2022), which has been pretrained on a large corpus of image-caption pairs. We then fine-tune it separately on the VQA-X and A-OKVQA training sets, ensuring it can integrate external knowledge effectively.

Image Feature Extraction. Following prior work, we represent each image with features extracted via ViT-B/16 from CLIP (Radford et al., 2021). This encoder converts images into patch-level feature embeddings, which are then fed (alongside question and/or knowledge embeddings) into our models.

Baselines. We compare our attacks against adversarial attacks targeting different modalities. For image-based attacks, we adopt DR (Lu et al., 2020) and SSP (Naseer et al., 2020) as baselines. These approaches are designed to perturb image features exclusively and can be readily adapted to our setting. In contrast, other methods either rely entirely on classifier outputs (Wei et al., 2022b; Wu et al., 2020; Madry et al., 2018; Dong et al., 2019; Xie et al., 2019; Wang et al., 2024) or combine feature perturbation with classification loss (Huang et al., 2019; Inkawhich et al., 2020b,a). Such methods are incompatible with our problem setup, as pre-trained and fine-tuned models often employ dif-

	Method	VQA-X							A-OKVQA						
		B1	B2	B3	B4	RL	M	BS	B1	B2	B3	B4	RL	M	BS
Text-based	Plural	62.6	45.6	33.2	24.4	45.9	40.6	75.4	64.9	39.6	26.8	17.8	41.9	39.9	75.7
	Our Attack	**59.0**	**41.6**	**29.3**	**21.0**	**43.6**	**37.6**	**73.3**	**64.7**	**38.6**	**25.6**	**16.6**	**41.7**	**38.0**	**74.0**
	Alleviation	59.2	41.8	29.5	21.1	43.9	37.8	73.7	65.0	39.1	25.7	16.9	41.8	38.5	74.4
Image-based	DR	64.1	46.9	34.5	25.3	47.0	41.5	75.7	63.9	38.9	26.0	16.9	41.0	37.9	74.0
	SSP	65.3	48.2	35.4	25.7	47.1	42.3	76.0	65.4	40.4	27.4	18.2	42.9	39.6	74.8
	Our Attack	**59.9**	**41.6**	**29.6**	**21.2**	**43.0**	**36.9**	**73.1**	**62.0**	**37.4**	**24.3**	**15.3**	**40.4**	**36.9**	**73.8**

Table 1: Comparison with the baselines on VQA-X and A-OKVQA datasets in the scenario of "unfiltered" scores.

	Method	VQA-X								A-OKVQA							
		B1	B2	B3	B4	RL	M	BS	Acc	B1	B2	B3	B4	RL	M	BS	Acc
Text-based	Plural	65.9	49.1	36.5	27.2	48.5	43.3	77.1	74.6	69.8	45.5	32.3	23.0	46.4	45.2	78.5	37.5
	Our Attack	**63.5**	**46.5**	**33.9**	**24.8**	**46.9**	**41.2**	**75.3**	**67.2**	**69.4**	**45.4**	**31.9**	**22.4**	**46.0**	**44.0**	**77.4**	**31.7**
	Alleviation	63.8	46.7	34.0	24.9	46.9	41.6	75.7	68.1	70.7	46.1	32.2	22.7	46.6	44.7	77.5	30.1
Image-based	DR	67.4	50.4	38.0	28.3	49.8	44.9	77.4	74.1	69.2	46.1	32.8	22.5	45.0	44.2	76.8	36.8
	SSP	69.4	52.6	39.4	29.1	50.2	45.5	77.7	72.6	69.3	46.7	33.8	24.6	47.9	46.1	77.5	38.7
	Our Attack	**63.3**	**45.5**	**33.2**	**24.1**	**46.1**	**40.3**	**75.0**	**70.1**	**66.8**	**43.6**	**30.0**	**19.7**	**43.9**	**41.9**	**76.4**	**33.3**

Table 2: Comparison with the baselines on VQA-X and A-OKVQA datasets in the scenario of "filtered" scores.

ferent prediction heads and are optimized for distinct tasks. For text-based attacks, we refer to the method from (Ravichander et al., 2020) as Plural, since the original work does not assign it a specific name. This approach converts singular nouns into their plural forms. To minimize semantic drift and avoid introducing contradictions, we modify only one noun per question. We adopt this method as one of our baselines for evaluating textual robustness.

