

Adversarial Textual Robustness of Visual Dialog

Lu Yu^{1*}, Verena Rieser^{2,3†}

¹Tianjin University of Technology, Tian jin, China

²Heriot-Watt University, Edinburgh, United Kingdom

³Alana AI

luyu@email.tjut.edu.cn, v.t.rieser@hw.ac.uk

Abstract

Adversarial robustness evaluates the worst-case performance scenario of a machine learning model to ensure its safety and reliability. For example, cases where the user input contains a minimal change, e.g. a synonym, which causes the previously correct model to return a wrong answer. Using this scenario, this study is the first to investigate the robustness of visually grounded dialog models towards textual attacks. We first aim to understand how multimodal input components contribute to model robustness. Our results show that models which encode dialog history are more robust by providing redundant information. This is in contrast to prior work which finds that dialog history is negligible for model performance on this task. We also evaluate how to generate adversarial test examples which successfully fool the model but remain undetected by the user/software designer. Our analysis shows that the textual, as well as the visual context are important to generate plausible attacks.

1 Introduction

Adversarial robustness has recently gained increased attention within the NLP community (e.g. Moradi and Samwald, 2021; Chang et al., 2021; Goel et al., 2021; Wang et al., 2021). In contrast to this previous work, which focuses on text-only models, we evaluate the adversarial robustness of Visual Dialog (VisDial) models with the aim to understand how different input components contribute to robustness. For example, it has previously been established that multiple input modalities increase robustness of pre-neural conversational interfaces, e.g. (Oviatt, 2002; Bangalore and Johnston, 2009). Here, we want to know which modalities can mitigate input attacks on neural visual dialog systems, and to what extent. This is important, since worst-

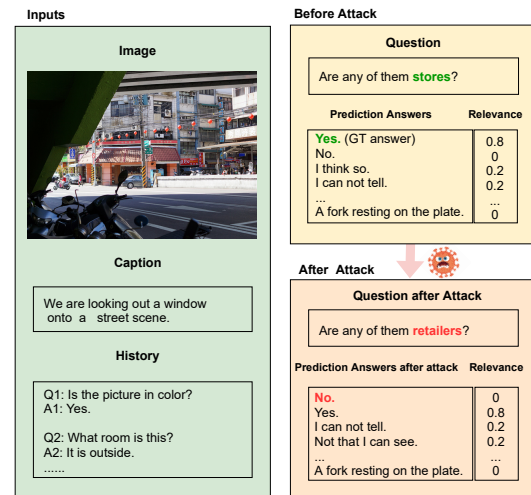


Figure 1: A VisDial agent aims to answer an image-related question by ranking a list of candidate answers, given the dialog history. The attacker permutes the text by replacing a word with its synonym so that the ranking of the predicted answers changes.

case input permutations can also happen spontaneously during user interaction.

A successful attack in our setting is a perturbation which changes the model prediction, but, at the same time, remains unnoticed by the user/software developer. While this is relatively simple for pixel-level visual attacks, textual perturbations are substantially harder to conceal. In the following we thus investigate the trade-off between effectiveness and linguistic quality for text-based attacks. The focus on text-only attacks is further motivated by our preliminary experiments, which show that encoding visual information has very little gain for performance and robustness (around 5%, cf. Table 1), confirming previous results (Massiceti et al., 2018; Agarwal et al., 2020). In addition, we argue that black box visual attacks are very unlikely (if not impossible) to occur in the real-world: They are highly inefficient and expensive to run at real-time (mainly due to the way features are pre-computed using FasterRCNN).

* Work done at Heriot-Watt University.

† Now at Google DeepMind.

To the best of our knowledge, we are the first to explore adversarial attacks on VisDial, which was introduced as a shared task by (Das et al., 2017a). A visual dialog system consists of three components: an image (with a caption), a question and the dialog history, i.e. previous user and system turns. In order to answer the question accurately, the AI agent has to ground the question in the image and infer the context from history, see Fig. 1. Most existing research has focused on improving the modelling performance on this task, (e.g. Das et al., 2017b; Kottur et al., 2018; Jain et al., 2018; Zheng et al., 2019; Niu et al., 2019; Yang et al., 2019; Qi et al., 2020; Murahari et al., 2020; Agarwal et al., 2020; Nguyen et al., 2020; Wang et al., 2020; Chen et al., 2022a), whereas our aim is to evaluate adversarial robustness. In addition, we use these attacks to improve our understanding of how the model works (i.e. interpretability). Our main contributions are:

- *We show that dialog history contributes to model robustness:* We attack ten VisDial models which represent a snapshot of current methods, including different encoding and attention mechanisms, as well as graphical networks and knowledge transfer using pretraining. We measure the performance change before and after attack and show that encoding history helps to increase the robustness against adversarial questions by providing redundant information. We also show that models become more uncertain when the history is attacked.
- *We evaluate adversarial text-generation within VisDial:* We leverage well-established *Synonym Substitution* methods for adversarial black-box attack (Jin et al., 2020; Li et al., 2020) and show that BERT-based models are able to generate more contextually coherent perturbations. We also conduct a detailed analysis to study the trade-off between the effectiveness of the attack versus the text quality.
- *We show that human evaluator are able to identify an attack from the textual and multimodal context:* We conduct a detailed human evaluation to investigate the trade-off between successful attacks and their ability to remain unnoticed by humans. In particular, we evaluate semantic similarity, fluency/grammaticality and label consistency.

All code will be made available.

2 Related Work

Adversarial attacks have been widely investigated within *uni-modal* applications, foremost for com-

puter vision (Narodytska and Kasiviswanathan, 2016; Dong et al., 2018; Xie et al., 2019; Mahmood et al., 2021). Adversarial attacks on text are more challenging due to its discrete nature, which makes it harder to stay undetected. Adversarial textual attack methods can be divided into three levels of granularity (Zhang et al., 2020; Wang et al., 2019): character-level, word-level and sentence-level attacks. While the former two are relatively easy to detect, recent word-level attack methods (e.g. Lee et al., 2022; Zang et al., 2020; Jin et al., 2020; Li et al., 2020; Ren et al., 2019), on the other hand, are more subtle: they are targeted towards ‘vulnerable’ words, which are substituted via their synonyms in order to preserve semantic meaning. In our paper, we explore word-level attack methods on VisDial.

Research on adversarial attacks for multi-modal tasks is relatively scarce, including studies for Optical Character Recognition (Song and Shmatikov, 2018), Scene Text Recognition (Yuan et al., 2020), Image Captioning (Chen et al., 2017, 2022b) and VQA (Li et al., 2021; Shi et al., 2018). Most of these works utilise white box attack, where the parameters, gradient and architecture of the model are available, e.g. by attacking attention (Xu et al., 2018; Sharma et al., 2018). *Whereas we follow a more realistic black-box setting which assumes that the attacker only has access to the model’s prediction on test data.*

Shi et al. (2018) is closest related to our work: they generate adversarial textual attacks for the VQA task using contrastive examples and thus don’t pay attention to semantic similarity. In contrast, we are interested in generating adversarial attacks which follow three desiderata, as outlined by (Morris et al., 2020): An adversarial text should (1) keep the same semantic meaning (*semantic similarity*); (2) guarantee fluency and grammar (*grammaticality*); (3) stay unnoticed by humans, i.e. the human still assigns the correct label, while the model prediction changes (*label consistency*).

