Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) have made significant progress, blending visual perception with linguistic comprehension to produce outputs that are both consistent and contextually relevant. Despite the promising progress achieved, LVLMs still suffer from the hallucination problem, *e.g.*, they tend to produce contents that are non-existent in the input images. We investigate that the cause for the model's hallucinatory output is the deficiency in fine-grained comprehension on the visual aspect, due to appearance or semantic similarities, *e.g.*, bicycle *vs.* motorcycles, baseball bat *vs.* baseball. In this work, we show such hallucination is naturally mitigated via a novel method called *visual evidence prompting*, utilizing small visual models to complement the LVLMs. Traditional visual models are not adept at interacting with humans, but they excel at perceiving the fine-grained image contents. Symbolizing the professional outputs of domain-expert models as prompts, the LVLM generalists are able to refer to these evidences as visual knowledge to generate more precise answers. Detailed analysis shows that visual evidence enables models to adjust and rectify the attribution on the images, diminishing the degree of visual confusion and suppressing false activation while augmenting correct activation. Extensive experiments reveal that our approach not only mitigates LVLM's hallucination, *e.g.*, obtaining 4.20% improvement in accuracy over the baseline GPT-4V, but also maintains and even improves their original strengths of general multimodal understanding. We hope our straightforward but insightful work enhances the comprehension of hallucination in LVLMs and offers valuable perspectives on addressing such challenges.

1 Introduction

The success of large vision-language models (LVLM) has resulted in significant advancements in overall comprehension of visual semantics (Chen et al., 2023; Li et al., 2023a). Pioneers like GPT4-V have spearheaded the movement to unprecedented levels, demonstrating exceptional capabilities across a wide array of activities (Yang et al., 2023), including generative tasks like fully describing the contents of a given image, and discriminative tasks like answering whether an object appears in the picture.

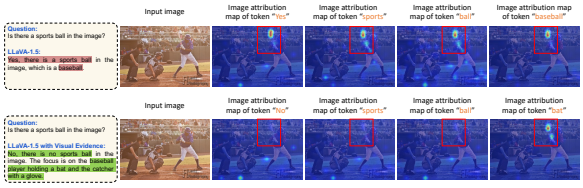
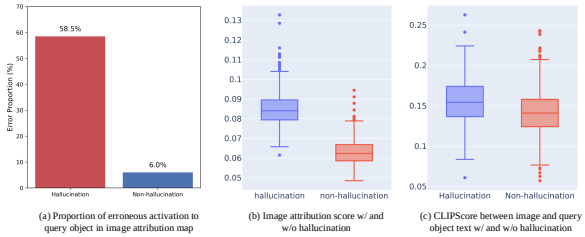


Figure 1: Visualization of the image attribution map for LLaVA-1.5-7B when hallucination happens and after the integration of visual evidence. More cases can be found in Appendix C.2 and C.3.

Despite the success of LVLMs, they still grapple with a notable issue called *multimodal hallucination*. They tend to produce non-existent objects (there is a "chair" in the image), relations (the dog is "behind" the cup), or attributes (the color of the cat is "white") in the image (Li et al., 2023b; Gunjal et al., 2023; Liu et al., 2023a). This issue reveals fundamental shortcomings of LVLMs, such as over-reliance on the co-occurrence bias with objects and insufficient perception ability of the image content (Agarwal et al., 2019; Goyal et al., 2016; Li et al., 2023b).

As shown in the upper section of Fig. 1, when the model is queried about the presence of a sports ball in the image, it hallucinates and responds with "yes". To further investigate why the model answers incorrectly, we use a widely used interpretability method (Chefer et al., 2021) to trace the attribution of each image patch when the model generates each token in the responses. The higher the attribution score of a particular image region, the more significant its contribution to the model's prediction. We find that when the model answers



(a) Proportion of erroneous activation to query object in image attribution map

(b) Image attribution score w/ and w/o hallucination

(c) CLIPScore between image and query object text w/ and w/o hallucination

Figure 2: Hallucination analysis on LLaVA-1.5. **(a)** The proportion of false activation to semantically similar regions during hallucination and non-hallucination. **(b)** The attribute scores of image tokens to final answers. **(c)** The CLIPScore (*i.e.*, image-text similarity) between the corresponding images and hallucinated/non-hallucinated objects. These statistic analyses are conducted on POPE.

with the first token "yes", it clearly attends to the baseball bat. As the model continues to generate the tokens "sports" and "ball", the relevance of the baseball bat area is further strengthened, indicating that the area causing the model's hallucination is the baseball bat. This is may due to the high semantic similarity of baseball bats and sports balls, leading the model to mistakenly assume the presence of a sports ball upon seeing a baseball bat.

Following a thorough statistical analysis, we observe that such phenomena are prevalent when hallucination happens and we further analyze the underlying causes of this phenomenon. In Fig. 2(a), we carefully analyze the proportion of erroneous activation in the image attribution maps corresponding to regions that are semantically or visually similar to the query object, under conditions where the model either hallucinates the object or does not. Our results show that, when hallucinations occur, the model incorrectly activates regions similar to the query object in terms of semantics or appearance at a rate of 58.5%, as exemplified by the case discussed in Fig. 1. In Fig. 2(b), we present the attribution scores of image tokens to the final prediction both in the presence and absence of hallucinations. The detailed calculation method for the attribution score can be found in Sec. A of the Appendix. The results indicate that when hallucinations occur, the image attribution is significantly higher, which aligns with the findings in Fig. 2(a). In Fig. 2(c), we further compute the feature similarity between hallucinated and non-hallucinated objects (using the prompt template "A photo of {query_object}.") and their corresponding images using CLIP (ViT-L-14@336, which also serves as the vision encoder in LLaVA-1.5). We observe that hallucinated objects exhibit a notably higher CLIPScore with their images, suggesting a stronger semantic alignment between these objects and the images. This explains why the model tends to incorrectly activate regions of the image that are semantically or visually closer to the query object when hallucinations occur. These analyses and findings suggest that the causes for hallucination of LVLM are likely the deficiency in fine-grained context discrimination on the visual aspect, due to the appearance or semantic similarities, *e.g.*, bus *vs.* truck, baseball *vs.* baseball bat.