4.2 Results on Text-based Attack

The results of our text-based adversarial attacks on the VQA-X and A-OKVQA datasets are detailed in Tables 1 and 2. These tables differentiate between "unfiltered" evaluations, which assess all explanations regardless of the accuracy of the corresponding answers, and "filtered" evaluations, which consider only explanations linked to correct answers. In Table 1, we illustrate that our attack not only compromises the integrity of the original model but also induces a marked reduction in the consistency of the explanations when compared to the baseline method. This effect is quantitatively substantial, with our method resulting in a 4% decrease in BLEU-2 scores on VQA-X and a 1.9% decrease in METEOR scores on A-OKVQA. This highlights the effectiveness of our attack in disrupting the model's ability to generate coherent and contextually appropriate explanations, thereby revealing the model's vulnerability to linguistic perturbations.

Furthermore, the incorporation of external knowledge into the model's framework has demonstrated a capability to alleviate these inconsistencies. By enhancing the contextual grounding of the explanations, this strategy not only restores but also improves their reliability, suggesting that external knowledge can serve as a countermeasure to adversarial attacks. Moving to the filtered results showcased in Table 2, our attack methodology continues to outperform the baseline in terms of diminishing explanation consistency, thereby reinforcing the attack's effectiveness. Concurrently, our defense mechanism again proves beneficial, enhancing the consistency of explanations even when considering only correct answer contexts. This dual success underscores the comprehensive strength of our approach in both compromising and subsequently reinforcing the model's explanatory capabilities. The reduction in consistency, driven by our effective adversarial attacks, correlates strongly with a decline in accuracy, as recorded in the "Acc" column of both tables. This decline emphasizes the direct impact of our attacks on the model's overall performance, highlighting the critical link between the accuracy of answers and the coherence of explanations. These results affirm the necessity of developing more resilient models that can withstand such linguistic adversarial challenges while maintaining high standards of accuracy and explanatory depth.

Method	VQA-X			A-OKVQA		
	Correctness ↓	Detail ↓	Context ↓	Correctness ↓	Detail ↓	Context ↓
DR	2.79	1.90	3.11	2.22	1.78	2.69
SSP	2.72	1.88	3.01	2.26	1.75	2.77
Our Attack	**2.20**	**1.73**	**2.47**	**1.93**	**1.56**	**2.35**

Table 3: Comparison with the baselines on VQA-X and A-OKVQA datasets in the scenario of "unfiltered" scores for image-based attack.

Method	VQA-X				A-OKVQA			
	Correctness ↓	Detail ↓	Context ↓	Acc	Correctness ↓	Detail ↓	Context ↓	Acc
DR	3.03	1.97	3.26	74.1	3.09	2.17	3.44	36.8
SSP	2.98	1.97	3.19	72.6	3.00	2.05	3.36	38.7
Our Attack	**2.39**	**1.77**	**2.57**	**70.1**	**2.72**	**1.89**	**3.04**	**33.3**

Table 4: Comparison with the baselines on VQA-X and A-OKVQA datasets in the scenario of "filtered" scores for image-based attack.

4.3 Results on Image-based Attack

We systematically modify images in a controlled fashion, expecting that the model's explanations remain consistent despite minor visual differences. By strategically removing objects that do not directly answer the question but influence the generation of explanations, we evaluate the robustness of the model in maintaining coherent explanations that align with the altered image content. Importantly, even after our proposed image edits, the attacked images maintain a high degree of similarity to their original counterparts, with average cosine similarities of 81.7% and 82.1% on VQA-X and A-OKVQA datasets, respectively, as measured by CLIP embeddings. This ensures that the visual changes are minimal and localized. Tables 3 and 4 demonstrate the effectiveness of our proposed adversarial attacks, which significantly reduce the consistency of the model's responses compared to those affected by random noise. Notably, our targeted attack leads to a marked decrease in the accuracy scores on both VQA-X and A-OKVQA datasets, with the lowest recorded accuracies being 70.1% and 33.3%, respectively. This substantial drop indicates that the models heavily rely on the presence of specific objects to generate their answers and explanations, resulting in substantial misinterpretations of the image context when these objects are absent.

