Evaluation of the Adversarial Robustness in LLM-based Visual Dialog System

Yahan Yu Fei Cheng Chenhui Chu
Kyoto University
yahan@nlp.ist.i.kyoto-u.ac.jp {feicheng, chu}@i.kyoto-u.ac.jp

Abstract

Large Language Models (LLMs) are widely employed. However, their susceptibility to adversarial attacks poses a significant security concern. In this paper, we focus on LLM-based visual dialog systems and delve into their sensitivity in the aspect of both visual attack and textual attack. Our work aims to investigate the robustness of these systems, and give researchers the understanding of the security challenges that LLMs may face in practical applications.

1 Introduction

In recent years, the field of natural language processing has witnessed a surge in the utilization of Large Language Models (LLMs) [1] and their multi-modal extensions [2, 3, 4], underscoring their pivotal role in various applications. In contrast to non-LLM-based multi-modal systems, such as ViLBERT-based and CLIP-based, the LLM-based approaches often have more powerful abilities for contextual understanding and transfer learning because of their large-scale training. Despite the success of LLM-based multi-modal systems, an escalating concern surrounds the security robustness [5] of these models, particularly in the face of adversarial attacks [6].

Adversarial attacks [7], characterized by purposeful manipulations of input data, exploit the inherent vulnerabilities in model architectures, posing formidable challenges to the reliability and security of LLMs. In multi-modal scenarios, the system may exhibit varying sensitivities to inputs from different modalities. Attackers can exploit this by selectively targeting the most sensitive modality during adversarial attacks to achieve more effective manipulation. It is necessary to consider the adversarial robustness of the system on multiple modalities simultaneously to obtain a comprehensive evaluation.

In response to this security risk and insufficient research in LLM-based multi-modal systems, our work directs its focus toward the specific domain of LLM-based visual dialog systems [8]. Given that Chatbot [9] represents a foundational function of the application of LLMs, our work aims to evaluate their robustness boundaries. To emulate real-world scenarios, we introduce a zero-shot setting system tailored for the visual dialog task. This system serves to evaluate the robustness of LLM-based multi-modal systems against adversarial attacks, incorporating assessments in both text and visual modalities. Specifically, we scrutinize robustness through Fast Gradient Sign Method (FGSM) attacks [7], introducing adversarial noise to images, and coreference attacks [10], manipulating the textual input.

Our contributions are summarised as below:

1. To reflect the real robustness of LLMs, we construct the LLMs-based visual dialog system in zero-shot settings.
2. Our successful execution of FGSM and coreference attacks sheds light on the nuanced vulnerabilities inherent in LLM-based visual dialog systems.

2 Related Work

2.1 Visual Dialog Task

Visual dialog [8] has been introduced as an extended task of Visual Question Answering (VQA) [11]. In Visual dialog, the system is tasked with responding to a sequence of interconnected questions, leveraging both an image and a dialog history. Prior works have explored attention mechanisms [12] that account for the intricate interactions among modalities of the image, dialog history, and question. Some investigations [13] have delved into uncovering the semantic structures within the dialog through graph neural net-

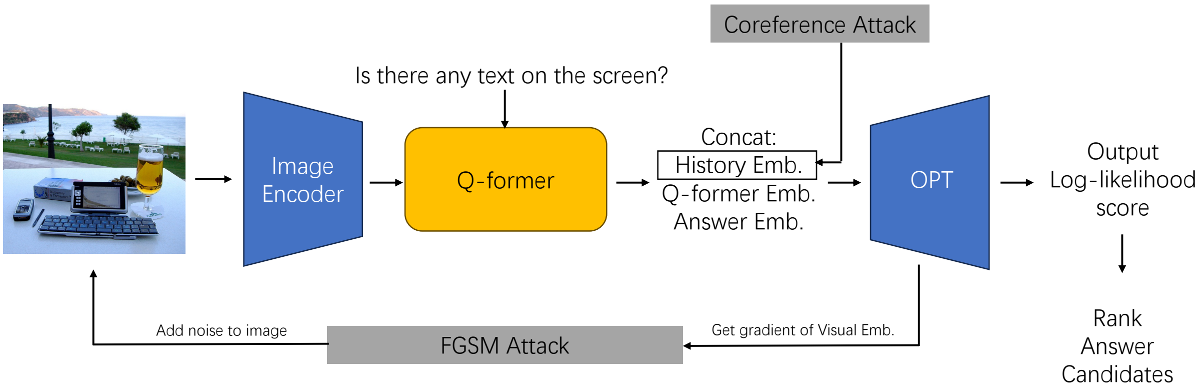

Figure 1 The proposed structure of LLM-based zero-shot visual dialog system. We evaluate its adversarial robustness against FGSM attack and coreference attack. Example of coreference attack is shown in Table 1.

works. Our work takes into account the excellent zero-shot generalization capability of LLMs, and leverages this capability to facilitate interactions and inference across modalities within the visual dialog system.