However, the existing methods to alleviate hallucination are not designed to enhance the understanding capabilities of the visual aspect. Early works tried to instruction-tune the models to negate descriptions in the question that do not match the image contents by annotating negative instructions or unfaithful object descriptions and relations (Gunjal et al., 2023; Liu et al., 2023a). The instruct tuning process does not bring new knowledge to the model, but encourages the model to learn the style of answering (Zhang et al., 2023). In addition to the huge cost, there is a risk of overly optimizing the model to fit a specific problem or dataset, leading to catastrophic forgetting (Zhai et al., 2023). Several recent studies address this issue by applying an additional large language model to amend responses, training a post-hoc corrector to reconstruct less hallucinatory outputs (Yin et al., 2023), or adjusting the output distribution via optimizing the decoding strategy (Leng et al., 2023; Huang et al., 2023). Regardless of whether employing instruction tuning, RLHF, or methods via optimizing decoding strategies, existing approaches do not endow the model with more visual knowledge to utilize in the process of generating answers, meaning they do not enhance the model's own ability of fine-grained visual content perception.

Simultaneously inspired by the observation that humans refer to the key contents of a picture when conversing (Henderson and Ferreira, 2013) , and the solutions of retrieving knowledge evidence to tackle the hallucination problem in language models (Ren et al., 2023; Mialon et al., 2023), one promising yet under-explored cure for hallucinations is to refer to additional visual knowledge from the image. Generally, traditional small visual models excel at the tasks they are trained for. For instance, in the task of object detection, small visual models can efficiently identify and locate objects within an image (Fang et al., 2021; Carion et al., 2020). In the task of scene graph generation (SGG)

(Zellers et al., 2018; Cong et al., 2023), small visual models can generate detailed descriptions of objects and their visual relations within a given scene, such as "dog near cup, newspaper on table". Small visual models are better characterized as narrow experts who focus on the processing and understanding of visual content, while LVLMs are competent generalists who have strong semantic understanding and generalization capabilities. Naturally, the small visual models complement the LVLMs by effectively extracting contextual information from images to generate more precise answers.

This work explores how the hallucinations of LVLMs can be mitigated by referring to visual evidence from small visual models. An example is shown in Fig. 3. The original LVLM produces inaccurate answer when queried about the presence of a chair within the image. However, small visual models, *i.e.*, object detection and scene graph generation models, can output accurate objects and relations, *e.g.*, "dog", "cup", "dog near cup". After symbolizing the accurate and faithful output of small visual models as prompting texts, the model are able to generate correct responses. We refer to this training-free approach as *visual evidence prompting* (VE).

Through detailed analysis, we find that visual evidence enables models to adjust and rectify the attribution on the images, diminishing the degree of visual confusion and suppressing false activation while augmenting correct activation. Integrating visual evidence also corrects the prediction distribution for hallucination instances and elevates the confidence for non-hallucination samples. Extensive experiments on 9 LVLMs on 5 benchmarks show that visual evidence prompting mitigates the object, attribute, and relation hallucinations for the generative task and discriminative task simultaneously, and maintains or improves the ability of general multimodal understanding of LVLMs.

Our results serve as compelling evidence for the potential applicability and efficacy of visual evidence prompting. We aim for our work to not only establish a minimal yet robust baseline for the challenging benchmarks but also draw attention to the understanding and interpretation of hallucination in LVLMs.

2 Visual Evidence Prompting

The goal of this paper is to endow LVLMs with the ability to mitigate hallucinations with the help of visual evidence. Generating answer A with input image I and query question Q can be expressed in a probabilistic framework as estimating a conditional distribution $P(A|Q, I)$. The visual evidence prompting is formalized as $P(A|Q, I, E)$, where E is the key contents from the image.

Considering the internal process of a person when answering questions based on image content, it is typical to decompose the problem into two steps (Barsalou, 2008; Palmer, 1999). For example, as in Fig. 3, there is a question about "Is there a chair in the image?". Firstly identify the key elements in the image as evidence ("1 dog, 1 cup, 1 newspaper, dog near cup, dog on table, newspaper on table"). Then, symbolize and combine the relevant content within the evidence to answer the question. After this process, an answer is generated.

Extraction. The input image of the large vision-language model is fed into the small visual model, and the output is formulated as predefined formats. For the object detection models, the output is defined as the semantic label from the label map given the predicted class index. If the model detects multiple objects of the same category, we will merge these objects and formulate them in terms of numbers, such as "3 dogs, 1 cat". For the scene graph generation models, the output is composed of the ⟨subject, relation, object⟩ triplets. Each triplet is firstly formulated as `{subject}{relation}{object}`. Multiple triplets are joined with the ",". For example, ⟨man on surfboard⟩ and ⟨man has hair⟩ are formulated as "man on surfboard, man has hair".

Prompting. In the second step, we use symbolized visual evidence along with prompted questions to extract the final answer from the LVLM. To be concrete, we simply concatenate two elements as with "You can see `{evidence}` in the image. `{question}`?". The prompt for this step is self-augmented since the prompt contains the visual evidence generated by the visual model. This is one simple and effective formulation of visual evidence. More sophisticated formats may bring further improvement. Finally, LVLM is fed the prompted text and the original image as input to generate final answers.

