An Unified Hallucination Mitigation Framework for Large Vision-Language Models

Anonymous ACL submission

Abstract

Hallucination is a common problem for Large Vision-Language Models (LVLMs) with long generations which is difficult to eradicate. The generation with hallucinations is partially inconsistent with the image content. To mitigate hallucination, current studies either focus on the process of model inference or the results of model generation, but the solutions they design sometimes do not deal appropriately with various types of queries and the hallucinations of the generations about these queries. To accurately deal with various hallucinations, we present a zero-shot unified framework, **Dentist**, for hallucination mitigation. The core step is to first classify the queries, then perform different processes of hallucination mitigation based on the classification result, just like a dentist first observes the teeth and then makes a plan. In a simple deployment, Dentist can classify queries as perception or reasoning and easily mitigate hallucinations in answers which has been demonstrated in our experiments. On MMbench, we achieved a 13.44%/10.2%/15.8% improvement on Image Quality, a Coarse Perception visual question answering (VQA) task, over the baseline InstructBLIP/LLaVA/VisualGLM. Our source code will be released on GitHub.

1 Introduction

Hallucination in Large Vision-Language Models (LVLMs) is a critical issue [1], which manifests as the model's generated content partially deviating from the actual content of the image. For example, LVLMs inaccurately identify characters' actions and misinterpret relationships between characters, as illustrated in Fig. 1. Such inaccuracies can lead to misinformation, potentially degrading the user experience and misleading individuals. This issue underscores the necessity for ongoing improvements to enhance the reliability and accuracy of LVLMs, mitigating the risk of hallucinations and their consequent misinformation.

To tackle the above challenge, existing work either focuses on optimizing the training data and the parameters of the existing model, or correcting the hallucinations during the generation stage without model update. The former collects high-quality training data, such as adding negative instances to the training data to avoid overconfidence in the model. The latter mainly utilize the information from the model generation to design methods to detect hallucinations and eliminate them. For example, Woodpecker [2] extracts main objects from the generated response from LVLMs and get their bounding boxes. The hallucinations in the response are corrected with the bounding boxes as evidence. Similarly, HalluciDoctor [3] makes the description-oriented answer chunks extraction and formulates corresponding questions, uses answers for these questions which are gathered from various MLLMs to do the consistency cross-checking and remove hallucinations. However, the above methods have a common problem, that is, when faced with diverse model outputs, a set of verification methods may sometimes be inappropriate and ineffective. For example, the effect of a method that uses object

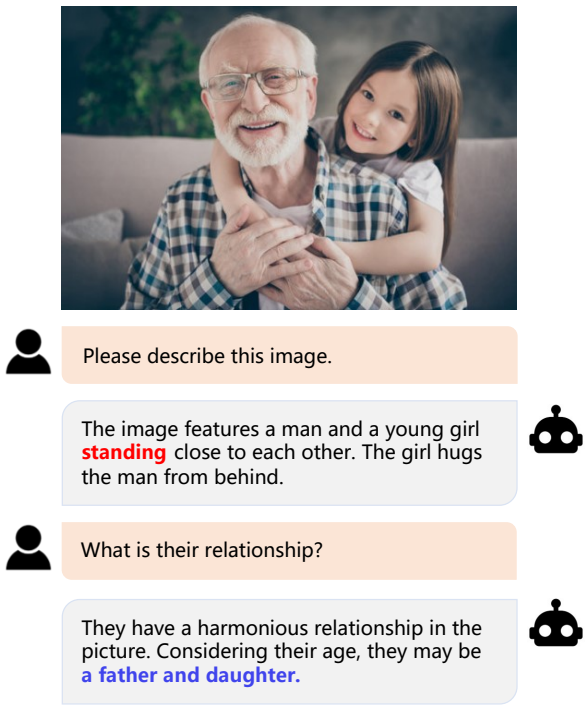

Figure 1: An example image of hallucination. The generation of the model is partially inconsistent with the image, which we call perception hallucination and reasoning hallucination respectively.

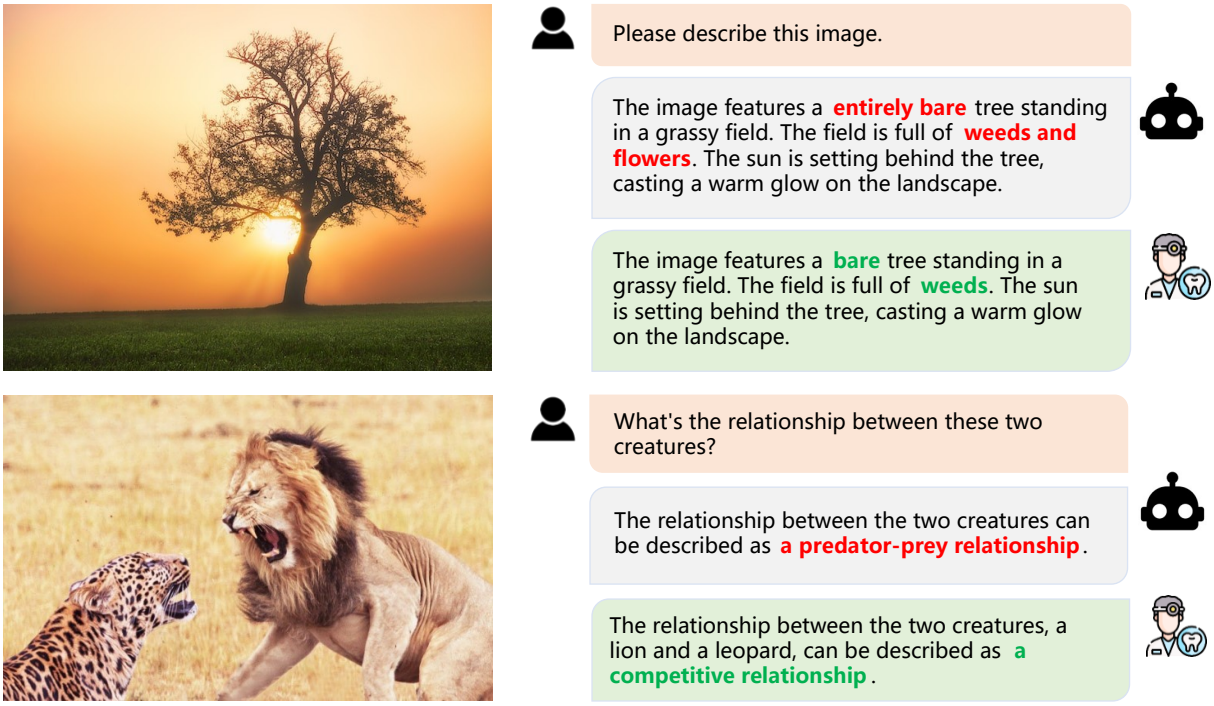

Figure 2: Example images of our hallucination mitigation. The part of the generation that conflicts with the content of the picture has been corrected.

detection on pictures to verify whether the object in the answer exists will be reduced when the query is the reasoning type.