3 Method

3.1 Problem Formulation

VisDial is formulated as a discriminative learning task, where the model is given an image I , the dialog history (including the image caption C) $H = (\underbrace{C}_{H_0}, \underbrace{(Q_1, A_1)}_{H_1}, \dots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$, the question Q_t , and $N = 100$ candidate answers $A_t = (A_t^1, A_t^2, \dots, A_t^{100})$ to rank, including the

ground truth (GT), which is labelled Y_t , where t indicates the round ID.

In the following, we focus on generating textual adversarial examples for the question and history (including the caption). That is, for a sentence $X \in \{Q, H\}$, and $F(X) = Y$, a successful adversarial attack sentence X_{adv} should result in $F(X_{adv}) \neq Y$, while meeting the following requirements:

- **Semantic Similarity:** $Sim(X, X_{adv}) \geq \varepsilon$, where $Sim(\cdot)$ is a semantic and syntactic similarity function. The semantic similarity between the original sentence X and the adversarial attack sentence X_{adv} should be above a similarity threshold ε ; Following (Jin et al., 2020), we use Universal Sentence Encoder (Cer et al., 2018) to encode the two sentences into high dimensional vectors and use their cosine similarity score as an approximation of semantic similarity.
- **Grammaticality:** The adversarial attack sentence X_{adv} should be fluent and grammatical.
- **Label Consistency:** Human annotators still assigns the correct GT label Y after the original sentence X changes to X_{adv} .

3.2 Visual Dialog Models

We explore a wide range of ten recent VisDial models to attack – representing a snapshot of current techniques and architectures popular for VisDial.¹

Agarwal et al. (2020) experiment with several multi-modal encodings based on **Modular Co-Attention (MCA)** networks (Yu et al., 2019b): MCA-I encodes the image and question representation using late fusion; MCA-H only encodes the textual history with late fusion; MCA-I-H encodes image and history with late fusion; MCA-I-HGQ encodes all three input modalities using early fusion between question and history; MCA-I-VGH is another early fusion variant which first grounds the image and history.

We also consider **Recursive Visual Attention (RvA)** (Niu et al., 2019) as an alternative to MCA, encoding history and image information.

In addition, we test two variants of causal graphs from (Qi et al., 2020) by adding to **causal principles P1/P2**: P1 removes the history input to the model to avoid a harmful shortcut bias; P2 adds one new (unobserved) node U and three new links to history, question and answer respectively.

¹Details on model architecture can be found in the original papers.

Finally, we test a **Knowledge Transfer (KT)** method based on **Sparse Graph Learning (SGL)** (Kang et al., 2021), which uses P1/P2 as pre-trained models.

3.3 Synonym-based Methods

For generating attacks, we explore two popular synonym-based methods, which first find the vulnerable words of the sentence, and then replace them with a semantically similar word.²

- **TextFooler** (Jin et al., 2020) performs embedding-similarity-based perturbations as a constraint to generate semantically consistent adversarial examples.
- **BERT-Attack** (Li et al., 2020) generates context-aware perturbations using BERT.

Following these previous works, we first detect vulnerable words by calculating prediction change before and after deleting a word. We then impose additional constraints to improve the quality (in particular the grammaticality) of our attacks, which we further analyse in Section 6. We apply a stop word list before synonym substitution, extending the list by (Jin et al., 2020; Li et al., 2020) for our domain. We also filter antonyms following the original BERT-Attack implementation (Li et al., 2020). We then apply additional quality checks for selecting synonym candidates: We filter by part-of-speech (POS)³ to maintain the grammar of the sentence. We then experiment with a semantic similarity threshold ε to choose the top k synonyms. Finally, we iteratively select the word with the highest similarity until the attack is successful. See Appendix B for further details.

3.4 Adversarial Attack on Visual Dialog Models

We perform attacks on 2 textual inputs to the model: The current question and the previous history.

3.4.1 Question Attack

We first attack the current input question in VisDial. In order to do this, we have to adapt textual attacks methods to the unique challenges of this setting, which differs from other common textual attacks (such as sentiment classification, image captioning or news classification) both in terms of textual

²Note that previous work refers to these methods as “synonym-based”, e.g. (Morris et al., 2020), but not all of the substitutions are synonyms. They can also include different lemmas of the same lexeme, such as singular and plural, as well as different spellings, etc. Also see Table 8.

³Using SpaCy <https://spacy.io/api/tagger>.

input, as well as in predicted output: First, the question in VisDial is generally much shorter than a typical declarative sentence in the above tasks. The average length of the question in the VisDial dataset is 6.2 words, which makes it harder to find a word to attack. For instance, “*Is it sunny?*”, “*What color?*”, “*How many?*”, there is only one word left to attack after filtering out the stop words, i.e. {*is, it, what, how*}.

For the VisDial task, the model ranks N possible candidate answers according to its log-likelihood scores. The attack is considered successful once the *top ranked answer* differs from the GT. However, there can be several candidate answers which are semantically similar or equivalent, such as “*yes/yep/yeah*”. This is different from other labelling tasks, such as “*positive/neutral/negative*” sentiment. We account for this fact by considering several common retrieval metrics before and after the attack, including $R@k$ ($k=1,5,10$), *Mean Reciprocal Rank (MRR)*, and *Normalized Discounted Cumulative Gain (NDCG)* – a measure of ranking quality according to manually annotated semantic relevance scores in a 2k subset of VisDial.

3.4.2 History Attack

We also attack the textual history using the same procedure. The use of history is the main distinguishing feature between the VisDial and the VQA task, and thus of central interest to our work. History is mainly used for contextual question understanding, including co-reference resolution, e.g. “*What color are they?*”, and ellipsis, e.g. “*Any others?*” (Yu et al., 2019a; Li and Moens, 2021).

Our preliminary results indicate that attacking history is hardly ever successful, i.e. does not result in label change. This is in line with previous work, which suggests that history only plays a negligible role for improving model performance on the VisDial task, e.g. (Massiceti et al., 2018; Agarwal et al., 2020). However, there is also some evidence that history helps, but to a smaller extent. For example, Yang et al. (2019) show that accuracy can be improved when forcing the model to pay attention to history. Similarly, Agarwal et al. (2020) show that history matters for a sub-set of the data.

In a similar vein, we investigate how history contributes to the model’s robustness and, in particular, can increase the model’s certainty in making a prediction. We adopt the *perplexity* metric, following (Sankar et al., 2019), to measure the change of prediction distribution after (unsuccessfully) attack-

ing the history, i.e. after adding the perturbation to the history while the top-1 prediction is unchanged. The difference between the perplexity before and after the attack reflects the uncertainty change of the model. The perplexity with the original history input is calculated with the following equation:

$$PPL(F(X), Y) = - \sum_X F(X) \log_2 Y \quad (1)$$

And the perplexity after attack is:

$$PPL(F(X_{adv}), Y) = - \sum_{X_{adv}} F(X_{adv}) \log_2 Y \quad (2)$$

4 Experimental Setup

4.1 Dataset

We use the VisDial v1.0 dataset, which contains 123,287 dialogs for training and 2,064 dialogs for validation. The ten target models are trained on the training set and the adversarial attacks are generated for *validation* set (as the test set is only available to challenge participants).