2.2 Adversarial Attacks on Multi-modal Models

In the domain of multi-modal models, a few studies have recently delved into examining their susceptibility to adversarial attacks. In [14], this work attacked textual inputs using the methods of BERT-Attack and TextFooler, providing the first investigation into the robustness of visual dialog models against textual attacks. In [15], this work demonstrated that imperceptible attacks on images, altering the output of an image caption model, can be exploited by malicious content providers to harm users. In [10], both image and text-level attacks were imposed on a ViLBERT-based visual dialog system, evaluating the adversarial robustness across multiple modalities. Our evaluation builds on these insights, concentrating on LLM-based multi-modal models and demonstrating the effectiveness of the attacks on the visual dialog task.

3 Task Definition

For the visual dialog task, we define the input image as V, the dialog history as $H = \{H_1, H_2, ..., H_t\}$ where t means dialog turns and $H_t = \{Q_t, A_t\}$ means the question Q_t and answer A_t in one turn, and the final question as Q_0. The objective of this task is to enable the model to select the appropriate answer Ans from a set containing N candidate answers $\{Ans_1, Ans_2, ..., Ans_N\}$ based on the input information.

As for the attacks, we define the visual attack as $\mathrm{VA}(\cdot)$ and the textual attack as $\mathrm{TA}(\cdot)$. After attacking, we input the perturbed image $V' = \mathrm{VA}(V)$ or the perturbed dialog history $H' = \mathrm{TA}(H)$ into the model, and expected the answer Ans' should be different from the ground truth answer.

4 Method

4.1 LLM-based Zero-shot Visual Dialog Sysytem

The proposed structure is shown in Figure 1.

Selection for LLM Considering the advanced performance and open-source implementation, we choose BLIP-2 [16] as the base model in the zero-shot visual dialog system, which is a recent multi-modal model that gained significant attention. BLIP-2 seamlessly integrates vision and language understanding by combining a pre-trained visual model with a LLM. Consequently, it possesses the capability to handle both visual and textual inputs, leading to the generation of coherent natural language outputs.

Zero-shot Setting We construct a zero-shot system for the evaluation of adversarial robustness. The zero-shot setting aims to closely mimic real-world application scenarios, where models may encounter previously unseen categories or tasks. This configuration better simulates the robustness observed in practical use, providing a more realistic assessment of the model's performance.

Scoring of Answers To select the appropriate answer from the candidate set, we calculate the log-likelihood score [8] of all the candidates because our visual dialog system contains a generative decoder. In the evaluation phase, the rank of the candidate set is given by their log-likelihood scores.

4.2 Adversarial Visual Attack

We make the assumption that an attacker possesses the capability to introduce minor perturbations to the visual inputs of the model. Additionally, we assume that the attacker has unrestricted access to all model weights.

In this setting, the Fast Gradient Sign Method [7] is an attack method that introduces perturbations to visual inputs guided by the gradients of the loss concerning the visual inputs:

$$\mathrm{FGSM}(V) = V + \epsilon \cdot \mathrm{sign}(\nabla_V Loss(V, label)), \quad (1)$$

where V and $label$ represent the visual inputs and their corresponding ground-truth labels, respectively. $\nabla Loss(\cdot)$ means the gradient of the model. The hyperparameter ϵ is utilized to modulate the intensity of perturbations. And the $\mathrm{sign}(\cdot)$ is a mathematical function that returns the sign of a real number. Specifically, it maps a positive number to 1, a negative number to -1, and zero to 0. The $\mathrm{sign}(\cdot)$ function is denoted as follows:

$$\mathrm{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases} \quad (2)$$

However, in contrast to the aforementioned configuration, each question in the visual dialog task may have one or more pertinent answers within the list of candidate answers. This is because some candidate answers may be semantically identical (e.g. 'yeah' and 'yes'). So the validation dataset of visual dialog task is annotated with the relevance score [8] between each of the candidate answers and ground truth, which ranges from 0 to 1. Therefore, we modify the FGSM attack as follows:

$$\mathrm{FGSM}(V) = V + \epsilon \cdot \mathrm{sign}(\sum_{n=1}^{N} Rel_n \cdot \nabla_V Loss((V, H, Q), Ans_n)), \quad (3)$$

where Ans_n means the $n-th$ candidate answer, Rel_n means the relevent score between ground truth answer and the $n-th$ candidate answer. Equation signifies that the gradients of the loss with respect to all relevant answers are taken into account in the FGSM attack.

Table 1 Example of adversarial textual attack. The red words indicate that the target word for the attack is *adults* which is replaced with *grownup* after the attack.

Original Text	Attacked Text
Q: what is in the background?	Q: what is in the background?
A: trees and buildings	A: trees and buildings
Q: how many are adults?	Q: how many are grownup?
A: 1 adults	A: 1 grownup

4.3 Adversarial Textual Attack

We also investigate adversarial robustness against textual attacks. As shown in Table 1, we employ the coreference attack [10], wherein noun phrases in the dialog history are replaced with their synonyms to deceive the models. Coreference means that the target words and the replacement words not only have semantic similarity but also refer to the same object. In this method, we utilize an off-the-shelf neural coreference resolution tool [17] to identify words in the dialog history referring to objects mentioned in a given question. We perform a greedy substitution of words with their synonyms, selecting those with the minimum cosine distance in the counter-fitting word embedding space [18]. After going through these steps, we obtained the attacked text samples.

5 Experiments

5.1 Dataset

VisDial v1.0 [8] is a version of the Visual Dialog dataset, which is widely used in the visual dialog task. It consists of a collection of dialogues between humans discussing images. The dataset is designed to facilitate research on the intersection of computer vision and natural language processing, particularly in tasks involving dialog understanding and visual reasoning. The VisDial v1.0 dataset contains 123k, 2k, and 8k dialogs as train, validation, and test split. In our zero-shot setting, we only use the validation set.

5.2 Implementation

As for the BLIP-2 model, in our work, we opted for CLIP (ViT-L/14) [19] as the Image Encoder and the decoder-based OPT model (2.7B and 6.7B) [20] as the LLM.

Table 2 Results on LLM-based zero-shot visual dialog system without attacks. The ViLBERT-based, CLIP-based and FROMAGe-based approaches are the existing methods, while two BLIP-2-based approaches are our baselines. We emphasize the current state-of-the-art results.

	w/o attacks				
Base Model	NDCG	MRR	R@1	R@5	R@10
ViLBERT [21]	11.6	6.9	2.6	7.2	11.3
CLIP (Vit-L/14) [19]	10.9	8.5	3.1	8.7	15.9
FROMAGe [22]	**16.5**	**22.0**	**17.6**	**20.1**	**25.1**
BLIP-2 (OPT2.7B)	13.9	20.8	15.9	18.4	23.2
BLIP-2 (OPT6.7B)	14.2	20.9	16.2	18.6	23.8

5.3 Metrics

We adhere to the standardized evaluation protocol introduced in [8]. Generative tasks for visual dialog models are appraised using retrieval-based evaluation metrics, including normalized discounted cumulative gain (NDCG), mean reciprocal rank (MRR) and recall@k (R@k, k = 1, 5, 10). Each dialogue comprises a list of 100 answer candidates for every question, with one ground-truth answer included in the candidates. The model arranges the answer candidates based on log-likelihood scores and is subsequently evaluated using the aforementioned three metrics. NDCG considers all relevant answers from the candidate set, while MRR and R@k take into account the rank of the single ground-truth answer. So, NDCG is considered the primary evaluation metric in current works.

5.4 Results on LLM-based Zero-shot Visual Dialog System

Table 2 shows the results on our proposed system without adversarial attacks in contrast with other zero-shot visual dialog systems. In details, BLIP-2 has 188M trainable parameters in the pretraining stage, while ViLBERT has 114M, CLIP has 300M and FROMAGe has 5.5M. The performance and trainable parameters of BLIP-2-based model are close to the existing zero-shot visual dialog models, which can illustrate that the structure of BLIP-2-based models is reasonable, and can be used in the attack process.