Discussion. The technical comparison between our approach and previous methods (LRV (Liu et al., 2023a), VCD (Leng et al., 2023)) is shown in Tab. 1. "Training-free" indicates no fine-tuning of the LVLMs. "Model-free" indicates not depen-

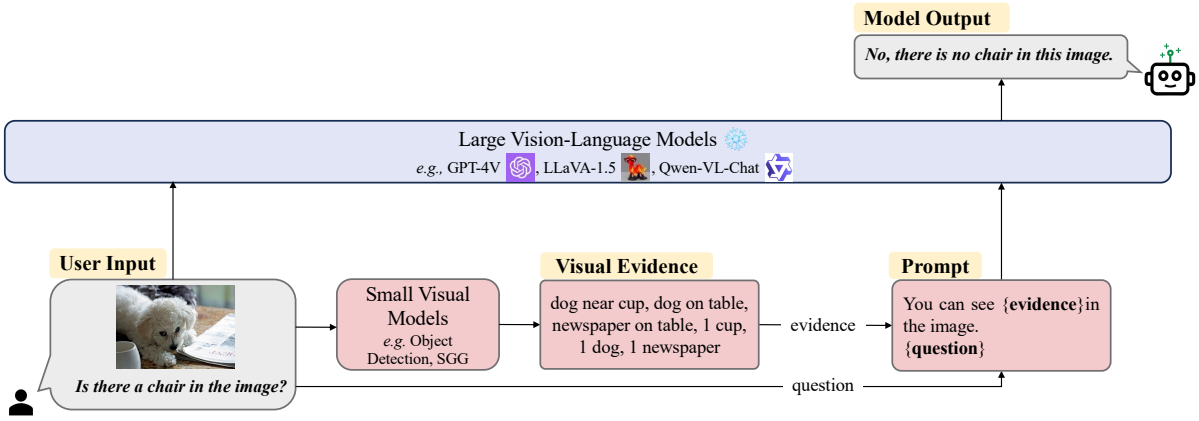


Figure 3: An overview of visual evidence prompting, which mitigates hallucinations in LVLMs via referring to visual evidence from small visual models. Given the input image, the small visual models generate visual evidence about different aspects of the image, *e.g.*, object categories, and relations between objects. Then the "visual evidence" prompts are used to extract the answer from the image and evidence texts.

Related works	Training-free	Model-free	Visual knowledge
LRV (Liu et al., 2023a)	✗	✗	✗
VCD (Leng et al., 2023)	✓	✗	✗
Ours	✓	✓	✓

Table 1: Comparison with previous representative methods.

dent on the checkpoint parameters or logits of a model and also applicable to API services. "Visual knowledge" means referring to the visual evidence generated by small domain-specific visual models, *e.g.*, object detection or scene graph generation models. In this paper, visual models refer to the object detection and scene graph generation models. Other models such as semantic segmentation models, OCR models, and human-object interaction models can also be considered and may bring potential gains, but that is not the priority of this paper.

3 Experiments

3.1 Experimental Setup

Datasets. POPE (Li et al., 2023b) is dedicated to evaluating object hallucinations of LVLMs. It contains the settings of random, popular, and adversarial sampling, which mainly differ in the way negative samples are constructed. AMBER(Wang et al., 2023) provides a coverage of evaluations for both generative task and discriminative task including object, attribute and relation hallucination. The evaluation of relation hallucination in AMBER (AMBER use the prompt "`Is there direct contact between the {object 1} and {object 2} in this image?`" to probe relation hallucination) is relatively limited. In order to further verify the effectiveness of our method in mitigating relation hallucination, we meticulously follow the same recipe as POPE on Visual Genome (Krishna et al., 2017) to construct a new relation hallucination evaluation dataset named Relation Probing Evaluation (RPE).[1] More details about the three benchmarks are shown in the Appendix B.1.

Evaluation metrics. POPE converts the hallucination evaluation into a binary classification problem to probe the model's awareness of whether a specific object exists in the image, with the output of "Yes" or "No", *e.g.*, "`Is there a chair in this image?`". If the model's response includes neither "Yes" nor "No", it will be disregarded in the calculation of metrics. The accuracy reflects the proportion of correctly answered questions. For the generative task of AMBER, the Caption Hallucination Assessment with Image Relevance (CHAIR) is a specifically developed evaluation metric tailored to assess the extent of object hallucination in image captioning tasks. Specifically, CHAIR measures the level of object hallucination in a provided image description by calculating the proportion of referenced objects in the description that do not exist in the actual ground-truth label set. For the discriminative task of AMBER and RPE, the evaluation metric is the same as POPE.

Baselines. In order to conduct our experimental analysis, we incoporate a wide array of 7 popular open-source models and two black-box APIs as representatives, including MiniGPT-4 (Zhu

[1]Note that there is no overlap between the datasets used for training small models and all of the LVLM hallucination test sets.

Model	POPE		AMBER				RPE		Latency token/sec.
	Acc.	Δ	CHAIR$_{\downarrow}$	Δ	Acc.	Δ	Acc.	Δ	
LLaVA-1.5-7B	80.23	-	8.07	-	71.02	-	61.92	-	28.86
+ Visual Evidence	**87.43**	+7.20	**6.78**	-1.29	**75.28**	+4.26	**68.00**	+6.08	23.96
LLaVA-1.6-7B	84.93	-	8.59	-	70.09	-	70.20	-	22.71
+ Visual Evidence	**89.43**	+4.50	**7.73**	-0.86	**76.08**	+5.99	**70.46**	+0.26	20.61
MiniGPT-4-v2	75.33	-	8.67	-	61.16	-	60.75	-	18.06
+ Visual Evidence	**83.17**	+7.84	**8.39**	-0.28	**70.06**	+8.90	**68.38**	+7.63	14.27
GPT4-V (API)	82.21	-	6.97	-	85.50	-	75.56	-	12.26
+ Visual Evidence	**86.41**	+4.20	**6.76**	-0.21	**86.73**	+1.23	**76.05**	+0.49	11.07

Table 2: The main results on POPE, AMBER and RPE dataset.

et al., 2023), LLaVA (state-of-the-art open-source model), and GPT4-V. More details can be refer to Appendix B.2.