4.2 Automatic Evaluation Metrics

In order to assess the impact of an attack, we use the automatic evaluation metrics from (Jin et al., 2020): The accuracy of the model tested on the original validation data is indicated as *original accuracy* and *after accuracy* on the adversarial samples – the larger gap between these two accuracy means the more successful of our attack (cf. relative performance drop $[\Delta]$). The *perturbed word percentage* is the ratio of the perturbed words and the length of the text. The *semantic similarity* measures the similarity between the original text and the adversarial text by cosine similarity score. The *number of queries* shows the efficiency of the attack (lower better). In addition, we use retrieval based metrics to account for the fact that VisDial is a ranking task: *original/after R@{5, 10}* measures the performance of top 5/10 results before and after attack (where R@1 corresponds to accuracy); we also report *original/after mean reciprocal rank (MRR)* and *original/after Normalized Discounted Cumulative Gain (NDCG)* which measure the quality of the ranking. Detailed results with $R@k$ ($k=10$) are shown in Appendix B and C. Further implementation details are given in Appendix A.

5 Results

5.1 Question Attack

Table 1 summarises the results for attacking the question. Across the ten target models, we first

Question Attack												
Inputs	Methods	Orig.R@1	Aft.R@1 [Δ]	Orig.R@5	Aft.R@5 [Δ]	Orig.NDCG	Aft.NDCG [Δ]	Orig.MRR	Aft.MRR [Δ]	Pert.	S.S.	Quer.
BERT-Attack												
I-only	MCA-I	46.6	38.2 [-18.0]	76.3	62.7 [-17.8]	61.5	54.9 [-10.7]	60.0	47.7 [-20.5]	16.7	74.4	5.2
H-only	MCA-H	45.9	40.0 [-12.9]	76.8	67.3 [-12.4]	52.2	48.4 [-7.3]	60.0	51.1 [-14.8]	16.7	75.4	5.2
	MCA-I-HGQ	50.8	45.6 [-10.2]	81.7	71.4 [-12.6]	60.0	55.2 [-8.0]	64.3	55.6 [-13.5]	17.1	74.1	5.2
I+H	MCA-I-VGH	48.6	43.3 [-10.9]	78.7	68.0 [-13.6]	62.6	57.3 [-8.5]	62.2	53.3 [-14.3]	16.7	74.3	5.2
	MCA-I-H	50.0	45.2 [-9.6]	81.4	69.5 [-14.6]	59.6	54.6 [-8.4]	63.8	54.6 [-14.4]	16.7	74.8	5.2
I+H	RvA	49.9	43.9 [-12.0]	82.2	72.2 [-12.2]	56.3	50.9 [-9.6]	64.2	54.5 [-15.1]	17.0	74.4	5.2
I-only	P1	48.8	43.5 [-10.9]	80.2	69.2 [-13.7]	60.0	54.2 [-9.7]	62.9	54.1 [-14.0]	17.4	74.2	5.2
I+H	P1+P2	41.9	37.1 [-11.5]	66.9	57.8 [-13.6]	73.4	67.9 [-7.5]	54.0	46.2 [-14.4]	17.0	73.7	5.2
	SLG	49.1	43.9 [-10.6]	81.1	72.1 [-11.1]	63.4	58.4 [-7.9]	63.4	55.0 [-13.2]	17.5	73.4	5.2
I+H	SLG+KT	48.7	42.6 [-12.5]	71.3	60.8 [-14.7]	74.5	68.2 [-8.5]	59.9	50.3 [-16.0]	17.3	74.6	5.2
TextFooler												
I-only	MCA-I	46.6	36.1 [-22.5]	76.3	63.9 [-16.3]	61.5	53.9 [-12.4]	60.0	47.1 [-20.5]	16.8	74.4	19.7
H-only	MCA-H	45.9	39.1 [-14.8]	76.8	68.5 [-10.8]	52.2	48.0 [-8.0]	60.0	51.1 [-14.8]	17.1	74.6	19.7
	MCA-I-HGQ	50.8	44.2 [-13.0]	81.7	71.6 [-12.4]	60.0	54.4 [-9.3]	64.3	54.8 [-14.8]	17.0	74.4	19.9
I+H	MCA-I-VGH	48.6	41.5 [-14.6]	78.7	68.2 [-13.3]	62.6	56.5 [-9.7]	62.2	52.3 [-15.9]	16.5	74.4	19.8
	MCA-I-H	50.0	43.1 [-13.8]	81.4	71.2 [-12.5]	59.6	53.7 [-9.9]	63.8	54.0 [-15.4]	16.9	74.7	19.8
I+H	RvA	49.9	43.6 [-12.6]	82.2	73.2 [-10.9]	56.3	50.2 [-10.8]	64.2	55.3 [-13.9]	16.9	74.9	19.9
I-only	P1	48.8	42.6 [-12.7]	80.2	71.1 [-11.3]	60.0	53.5 [-10.8]	62.9	54.4 [-13.5]	17.3	74.3	20.1
I+H	P1+P2	41.9	35.8 [-14.6]	66.9	56.9 [-14.9]	73.4	66.9 [-8.9]	54.0	45.1 [-16.5]	17.1	73.7	19.8
	SLG	49.1	43.1 [-12.2]	81.1	73.4 [-9.5]	63.4	57.8 [-8.8]	63.4	55.3 [-12.8]	17.3	74.2	19.9
I+H	SLG+KT	48.7	41.6 [-14.6]	71.3	59.7 [-16.3]	74.5	67.6 [-9.3]	59.9	49.8 [-16.9]	17.1	74.6	19.9

Table 1: VisDial model performance before attacking question (Orig.) and after (Aft.). In addition to standard metrics, we measure the perturbed word percentage (Pert.), semantic similarity (S.S) and the number of queries (Quer.) to assess BERT-Attack vs. TextFooler. The *relative* performance drop is listed as [Δ]. Highlights indicate the **least robust** and **most robust** model.

compare different **input encodings and fusion mechanisms**, answering the question whether multiple inputs can help robustness. We find that MCA-I (with image input only) is the least robust model with a relative performance drop of over 22% on R@1 using TextFooler. MCA-H (with no image input) is vulnerable with respect to R@1, but does well on NDCG, suggesting that history helps to produce a semantically similar response despite the lack of encoding the input image. One possible explanation of these results is given by previous research claiming that VisDial models mainly pay attention to text while ignoring the image, e.g. (Massiceti et al., 2018). However, in contrast to claims by (Massiceti et al., 2018), we find that history is important for robustness: In general, models encoding history are more robust with the *MCA-I-H* model being the least vulnerable model. Note that this is also the best performing model in (Agarwal et al., 2020).