5.5 Results on Adversarial Visual Attack

Table 3 shows the results on LLM-based zero-shot visual dialog system with the visual attack. Hyperparameter ϵ adjusts the intensity of perturbations. As the ϵ rises, the perturbation applied to the image becomes stronger and NDCG drops, which means that the FGSM attack is successful on the BLIP-2-based model. In the ViLBERT-based visual dialog model [10], FGSM usually causes a drop by 30%. In our work, the slight drop in NDCG can be attributed to the robustness of BLIP-2 because of the pretraining on multiple vision datasets such as LAION [23].

Table 3 Results on LLM-based zero-shot visual dialog system with adversarial visual attack. We adjust the perturbation strength by varying the hyperparameter ϵ, observing changes in the model's performance.

	NDCG			
Base Model	$\epsilon = 0$	$\epsilon = 0.01$	$\epsilon = 0.03$	$\epsilon = 0.05$
BLIP-2 (OPT2.7b)	13.9	13.8	13.8	13.8
Model	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$
BLIP-2 (OPT2.7b)	13.8	13.8	13.7	13.7
Model	$\epsilon = 0.5$	$\epsilon = 0.6$	$\epsilon = 0.7$	
BLIP-2 (OPT2.7b)	13.7	13.6	13.5	

Table 4 Results on LLM-based zero-shot visual dialog system with adversarial textual attack.

	NDCG	
Base Model	w/o attacks	w/ textual attack
BLIP-2 (OPT2.7b)	13.9	16.8

5.6 Results on Adversarial Textual Attack

Table 4 shows the results on LLM-based zero-shot visual dialog system with the coreference attack. In the ViLBERT-based visual dialog model [10], NDCG usually drops by 4%. But in our work, NDCG improves. We hypothesize that, based on the language understanding and reasoning capabilities of LLMs, the substitution of synonymous words in the text contributes to the semantic completion of the dialog history. This indicates that LLM-based models are robust against our coreference attack.

6 Conclusion

In conclusion, our study underscores the widespread utilization of LLMs while highlighting a critical security concern due to their vulnerability to adversarial attacks. The identified weaknesses in security extend not only to LLMs but also to LLM-based multi-modal models. Specifically focusing on LLM-based visual dialog systems, our research finds their sensitivity to visual attacks, and the opposite impact to textual attacks. Future work could involve the design of more advanced adversarial attacks to attain stronger visual attacks and more effective textual attacks. By developing novel adversarial strategies, researchers can gain a deeper understanding of the security boundary in LLM-based visual dialog systems, contributing to robust defense mechanisms.

Acknowledgements

This work was supported by 京都大学大学院情報学研究科 博士後期課程の学生に対する支援事業.

References

[1] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. **arXiv preprint arXiv:2304.08485**, 2023.

[3] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. **arXiv preprint arXiv:2305.18565**, 2023.

[4] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. **arXiv preprint arXiv:2306.17107**, 2023.

[5] Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Virtual prompt injection for instruction-tuned large language models. **arXiv preprint arXiv:2307.16888**, 2023.

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

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

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In **Proceedings of the IEEE conference on computer vision and pattern recognition**, pp. 326–335, 2017.

[9] Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. Data decentralisation of llm-based chatbot systems in chronic disease self-management. In **Proceedings of the 2023 ACM Conference on Information Technology for Social Good**, pp. 205–212, 2023.

[10] Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The dialog must go on: Improving visual dialog via generative self-training. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 6746–6756, 2023.

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

[12] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 2039–2048, 2019.

[13] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 6669–6678, 2019.

[14] Lu Yu and Verena Rieser. Adversarial textual robustness of visual dialog. In **61st Annual Meeting of the Association for Computational Linguistics 2023**, pp. 3422–3438. Association for Computational Linguistics, 2023.

[15] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 3677–3685, 2023.

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

[17] Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coreference models. **arXiv preprint arXiv:1609.08667**, 2016.

[18] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 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**, Vol. 34, pp. 8018–8025, 2020.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning**, pp. 8748–8763. PMLR, 2021.

[20] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. **arXiv preprint arXiv:2205.01068**, 2022.

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. **Advances in neural information processing systems**, Vol. 32, , 2019.

[22] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In **International Conference on Machine Learning**, pp. 17283–17300. PMLR, 2023.

[23] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 25278–25294, 2022.