CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning

Hongyu Hu[*]
ByteDance Inc
Shanghai
huhongyu.123@bytedance.com

Jiyuan Zhang[*]
ByteDance Inc
Shanghai
zhangjiyuan@bytedance.com

Minyi Zhao
ByteDance Inc
Shanghai
minyi.zhao@bytedance.com

Zhenbang Sun[†]
ByteDance Inc
Shanghai
sunzhenbang@bytedance.com

Abstract

Nowadays, the research on Large Vision-Language Models (LVLMs) has been significantly promoted thanks to the success of Large Language Models (LLM). Nevertheless, these Vision-Language Models (VLMs) are suffering from the drawback of hallucination – due to insufficient understanding of vision and language modalities, VLMs may generate incorrect perception information when doing downstream applications, for example, captioning a non-existent entity. To address the hallucination phenomenon, on the one hand, we introduce a **C**ontrastive **I**nstruction **E**valuation **M**ethod (CIEM), which is an automatic pipeline that leverages an annotated image-text dataset coupled with an LLM to generate factual/contrastive question-answer pairs for the evaluation of the hallucination of VLMs. On the other hand, based on CIEM, we further propose a new instruction tuning method called CIT (the abbreviation of **C**ontrastive **I**nstruction **T**uning) to alleviate the hallucination of VLMs by automatically producing high-quality factual/contrastive question-answer pairs and corresponding justifications for model tuning. Through extensive experiments on CIEM and CIT, we pinpoint the hallucination issues commonly present in existing VLMs, the disability of the current instruction-tuning dataset to handle the hallucination phenomenon and the superiority of CIT-tuned VLMs over both CIEM and public datasets. Please contact the authors for code and generated dataset.

1 Introduction

Based on the revolutionary advancement of various Large Language Models (LLM) [1; 2; 3; 2; 4; 5], pre-training [6; 7] and fine-tuning [8; 9] techniques, and adapter solutions [10], a large number of Vision-Language Models (VLM) have emerged, like BLIP-2 [11], MiniGPT-4 [12], LLaVA [13], Otter [14], InstructBLIP [15] and *etc*.

Although these VLMs succeed in significantly facilitating various vision-language downstream tasks, *e.g.*, visual captioning [16] and visual question answering [17], VLMs are also suffering from the hallucination issue [18]. Taking Fig. 1(a) for instance, when doing captioning and question answering, VLM may mistakenly recognize the objects or incorrectly perceive the color of an object. What makes

[*] Equal contribution

[†] Corresponding author

Workshop on Instruction Tuning and Instruction Following at NeurIPS 2023.