Implementation details. Firstly, we use the corresponding visual small model (*i.e.*, object detection model and scene graph generation model) to process the images in the evaluation datasets and obtain the corresponding visual evidence. As our primary focus is to demonstrate the efficacy of our proposed framework, unless specified otherwise, we employ detr-resnet-101 (Carion et al., 2020) as the default model for extracting object evidence, while RelTR (Cong et al., 2023) is our default choice for obtaining relation evidence, though models with superior performance and open-vocabulary models are also applicable. We employed the default parameter settings provided in the official repository for each model, respectively.

3.2 Results

Baselines performance. We firstly evaluate the hallucination performance on the datasets of POPE and AMBER of the 7 open-source models and two black-box APIs. The extensive evaluation results are presented in Tab. 2 and Tab. 9 in Appendix. Due to limited space, we report the results of POPE on the most challenging subset COCO-Adversarial. The results for the other two subsets are included in the Tab. 10 in Appendix. There are several interesting observations: 1) The LVLM with more parameters does not necessarily have fewer hallucinations, e.g., 80.23% of LLaVA-1.5-7B on POPE vs. 78.70% on LLaVA-1.5-13B. 2) Models with updated versions typically have fewer hallucinations. For example, 71.16% of MiniGPT-4 on POPE vs. 75.33% on MiniGPT-4-v2, and 82.33% on Qwen-VL-Chat vs. 87.90% on Qwen-VL-Max.

Effect of visual evidence. After incorporating visual evidence prompting, without bells and whistles, almost all models including black-box APIs generate more precise discernment of the contents within the image. For example, the accuracy of LLaVA-1.5-7B on POPE increases from 80.23% to 87.43% (+7.20%), and the accuracy on AMBER increases from 71.02% to 75.28% (+4.26%). It's worth noting that we also achieve non-trivial improvements on the black-box APIs. For example, on POPE, there is a 4.20% and 2.76% improvement for GPT-4V when combined with visual evidence. This demonstrates the superiority of our approach, which enhances the performance of both open-source and proprietary models. The results on AMBER show that visual evidence helps mitigate the hallucination of objects, relations, and attributes on discriminative tasks, *e.g.*, "Is there a cat in the image?". The CHAIR (lower is better) of LLaVA-1.5-7B decreases from 8.07 to 6.78 (−1.29), showing that visual evidence also reduces the hallucination of generative tasks, *e.g.*, "Describe this image". Qualitative results are shown in Appendix C.1.

3.3 Integrated with Existing Methods

To further validate the general applicability of our method, we choose to integrate it with two representative methods, LRV-Instruction (Liu et al., 2023a) and VCD (Leng et al., 2023). LRV-Instruction fine-tunes the model while VCD optimizes the decoding process of the model. We use the officially released checkpoint, codes and reproduce the results. The experiment results are displayed in Tab. 3. Compared with the baseline, LRV shows a decrease (−8.09%) and VCD presents a marginal improvement of less than 2%. Compared with LRV or VCD alone, the combination of VE with LRV or VCD further enhances the model's performance, resulting in fewer hallucinations. Notably, VCD

Model	POPE		AMBER			
	Acc.	Δ	CHAIR↓	Δ	Acc.	Δ
MiniGPT-4	71.16	-	14.13	-	64.28	-
+ LRV	63.07	-8.09	14.52	+3.90	50.04	-14.24
+ VE	**80.47**	+9.31	**13.63**	-0.5	**69.68**	+5.40
+ LRV + VE	**72.59**	+1.43	**11.13**	-3.00	**54.79**	-9.49
LLaVA-1.5-7B	79.73	-	11.84	-	74.94	-
+ VCD	81.10	+1.37	8.02	-3.82	76.79	+1.85
+ VE	**86.23**	+6.50	**10.25**	-1.59	**75.53**	+0.59
+ VCD + VE	**87.27**	+7.54	**7.39**	-4.45	**77.31**	+2.37

Table 3: Integrating VE with LRV and VCD.

Evaluation	Model	Acc. (%)
OOD Object	LLaVA-1.5-7B	66.22±0.38
	+ VE	**73.47±0.52**
	MiniGPT-4-v2	60.56±0.25
	+ VE	**66.10±0.24**
OOD Relation	LLaVA-1.5-7B	60.82±0.92
	+ VE	**62.81±0.49**
	MiniGPT-4-v2	61.84±0.48
	+ VE	**68.26±0.38**

Table 4: Hallucination evaluation on out-of-domain datasets.

+ VE achieves the best performance among three benchmarks, verifying the effectiveness and plug-and-play attribute of VE.

3.4 Ablation of Visual Models

Given that the small visual models used in this work are contrasted with large vision-language models, presenting results using various versions of visual models will ensure a fair and comprehensive comparison. We employ 6 object detection models with different architectures (Fang et al., 2021; Carion et al., 2020; Zhang et al., 2022), including open-vocabulary detection model (Minderer et al., 2022). The experimental results w.r.t. object evidence prompting are presented in Tab. 6. The experimental results of relation hallucination are presented in Tab. 11 in Appendix. Notably, the results demonstrate a positive correlation between the detection abilities of small visual models and the reduction of object hallucinations in LVLMs. This phenomenon can be ascribed to the fact that a good detection model provides high-quality object labels, which provides more accurate evidence. More analysis and ablation studies of the visual models, prompt templates are presented in the Appendix E.2, E.3 and E.4.