In order to solve this problem, we propose a zero-shot unified framework for all kinds of hallucination mitigation. As shown in Fig. 2, regardless of whether it is a descriptive answer or a logical reasoning answer, any part of the answer that does not match the content of the picture will be corrected by our framework Dentist. Specifically, the framework we proposed is a verification cycle, and each cycle is divided into two core steps: (1) Query classification divides the query into two categories: perception and reasoning. As shown in Fig. 1, We abstract these two situations into two hallucination categories, namely perception hallucination and reasoning hallucination, corresponds to the two categories of queries. The former is manifested by incorrectly describing image content in model generation, such as errors when describing object attributes, while the latter refers to the model producing fallacies in logical reasoning answers. (2) Differential processing makes the mitigation based on the classification. The generation for the perception query will be verified by the sub-questions, while the generation for the reasoning query will be verified with the help of Chain-of-Thought (CoT). To ensure that hallucinations are mitigated as much as possible, the above verification cycle will loop until the revised generation no longer changes significantly or the loop limit is reached.

We completed quantitative experiments on MM-bench [4] and LLaVA-QA90 [5] using three models: InstructBLIP [6], VisualGLM [7, 8] and LLaVA [5], respectively, to test the effectiveness of our proposed method. Our method demonstrates the effectiveness and superiority in many visual language tasks, and promotes the performance of the baseline models. In particular, in the experiment, we achieve a 13.44%/10.2%/15.8% improvement in the visual language task of Image Quality, compared with the baseline model InstructBLIP/LLaVA/VisualGLM.

Our main contributions are summarized as follows: (1) We propose a zero-shot unified framework called Dentist for hallucination classification and mitigation. To the best of our knowledge, we are the first to distinguish treatment based on classification of hallucinations and moreover use a validation cycle for complete removal of hallucinations. (2) Our unified framework is easily integrated into various LVLMs. The clear design of the framework also provides convenience for new classifications and treatments to access the framework. (3) We comprehensively evaluated our method on MM-bench and LLaVA-QA90 with a detailed superior-

ity analysis.

2 Related Work

2.1 Large Vision-Language Model

Inspired by the success of Large Language Models (LLMs) in zero-shot / few-shot learning [9, 10], the multimodal learning community shifted research attention to LVLMs. LVLMs mainly use the cross-modality aligner to connect the visual encoder (such as CLIP [11]) and LLMs (such as LLaMA [12]) to tackle vision-language tasks. For example, LLaVA [5] connects a vision encoder and a LLM for general-purpose visual and language understanding, suggesting practical tips for building a general-purpose visual agent. Meanwhile, InstructBLIP [6] introduce an instruction-aware Query Transformer which extracts visual features from the output embeddings of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM. In addition, VisualGLM [7, 8] use Qformer [13] which builds a bridge between the visual model and the language model. Though these above LVLMs have powerful visual language understanding ability on the generation task, sometimes their outputs still contain hallucinations that need to be corrected.

2.2 Hallucination

With the progress of research on LVLMs, the problem of hallucination has gradually been exposed, and it has attracted more and more attention. Research around hallucination focuses on three aspects: detecting [14, 15], mitigating [16, 17, 18, 2, 3], and evaluating hallucinations [1, 19]. In this paper, we mainly focus on hallucination mitigation. Previous works on hallucination mitigation can be divided into two categories: model inference optimization and model generation verification. The first category focuses on the process of the training and inference of the LVLMs. Ever [16] points out that mitigating hallucinations in real time during model inference is more appropriate than generating corrections from the model outputs, as the latter is subject to snowballing effects. VIGC [18] uses an iterative method to concatenate the short sentences generated each time, and ensures accuracy by controlling the length of the generation. On the other hand, the second category focuses on the aspect of the generation of LVLMs, designing methods to obtain hallucination information from the output of the model and do the mitigation. For example, Woodpecker [2] makes the question formulation and visual knowledge validation base on the keywords which are extracted from the output of the model and hires an LLM to modify the hallucinations in the generated responses.

Despite the success of the existing method, they overlook the diversity of hallucinations which results in a fixed hallucination elimination method that cannot be applied to all hallucination situations well. To solve this problem, we propose a zero-shot unified framework for mitigating hallucinations, the core step of which is to classify hallucinations caused by different queries.

3 Method

Existing hallucination elimination methods cannot handle all types of queries well. In order to solve this problem, our objective is to propose a zero-shot unified framework for hallucination mitigation. The difficulty lies in how to classify various queries and provide appropriate processing methods for the hallucinations caused by different types of queries. We divide the entire framework into three major sections: query classification(Sec. 3.1), differential processing(Sec. 3.2), and Validation Cycle (Sec. 3.3). The differential processing section is further divided into perception processing(Sec. 3.2.1) and reasoning processing(Sec. 3.2.2).

Fig. 3 is the overall block diagram including the above sections. Later in each section we will elaborate on our method.

3.1 Query Classification

Since we are committed to mitigating the hallucinations obtained by various types of queries in a targeted manner, we need a suitable classification standard to classify queries into several major categories and then handle them differently. Considering that when classifying queries, we need to simplify the classification as much as possible while retaining the fine-grainedness of the classification, we reviewed a large number of LVLMs benchmark papers. In the end, we think it is appropriate to divide the query into two categories: perception and reasoning [4]. This classification method covers almost all queries and is sufficiently fine-grained while having few types. The perception query feature mainly requires the model to have the ability to perceive visual features, such as attribute recognition, scene description, etc. The hallucinations