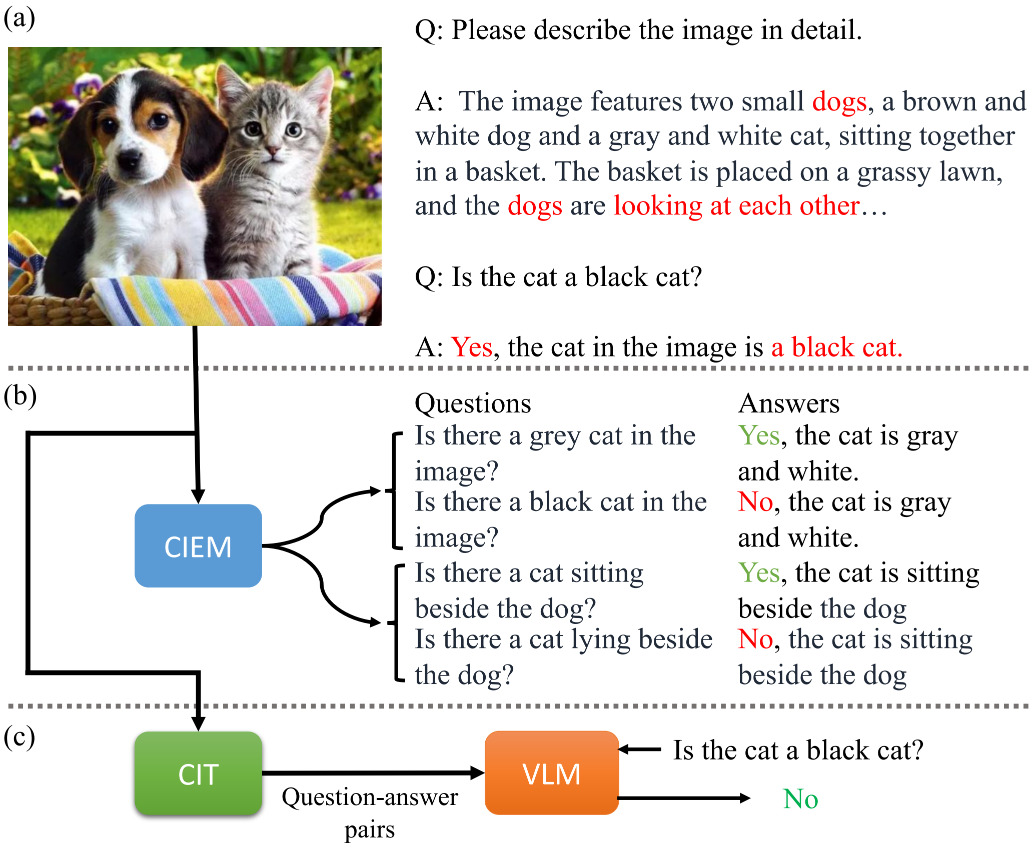

(a)
Q: Please describe the image in detail.
A: The image features two small dogs, a brown and white dog and a gray and white cat, sitting together in a basket. The basket is placed on a grassy lawn, and the dogs are looking at each other…
Q: Is the cat a black cat?
A: Yes, the cat in the image is a black cat.
(b)
Questions
Answers
Is there a grey cat in the image?
Yes, the cat is gray and white.
Is there a black cat in the image?
No, the cat is gray and white.
Is there a cat sitting beside the dog?
Yes, the cat is sitting beside the dog
Is there a cat lying beside the dog?
No, the cat is sitting beside the dog
CIEM
(c)
CIT
Question-answer pairs
VLM
Is the cat a black cat?
No

Figure 1: Illustration of (a) Hallucination phenomenon; (b) Our proposed CIEM method used to evaluate hallucination; and (c) Our CIT method used to boost VLMs.

matters worse is that some VLMs will answer "Yes" without hesitation and groundlessly explain why the answer is "Yes" because of the distribution bias in the turning dataset [13]. Obviously, the hallucination phenomenon severely impairs the performance of VLMs.

Recently, some attempts have been made to measure and evaluate the models' hallucinations. In particular, POPE [19] and MME [20] propose to collect datasets to check the hallucination by constructing question-answer (QA) pairs. However, these methods have the following drawbacks: 1) They introduce human resources to annotate data, which is inconvenient and time-consuming when applied to other datasets and settings. 2) They only focus on the hallucination measurement but fail to provide a technique to tackle the hallucination issue. 3) Their generated data only offer a "Yes/No" answer and lack detailed justification for the answer.

To solve the issues as mentioned above, in this paper, we first present a new **C**ontrastive **I**nstruction **E**valuation **M**ethod (CIEM) to evaluate the hallucination of VLMs. In particular, as an automatic pipeline, the core idea of CIEM is to generate factual/contrastive QA pairs (as shown in the first pair in Fig. 1(b), "grey cat" is based on the fact, while "black cat" is contrastive to the fact) and the corresponding Chain-of-Thought (CoT) justification based on the labeled caption of an image. To this end, we explore ChatGPT [5] as the LLM to automatically generate data by feeding a well-designed prompt that includes the gold caption, the definition of contrastive, and the CoT guidance. Then, QA metrics, like accuracy and recall, can be used to directly measure the model hallucination. Moreover, we propose an instruction tuning method (CIT) to mitigate the hallucination issue. As shown in Fig. 1(c), CIT automatically generates numerous and inexpensive factual/contrastive QA pairs and CoT justifications, making the VLMs understand the answers and the detailed CoT explanations.

The contributions of this paper are summarized as follows: 1) We propose a benchmark, **C**ontrastive **I**nstruction **E**valuation **M**ethod (CIEM), to systematically evaluate the perception ability of VLMs.

CIEM can automatically construct question-answer pairs based on any dataset with caption annotations and thus can evaluate the visual hallucination degree via question-answering accuracy. 2) We propose **C**ontrastive **I**nstruction **T**uning (CIT), which can automatically generate training data in a contrastive-instruction manner based on raw caption annotations. 3) We implement several VLMs on our CIEM benchmark, checking and illustrating their ability for visual hallucination. Furthermore, we apply CIT to some representative VLMs. Experimental results show the advantages of CIT-tuned VLMs on both CIEM setting and public datasets.

2 Related works

2.1 Large Vision-Language Models

Motivated by the recent success of large language models (LLM), recent studies focus on improving vision-language models (VLMs) by integrating powerful language model for broader knowledge and better language understanding. BLIP-2 [11] proposes a generic and efficient pre-training strategy that bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models. A Querying Transformer is proposed to bridge the modality gap between vision and language models. Similarly, Mini-GPT4 [12] aligns a frozen visual encoder with a frozen LLM, Vicuna [21], using just one projection layer to achieve performance close to GPT4. The introduction of powerful LLMs further enhances the VLM's ability on various downstream vision-language tasks (image captioning, visual question answering, visual reasoning) since it serves as a knowledge base to better process the multi-modality information. However, hallucination and inaccurate information are also introduced in this way, which restricts the further application of VLMs.