Next, we compare attention mechanisms on the input encodings. Recursive visual Attention (RvA) in general shows lower robustness than MCA-based methods. Causal encodings using graphs lead to comparable robustness results for P1. Adding P2 results in a slight drop in robustness. This is interesting, because P2 adds an unobserved node to represent history while avoiding spurious

Question	R@1 Answer
Orig.: Is the mannequin a woman ?	Orig.: No.
Aft.: Is the mannequin a girl ?	Aft.: Yes.
Orig.: Are there any pets in the photo?	Orig.: No pets or people.
Aft.: Are there any animals in the photo?	Aft.: No.
Orig.: What color is the plane?	Orig.: White.
Aft.: What colour is the plane?	Aft.: Not sure.

Figure 2: Examples of answer change after question attack on MCA-I-H model with BERT-Attack.

correlations from training data. This drop thus might suggest that previous robustness is due to the very same bias. Additionally, we observe that knowledge transfer (KT) via pre-training for the SLG method helps to boost the performance of NDCG, however not the robustness.

We further perform an example based analysis of the top-1 predicted answer changes after a successful question attack, see Fig. 2. We observe answer changes to the opposite meaning (e.g. from “no” to “yes”), which can be considered as a maximum successful attack. Some answers, however, change to a very similar meaning in context (e.g. from “No pets or people” to “No”), which is reflected in fewer NDCG changes. In some cases, the answer changes from certain / definite to uncertain / noncommittal and the other way round (e.g. from “white” to “Not sure”).

BERT-Attack	TextFooler
N/A	Orig.: Is it a flat screen? Aft.: Is it a loft screen?
Orig.: Is it a close up of their faces or their bodies? Aft.: Is it a close up of their face or their bodies?	Orig.: Is it a close up of their faces or their bodies? Aft.: Is it a close up of their confront or their bodies?
Orig.: What color is the house ? Aft.: What color is the home ?	Orig.: What color is the house ? Aft.: What color is the residence ?
Orig.: Are there trees no the mountain? Aft.: Are there woods on the mountain?	Orig.: Are there trees no the mountain? Aft.: Are there sapling on the mountain?

Figure 3: Example attacks on the MCA-I-H target model generated by BERT-Attack and TextFooler.

	History Attack	
	Orig.PPL	Aft.PPL [Δ]
MCA-I	-	-
MCA-H	53.2	60.0 [+6.8]
MCA-I-HGQ	49.4	52.2 [+2.8]
MCA-I-VGH	52.3	52.3 [0]
MCA-I-H	49.5	51.9 [+2.4]
RvA	53.4	56.4 [+3.0]
P1	-	-
P1+P2	77.0	77.0 [0]
SLG	52.7	53.4 [+0.7]
SLG+KT	65.0	65.3 [+0.3]

Table 2: Comparison of perplexity increase [Δ] when attacking the history of different VisDial models with BERT-Attack.

Next, we **compare the two attack methods**. We find that TextFooler is more effective: It achieves up to 4.5% higher drop than BERT-Attack. However, BERT-Attack is more efficient: It reduces the number of queries (Quer.) about four times compared to TextFooler. Efficiency is important in attack settings, as attackers always run into danger of being discovered. Furthermore, the perturbed word percentage (Pert.) for both methods is around 17%, which means the average perturbation is about one word for each question (since the average length of the question is 6.2). Similarly, the semantic similarity (S.S.) is over 70% which is about the same across all models.

We further compare TextFooler and BERT-Attack using an example-based analysis, see Fig. 3. We find that TextFooler is not able to distinguish words with multiple meanings (homonyms), whereas BERT-Attack is able to use BERT context-embeddings to disambiguate. Consider the examples where TextFooler replaces “flat” (adverb) with “loft” (noun) and “faces” (noun) with “confront” (verb), which POS tagger failed to catch.

Based on the above results, we use BERT-Attack to attack the MCA-I-H model in the following experiments.

	Caption	User (question)	System (answer)
Attack	44.9%	30.8%	24.3%

Table 3: Comparing which part of History was chosen for an attack on MCA-I-H model with BERT-Attack.

	$\Delta R@1$	$\Delta NDCG$	ΔMRR
Random	-7.6	-6.0	-12.4
Ours	-9.6	-8.4	-14.4

Table 4: Effect of vulnerable word attack on MCA-I-H model with BERT-Attack.

5.2 History Attack

Next, we analyse the results for attacking dialog history. As explained in Section 3.4.2, we consider an attack ‘successful’ once the probability of the corresponding GT decreases and we use perplexity to measure the uncertainty of the prediction. The results in Table 2 show that attacking history increases the uncertainty of almost all the models, especially when the history is the unique input component (MCA-H model).⁴ This confirms our preliminary results that encoding history increases robustness.

We then analyse which part of history gets attacked the most, see Table 3. We find that 44.9% of the time the image caption was attacked, followed by system answer 30.8% and user question 24.3%. We thus conclude that the image caption, i.e. a description of what can be seen in the picture, is the most vulnerable part (and ergo the most informative) compared to the rest of history. We hypothesise that the image caption can “replace” information corrupted by the attack. Thus, history contributes to robustness by providing redundant information.

6 Detailed Analysis of Linguistic Quality Constraints

Next, we analyze the impact of the linguistic quality constraints. We are interested in the trade-off between using these constraints to produce high quality text (which increases the chance of the attack to remain unnoticed by humans) versus an effective attack (which increases the chance of the model changing its prediction).

⁴Attacking the history of MCA-I-VGH model doesn’t change the prediction distribution because its encoder only uses a single round of history following (Agarwal et al., 2020).

	$\Delta R@1$	$\Delta NDCG$	ΔMRR
All	-12.6	-9.2	-10.3
Ours	-9.6	-8.4	-14.4

Table 5: Effect of stop words set on MCA-I-H model with BERT-Attack.

ε	Num./(%)	$\Delta R@1$	$\Delta NDCG$	ΔMRR
0.1	219 (10.6%)	-10.8	-9.6	-14.1
0.3	215 (10.4%)	-10.8	-9.2	-14.1
0.5	198 (9.6%)	-9.6	-8.4	-14.4
0.7	135 (6.5%)	-6.0	-6.7	-15.2

Table 6: Comparison of number of successful attacks (*total val set n=2064*) with different semantic similarity thresholds ε on MCA-I-H model with BERT-Attack.

Effect of Selecting Vulnerable Words First, we compare the results of choosing a random word in text to attack and our vulnerable word attack. The results in Table 4 confirm that attacking the vulnerable word achieves a 2.0% higher relative drop for R@1, NDCG and MRR.

Effect of Stop Words Next, we compare the results with/without stop words. The results in Table 5 show that attacking all words leads to more successful attack in terms of R@1 and NDCG, while attacking with stopwords leads more successful attacks for MRR. We use stop words list for all the experiments since attacking question words, preposition or pronouns result in highly ungrammatical sentences.