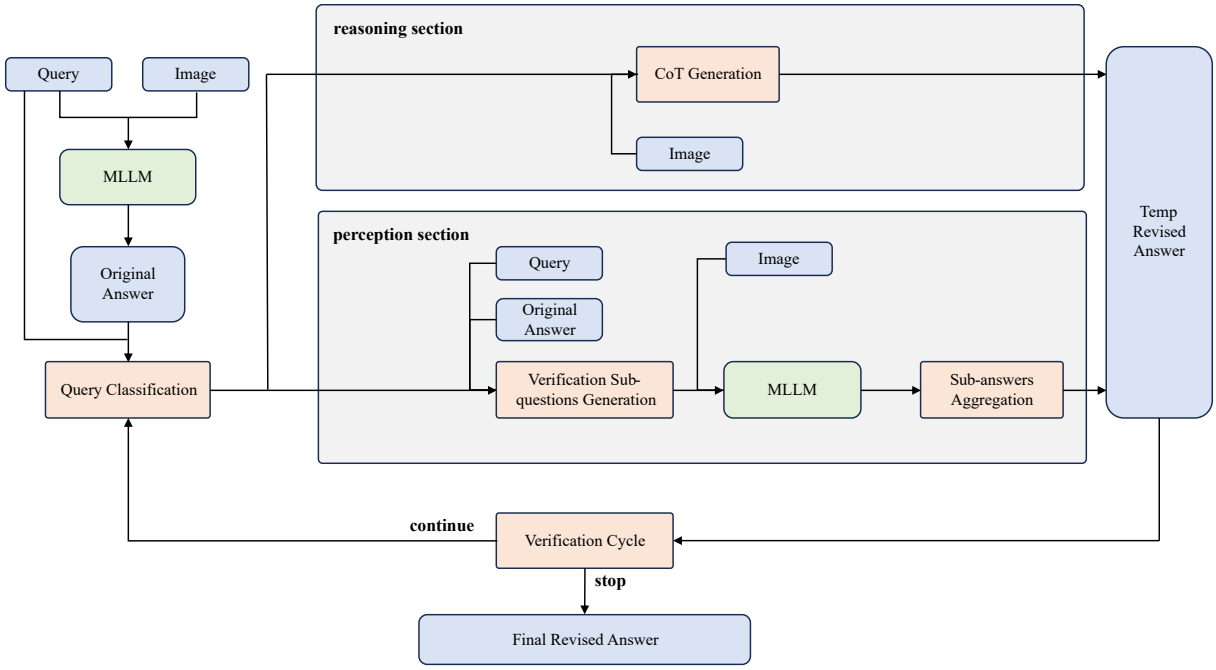

Figure 3: An overview of the proposed method. The core point is to customize different methods of mitigating hallucinations by classifying the query. The reasoning section is used to mitigate the hallucinations caused by reasoning queries, while the perception section is used to mitigate the hallucinations caused by descriptive queries.

caused by this type of queries can be effectively alleviated through visual methods. The reasoning query mainly tests the understanding and reasoning ability of the model and it's hallucinations should be mitigating by other appropriate method such as CoT.

We employ ChatGPT to complete query classification through the prompt. The prompt template used for classification is shown in Appendix A.1.

3.2 Differential Processing

After query classification, the model generation for perception or reasoning queries need to be processed differently. To deal with the perception query, we need to generate verification sub-questions based on the original query and the original answer with hallucinations generated by the LVLMs. The model answers these sub-questions to obtain sub-answers, and finally aggregates this answers to form the output with less hallucinations. Refer to section Sec. 3.2.1 for specific steps.

For reasoning query, model output does not necessarily give the inference process, and it is difficult to deal with the reasoning problem by raising verification sub-problems like what we have done to perception query. In response to this situation, the method we propose is to use the CoT. Refer to section Sec. 3.2.2 for specific steps.

3.2.1 Perception Processing

LVLMs is prone to hallucinations when generating long descriptive texts [20]. This exactly corresponds to the situation of perception queries. We were inspired that when the long answer to perception queries contains hallucinations, we can split the long answer into short sentences and design verification sub-questions based on the key points in the sentences. After generating verification sub-questions, these sub-questions are fed to LVLMs along with the original image, and the verification sub-answers are obtained from LVLMs.

It is worth mentioning that the LVLMs we used for generating the sub-answers are just the original model which has hallucinations need to be revise. It can probably be replaced with any visual question answering (VQA) model, but would be accompanied by the suspicion of using a better model for better work. To demonstrate the ability of our approach to mitigate hallucinations rather than the ability of the rectified models, we chose to use the original LVLMs.

The final step is to aggregate the sub-answers and correct the errors caused by hallucinations. For the original answer and the verification sub-answers, we consider the latter to be the ground truth. So the way to aggregate answers is to find the contradiction between the two answers and replace

them with the corresponding contents of ground-truth sub-answers.

In order to automatically implement the above two steps, we use prompt to hire ChatGPT to complete the process of generating sub-questions and aggregating sub-answers. Refer to Appendix A.2 for the prompt template used for generating sub-questions and Appendix A.3 for the prompt template used for aggregating sub-answers.

3.2.2 Reasoning Processing

The answers generated by the model from reasoning queries are not as straightforward as the answers to perception questions. The details of visual perception and logical reasoning based on the perception are easily hidden and only the reasoning results are output. Therefore, the method of the perception section is no longer applicable. In order to solve this problem, we use the CoT prompt method to obtain the answer which contains more reasoning details. Add "Let's think step by step" to the start of the original query to do the CoT and use ChatGPT by prompt to obtain the revised answer. Refer to Appendix A.4 for the prompt template.

3.3 Validation Cycle

After the above steps, we have obtained the preliminary verified answer which may still contain hallucinations that have not been eliminated because of the imperfections of the verification sub-questions generation. In order to solve this problem, we propose to regard the entire verification framework as a repeated block in the verification cycle chain. The verified answer is treated as the original answer and re-verified. The difficulty of loop verification is how to judge when the hallucinations in the answer has been completely removed so the loop can be stopped. We believe that if and only if the verified answer does not change significantly after a new round of verification, it means that all the hallucinations that can be eliminated have been eliminated. On the other hand, if the answer still changes significantly after a specific number of rounds of verification, we believe that there is a snowball error phenomenon in the verification cycle. We will stop the loop and use only the answer from the first verification as the final revised answer. Regarding automatically determining whether the answer is no longer changing, we use a prompt to make ChatGPT automatically complete it instead of manual work. Refer to Appendix A.5 for the prompt template.

4 Experiment

In this section, we present the experiment details, including experiment settings and result analysis.

4.1 Experiment Settings

Dataset. **MMBench** is a novel multi-modality benchmark, which develops a fine-grained ability assessment for LVLMs. The MMBench evaluation standard is divided into three levels. The L-1 ability dimension incorporates Perception and Reasoning, L-2 ability dimension consists of Coarse Perception, Fine-grained, etc. and L-3 ability dimension covers Image Style, Image Scene, Image Emotion, etc. We conduct experiments under the setting of the L-3 level abilities, which achieve the most fine-grained evaluation.

LLaVA-QA90 is also a dataset used to evaluate LVLMs. LLaVA-QA90 contains 90 VQA tasks and 30 images taken from COCO Val 2014 [21]. To evaluate the generated response, we feed the query, image, and model response to GPT-4V [22] to get a score of a scale of 1 to 10. The prompt template is available in Appendix A.7

Baselines. We select 3 currently mainstream LVLMs as our baseline models, including InstructBLIP, LLaVA, and VisualGLM.