2.2 Evaluation of VLMs

Since the large vision language model has shown superb ability to understand and process multi-modality information, traditional vision-language benchmarks and datasets are widely adopted to evaluate the VLMs, such as MSCOCO [22], NoCaps [23] for image captioning, VQAv2 [24] and ScienceQA [25] for vision question answering. Evaluation on these benchmarks is limited to a small range of selected tasks or datasets, which needs comprehensive quantitative comparison. Later works are devoted to developing new benchmarks designed for VLMs. Fu *et al.* [20] design a comprehensive evaluation benchmark called MME, which includes 14 perception and cognition tasks. LAMM-Benchmark [26] is also proposed for the systematic evaluation on 2D/3D vision tasks.

However, the works mentioned above all focus on evaluating how well the VLMs can perceive and understand, ignoring the hallucination issues. Concerning this issue, Li *et al.* [19] throw lights into object hallucination through a query method POPE but leave the hallucination of fine-grained object attributes unexplored and fail to provide a solution to address the issue. To this end, we first propose a new Contrastive Instruction Evaluation Method (CIEM), which is an automatic pipeline to assess visual hallucination and considers both the existence and fine-grained attributes of objects. Contrastive Instruction Tuning (CIT) is further designed to alleviate visual hallucinations.

2.3 Instruction Tuning

Originating from the natural language processing (NLP) domain, instruction tuning is introduced to enable large language models, such as GPT-3 [27] and FLAN-T5 [28], to follow natural language instructions and complete real-world tasks. By unifying massive training corpora into an integrated format, instruction tuning can effectively improve the zero- and few-shot generalization abilities of LLM. Inspired by the development of instruction tuning in the NLP domain, researchers focus on introducing it to the multi-modality field. Early works first adapt instruction-tuned LLMs to VLMs by injecting visual information into the LLMs. BLIP-2 [6] uses off-shelf FlanT5 models and closes the modality gap by training a Q-Former to etract visual features as input to the LLMs. MiniGPT4 [12] adopts the instruction-tuned Vicuna [21] as the LLM and a single projection head to bridge the vision and language modalities. While promising task transfer generalization performance is presented, these models are not explicitly trained with multi-modality instruction data. To address this issue, LLaVA [29] uses language-only GPT-4 to generate multi-modality language-image instruction-following data. InstructBLIP [15] gathers a wide variety of publicly available datasets and transforms them into instruction format. Better downstream performance is achieved by introducing instruction-aware visual feature extraction of InstructBLIP. However, the synthesized multimodal instruction data only provides positive samples, easily introducing factual bias to the VLMs.