4 Analysis

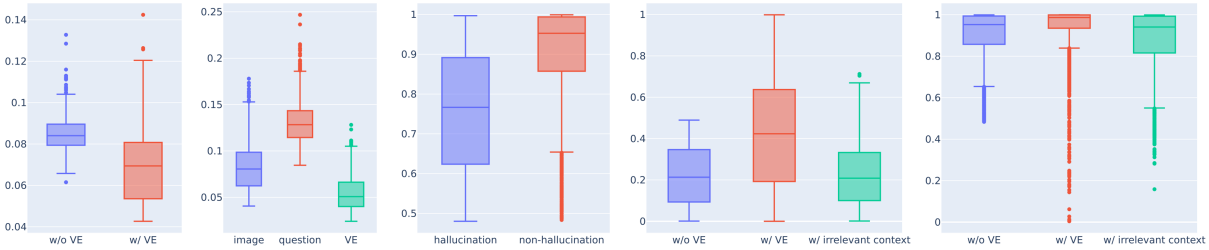
4.1 How Does VEP Work?

Firstly, we conduct analysis on the answer attribution. In Fig. 4(a), we delineate the image attribution scores before and after the incorporation of visual evidence, for hallucination samples where the query object is absent in the image. It is observed that the image attribution score significantly diminishes with visual evidence, implying a reduction in the model's incorrect focus on certain confusion objects/regions within the image, as analyzed in the introduction section. As depicted in Fig. 4 (b), the model attributes a lower score to the visual evidence compared to the image and the question, which suggests that the model synthesizes information from the image, question, and evidence in a comprehensive manner to arrive at the final answer. Upon further analysis of the attribution map after the integration of visual evidence, it is intriguing to note that the model ceases to erroneously focus on previously incorrectly attended object region. As illustrated in the bottom half of Fig. 1, following the incorporation of visual evidence, the model no longer misdirects its attention to the baseball bat during answer generation. Instead, it accurately attends to the region containing the baseball bat when predicting the token "bat". More cases can be found in Appendix C.3. Subsequently, we delves into the model's prediction confidence. We use the prediction probabilities from the model to measure the confidence of the model's predictions, which is a common method (Geng et al., 2024; Hinton, 2015). As show in Fig. 4(c), we observe that, for hallucination samples, the model's confidence is notably lower compared to non-hallucination ones (0.76 vs 0.90). Upon incorporating visual evidence, there is a significant enhancement in the model's confidence in predicting right answers for hallucination (Fig. 4(d)). Intriguingly, even in non-hallucination samples, the model's confidence in predicting the right answers witnesses considerable improvement (Fig. 4(e)).

These intriguing phenomena suggest that the incorporation of visual evidence can guide the model in dynamically adjusting and rectifying the focus on image regions, diminishing the degree of confusion in visual information (suppressing erroneous activations while reinforcing correct activations). This process enables the model to acquire visual context or knowledge with higher confidence prior to generating the final answers.

4.2 Extra Hints or Complete Replacement for Visual Information?

To further investigate that providing visual evidence whether offers extra visual hints to the model, or dominates the image itself, we randomly replaced the original input image with a different



(a) Image attribution score w/ and w/o VE (b) Attribution score of image, question and visual evidence (c) Prediction probabilities of the first token w/ and w/o hallucination (d) Prediction probabilities for true label w/ and w/o VE on hallucination samples (e) Prediction probabilities for true label w/ and w/o VE on non-hallucination samples

Figure 4: Analysis about how VEP works.

LVLM	Acc.(%) Original image	Acc.(%) Random image
LLaVA-1.5-7B	87.43	80.40
LLaVA-1.6-7B	89.43	69.30

Table 5: Results of random images.

image from the COCO2014 validation set.[2] The results are shown in Tab. 5, reveal a significant decrease in the model's performance when the images are substituted with unrelated random images. This suggests that upon integration of visual evidence, the model does not merely rely on textual information but synthesizes both textual and visual data to arrive at the final decision. This observation aligns with the phenomena analyzed in Section 4.1.

4.3 Mitigating Hallucinations beyond Objects

In LVLM, hallucinations not only include object hallucinations but also relation hallucinations. The evaluation for relation hallucinations in AMBER is somehow coarse-grained. To further verify whether our method can also address fine-grained relation hallucinations, we conduct extensive experiments on a new relation hallucination benchmark RPE using the RelTR (Cong et al., 2023) vision model. The results are shown in the second last column of Tab. 2 and Tab. 9 in Appendix. Firstly, our method effectively reduces the relation hallucination for most of the models. Secondly, compared with object, it is evident that the performance of all models including GPT-4V is inadequate in terms of relation. This may be because understanding the fine-grained visual relations in an image first requires the comprehension of objects, which is a more difficult capability for the model (limitation of GPT-4V). The fine-grained results across different relationship categories are shown in Appendix D.2. In Appendix C.7 we also demonstrate that other fine-grained tasks like object counting and OCR can also be enhanced effectively by our method. These outcomes strongly validate the effectiveness and versatility of our proposed framework.

[2]Employing purely language-based models such as GPT-4 for responding to visual queries leads a refusal response such as "I'm sorry, but I can't help with identifying or making assumptions about content in images or assist with questions about unseen images."

4.4 Results on Open-world Scenario

An important concern is whether visual evidence prompting is able to generalize to the open-vocabulary objects and relations. As a first step toward investigating this question, we conduct evaluations on out-of-domain datasets. Specifically, we collect 2,540 samples from another two object detection (Object365 (Shao et al., 2019)) and scene graph generation (OpenImage (Kuznetsova et al., 2020)) datasets for quantitative analysis. Tab. 4 presents the comparison with baseline results for the evaluation on out-of-domain datasets. We also follow CLIP (Radford et al., 2021) and randomly select 2 samples (one for object hallucination and another for relation hallucination) from 10 open-world out-of-domain datasets for qualitative analysis. These 20 cases are in Appendix C.6. The results indicate that using incomplete visual evidence can still mitigate the hallucination of objects and relations effectively. Compared to the *recall* of the objects and relations in the visual evidence, the *precision* of small visual models is more important for the proposed method.