Effect of Semantic Similarity The semantic similarity threshold between the original text and adversarial text is used to guarantee the similar meaning of the attack. In the previous experiments, we set 0.5 as default threshold. Table 6 shows results with different semantic similarity thresholds (0.1, 0.3, 0.5 and 0.7) respectively. The results show that when increasing the threshold ε from 0.1 to 0.7, the number of successful attack decreases 4.1%, while R@1 and NDCG drop around 3% after attack, which means there are more successful attacks if we loosen the semantic similarity constraint. In addition, the examples in Fig. 4 illustrate that a lower semantic similarity threshold comes at the cost of lower fluency and grammaticality, i.e. at the price of being more easily detectable by humans. We will explore this trade-off in more detail in the human evaluation study.

Next, we analyze the combined effect of adding POS, semantic similarity constraint and grammar

Constraints	Examples
ε (0.7)	Orig.: Is it a large church? Aft.: Is it a big church?
	Orig.: What color is the wine? Aft.: What colour is the wine?
ε (0.5)	+ Orig.: Is her hair pulled back? Aft.: Is her wig pulled back?
	+ Orig.: Is the fireplace lit? Aft.: Is the furnace lit?
ε (0.3)	+ Orig.: What is the adult doing ? Aft.: What is the adult done ?
	+ Orig.: Is there buildings ? Aft.: Is there houses ?
ε (0.1)	+ Orig.: Is the picture outside ? Aft.: Is the picture beyond ?
	+ Orig.: Are they titled ? Aft.: Are they untitled ?

Figure 4: Attack examples with different semantic similarity thresholds ε on MCA-I-H model with BERT-Attack.

	Num./(%)	$\Delta R@1$	$\Delta NDCG$	ΔMRR
Raw Attack	224 (10.9%)	-11.6	-9.9	-13.9
+POS	221 (10.7%)	-11.0	-9.7	-14.1
+POS+ ε (0.5)	198 (9.6%)	-9.6	-8.4	-14.4
+POS+ ε (0.5)+Gram.	190 (9.2%)	-9.2	-6.2	-13.6

Table 7: Effect of different quality constraints on MCA-I-H model with BERT-Attack.

check modules – using the same grammar tool as in (Morris et al., 2020). The results in Table 7 show that an attack is less successful as the number of constraints increases – also see examples in Appendix D Fig. 6. The success from raw attack to linguistically ‘disguised’ attack decreases 2.4% on R@1, 3.7% on NDCG, but there is little effect on MRR.

7 Human Evaluation Study

We evaluate the quality of our generated adversarial question attack by asking human judges on Amazon Mechanical Turk (AMT) to rate three aspects: if the generated question preserve the semantic similarity (*semantic similarity with/without given image*); if the generated question is natural

Attack Types	Percentage	Gram. Score
British vs. American English	34.9%	4.923
Synonyms/near synonyms	34.3%	4.417
Singular vs. Plural	19.7%	3.974
Comparatives and Superlatives	4.0%	4.208
Others	7.1%	3.452

Table 8: Percentage and grammaticality score of different types of attack on MCA-I-H model with BERT-Attack.

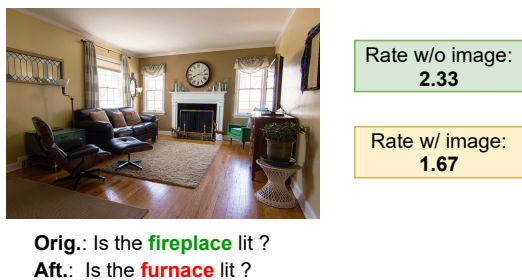


Figure 5: The visual context changes the perceived similarity rating by humans: ‘furnace’ becomes more dissimilar to ‘fireplace’ in a living room context.

and grammatical (*grammaticality*); if the human’s prediction is unchanged for the generated question (*label consistency*). We evaluate a total of 198 generated attacks, randomly sampled from the development set, where three users are asked to rate each instance, following best practices for human evaluation in NLP (van der Lee et al., 2019).⁵ See Appendix for further details.

Evaluation of Semantics We first ask crowd workers to evaluate whether the original and the adversarial question still have the same meaning on a scale from 1 to 4, where 1 is “One text means something completely different” and 4 is “They have exactly the same meaning”. Appendix E shows the interface and instructions. We elicit ratings with and without showing the image in order to measure the effects of multimodal grounding. Our results show that the semantic similarity is rated slightly lower when shown together with the original image (average score **3.518 / 4**) than without image (average score **3.564 / 4**). This indicates that the visual context can change the semantic similarity ratings, as illustrated in Fig. 5. Therefore, a promising future avenue is to use visually grounded word embeddings for generating synonyms for V+L tasks.

Evaluation of Grammaticality Next, we evaluate whether the utterance is fluent and grammatical (as defined in Appendix E) on a scale from 1-5, where 1 is “Not understandable” and 5 is “Everything is perfect; could have been produced by a native speaker”. Overall, our attacks are rated as highly grammatical (average score 4.429 / 5). We furthermore investigate the effect of different attacks, where we manually label five common types of successful attacks, see Table 8. We find that

⁵For example, van der Lee et al. (2019) report that standard evaluations only include an average of 100 samples rated by a median of 4 annotators. We chose to increase the sample size in order to increase the effect and chose 3 annotators in order to allow for ties.

synonyms/near synonyms is the main type of attack, closely followed by *British vs. American English* (e.g. “color” vs. “colour”). In addition, we identify *Singular vs. Plural*, and *Comparatives and Superlatives* (e.g. “great/greater/greatest”), as well as *Others* which mainly include grammar operations like uncaught POS change (e.g. “sunny” vs. “sun”) and tense change (e.g. “eat” vs. “ate”). Looking at the grammar ratings, we find that substituting *British vs. American English* has the least impact on grammaticality, whereas grammatical operations, such as replacing singular with plural, as well as changes classified under *Others* have the highest impact on the perceived linguistic quality.

Evaluation of Label Consistency Finally, we evaluate label consistency by asking users to judge whether the answer remains unchanged for the adversarial question by selecting among “1 - Yes, answer is correct”, “2 - No, answer is incorrect” and “3 - Unsure” as shown in Appendix E. We ask three judges to rate each instance and describe results by averaging and by (a more conservative) majority vote to assign a gold label. The results show that most (**82.0%** by averaging and **86.4%** by majority vote) crowdworkers think the answer is unchanged, few (9.6% and 8.1%) think the answer changes, and the rest (8.4% and 5.5%) are not sure about the change. We thus conclude that synonym-based attacks mostly remain undetected by humans.

8 Conclusions and Limitations

We present a detailed study investigating adversarial robustness of visual dialog models. We find that multiple inputs increase robustness, and in particular dialog history contributes to robustness, despite previous results which suggest that history has negligible effect on model performance, e.g. (Massiceti et al., 2018; Agarwal et al., 2020). We investigate the trade-off between effectiveness and linguistic quality, where we show limitations of current synonym-based textual attack models, and stress the importance of context (both textual as well as multi-modal) to generate semantically coherent and grammatically fluent adversarial attacks, which are likely remain undetected. While the observed effects of visually-grounded interpretations in our human evaluation were relatively small, we do believe that it is an important future direction. For example, we expect improved results by using synonym substitution methods based on visually-grounded word embeddings.