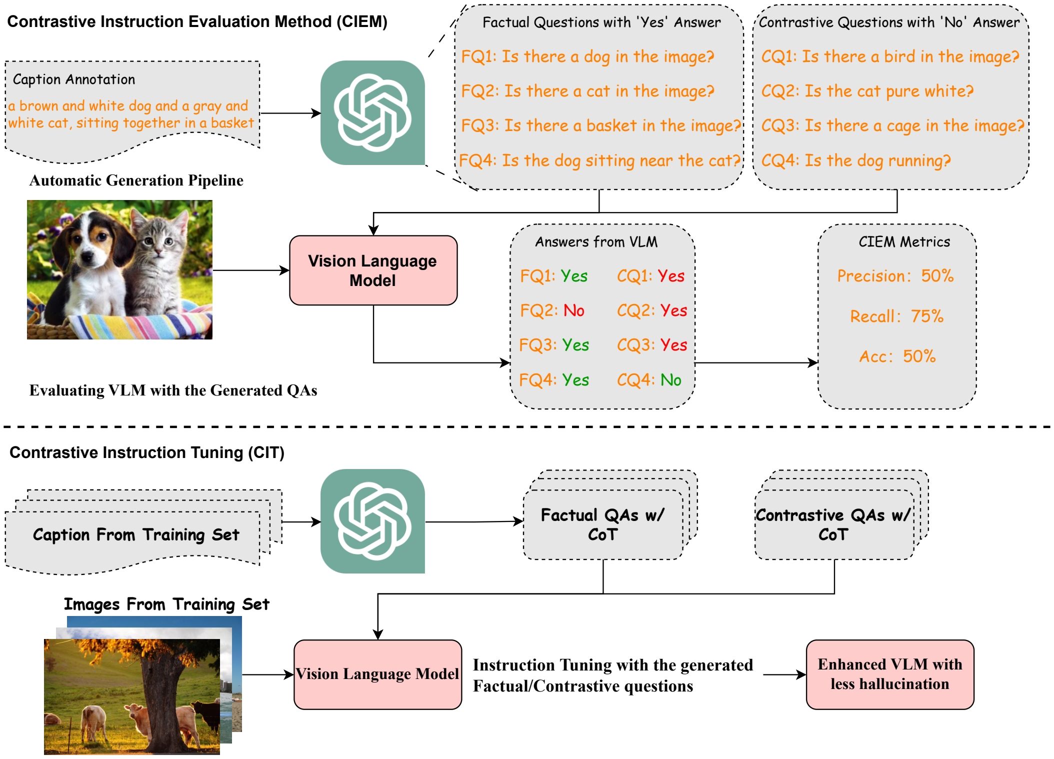

Contrastive Instruction Evaluation Method (CIEM)
Caption Annotation
a brown and white dog and a gray and white cat, sitting together in a basket
Automatic Generation Pipeline
Factual Questions with 'Yes' Answer
FQ1: Is there a dog in the image?
FQ2: Is there a cat in the image?
FQ3: Is there a basket in the image?
FQ4: Is the dog sitting near the cat?
Contrastive Questions with 'No' Answer
CQ1: Is there a bird in the image?
CQ2: Is the cat pure white?
CQ3: Is there a cage in the image?
CQ4: Is the dog running?
Vision Language Model
Answers from VLM
FQ1: Yes
CQ1: Yes
FQ2: No
CQ2: Yes
FQ3: Yes
CQ3: Yes
FQ4: Yes
CQ4: No
CIEM Metrics
Precision: 50%
Recall: 75%
Acc: 50%
Evaluating VLM with the Generated QAs
Contrastive Instruction Tuning (CIT)
Caption From Training Set
Factual QAs w/ CoT
Contrastive QAs w/ CoT
Images From Training Set
Vision Language Model
Instruction Tuning with the generated Factual/Contrastive questions
Enhanced VLM with less hallucination

Figure 2: The overall framework of the proposed CIEM and CIT. CIEM is an automatic pipeline to evaluate visual hallucination issue, and CIT focuses on alleviating this problem.

3 Methodology

The overall framework of the proposed method is shown in Fig. 2. In stage 1, Contrastive Instruction Evaluation Method (CIEM) leverages annotated image-text datasets with an off-shelf LLM to generate factual/contrastive question-answer pairs to evaluate the hallucination of VLMs. Moreover, to alleviate this issue, we generate more factual/contrastive data with justification from the training set by Contrastive Instruction Tuning (CIT) in stage 2. Note that CIEM works on the test set of the annotated dataset, while CIT is adopted on the training set; thus there is no data leakage in the proposed framework.

3.1 Contrastive Instruction Evaluation Method

3.1.1 Automatic Generation of CIEM

Given the test dataset with the image caption annotations, the primary information from the ground truth captions is the entities (object entities), attributes (color/size/shape. *etc*.), and the relation between entities (actions or spatial relations). Based on the provided factual information, we can design a series of questions to query about the image's content, which can be efficiently done with the help of LLM, *e.g.*, ChatGPT. In an automatic pipeline, the image caption is fed into the LLM with the prompt as follows:

You are provided with the sentence which describes an image. You need to finish the following tasks: design questions based on the objects/attributes/actions mentioned in the sentence. The answer to the question should be "yes" because the objects/attributes/actions are mentioned in the sentence.

Table 1: Verification on generated QA pairs from COCO test set.

	Factual QA	Contrastive QA	Total
Num of #	40367	37753	78120
Error QAs	2051	1596	3647
Error Rate	5.1%	4.2%	4.6%

The generated questions, together with the positive answer "Yes", are further formulated as factual QA pairs.

In contrast, for visual hallucination, we further generate some non-existent information for contrastive QAs in the same manner. The key prompt is :

You are provided with the sentence which describes an image. You need to finish the following tasks: design questions based on the contrastive objects/attributes/actions. The contrastive object/attributes/actions are defined as having similar features, easy to confuse or always co-occur. The answer to the questions should be "no" because the contrastive objects/attributes/actions are not mentioned in the sentence.

Based on the method mentioned above, factual questions with a positive answer "Yes" and contrastive questions with a negative answer "No" are generated for downstream evaluation. For instance, questions about the existence of the dog, cat and basket, and the sitting action are all factual questions. In contrast, questions about the existence of the bird, cage, the pure white color of the cat, and the running action are contrastive questions. The generated QA pairs cover a wide variety of aspects, ranging from existence of the objects to the fine-grained attributes such as color, shape, and actions. Note that the proposed automatic pipeline is agnostic to external large language models and does not require human labeling, which is flexible and applicable to different downstream datasets.