In the filtered scenario, where only correct answers contribute to the evaluation metrics, the attacked models on VQA-X and A-OKVQA exhibit notable declines in the "Correctness," "Detail," and "Context" scores compared to their performances under noise and original conditions. Specifically, "Detail" score reduction from 1.97 to 1.77 on VQA-X under attack conditions suggests that the removal of contextually relevant objects disrupts the model's ability to provide detailed and relevant explanations. Similarly, compared with the best baseline SSP with 3.19, our attack reduces the "Context" score to 2.57 on VQA-X, highlighting a deterioration in the model's capacity to comprehend and relate to the altered visual context. This is exacerbated on the A-OKVQA dataset, where the "Context" score plummets to 3.04, underscoring the model's increased sensitivity to visual manipulations.

These findings underscore the vulnerability of current VQA models to targeted adversarial attacks that remove non-answering yet contextually significant objects. They also emphasize the need for developing more robust VQA systems that do not merely focus on detectable objects but comprehend the holistic scene to maintain explanation integrity under adversarial conditions.

4.4 Case Study

Text Attack As illustrated in Figure 3 (right), the slightly changed question: *"Is this at an invitational?"* yields the answer and explanation: *"no because there are no people in the picture."* This explanation contradicts the visual evidence in the image, suggesting that the model's visual grounding has been disrupted by the subtle rephrasing. The shift leads the model to overlook relevant visual cues it had initially attended to. However, after being provided with the knowledge describ-

Ground Truth

Q: Is this room clean?
A: yes
E: Everything is spotless and in its place.

Plural

Q: Is this **rooms** clean?
A: yes
E: There is no dirt or stains on the walls

Our Attack

Q: Is this room **neat**?
A: yes
E: There is a toilet and a sink.

Our Alleviation

Q: Is this room **neat**?
K: Cleanliness is determined by the absence of dirt, dust, or clutter. It's subjective and can vary.
A: yes
E: there is no dirt or clutter on the floor.

Ground Truth

Q: Is this at an event?
A: yes
E: There are people gathered in cosplay

Plural

Q: Is this at an **events**?
A: yes
E: There are people dressed up and talking on cell phones

Our Attack

Q: Is this at an **invitational**?
A: no
E: There are no people in the picture.

Our Alleviation

Q: Is this at an **invitational**?
K: Events are organized gatherings of people for specific purposes, such as conferences, concerts, or sports.
A: yes
E: There are people dressed up.

Figure 3: Qualitative examples of our text-based attack.

ing what "event" is, the model is able to grasp the context of the specific question, thus generating the answer and explanation that align with the question. Similarly, in Figure 3 (left), for the question: *"Is this room neat?"*, the model correctly answers: *"yes"*, but the accompanying explanation: *"there is a toilet and a sink."* is semantically misaligned. This explanation is inadequate as it merely lists objects presented in the room without addressing the notion of neatness or cleanliness, and thus fails to justify the answer. In our alleviation strategy, we guide the model by explicitly associating cleanliness with the absence of dirt. This helps steer its reasoning toward evaluating the actual tidiness of the room, resulting in a more contextually appropriate explanation.

Image Attack Figure 4 (top) exemplifies how VQA-NLE models rely on spurious correlations rather than genuine scene understanding. Initially, the model correctly identifies it is not the ocean based on the presence of the dog in the water. However, after removing the dog from the image, the model shifts its prediction to *"yes"*. This suggests that the model's decision-making process is heavily

Figure 4: Qualitative examples of our image-based attack. The "Q" is omitted if it matches the ground truth.

influenced by particular objects rather than reasoning holistically about the scene. In Figure 4 (bottom), the model correctly explains the image is not an old photo by referring to the modern clothing worn by the man. However, after removing people from the image, the model still answers correctly with *"no"*, yet the explanation becomes *"it is in black and white"*, which obviously contradicts the visual evidence, suggesting a disconnect between the model's generation and its visual grounding capabilities. Overall, the model heavily relies on superficial correlations rather than deep reasoning and contextual understanding. Instead of accurately grounding its explanations in the image and question, the model often justifies its answers using spurious associations. The VQA-NLE models struggle to adapt to minor question variations and image modifications, leading to explanations that either misalign with the question or contradict visual evidence. This indicates a fundamental gap between the model's answer generation and its ability to provide logically sound explanations.