Ethics Statement

We use adversarial attack as a tool to evaluate the robustness of visual dialog models. However, the same techniques can also be used to maliciously attack the system. Our experiments demonstrate that most synonym-based attacks are successful in remaining undetected by humans. However, our results also show that the most effective attacks are also the ones which are easiest for humans to detect. Further work is thus needed to automatically detect malicious attacks, e.g. using our proposed grammaticality and contextual multimodal methods.

Acknowledgments

Lu Yu and Verena Rieser were supported by the EPSRC project ‘AISEC: AI Secure and Explainable by Construction’ (EP/T026952/1), Lu Yu was also supported by NSFC (NO.62202331), Verena Rieser was also supported by ‘Gender Bias in Conversational AI’ (EP/T023767/1) and ‘Equally Safe Online’ (EP/W025493/1), as well as by a Leverhulme Trust Senior Research Fellowship (SRF/R1/201100).

References

- Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. [History for visual dialog: Do we really need it?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.
- Srinivas Bangalore and Michael Johnston. 2009. Robust understanding in multimodal interfaces. *Computational Linguistics*, 35(3):345–397.
- Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
- Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. [Robustness and adversarial examples in natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26, Punta Cana, Dominican Republic & Online. Association for Computational Linguistics.
- Cheng Chen, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, Yudong Zhu, and Xiaodong Gu. 2022a. Utc: A unified transformer with inter-task contrastive learning for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18103–18112.
- Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2017. [Attacking visual language grounding with adversarial examples: A case study on neural image captioning](#). *arXiv preprint arXiv:1712.02051*.
- Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, and Wei Yang. 2022b. [Nicslowdown: Evaluating the efficiency robustness of neural image caption generation models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15365–15374.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. [Visual dialog](#). In *CVPR*, pages 326–335.
- Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017b. [Learning cooperative visual dialog agents with deep reinforcement learning](#). In *ICCV*, pages 2951–2960.
- Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. [Boosting adversarial attacks with momentum](#). In *CVPR*, pages 9185–9193.
- Karan Goel, Nazneen Fatema Rajani, Jesse Vig, Zachary Taschdjian, Mohit Bansal, and Christopher Ré. 2021. [Robustness gym: Unifying the NLP evaluation landscape](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 42–55, Online. Association for Computational Linguistics.
- Unnat Jain, Svetlana Lazebnik, and Alexander G Schwing. 2018. [Two can play this game: visual dialog with discriminative question generation and answering](#). In *CVPR*, pages 5754–5763.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. [Is bert really robust? a strong baseline for natural language attack on text classification and entailment](#). In *Proceedings of the AAAI conference on artificial intelligence*, pages 8018–8025.
- Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. 2021. [Reasoning visual dialog with sparse graph learning and knowledge transfer](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 327–339.
- Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. [Visual coreference resolution in visual dialog using neural module networks](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.
- Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. 2022. [Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization](#). In *International Conference on Machine Learning*, pages 12478–12497. PMLR.

- Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051.
- Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.
- Mingxiao Li and Marie-Francine Moens. 2021. [Modeling coreference relations in visual dialog](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3306–3318, Online. Association for Computational Linguistics.
- Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. 2021. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847.
- Daniela Massiceti, P. Dokania, N. Siddharth, and P. Torr. 2018. Visual dialogue without vision or dialogue. In *Proceedings of the NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*.
- Milad Moradi and Matthias Samwald. 2021. [Evaluating the robustness of neural language models to input perturbations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. [Reevaluating adversarial examples in natural language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3829–3839, Online. Association for Computational Linguistics.
- Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. 2020. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *European Conference on Computer Vision*, pages 336–352. Springer.
- Nina Narodytska and Shiva Prasad Kasiviswanathan. 2016. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*.
- Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *European Conference on Computer Vision*, pages 223–240. Springer.
- Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6679–6688.
- Sharon Oviatt. 2002. Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. *Advances in computers*, 56:305–341.
- Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869.
- Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.
- Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. [Do neural dialog systems use the conversation history effectively? an empirical study](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.
- Vasu Sharma, Ankita Kalra, Sumedha Chaudhary Vaibhav, Labhesh Patel, and Louis-Phillippe Morency. 2018. Attend and attack: Attention guided adversarial attacks on visual question answering models. In *Proc. Conf. Neural Inf. Process. Syst. Workshop Secur. Mach. Learn.*, volume 4323.
- Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3715–3727.
- Congzheng Song and Vitaly Shmatikov. 2018. Fooling ocr systems with adversarial text images. *arXiv preprint arXiv:1802.05385*.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.
- Boxin Wang, Chejian Xu, Shuohang Wang, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Awadallah, and Bo Li. 2021. [Adversarial glue: A multi-task benchmark for robustness evaluation of language models](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Wenqi Wang, Lina Wang, Run Wang, Zhibo Wang, and Aoshuang Ye. 2019. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*.
- Yue Wang, Shafiq Joty, Michael Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. 2020. Vd-bert: A unified vision and dialog transformer with bert. In

- Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3325–3338.
- 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.
- Xiaojun Xu, Xinyun Chen, Chang Liu, Anna Rohrbach, Trevor Darrell, and Dawn Song. 2018. Fooling vision and language models despite localization and attention mechanism. In *CVPR*, pages 4951–4961.
- Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *ICCV*, pages 2561–2569.
- Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019a. [What you see is what you get: Visual pronoun coreference resolution in dialogues](#). 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)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.
- Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019b. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.
- Xiaoyong Yuan, Pan He, Xiaolin Lit, and Dapeng Wu. 2020. Adaptive adversarial attack on scene text recognition. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 358–363. IEEE.
- Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080.
- Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.
- Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. 2019. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6669–6678.

A Licence

Visual Dialog annotations and this website are licensed under a Creative Commons Attribution 4.0 International License.

B Implementation Details

All models are implemented with Pytorch. We embedded BERT-Attack and TextFooler to our VisDial system⁶. We initially set the semantic similarity threshold 0.5 for attacking both question and history (but see detailed study of different threshold in Table 6).

To ensure the high quality of generated attack, We did the following modifications for original BERT-attack method:

- We updated the stop-word list by adding some new words in and removing some words out due to the visual dialogue domain, compared to the original stop-word list⁷.
- We filter by part-of-speech (POS) to maintain the grammar of the sentence for selecting synonym candidates.
- We used the grammar tool from (Morris et al., 2020) to check the sentence after the attack.

⁶BERT-Attack code from <https://github.com/LinyangLee/BERT-Attack> and TextFooler code from <https://github.com/jind11/TextFooler>.