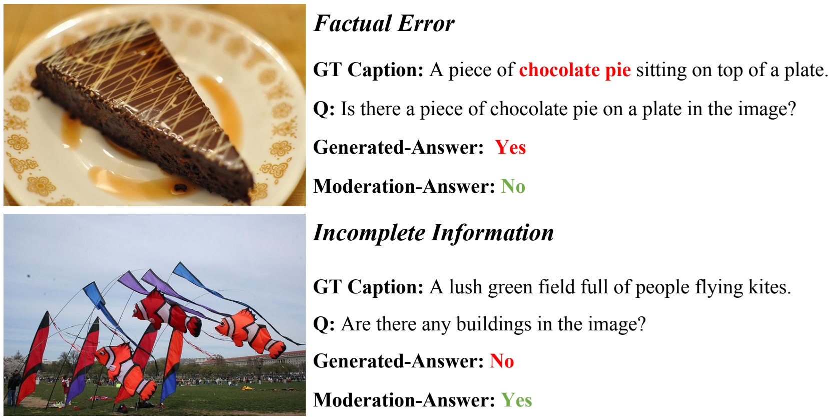
3.1.2 Verifying CIEM

In order to verify the accuracy of the factual/contrastive QA pairs automatically generated by CIEM, we adopt a three-round blind review strategy. Three different moderators will verify whether the generated QA pairs are correct. The two moderators' consistent results will be considered the final result. In particular, we apply the auto-generation method on the test set of COCO caption[22] with ChatGPT [5]. Table 1 shows that the error rate of the generated QA pairs is around 5%. Without human annotation, the automatic pipeline of CIEM is capable of generating accurate QA pairs, which is easy and flexible to deploy on various downstream datasets.

Upon further examination of the inaccurate QA pairs, Fig. 3 shows that the main reasons for the inaccuracy are *factual errors* and *incomplete information* in the annotations. The automatic pipeline would generate non-existent objects due to factual errors in the annotations. At the same time, incomplete information would miss some entities (there are buildings in the background, but they are not mentioned in the caption annotation). Given more accurate annotations, it is expected that CIEM would generate factual/contrastive QA pairs of better quality.

3.1.3 Evaluating VLMs on CIEM

Given the image caption dataset, we can construct a series of factual and contrastive QA pairs of the image via the automatic generation pipeline of CIEM. We further query the VLMs about the factual and contrastive questions of the image and compare the answers from VLMs with the answers generated by CIEM. As the answers to factual/contrastive questions are "Yes/No", the evaluation on CIEM is further formulated as a binary classification task; thus we adopt Accuracy, Precision, Recall, Specificity (recall of negative samples), and F1-score as the evaluation metrics. In the setting of CIEM, Accuracy reflects the overall proportion of the correct answers. Recall and Specificity represent the ratio of correctly answered questions whose ground truth answers are "Yes" or "No", respectively. Moreover, precision is the accuracy of positive predictions, which directly reflects the hallucination issue of the VLMs, and the F1-score is a trade-off metric to balance Precision and Recall.



Factual Error

GT Caption: A piece of **chocolate pie** sitting on top of a plate.

Q: Is there a piece of chocolate pie on a plate in the image?

Generated-Answer: Yes

Moderation-Answer: No

Incomplete Information

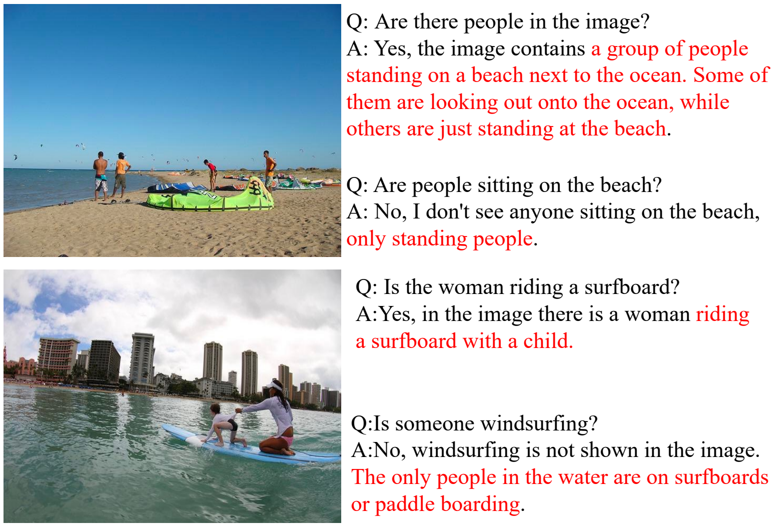
GT Caption: A lush green field full of people flying kites.

Q: Are there any buildings in the image?

Generated-Answer: No

Moderation-Answer: Yes

Figure 3: Inaccurate QA pairs caused by *factual error* and *incomplete information* in the caption annotation.



Q: Are there people in the image?
A: Yes, the image contains a group of people standing on a beach next to the ocean. Some of them are looking out onto the ocean, while others are just standing at the beach.

Q: Are people sitting on the beach?
A: No, I don't see anyone sitting on the beach, only standing people.

Q: Is the woman riding a surfboard?
A:Yes, in the image there is a woman riding a surfboard with a child.

Q:Is someone windsurfing?
A:No, windsurfing is not shown in the image. The only people in the water are on surfboards or paddle boarding.