Implementation Details. We utilize GPT-3.5-turbo [23] to assist in keyword extraction, sub-question generation, validation cycle, and verification answer integration. Experiments have proven that GPT-3.5-turbo can tackle these tasks. On MMBench, we set the experiment rounds to 2: (1) In the first round of evaluation, we have the model generate raw predictions according to MMBench's evaluation rules and submit them to MMBench's official platform to obtain various accuracy rates; (2) In the second round of evaluation, based on the original prediction of the model, query classification, different verification processes and answer integration are carried out using GPT-3.5-turbo (specific details can be found in Section 3). Similarly, we upload the results of the second round of evaluation to the official MMBench platform to obtain various accuracy rates; (3) Finally, we jointly analyze the results of two rounds of evaluation to demonstrate the effectiveness and superiority of Dentist .

4.2 Experiment Results

Results on MMBench. The results on MMBench are summarized in Tab. 1. From this table, we have several observations. (1) The largest accuracy im-

LVLM / L-3 Ability	InstructBLIP-7B		LLaVA-V1.5-7B		VisualGLM-6B	
	Baseline	Dentist	Baseline	Dentist	Baseline	Dentist
Action Recognition	58.45%	57.45%	87.5%	85.2%	35.2%	**38.6%**
Attribute Comparison	2.56%	2.56%	21.2%	**25.8%**	8.8%	**10.8%**
Attribute Recognition	46.46%	**51.41%**	66.7%	**70.6%**	40.0%	**43.7%**
Celebrity Recognition	40.79%	**49.15%**	60.2%	**68.6%**	52.5%	**55.2%**
Function Reasoning	46.22%	**49.62%**	74.5%	**77.8%**	44.9%	**50.6%**
Future Prediction	46.67%	**55.0%**	43.1%	**52.1%**	21.6%	**31.0%**
Identity Reasoning	68.29%	68.29%	86.6%	86.6%	81.7%	**88.4%**
Image Emotion	32.7%	**41.31%**	67.5%	**76.4%**	41.7%	**50.6%**
Image Quality	0.00%	**10.58%**	0.00%	**10.2%**	0.00%	**15.6%**
Image Scene	58.13%	**60.31%**	85.3%	**88.3%**	68.5%	**70.3%**
Image Style	38.82%	37.05%	58.8%	55.3%	30.6%	**35.8%**
Image Topic	60.0%	60.0%	97.6%	96.4%	52.9%	50.2%
Nature Relation	22.22%	**27.28%**	38.3%	38.3%	24.7%	**30.6%**
OCR	51.90%	**58.57%**	70.1%	**78.3%**	41.6%	**43.6%**
Object Localization	3.90%	**14.42%**	16.3%	11.54%	8.6%	**10.9%**
Physical Property Reasoning	21.0%	**21.9%**	55.0%	50.0%	26.0%	**30.3%**
Physical Relation	11.30%	**17.3%**	28.85%	28.85%	3.80%	**9.60%**
Social Relation	27.58%	**41.02%**	62.8%	**69.6%**	46.2%	45.3%
Spatial Relationship	11.11%	8.64%	11.11%	**15.33%**	7.30%	**10.9%**
Structuralized Image-Text Understanding	5.94%	**6.97%**	11.89%	10.9%	3.9%	**5.0%**
Overall	33.9%	**36.9%**	51.0%	**54.8%**	32.0%	**36.35%**

Table 1: Results on MMBench. Dentist denotes the performance after the hallucination is corrected by our verification method. The performance is measured by accuracy, where the better performance for each partition is highlighted in bold.

provement among the three LVLMs exceeds 15.6%, showing that Dentist have excellent correction effects, making obvious improvements in various metrics for the baselines. (2) Dentist performs outstandingly in Image Emotion, Image Quality, Future Prediction, Attribute Recognition, etc., which indicates that Dentist is capable of mitigating hallucination in coarse perception, fine-grained perception and logic reasoning. (3) Among all metrics, Image Quality shows the highest improvement, which indicates that dentist is particularly effective for hallucinations in such problems.

Results on LLaVA-QA90. If manual verification is required, the evaluation on LLaVA-QA90 is labor-intensive and somewhat subjective. Therefore, it is necessary to use a powerful evaluation tool to ensure consistency in evaluation standards while also possessing strong visual language task answering and instruction following abilities. Therefore, we consider utilizing the powerful LVLM, GPT-4V. Specifically, we involve GPT-4V in scoring and evaluating model responses by setting appropriate prompt words. We have designed the following three metrics:

- Accuracy: how accurate is the model response about the image content.
- Detail description: level of details of the responses.
- Complex Reasoning: whether the reasoning content of response is reasonable.

Tab. 2 shows the results. Obviously, equipped with our verification method, the models' performance has been comprehensively improved across the three metrics. On average, there is an improvement of over 0.5 points (relative improvement exceeding 13.6%). This indicates that Dentist not only improves the accuracy of LVLMs in describing image content, but also promotes the detail of image description and the rationality of inference content.

LVLM		Accuracy	Detail description	Complex reasoning
InstructBLIP	Baseline	6.5	4.9	4.3
	Dentist	**7.0** (+0.5)	**5.5** (+0.6)	**4.8** (+0.5)
LLaVA	Baseline	6.0	5.3	4.4
	Dentist	**6.6** (+0.6)	**5.8** (+0.5)	**5.0** (+0.6)
VisualGLM	Baseline	5.6	5.0	4.0
	Dentist	**6.2** (+0.6)	**5.8** (+0.8)	**4.7** (+0.7)

Table 2: Results on LLaVA-QA90. The accuracy detail description and complex reasoning metrics are on a scale of 10, and a higher score indicates the better performance.The better performance for each partition is highlighted in bold.