4.5 Evaluation on General Multimodal Understanding Tasks

As the hallucination is mitigated, it is expected that the LVLMs will exhibit enhanced precision in responding to user questions related to the images. Consequently, an LVLM incorporated with visual evidence is likely to achieve superior results on benchmarks designed to assess the ability of general multimodal understanding. To verify this, we evaluate baseline models and models

LVLM	Visual model Model name	mAP	Acc. (%)
LLaVA-1.5-7B	-	-	80.23
	yolos-tiny	28.7	**84.13**
	owlvit-base-patch16	30.3	**84.63**
	yolos-small	36.1	**85.50**
	detr-resnet-50	42.0	**87.50**
	detr-resnet-101	43.5	**87.43**
	DINO-4scale-swin	58.0	**88.00**

Table 6: Object hallucination results of incorporating visual evidence from different object detection models.

Model	MME		MMBench	
	Scores↑	Δ	Acc.	Δ
MiniGPT-4	904.7	-	53.95	-
+ VE	**1086.4**	+181.7	**56.10**	+2.15
LLaVA-1.5-7B	1756.9	-	74.48	-
+ VE	**1819.6**	+62.7	**75.34**	+0.86
LLaVA-1.6-7B	1660.4	-	75.60	-
+ VE	**1729.5**	+69.1	**76.63**	+1.03

Table 7: Results on general multimodal understanding benchmarks.

incorporated with visual evidence on two multimodal benchmarks, *i.e.*, MME (Fu et al., 2023) and MMBench (Liu et al., 2023b), which measure comprehensive VQA capabilities and perceptual and reasoning abilities. The results are shown in Tab. 7, which demonstrate that the incorporation of visual evidence into the model yields a modest enhancement in the model's overall general multimodal capabilities. This relatively minor improvement could be attributed to the fact that generic multimodal evaluation datasets typically include a wide array of assessing dimensions, and the visual evidence acquired by small vision models is unlikely to assist in every aspect. Nonetheless, the experiment results suggest that our approach can mitigate hallucinations without compromising the model's comprehension abilities across other dimensions.

5 Related Works

5.1 Hallucinations in Large Language Models

The extraordinary capabilities of large language models (LLMs) come with a significant drawback: their potential to generate unsupported text due to their lack of understanding of what is factual and what is not (Maynez et al., 2020; Krishna et al., 2021; Longpre et al., 2021). As a result, there has been a surge of interest in addressing LLM hallucination through knowledge-grounded neural language generation. To address this limitation, various works augment LLMs with knowledge consisting of personalized recommendations (Ghazvininejad et al., 2017), Wikipedia articles and web search (Dinan et al., 2018; Shuster et al., 2022), structured and unstructured knowledge of task-oriented dialog (Peng et al., 2022). In the LVLMs, it is infeasible to acquire grounded knowledge from a general knowledge base.

5.2 Hallucinations in Large Vision-Language Models

Similar to LLMs, LVLMs tend to generate nonexistent contents in a target image. In the literature of computer vision field (Rohrbach et al., 2018; Biten et al., 2021). object hallucination refers to the model generating descriptions or captions that contain objects that are inconsistent with or even absent from the target image. In general, object hallucination can be defined at different semantic levels. In this work, we focus on coarse-grained object hallucinations and fine-grained relation hallucinations at the same time, other hallucinations such as the number of the object are left for future work. In previous works (Li et al., 2023b), POPE is proposed to evaluate object hallucinations in LVLMs by polling questions about object existence. (Gunjal et al., 2023) created a hallucination dataset and optimized the LVLM over the dataset with a variation of Direct Preference Optimization (Rafailov et al., 2023). These studies collectively contribute to the understanding and mitigation of hallucination-related challenges in LVLMs, by providing evaluation metrics, datasets, and tuning methods that enhance the reliability and consistency of the generated answers. Yet, there is a risk of overly optimizing the model to fit a specific problem or dataset, leading to catastrophic forgetting and lack of generalization ability (Zhai et al., 2023).

6 Conclusion

We have explored visual evidence prompting as a simple and broadly applicable method for mitigating hallucinations in large vision-language models. Through experiments on 9 models and various benchmarks, we find that visual evidence is an effective, robust, and general cure for large vision-language models. Broadening the range of visual evidence will hopefully inspire further work on external-knowledge-based approaches for improving the behaviors of LVLMs.

7 Limitation

While our work sheds light on hallucination reduction, there are several limitations to our work. 1) Unlike fine-tuning, the prompting strategy is not able to embed new knowledge into the foundation models. Previous work (Zhai et al., 2023) showed that as the fine-tuning progresses, model starts to hallucinate by disregarding the questions and exclusively generating text based on the examples in the fine-tuning datasets. On the other hand, prompt-based method does not modify the parameters of the model, and offer greater controllability, which is advantageous for preserving the model's original generalization capability. 2) The introduction of new visual models leads to somewhat efficiency degradation. The extra cost of our method is much less than which of other methods. For example, the instruction-tuning-based methods are data- and computation-intensive, the multi-step-generation-based schemes involves additional cost of inferring large visual-language models or calling APIs. Moreover, the extra cost of our method can be further alleviated by paralleling the inference of visual models, as well as the vision encoder of LVLM. 3) The quality of the visual evidence affects the effectiveness of our method. In addition to the 9 small models mentioned previously, we recommend utilizing popular and thoroughly validated small models. 4) It is known that the design of prompt is a delicate and experience-based process. Although we have conducted experiments to verify the robustness against prompt templates, different prompts still inevitably perturb the effectiveness of the proposed method. 5) The verification of the effectiveness of visual evidence prompting is mainly conducted on mitigating hallucinations of objects, attributes and relations. We only conduct proof-of-concept experiment to evaluate the effectiveness on for OCR and object counting of LVLMs the with the same methodology. In summary, we do not view our findings as belief that our specific method can solve all instances, but as demonstration of the general potential of combining the open capacity and robustness of large models with the expertise of small models, and utilizing the domain-specific knowledge of small models to complement the large models.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2019. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. *CVPR*, pages 9687–9695.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, pages 2011–2018.