Figure 4: Examples of the generated samples for Contrastive Instruction Tuning (CIT). Apart from the correct answers, further explanation is also provided.

3.2 Contrastive Instruction Tuning

As CIEM provides an automatic pipeline to evaluate the hallucination issue of the VLMs, we further propose Contrastive Instruction Tuning (CIT) to alleviate the hallucination problem.

Current arts mainly adopt instruction tuning or in-context learning to tune the VLMs, which focus on unifying diverse vision-language data into an integrated Question-Answer format to learn the multi-modality factual information better. However, this manner would lead to a bias or deviation because the VLMs are more likely to give the positive answer "Yes" regardless of the image content. CIT is proposed to tackle this problem by automatically generating corresponding contrastive or adversarial samples, which can be easily achieved by the existing CIEM pipeline. Furthermore, contrastive samples are generated with a concise negative response and provided with a factual basis for further explanation as an effective reasoning path to enhance the VLMs. Similar to chain of thought (CoT), this paradigm matches how humans think of and answer questions. Specifically, we further modify the prompts in CIEM by adding more rules for generating corresponding explanations to factual/contrastive questions, and the prompt is shown as follows:

You are provided with a sentence which describes an image. You need to finish the following tasks: 1) design "yes or no" questions based on the objects/attributes/actions mentioned in the sentence. The answer to the question must start with "yes" because the objects/attributes/actions are in the image. 2) design "yes or no" questions based on the contrastive objects/attributes/actions. The contrastive object/attributes/actions are defined as having similar features, easy to confuse or always co-occur. The answer to the question must start with "no" because the contrastive action is not in the image. Rule: 1) prohibit just answering yes or no, the answer should be detailed and explain the reason. 2) pretend you are looking at the image when answering the questions, do not mention your knowledge is from the sentence.

By adding the sentence *"the answer should be detailed and explaining the reason"* in the prompt, the generated answer will contain further explanation related to the question rather than simply replying "yes" or "no". The detailed explanation or reasoning path would provide additional information to boost the VLMs.

To alleviate the visual hallucination issues, contrastive instruction tuning is performed by further tuning the VLMs with the contrastive samples.

4 Experiments

4.1 CIEM & CIT on COCO

We perform the proposed CIEM with the test set of COCO Caption to evaluate the visual hallucination problem of the VLMs. Specifically, we adopt the automatic pipeline with GPT-3.5 to generate the QA pairs from 4929 images. After the three-round blind review strategy, manually revised QA pairs are counted for downstream evaluation. Generally, there are 37193 factual QA pairs with positive answers and 35748 contrastive QA pairs with negative responses, and we evaluate the representative VLMs on these QA pairs, *i.e.*, LLaVA [29], Mini-GPT4 [12], BLIP-2 [6], and InstructBLIP [15]. Table.2 shows that LLaVA and Mini-GPT4 suffer from more severe visual hallucination, as the two VLMs have high Recall but poor Precision and F1-score, indicating that the models are prone to give positive responses regardless of the question. In contrast, InstructBLIP performs the best on F1-score, the trade-off between Precision (visual hallucination) and Recall (perception). We infer that the instruction tuning dataset for InstructBLIP covers a broader range with more diversity.

Table 2: Evaluating VLMs with CIEM on the test set of COCO Caption. (Pre:Precision, Rec: Recall, Spec: Specificity, F1: F1-score, Acc: Accuracy)

Model	Model Structure			Pre	Rec	Spec	F1	Acc
	Visual	Q-Former	LLM					
LLaVA	CLIP_L	Linear	MPT-7B	55.42	**95.59**	20.84	70.16	58.76
Mini-GPT4	EVA_G	Linear	Vicuna7B	58.95	94.14	32.51	72.50	63.77
BLIP2	CLIP_G	Q-Former	T5xxl	**82.07**	65.27	**85.37**	72.71	**75.20**
InstructBLIP	CLIP_G	Q-Former	T5xxl	71.11	81.75	65.94	**76.06**	73.95

As CIEM throws light on evaluating visual hallucination, CIT is a step forward to the solution to alleviating this issue. We further apply CIT on InstructBLIP with both Vicuna-7B and FLAN-T5 XL as the LLM. Particularly, to avoid data leakage, the training split of COCO caption is adopted to generate the factual/contrastive QA pairs, and there are 1.5 million pairs for 110 thousand images in total.

It is displayed in Table.3 that without contrastive instruction tuning, both the pre-trained version and the one tuned with LLaVA dataset show severe visual hallucination with Recall higher than 90% and relatively poor Precision and F1-score, which is consistent with the results in Table.2. This phenomenon also indicates that the current fashion of instruction tuning and dataset is likely to introduce hallucination. Table.3 further shows that the proposed CIT improves Precision, Specificity, F1-score and Accuracy by a large margin, and there is only a slight drop in Recall, which demonstrates that the VLM is now more *sane* and *conservative* to give positive answers, thus the hallucination is alleviated. Moreover, we find out that removing the CoT (detailed explanation and reasoning path) harms the VLM's overall performance despite an improvement on Precision. It further proves the necessity and effectiveness of the CoT because the VLM can learn from the details rather than just memorizing the final answer.