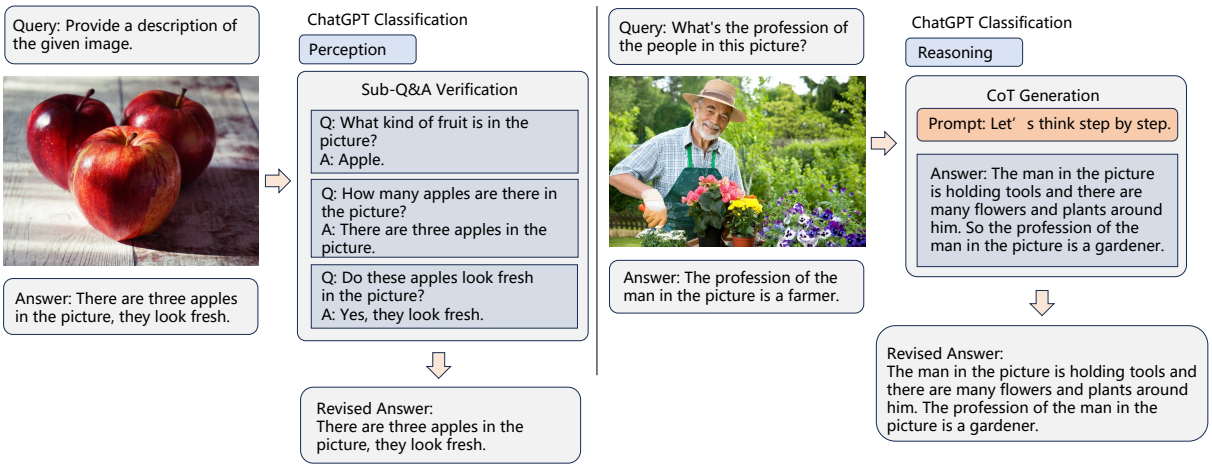

Figure 4: Examples of verification.

4.3 Ablation Studies

To explore the effect of the query classification and validation cycle, we conduct ablation studies in this section.

Variants	Perception Accuracy	Reasoning Accuracy
Baseline	33.74%	31.15%
Dentist	37.62%	35.92%
Dentist/N	34.86%	32.73%
Dentist/P	38.94%	25.48%
Dentist/R	28.34%	38.44%

Table 3: Results on MMBench with different variants of InstructBLIP. For more comprehensive evaluation results on LLaVA and VisualGLM, please refer to the Appendix A.8

Query Classification. We study three different variants and evaluate their performance on MMBench. (1) **Dentist /N**: we disable the query classification module of Dentist ; (2) **Dentist /P**: we classify all queries into perception for verification; (3) **Dentist /R**: we classify all queries into reasoning for verification;

Tab. 3 shows the results of InstructBLIP. We can see that: (1) If query classification is not performed and verification is performed directly (Dentist /N), the accuracy is not much higher than the baseline, and in some cases there is even a problem of reduced accuracy. Because at this point, the way Dentist corrects the model's answers is completely random, which largely depends on the performance and habits of GPT-3.5-turbo: it can be seen that the perception accuracy may not differ much from the baseline, or slightly higher than the baseline, while the reasoning accuracy may decrease. This is because the query classification module tends to treat the problem as perception for processing. (2) If all queries are classified into perception (Dentist /P), it can be seen that the perception accuracy is greatly improved, while the reasoning accuracy is greatly attenuated. This is because Dentist also verifies the reasoning problem as perception, so the verification method is not appropriate, resulting in a decrease in accuracy; (3) In the same way, if all problems are classified as reasoning (Dentist

/R), the reasoning accuracy is greatly improved, and correspondingly, the perception accuracy is reduced; (4) It can also be found that the perception accuracy of Dentist /P may even be slightly higher than that of Dentist . We speculate that this is due to the misjudgment by GPT-3.5-turbo when classifying queries, such as mistakenly categorizing a very small number of perception queries as reasoning, while Dentist /P precisely corrects this part of the misjudged perception queries. The same goes for reasoning queries.

Validation Cycle. Validation cycle is also a component that we need to study. We conduct additional experiments by varying the number of validation cycles in our framework and evaluating it on MMBench to demonstrate its effectiveness.

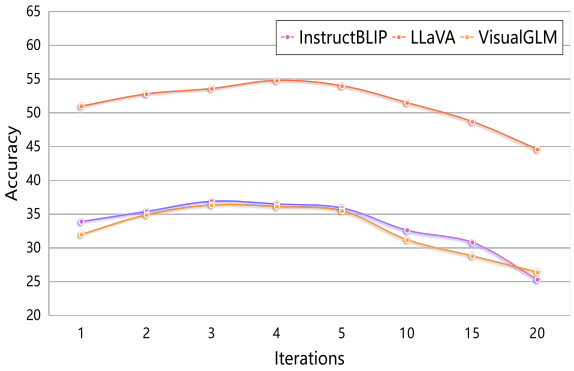

Figure 5: Results of validation cycle

Fig. 5 shows the results. We can see that when the number of validation cycle is small, there is a slight improvement in accuracy as the number of loops increases. However, when the number of cycles is large, the accuracy actually decrease as the number of cycles increases. We separately take out one of the cases for observation and found that when the number of cycles is large, the output of the LVLM and GPT gradually become chaotic and uncontrollable, which may lead to an avalanche of decrease in the accuracy of the model when the number of cycles is large enough. Therefore, we conclude that validation cycle is effective, but special attention needs to be paid to limiting its frequency. When the model answer matches the validation answer, it is important to exit the loop validation in a timely manner.

4.4 Case Study

We provide two testing examples in Figure 4 to conduct qualitative analysis. It is obvious from the above example that:

In the first example, Dentist classifies the query as "Perception". Then, Dentist answers the sub-questions one by one. The verification answer generated by Dentist is consistent with the original output of the model. Hence, the response of the model is faithful and we adopt it as the final answer.

In the second example, Dentist classifies the query as "Reasoning" and refines the hallucinated answer according to the content of the CoT and the original output of the model. Dentist catches some conflicting contents between original answer and verified answer, so it enters validation cycle module until the verified answer does not change significantly after a new round of verification, hence obtains the final answer.

4.5 Performance Visualization

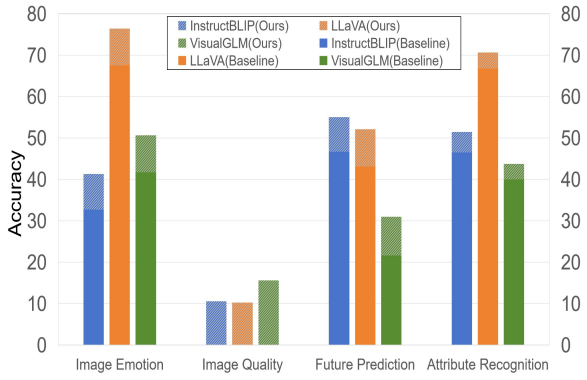

Figure 6: Four typical tasks on MMBench cover Perception and Reasoning. This figure reflects the improvement Dentist has brought to the baseline models.

Fig. 6 shows the performance differences of the baseline InstructBLIP/LLaVA/VisualGLM in Image Emotion, Image Quality, Future Prediction and Attribute Recognition and verified by Dentist. For more comprehensive performance visualization analysis, please refer to the Appendix A.6.