Ali Furkan Biten, Lluís Gómez i Bigorda, and Dimosthenis Karatzas. 2021. Let there be a clock on the beach: Reducing object hallucination in image captioning. *WACV*, pages 2473–2482.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. *CoRR*, abs/2005.12872.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023. Valor: Vision-audio-language omni-perception pretraining model and dataset. *ArXiv*, abs/2304.08345.

Adam Coates, A. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*.

Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. 2023. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *ArXiv*, abs/1811.01241.

Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. You only look at one sequence: Rethinking transformer in vision through object detection. *CoRR*, abs/2106.00666.

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, William B. Dolan, Jianfeng Gao, Wen tau Yih, and Michel Galley. 2017. A knowledge-grounded neural conversation model. In *AAAI*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Anish Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. *ArXiv*, abs/2308.06394.

John Henderson and Fernanda Ferreira. 2013. *The interface of language, vision, and action: Eye movements and the visual world*. Psychology Press.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS*, 33:2611–2624.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *North American Chapter of the Association for Computational Linguistics*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Li Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *ArXiv*, abs/2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? *ArXiv*, abs/2307.06281.

Shayne Longpre, Kartik Kumar Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *ArXiv*, abs/2109.05052.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *ArXiv*, abs/2302.07842.

Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers. *CoRR*, abs/2205.06230.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

Stephen E Palmer. 1999. *Vision science: Photons to phenomenology*. MIT press.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Lidén, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. Godel: Large-scale pre-training for goal-directed dialog. *ArXiv*, abs/2206.11309.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *ArXiv*, abs/2305.18290.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, W.K.F. Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *ArXiv*, abs/2208.03188.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:652–663.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *ArXiv*, abs/2311.07397.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE.

Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic scene graph generation. In *ECCV*, pages 178–196. Springer.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *ArXiv*, abs/2309.17421.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xingguo Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *ArXiv*, abs/2310.16045.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Y. Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *Preprint*, arXiv:2203.03605.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Definition of Attribution Score

The attribution score is computed following the interpretability method (Chefer et al., 2021), similar to the commonly used Grad-CAM (Selvaraju et al., 2017) in computer vision. It indicates the importance of each preceding token in predicting the current token, with a higher score denoting greater importance. The specific calculation method is as follows.

Firstly, we initialize the token attribution map R as an identity matrix, the dimensions of which correspond to the size of the attention matrix at each layer of the text encoder. Subsequently, we compute the gradients of the attention weights by leveraging the prediction logit of current output token and average them across all attention heads. This procedure yields an attribution map $\bar{\mathbf{E}}_\mathbf{i}$ of current output token for each layer i.

$$\bar{\mathbf{E}}_\mathbf{i} = \sum_{j=1}^{h} (\nabla \mathbf{A}_\mathbf{j}^\mathbf{i} \odot \mathbf{A}_\mathbf{j}^\mathbf{i})^+, \tag{1}$$

where $\odot$ is the Hadamard product, $\mathbf{A}_\mathbf{j}^\mathbf{i}$ denote the attention matrix of the head j in layer i, $\nabla \mathbf{A}_\mathbf{j}^\mathbf{i} := \frac{\partial logit_current_token}{\partial \mathbf{A}_\mathbf{j}^\mathbf{i}}$ for $logit_current_token$ which is the prediction logit of current output token such as the first output token "Yes".

Finally, we aggregate the explainability maps of all layers using the propagation rule as presented in (Chefer et al., 2021) to derive the final text attribution map.

$$\mathbf{R} \leftarrow \mathbf{R} + \bar{\mathbf{E}}_\mathbf{i} \cdot \mathbf{R}. \tag{2}$$

Then we can extract the attribution scores of every image token, which also formulate a image attribution map.

B Further Details for Experimental Setup

B.1 Datasets

POPE. In POPE (Li et al., 2023b), 500 images were randomly selected from the validation set of the MSCOCO (Vinyals et al., 2016), each containing more than three ground-truth objects in the annotations. For every image, six questions were formulated based on the annotations, with answers limited to "Yes" or "No". For questions with the answer "No", three strategies, *i.e.*, Random, Popular, and Adversarial, are used to sample their probing objects. The difficulty of the questions increased progressively from Random to Adversarial. For MSCOCO-Random, objects not present in the image are randomly chosen. For MSCOCO-Popular, the top-3% most frequent objects in the MSCOCO are selected. For MSCOCO-Adversarial, objects are ranked based on their co-occurring frequencies with the ground-truth objects, and the top-k most frequent objects not present in the image were chosen.

AMBER. AMBER (Wang et al., 2023) is a comprehensive benchmark for assessing hallucination in LVLMs free from LLMs, which includes a variety of tasks that gauge the models' abilities in both discriminative (*e.g.*, "Is there a dog in the image?") and generative questions (*e.g.*, "Describe this image."), comprising a total of $15,220$ questions. The images in AMBER are carefully collected by the authors, which have not been used for training LVLMs, featuring clear content with well-defined objects. The authors have thoroughly annotated these images and have manually constructed some prompt templates to generate questions for evaluating different types of hallucinations, including object hallucination (*e.g.*, "Is there a {object} in this image?"), attribute hallucination (*e.g.*, "Does the {object} {action} in this image?"), and relation hallucination (*e.g.*, "Is there direct contact between the {object 1} and {object 2} in this image?"). Among the total questions, $1,004$ are generative questions, and the remaining are discriminative questions.