Table 3: Results of Contrastive Instruction Tuning(CIT) on InstructBLIP. The baseline method is the zero-shot InstrcutBLIP (**Pretrain**) and the one tuned with the instruction-following data collected in **LLaVA**. Contrastive Instruction Tuning (CIT) alleviates visual hallucination issue and improves the overall performance of the VLM.

Model	Dataset	Precision	Recall	Specificity	F1	Acc
Vicuna-7B	Pretrain	73.9	93.4	66.2	82.5	79.9
	LLaVA	71.6	**93.9**	61.9	81.3	78.1
	CIT w/o CoT	**93.7**	44.2	**96.9**	60.1	70.2
	CIT w/ CoT	85.5	87.9	84.7	**86.7**	**86.3**
T5-XL	Pretrain	67.4	93.3	53.7	78.2	73.7
	LLaVA	66.9	**93.8**	52.2	78.1	73.3
	CIT w/o CoT	**79.7**	86.5	**77.4**	83.0	82.0
	CIT w/ CoT	78.7	91.8	74.5	**84.7**	**83.3**

4.2 How does CIT Affect Downstream Tasks?

A major concern about the CIEM and CIT is that CIT is specially designed for alleviating hallucination by tuning the VLMs with the generated instruction-following data with "Yes or No" QAs. However, how CIT would affect the VLM's original ability on multi-modality downstream tasks remains unknown. To address this issue, we also evaluate how the VLM performs on image captioning and VQA tasks after tuning with the CIT strategy. Table.4 further displays a slight improvement in image captioning tasks with NoCaps and COCO Caption datasets, but a slight drop on VQA with VQAv2. We infer that the generated QA pairs mainly focus on the primary objects, attributes and relations of the image content, which is beneficial to image captioning. On the other hand, the generated QAs are only based on the image caption, which might miss some more informative details for VQA tasks. Generally, CIT would hardly affect the original multimodal ability of the VLM; thus it is an effective and safe method to alleviate visual hallucination.

4.3 Qualitative Results

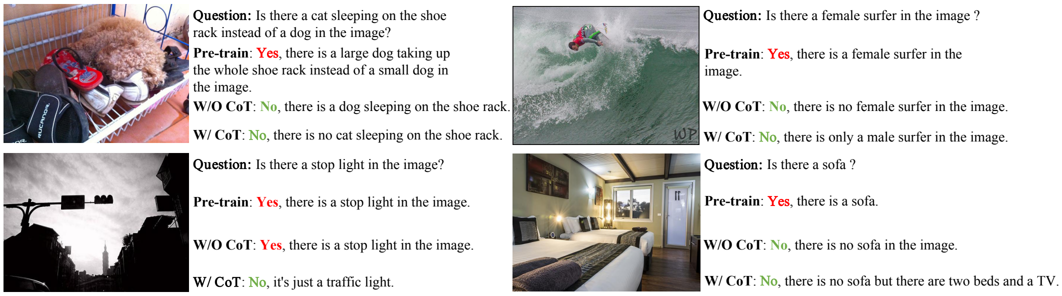
Fig.5 further shows the qualitative results of CIT. The open-source pre-trained VLM is prone to say "yes" without hesitation. With CIT, the VLM can judge objects that do not appear in the image, and CoT further enables the VLMs to give detailed explanations and justification regarding the question. The qualitative results also demonstrate the effectiveness of the proposed CIT.

4.4 Limitation and Future Work

Although the proposed CIEM and CIT is considered an effective method to evaluate and alleviate visual hallucination, there are still some limitations of the proposed method, which motivate our

Table 4: Experimental results on how Contrastive Instruction Tuning (CIT) affects multi-modality downstream tasks, such as image captioning and visual question answering.

Model	Dataset	NoCaps-Val		COCO Cap		VQAv2
		B@4	CIDEr	B@4	CIDEr	
Vicuna-7B	Pretrain	31.6	113.7	**38.1**	131.7	**76.6**
	LLaVA	32.2	115.4	37.7	132.0	74.9
	CIT w/o CoT	32.0	114.9	37.7	131.8	74.4
	CIT w/ CoT	**32.8**	**115.7**	37.5	**132.6**	74.4
T5-XL	Pretrain	37.0	121.2	40.8	140.7	73.4
	LLaVA	36.9	121.2	40.7	140.6	**73.4**
	CIT w/o CoT	36.9	121.4	40.9	140.8	73.1
	CIT w/ CoT	**37.3**	**121.7**	**41.0**	**141.1**	73.0



Question: Is there a cat sleeping on the shoe rack instead of a dog in the image?