5 Conclusion

In this work, we propose a unified zero-shot framework for hallucination classification and mitigation. We are the first to distinguish treatment based on the classification of hallucinations and use a validation cycle for the removal of hallucinations. Our framework has a clear design which is easily integrated into various LVLMs, and provides convenience for new classifications and treatments to integrate into the framework. To evaluate the effectiveness of our framework, we conduct a experiment on three models on MMbench.

Limitations

This study acknowledges limitation in the Dentist framework. When performing verification, we take the answer to the verification question as the ground truth which actually may still contain hallucinations. In addition, our Chain-of-Though (CoT) is relatively simple.

References

[1] Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. FAITHSCORE: evaluating hallucinations in large vision-language models. *CoRR*, abs/2311.01477, 2023.

[2] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[3] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*, 2023.

[4] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv*, abs/2305.06500, 2023.

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[8] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[9] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[10] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. A survey of large language models. *ArXiv*, abs/2303.18223, 2023.

[11] 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[12] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023.

[14] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

[15] Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*, 2024.

[16] Haoqiang Kang, Juntong Ni, and Huaxiu Yao. Ever: Mitigating hallucination in large language models through real-time verification and rectification. *arXiv preprint arXiv:2311.09114*, 2023.

[17] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*, 2023.

[18] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiao wen Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. Vigc: Visual instruction generation and correction. *ArXiv*, abs/2308.12714, 2023.

[19] Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*, 2024.

[20] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[22] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

A Appendix

In this section, we will display all prompt templates used in this framework. All the following prompt templates are used in GPT-3.5-turbo [23].

A.1 Query Classification

The prompt template is in Fig. 7.

A.2 Sub-questions Generation

The prompt template is in Fig. 8.

A.3 Sub-answers Aggregation

The prompt template is in Fig. 9.

A.4 CoT Verification

The prompt template is in Fig. 10.

A.5 Verification Cycle

The prompt template is in Fig. 11.

A.6 Results on MMBench

The results on MMBench are in Fig. 13, Fig. 14 and Fig. 15.

A.7 Prompt for GPT-4V-aided evaluation.

The prompt template for GPT-4V-aided evaluation is in Fig. 12

A.8 Results of ablation study

The results of LLaVA and VisualGLM are in Table 4 and Table 5.

Variants	Perception Accuracy	Reasoning Accuracy
Baseline	53.52%	50.13%
Dentist	56.83%	51.77%
Dentist/N	54.39%	48.76%
Dentist/P	57.90%	42.63%
Dentist/R	50.43%	52.11%

Table 4: Results on MMBench with different variants of LLaVA

Prompt

Role:
You are now one of my question classification assistants.
Please help me classify the question into two categories: perception or reasoning(binary classification).

Rules:
1.The classification result is only "perception" or "reasoning". Choose one of the two to output.
2.If the question focuses on perception ability, answer "perception"; if the question focuses on logical reasoning ability, answer "reasoning".
3.Don't answer anything else, your answer can only contain "perception" or "reasoning".

Examples:
1.my input: "How many people are there in this picture?"
your answer: "perception"

2.my input: "The person in the picture may do what soon?"
your answer: "reasoning "

{add more examples}

Now please classify the following question according to the example and then answer "perception" or "reasoning":

Figure 7: Prompt template for classification

Prompt

Role:
You are my language assistant for generating sub-questions.
Please generate sub-questions to verify the caption of the picture based on QA-examples below.

Rules:
1.The number of sub-questions cannot exceed three.
2.Extract keywords such as objects, quantities, and locations to generate sub-questions.
3.Each sub-question should have a different focus.
4.Don't ask repeated questions in different sub-questions.
5.If my input contains multiple choice questions, please generate sub-questions based on the question, options and answers.

Examples:
1.my input:
"Question: Write a detailed description for this picture.
Answer: The picture shows a man standing on the back of the yellow taxi, with a yellow shirt and black pants, and a blue backpack on his back. The taxi is driving on a city street with cars and taxis in the background."
sub-questions you generated:
"1.Is there a man standing on the back of a taxi in this picture?
2.What color are the T-shirt and pants that man wear?
3.What's in the background? "

{add more examples}

Now please generate verification sub-questions based on my input below:

Figure 8: Prompt template for generating sub-questions

Variants	Perception Accuracy	Reasoning Accuracy
Baseline	32.31%	31.60%
Dentist	36.35%	36.35%
Dentist/N	35.06%	28.73%
Dentist/P	37.83%	22.60%
Dentist/R	28.26%	37.60%

Table 5: Results on MMBench with different variants of VisualGLM

Prompt

Role:
You are my language assistant for correcting or remaining my passage.
Below is a passage and some Q&A pairs. You need to modify the passage or just keep it unchanged based on the Q&A pairs.

Rules:
1.The information provided by the Q&A pairs is the ground truth, and the information in the passage may contain errors.
2.If the passage conflict with the Q&A pairs, find them and correct the passage based on the Q&A pairs. Try to make minimal changes to retain the original sentence. Then give me the passage which have been corrected by you.
3.If the passage has no confliction with the Q&A pairs, just keep the original passage and give me that.
4.At any time your output should only be a passage.

Examples:
1.Passage:"There are two apples in the picture, they look stale."
Q&A pairs:
Q:How many apples are there in the picture? A:There are three apples in the picture.
Q:Do these apples look fresh in the picture? A:No, they look stale.
Your output:"There are three apples in the picture, they look stale."

{add more examples}

Now I give you the passage and some Q&A pairs, please follow the examples and give me the passage you modified:

Figure 9: Prompt template for aggregating sub-answers to form the output after alleviating the hallucination

Prompt

Role:
You are my language assistant for correcting or remaining my passage.
Below are two passages.
Rules:
1.The second passage is the ground truth, the first passage may contain some errors.
2.If the first passage conflict with the second passage, find them and correct the first passage based on the second passage. Try to make minimal changes to retain the original sentence. Then give me the first passage which has been corrected by you.
3.If the first passage has no confliction with the second passage, just remain the first passage and give me that.
4.At any time your output should only be a passage.

Examples:
1.The first passage: " The profession of the man in the picture is a farmer."
The second passage: "The man in the picture is holding tools and there are many flowers and plants around him. So the profession of the man in the picture is a gardener."
Your output: "The profession of the man in the picture is a gardener."

{add more examples}

Now I give you two passages, please follow the rules and examples and give me your output:

Figure 10: Prompt template for CoT verification

Prompt

Role:
You are my language assistant for determining whether there is a conflict between two passages.