⁷The stop-word list is updated to {does, do, see, look, seem, be, some, sort, have, has, had, so, a, about, above, across, after, afterwards, again, against, ain, all, almost, alone, along, already, also, although, am, among, amongst, an, and, another, any, anyhow, anyone, anything, anyway, anywhere, are, aren, aren't, around, as, at, back, been, before, beforehand, behind, being, below, beside, besides, between, beyond, both, but, by, can, cannot, could, couldn, couldn't, d, didn, didn't, doesn, doesn't, don, don't, down, due, during, either, else, elsewhere, even, ever, everyone, everything, everywhere, except, for, former, formerly, from, hadn, hadn't, hasn, hasn't, haven, haven't, he, hence, her, here, hereafter, hereby, herein, hereupon, hers, herself, him, himself, his, how, however, i, if, in, indeed, into, is, isn, isn't, it, it's, its, itself, just, latter, latterly, least, ll, may, me, meanwhile, mightn, mightn't, mine, moreover, must, mustn, mustn't, my, myself, namely, needn, needn't, neither, never, nevertheless, next, no, nobody, none, noone, nor, not, nothing, now, nowhere, o, of, off, on, once, only, onto, or, other, others, otherwise, our, ours, ourselves, out, over, per, please, s, shan, shan't, she, she's, should've, shouldn, shouldn't, somehow, something, sometime, somewhere, such, t, than, that, that'll, the, their, theirs, them, themselves, then, thence, there, thereafter, thereby, therefore, therein, thereupon, these, they, this, those, through, throughout, thru, thus, to, too, toward, towards, under, unless, until, up, upon, used, ve, was, wasn, wasn't, we, were, weren, weren't, what, whatever, when, whence, whenever, where, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, whoever, whole, whom, whose, why, with, within, without, won, won't, would, wouldn, wouldn't, y, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves. }

Constraints	Examples
Raw + POS + $\mathcal{E}(0.5)$ + Gram	Orig.: Is it a large church? Aft.: Is it a big church? ----- Orig.: Can you see the sun ? Aft.: Can you see the sunlight ?
Raw + POS + $\mathcal{E}(0.5)$	+ Orig.: What color is the tennis court? + Aft.: What colour is the tennis court? ----- + Orig.: Does the snow appear fresh? + Aft.: Does the snow appears fresh?
Raw + POS	+ Orig.: Are they indoors ? + Aft.: Are they outdoors ? ----- + Orig.: Is this inside ? + Aft.: Is this interior ?
Raw	+ Orig.: Is it red ? + Aft.: Is it reds ? ----- + Orig.: How tall is the man? + Aft.: How big is the man?

Figure 6: Generated adversarial examples under different quality constraints on MCA-I-H model with BERT-Attack.

C Full Table of Question Attack

We show the full table of question attack results including R@10 in Table 9 as supplement of Table 1.

D Detailed Results for Linguistic Analysis

The full tables of the in-depths studies are shown in Table 10, Table 11, Table 12 and Table 13, as supplement Table 4, Table 5, Table 6, Table 7 respectively.

E Details of Human Evaluation Study

Here, we provide more details on the human study. We show the interface of semantic similarity experiment for AMT task in Figure 7, including the instruction (top). Two versions of this interface are conducted, where one is provided with image, one is without image. The interface of fluency/grammaticality experiment for AMT task is shown in Figure 8. Two versions of this interface are done as well, where one is with grammar checker and one is without. Finally, the interface of label consistency experiment is shown in Figure 9.

Question Attack													
	Orig.R@1	Aft.R@1 [Δ]	Orig.R@5	Aft.R@5 [Δ]	Orig.R@10	Aft.R@10 [Δ]	Orig.NDCG	Aft.NDCG [Δ]	Orig.MRR [Δ]	Aft.MRR	Pert.	S.S.	Quer.
BERT-Attack													
MCA-I	46.6	38.2 [-18.0]	76.3	62.7 [-17.8]	86.6	74.1 [-14.4]	61.5	54.9 [-10.7]	60.0	47.7 [-20.5]	16.7	74.4	5.2
MCA-H	45.9	40.0 [-12.9]	76.8	67.3 [-12.4]	86.8	76.6 [-11.8]	52.2	48.4 [-7.3]	60.0	51.1 [-14.8]	16.7	75.4	5.2
MCA-I-HGQ	50.8	45.6 [-10.2]	81.7	71.4 [-12.6]	90.2	80.3 [-11.0]	60.0	55.2 [-8.0]	64.3	55.6 [-13.5]	17.1	74.1	5.2
MCA-I-VGH	48.6	43.3 [-10.9]	78.7	68.0 [-13.6]	88.6	78.4 [-11.5]	62.6	57.3 [-8.5]	62.2	53.3 [-14.3]	16.7	74.3	5.2
MCA-I-H	50.0	45.2 [-9.6]	81.4	69.5 [-14.6]	90.8	80.0 [-11.9]	59.6	54.6 [-8.4]	63.8	54.6 [-14.4]	16.7	74.8	5.2
RvA	49.9	43.9 [-12.0]	82.2	72.2 [-12.2]	91.1	82.6 [-9.3]	56.3	50.9 [-9.6]	64.2	54.5 [-15.1]	17.0	74.4	5.2
P1	48.8	43.5 [-10.9]	80.2	69.2 [-13.7]	89.7	80.7 [-10.0]	60.0	54.2 [-9.7]	62.9	54.1 [-14.0]	17.4	74.2	5.2
P1+P2	41.9	37.1 [-11.5]	66.9	57.8 [-13.6]	80.2	71.1 [-11.3]	73.4	67.9 [-7.5]	54.0	46.2 [-14.4]	17.0	73.7	5.2
SLG	49.1	43.9 [-10.6]	81.1	72.1 [-11.1]	90.4	81.2 [-10.2]	63.4	58.4 [-7.9]	63.4	55.0 [-13.2]	17.5	73.4	5.2
SLG+KT	48.7	42.6 [-12.5]	71.3	60.8 [-14.7]	83.4	74.4 [-10.8]	74.5	68.2 [-8.5]	59.9	50.3 [-16.0]	17.3	74.6	5.2
TextFooler													
MCA-I	46.6	36.1 [-22.5]	76.3	63.9 [-16.3]	86.6	74.9 [-13.5]	61.5	53.9 [-12.4]	60.0	47.1 [-20.5]	16.8	74.4	19.7
MCA-H	45.9	39.1 [-14.8]	76.8	68.5 [-10.8]	86.8	78.3 [-9.8]	52.2	48.0 [-8.0]	60.0	51.1 [-14.8]	17.1	74.6	19.7
MCA-I-HGQ	50.8	44.2 [-13.0]	81.7	71.6 [-12.4]	90.2	81.2 [-10.0]	60.0	54.4 [-9.3]	64.3	54.8 [-14.8]	17.0	74.4	19.9
MCA-I-VGH	48.6	41.5 [-14.6]	78.7	68.2 [-13.3]	88.6	78.9 [-10.9]	62.6	56.5 [-9.7]	62.2	52.3 [-15.9]	16.5	74.4	19.8
MCA-I-H	50.0	43.1 [-13.8]	81.4	71.2 [-12.5]	90.8	81.3 [-10.5]	59.6	53.7 [-9.9]	63.8	54.0 [-15.4]	16.9	74.7	19.8
RvA	49.9	43.6 [-12.6]	82.2	73.2 [-10.9]	91.1	84.2 [-7.6]	56.3	50.2 [-10.8]	64.2	55.3 [-13.9]	16.9	74.9	19.9
P1	48.8	42.6 [-12.7]	80.2	71.1 [-11.3]	89.7	82.2 [-8.4]	60.0	53.5 [-10.8]	62.9	54.4 [-13.5]	17.3	74.3	20.1
P1+P2	41.9	35.8 [-14.6]	66.9	56.9 [-14.9]	80.2	71.8 [-10.5]	73.4	66.9 [-8.9]	54.0	45.1 [-16.5]	17.1	73.7	19.8
SLG	49.1	43.1 [-12.2]	81.1	73.4 [-9.5]	90.4	82.7 [-8.5]	63.4	57.8 [-8.8]	63.4	55.3 [-12.8]	17.3	74.2	19.9
SLG+KT	48.7	41.6 [-14.6]	71.3	59.7 [-16.3]	83.4	74.9 [-10.2]	74.5	67.6 [-9.3]	59.9	49.8 [-16.9]	17.1	74.6	19.9