5 Conclusion and Future Work

In this paper, we examine the robustness of VQA-NLE models, revealing their susceptibility to generating mutually inconsistent explanations in response to linguistic and semantic variations. To systematically evaluate these vulnerabilities, we implement BERT-Attack that perturbs input questions, and also propose a novel adversarial attack framework that modifies image content. Our ex-

periments show that VQA-NLE models exhibit sensitivity to these perturbations, indicating a reliance on spurious correlations rather than genuine reasoning. To mitigate these inconsistencies, we introduce a method that integrates external knowledge into adversarially perturbed questions. Our results demonstrate that this approach reduces contradictions, thereby enhancing the robustness of VQA-NLE models. For future work, we plan to extend our investigation to large vision-language models such as LLaVA (Liu et al., 2023) and Qwen-VL (Bai et al., 2023). Additionally, we aim to explore the effectiveness of prompting techniques, such as chain-of-thought reasoning, as a defense mechanism against adversarial attacks by improving step-by-step reasoning.

Limitations

Our alleviation method depends on the question-related knowledge, which may not be effective in certain cases. For example, the knowledge "Dresses can be sleeveless or have varying sleeve styles, such as short, long, or cap sleeves." extracted from the benign question "Does the dress have sleeves?" is helpful for the adversarial question "Does the gown have sleeves?" because it relates "gown" with "dress" while guiding models to focus on sleeves. Meanwhile, "The dress could refer to a specific dress that gained viral attention in 2015 due to the optical illusion of its colors." provides little relevant information for answering the question, making it a poor knowledge statement.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the International Conference on Computer Vision*, page 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz1, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. 2020. Make up your mind! Adversarial generation of inconsistent natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4157–4165.

Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual Samples Synthesizing for Robust Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. 2019. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, page 6174–6181.

Qian Huang, Isay Katsman, Horace He, Zeqi Gu, Serge Belongie, and Ser-Nam Lim. 2019. Enhancing Adversarial Example Transferability with an Intermediate Level Attack. In *Proceedings of the International Conference on Computer Vision*, pages 4732–4741.

Nathan Inkawhich, Kevin J Liang, Lawrence Carin, and Yiran Chen. 2020a. Transferable Perturbations of Deep Feature Distributions. In *International Conference on Learning Representations*.

Nathan Inkawhich, Kevin J Liang, Binghui Wang, Matthew Inkawhich, Lawrence Carin, and Yiran Chen. 2020b. Perturbing Across the Feature Hierarchy to Improve Standard and Strict Blackbox Attack Transferability. In *Advances in Neural Information Processing Systems*.

Myeongjun Jang, Bodhisattwa Prasad Majumder, Julian McAuley, Thomas Lukasiewicz, and Oana-Maria Camburu. 2023. KNOW How to Make Up Your Mind! Adversarially Detecting and Alleviating Inconsistencies in Natural Language Explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, page 540–553.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings of the International Conference on Computer Vision*, pages 1244–1254.

Chengen Lai, Shengli Song, Shiqi Meng, Jingyang Li, Sitong Yan, and Guangneng Hu. 2024. Towards More Faithful Natural Language Explanation Using Multi-Level Contrastive Learning in VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2849–2857.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, page 6193–6202.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, page 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, pages 740–755.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated Knowledge Prompting for Commonsense Reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, page 3154–3169.

Yantao Lu, Yunhan Jia, Jianyu Wang, Bai Li, Weiheng Chai, Lawrence Carin, and Senem Velipasalar. 2020. Enhancing Cross-task Black-Box Transferability of Adversarial Examples with Dispersion Reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 937–946.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, page 12585–12602.

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*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *Preprint*, arXiv:2004.14546.

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. 2020. A Self-supervised Approach for Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 259–268.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311–318.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 8779–8788.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and Dark Outside?! Improving Answer Consistency in VQA through Entailed Question Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, page 5860–5865.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8332.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. In *European Conference on Computer Vision*, pages 146–162.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-Consistency for Robust Visual Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6649–6658.

Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. 2023. S3C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2656.

Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. 2024. Boosting the Transferability of Adversarial Attacks with Global Momentum Initialization. In *Expert Systems with Applications*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. 2022b. Towards Transferable Adversarial Attacks on Vision Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2668–2676.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. 2020. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *International Conference on Learning Representations*.

Jialin Wu and Raymond J. Mooney. 2019. Faithful Multimodal Explanation for Visual Question Answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, page 103–112.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan 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.

Lei Xu, Alfredo Cuesta-Infante, Laure Berti-Equille, and Kalyan Veeramachaneni. 2022. R&R: Metric-guided Adversarial Sentence Generation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 438–452.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. 2024. A Task is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. In *European Conference on Computer Vision*.