Rules:
1.Below are two passages.
2.If there is conflicting content between the two passages, you should answer "yes"
3.If there is no conflicting content between the two passages, you should answer "no"
4.At any time You can only answer yes or no.

Examples:
1.Passage 1: "There are two apples in the picture, they look stale."
Passage 2: "There are three apples in the picture, they look stale."
Your answer: "yes"

{add more examples}

Now I give you two passages, please follow the examples and give me your answer about whether there is a conflict between two passages.

Figure 11: Prompt template for judging when the validation cycle can be stopped

Prompt

You are my scoring assistant. You need to score two passages describing the picture based on the content of the picture.
What you need to pay special attention to is the hallucination, which refers to the conflict between the content of the passages and the content of the picture.
For example, the passage incorrectly describes the shape or color of the object in the picture, or makes wrong inferences based on the content of the picture.
Please rate the two passages on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1. Accuracy: Refers to whether the description of the picture by the passage is accurate. Passages with fewer hallucinations should be given higher scores.
2. Detail description: Refers to whether the description of the picture is detailed in the passage. Note that descriptions with hallucinations are not counted. Passages with more details should be given higher scores.
3. Complex reasoning: Refers to whether the logical reasoning made by the passage based on the picture content is complex and reasonable. Note that the logical reasoning with hallucinations are not counted. Passages with more logical reasoning should be given higher scores.

Please output a single line for each criterion, containing only two values indicating the scores for Passage 1 and 2, respectively.
The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Passage 1:
{Original Answer}

Passage 2:
{Revised Answer}

Figure 12: Prompt template for GPT-4V-aided evaluation

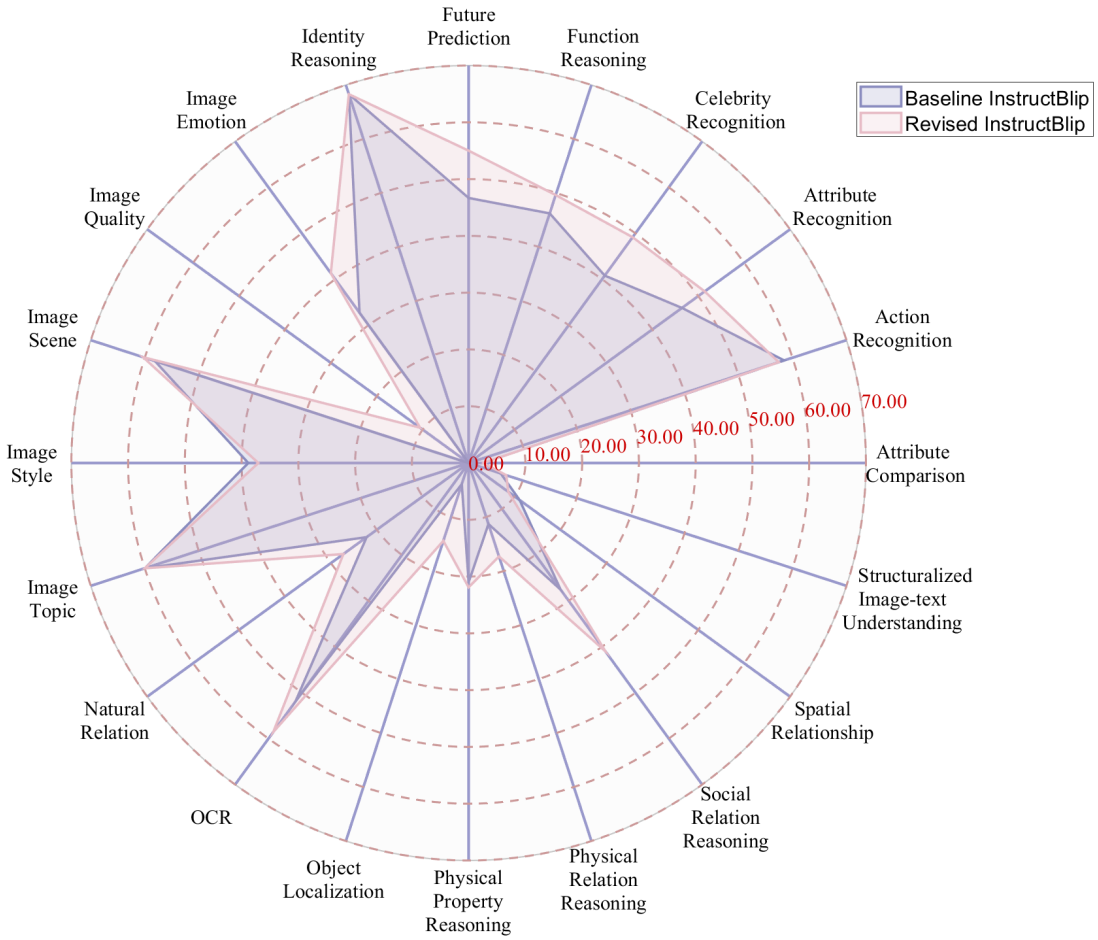

Figure 13: Results of InstructBLIP(Baseline and ours) across the 20 ability dimensions defined in MMBench. The blue area is the result of the baseline, and the red area is ours. See the legend. From this figure, we can intuitively see that our method can enhance the performance of baseline in terms of Image Impression, Image Quality and Future Prediction, etc. For more comprehensive evaluation results on LLaVA and VisualGLM, please refer to Fig.14 and Fig.15