Pre-train: Yes, there is a large dog taking up the whole shoe rack instead of a small dog in the image.

W/O CoT: No, there is a dog sleeping on the shoe rack.

W/ CoT: No, there is no cat sleeping on the shoe rack.

Question: Is there a female surfer in the image ?

Pre-train: Yes, there is a female surfer in the image.

W/O CoT: No, there is no female surfer in the image.

W/ CoT: No, there is only a male surfer in the image.

Question: Is there a stop light in the image?

Pre-train: Yes, there is a stop light in the image.

W/O CoT: Yes, there is a stop light in the image.

W/ CoT: No, it's just a traffic light.

Question: Is there a sofa ?

Pre-train: Yes, there is a sofa.

W/O CoT: No, there is no sofa in the image.

W/ CoT: No, there is no sofa but there are two beds and a TV.

Figure 5: Visualization of CIT: CIT can effectively alleviate the visual hallucination issue, and CoT further enables the VLM to correct the wrong information.

future works. Firstly, CIEM relies on the annotated multimodal dataset. The quality of the annotation itself might influence the accuracy of the generated QA pairs in CIEM, and CIEM is not able to work on raw image data without annotations. Involving other large models to generate the confident caption might be a direct and simple solution. Secondly, the generated QA pairs in CIEM and CIT are "Yes or No" questions.The questions are expected to be in more flexible and diverse formats for more general scenarios. Last but not the least, the current CIEM pipeline mainly focuses on the perception ability and the visual hallucination problem. Integrating the evaluation on more aspects of the VLMs, such as knowledge retrieval and reasoning, into a unified benchmark would also be our future work.

5 Conclusion

Despite showing superb performance on various multi-modality tasks, current VLMs are suffering from the drawback of hallucination. This paper proposes an automatic pipeline, Contrastive Instruction Evaluation Method (CIEM), to evaluate the visual hallucination issue in VLMs. With the assistance of external LLM, CIEM is capable of automatically generating high-quality factual/contrastive QA pairs to query the VLMs about the entities and attributes based on the image caption, and can be flexibly deployed on various downstream datasets. Contrastive Instruction Tuning (CIT) is further proposed to alleviate visual hallucination. By generating QA pairs with detailed explanations and reasoning path from the training data, the VLMs are expected to learn from the informative multi-modality data and hallucinate less. Experimental results reveal the hallucination issue of the VLMs, demonstrating the effectiveness of the proposed CIT.

6 Acknowledgement

We sincerely thanks Hanyu Wei, Lan Zhang, Wenjing Zhou and other colleagues from the AI Data Service and Operations Team for their support on data labelling.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[3] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[4] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.

[5] OpenAI, “Gpt-4 technical report,” 2023.

[6] J. Li, D. Li, C. Xiong, and S. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *International Conference on Machine Learning*, pp. 12888–12900, PMLR, 2022.

[7] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao, *et al.*, “Vision-language pre-training: Basics, recent advances, and future trends,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022.

[8] B. Lester, R. Al-Rfou, and N. Constant, “The power of scale for parameter-efficient prompt tuning,” *arXiv preprint arXiv:2104.08691*, 2021.

[9] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, “Visual prompt tuning,” in *European Conference on Computer Vision*, pp. 709–727, Springer, 2022.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” *arXiv preprint arXiv:2106.09685*, 2021.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” *arXiv preprint arXiv:2301.12597*, 2023.

[12] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.

[13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.

[14] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, “Otter: A multi-modal model with in-context instruction tuning,” *arXiv preprint arXiv:2305.03726*, 2023.

[15] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” *arXiv preprint arXiv:2305.06500*, 2023.

[16] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, “From show to tell: A survey on deep learning-based image captioning,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 539–559, 2022.

[17] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[18] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, “Object hallucination in image captioning,” *arXiv preprint arXiv:1809.02156*, 2018.

[19] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, “Evaluating object hallucination in large vision-language models,” *arXiv preprint arXiv:2305.10355*, 2023.

[20] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, *et al.*, “Mme: A comprehensive evaluation benchmark for multimodal large language models,” *arXiv preprint arXiv:2306.13394*, 2023.

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, “Judging llm-as-a-judge with mt-bench and chatbot arena,” *arXiv preprint arXiv:2306.05685*, 2023.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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*, pp. 740–755, Springer, 2014.

[23] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, “Nocaps: Novel object captioning at scale,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

[24] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

[25] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, “Learn to explain: Multimodal reasoning via thought chains for science question answering,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[26] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, *et al.*, “Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark,” *arXiv preprint arXiv:2306.06687*, 2023.

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[28] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, “Scaling instruction-finetuned language models,” *arXiv preprint arXiv:2210.11416*, 2022.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *arXiv preprint arXiv:2304.08485*, 2023.