Table 9: Comparison of performance before attacking question (Orig.) and after (Aft.) on different VisDial models. In addition to standard metrics, we measure the perturbed word percentage (Pert.), semantic similarity (S.S) and the number of queries (Quer.) to assess BERT-Attack vs. TextFooler. The *relative* performance drop is listed as [Δ]. Highlights indicate the **least robust** and **most robust** model, supplement of Table 1.

	Orig.R@1	Aft.R@1	Orig.R@5	Aft.R@5	Orig.R@10	Aft.R@10	Orig.NDCG	Aft.NDCG	Orig.MRR	Aft.MRR	Pert.	S.S.	Quer.
Random	50.0	46.2	81.4	71.7	90.8	81.4	59.6	56.0	63.8	55.9	17.0	73.4	5.2
Ours		45.2		69.5		80.0		54.6		54.6	16.7	74.8	5.2

Table 10: Effect of vulnerable word attack (full table) on MCA-I-H model with BERT-Attack, supplement of Table 4.

	Orig.R@1	Aft.R@1	Orig.R@5	Aft.R@5	Orig.R@10	Aft.R@10	Orig.NDCG	Aft.NDCG	Orig.MRR	Aft.MRR	Pert.	S.S.	Quer.
All	50.0	43.7	81.4	73.3	90.8	84.3	59.6	54.1	63.8	57.2	16.7	74.4	6.1
Ours		45.2		69.5		80.0		54.6		54.6	16.7	74.8	5.2

Table 11: Effect of stop words set (full table) on MCA-I-H model with BERT-Attack, supplement of Table 5.

ε	Orig.R@1	Aft.R@1	Orig.R@5	Aft.R@5	Orig.R@10	Aft.R@10	Orig.NDCG	Aft.NDCG	Orig.MRR	Aft.MRR	Pert.	S.S.	Quer.
0.7		47.0		69.2		79.4		55.6		54.1	16.1	82.0	5.8
0.5	50.0	45.2	81.4	69.5	90.8	80.0	59.6	54.6	63.8	54.6	16.7	74.8	5.2
0.3		44.6		69.5		79.9		54.1		54.8	16.9	71.8	5.1
0.1		44.6		69.5		80.0		53.9		54.8	17.1	70.9	5.1

Table 12: Effect of semantic similarity threshold ε (full table) on MCA-I-H model with BERT-Attack, supplement of Table 6.

	Orig.R@1	Aft.R@1	Orig.R@5	Aft.R@5	Orig.R@10	Aft.R@10	Orig.NDCG	Aft.NDCG	Orig.MRR	Aft.MRR	Pert.	S.S.	Quer.
Raw Attack		44.2		69.8		80.2		53.7		54.9	17.4	70.3	4.9
+POS		44.5		69.5		80.0		53.8		54.8	17.1	70.3	5.1
+POS+S.S.(0.5)	50.0	45.2	81.4	69.5	90.8	80.0	59.6	54.6	63.8	54.6	16.7	74.8	5.2
+POS+S.S.(0.5)+Gram.		45.4		70.9		81.2		55.9		55.1	13.0	71.4	5.2

Table 13: Effect of different constraints for adversarial attack (full table) on MCA-I-H model with BERT-Attack, supplement of Table 7.

Instructions

We give some examples for the different options.

A - One text means something completely different
 e.g. "Can you see big ben"/ Can you see huge ben? (Entity changes - independent of picture)
 e.g. "Are the planes close to each other" / "Are the planes close to any other?" (Question scope changes)

B - One text means something different
 e.g. "Is the dog/ dogs a Cocker Spaniel?" depends on whether there is more than 1 dog.


C - The meaning is somehow similar but one of texts means something slightly different.
 e.g. "Are any of them stores?"/"Are any of them retailers?" (Similar meaning)

D - They have exactly the same meaning
 e.g. "Does it have color?"/ "Does it have colour?" (Pretty much only applies to BE/ AE spelling?)

Text (and image)

Text 1: Is it night ?

Text 2: Is it evening ?



Question

How similar is the meaning of these two pieces of text ?

Select an option

1 - One text means something completely different.	1
2 - One text means something different, dependent on the context.	2
3 - The meaning is somehow similar but one of texts means something slightly different.	3
4 - They have exactly the same meaning	4

Figure 7: AMT task description and interface to evaluate semantic consistency before and after the attack w/o image.

Instructions

'Fluent - could this have been produced by a native speaker?'

'Grammatical - are there any grammar errors, such as verb agreement?'

Text

Is the blanket cleaned ?

Question

How fluent/grammatical is the text?

Select an option

1 - Not understandable	1
2- hard to understand because of grammar and fluency issues	2
3 - Somewhat hard to understand because of grammar and fluency issues	3
4 - One or two minor errors but still easy to understand	4
5 - Everything is perfect; could have been produced by a native speaker	5

Figure 8: Interface of 'Evaluation of Grammaticality' for AMT task.

Instructions

We give some examples for 'unsure' option.

"Unsure - the question doesn't make sense given the picture." (e.g. question asking about "a man" when there is only a child in the picture.)

"Unsure - I can't verify the answer given the picture." (e.g. question asking whether someone smiles, but it's hard to see.)

"Unsure - the question is difficult to understand because it's ungrammatical" (e.g. the question is highly ungrammatical and disfluent)

"Unsure - the question is ambiguous given the picture." (e.g. the question has more than one answer)

Text (and image)

Question: What colour is the train ?

Answer: Black and red.



Question

Is it a correct/resonable answer for the question given the image?

Select an option

- | | |
|-----------------------------|---|
| 1 - Yes, answer is correct | 1 |
| 2 - No, answer is incorrect | 2 |
| 3 - Unsure | 3 |

Figure 9: Interface of 'Evaluation of Label Consistency' for AMT task.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
section 8
- A2. Did you discuss any potential risks of your work?
Section Ethics Statement
- A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and section 1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
all
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
appendix A
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4.1
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
section 4.1

C Did you run computational experiments?

Section 5

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Appendix B

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

Section 5 and 7

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

Section 5 and appendix B

D Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

No response.

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

No response.

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

No response.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

No response.

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

No response.