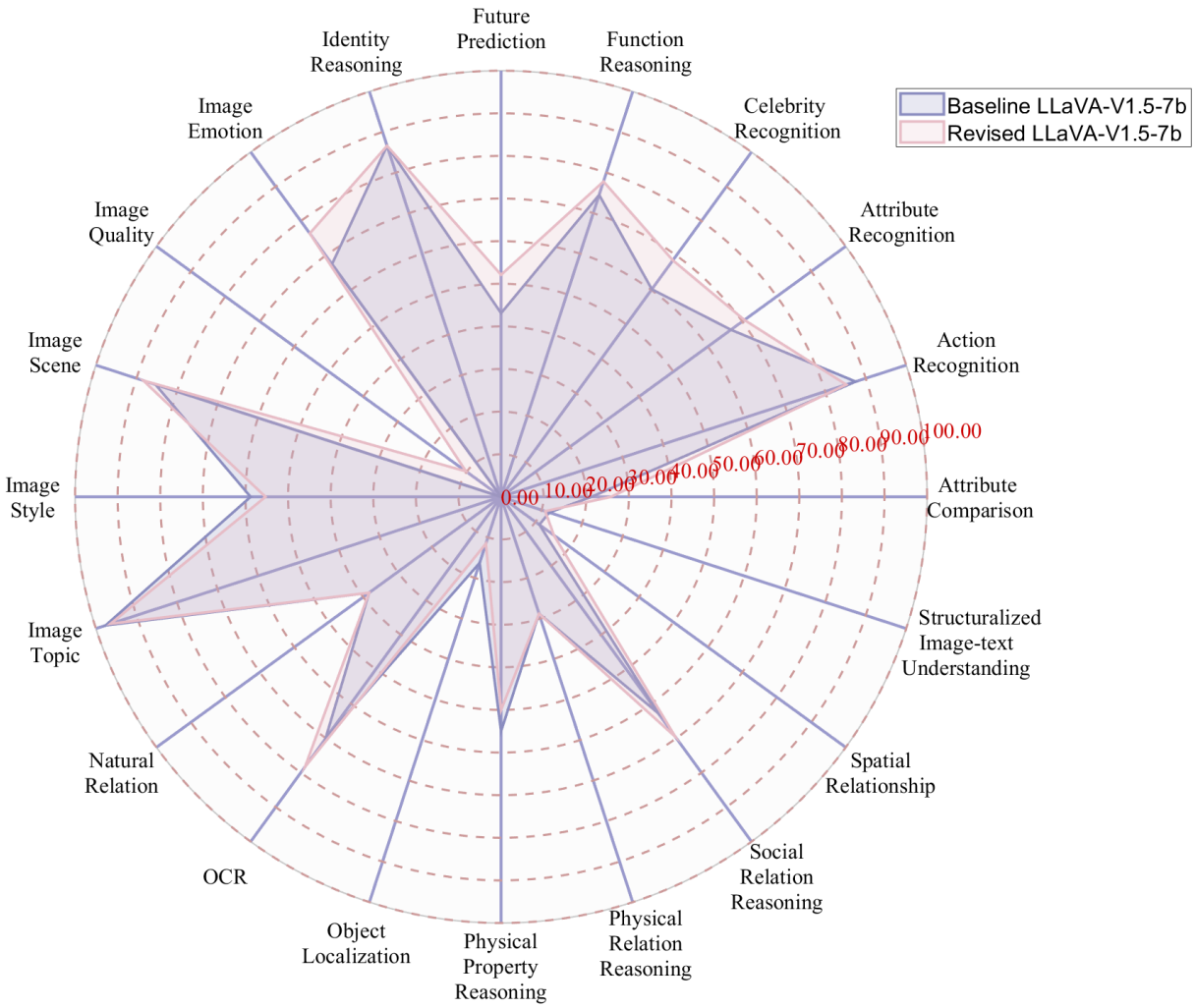

Figure 14: Results of LLaVA on MMBench.

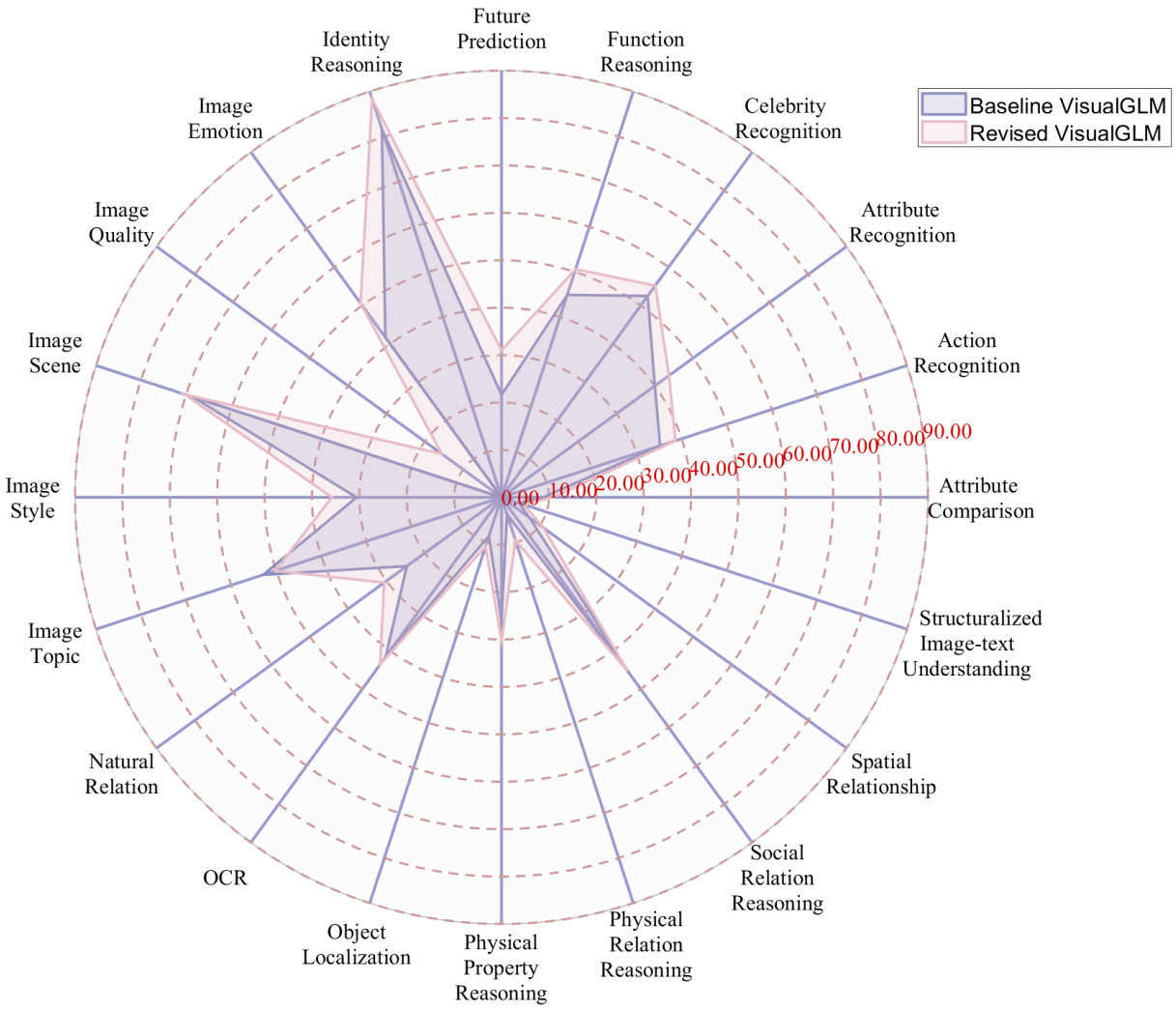

Figure 15: Results of VisualGLM on MMBench.