RPE. In order to further verify the effectiveness of our method in mitigating relation hallucination, we meticulously follow the same recipe as POPE on Visual Genome (Krishna et al., 2017) to construct a new relation hallucination evaluation dataset named Relation Probing Evaluation (RPE). Firstly, the 50 relation categories of VG are categorized into two groups, spatial and action relationships. Then we select 7 representative spatial relations (above, at, behind, in, in front of, on, and under) and 9 head action relations (carrying, eating, holding, lying on, looking at, riding, sitting on, standing on, and walking on), while the other tail relations are ignored. For each relation, we randomly select 75 images with questions whose answers are "Yes" and 75 images questions whose the answer are "No". Each "Yes" questions are constructed from annotations. For questions with the answer "No", the probing relations are randomly selected within the corresponding group of spatial or action relations with additional added negative relation, which is shown in the Tab. 8. To ensure not select

synonyms of the ground truth as probing relations, we carefully devise several pairs of synonymous relations as the "blacklist" as shown in the Tab. 8. The negative samples were assigned different relations from the positive samples, randomly chosen from the groups of spatial and action relations. To ensure accurate annotation of the negative samples, we devised pairs of synonymous relations and avoided selecting synonyms of the positive sample relation when choosing the relation for the negative sample. In summary, this dataset consists of 2400 triplets of image, question and answer, in which 1200 are "Yes" and 1200 are "No". In Fig. 5, we show some cases in our dataset.

B.2 Details of Evaluating SOTA Models

We tasked GPT-4V (model version: gpt-4-1106-vision-preview) to answer the questions in the three benchmarks via the official API provided by OpenAI. We use the official checkpoints for MiniGPT-4, MiniGPT-4-v2, Qwen-VL-Chat, LLaVA-1.5 and LLaVA-1.6. For LLaVA-1.5, we use the llava-v1.5-7b and llava-v1.5-13b. For LLaVA-1.6, we use the llava-v1.6-vicuna-7b and llava-v1.6-vicuna-13b. For Qwen-VL-Max, we access it via the official API provided by Alibaba Cloud.

C Qualitative Results

C.1 Effect of Visual Evidence

Fig. 6 shows a comparison of responses from the LLaVA-1.5-7B before and after the integration of visual evidence. It is apparent that, prior to incorporating visual evidence, the model may hallucinate non-existent objects or relations in the images, or fail to recognize objects that are present. However, after combining the visual evidence provided by small visual model, the model is able to correctly answer the questions. Additionally, for open-ended generative tasks, hallucinated objects present in the model's initial responses have disappeared after incorporating the visual evidence.

C.2 Comparison between hallucination and non-hallucination samples

In Fig. 7, we present more comparative of image attribution maps across both hallucination and non-hallucination samples. Observations suggest that the hallucination behaviors of model are not arbitrary but are triggered by regions within the images that possess a certain degree of resemblance in appearance or semantic meaning, leading to erroneous activations, such as the confusion between lampposts and traffic lights, backpacks and handbags, as well as cars and trucks.

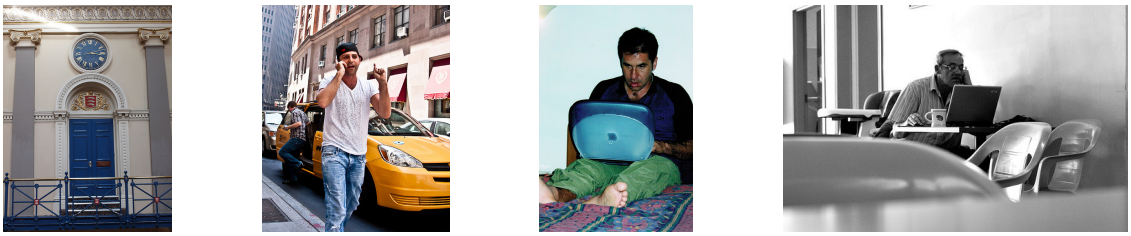
C.3 Comparison before and after the incorporation of VE

In Fig. 8, we present a comparison of the image attribution maps before and after the incorporation of visual evidence. It is evident from the comparison that, following the integration of visual evidence, the model no longer attends to the previously incorrectly focused regions, and then arriving at the correct answer.

C.4 Error Analysis

We conduct a detailed error analysis on POPE utilizing LLaVA-1.5 with visual evidence. We meticulously annotate each incorrect prediction produced by our method in the test set and systematically categorize these errors into 5 types (please refer to Appendix C.5 for error examples of each type): (1) *Annotation noise (*29.2%*)*: Objects being annotated as present in the image when they are not, or vice versa. (2) *Inconspicuous object undetected (*7.7%*)*: The object is present in the image but is not prominent enough to be detected by the visual model, and might be difficult to spot at first glance even by humans. (3) *Context ignored by LVLM (*58.0%*)*: The LVLM does not fully take into account the visual evidence provided to it. (4) *Spurious detection of non-existent object (*3.3%*)*: Visual model incorrectly detects something that does not actually exist. (5) *Semantically similar word confusion (*1.8%*)*: The visual model correctly identifies an object but outputs object names that are the synonyms to the objects appear in the questions (*e.g.*, confusing "oven" with "toaster", "skis" with "snowboard", "car" with "truck").

Among five types of errors, context ignored by LVLM (58.0%) represents the majority. A further analysis of the prediction confidence for these samples reveals that the initial error confidence of inaccurate responses is higher than that of samples whose errors are successfully corrected (0.83 vs. 0.66). Moreover, despite the incorporation of evidence failing to rectify the wrong answer, a notable decrease in error confidence level is observed (from 0.83 to 0.73). This indicates that the model displays certain deficiencies in integrating image information with external visual knowledge, particularly when handling hard samples. Inconspicuous object undetected (7.7%) reveal limitations



Positive question: Is the clock above the door?
Label: yes

Negative question: Is the clock behind the door?
Label: no

Positive question: Is the cap on the head?
Label: yes

Negative question: Is the cap under the head?
Label: no

Positive question: Is the man sitting on the bed?
Label: yes

Negative question: Is the man jumping over the bed?
Label: no

Positive question: Is the man looking at the laptop?
Label: yes

Negative question: Is the man holding the laptop?
Label: no

Figure 5: Several cases in RPE are depicted in this figure, with the two on the left representing spatial relations and the two on the right illustrating action relations.

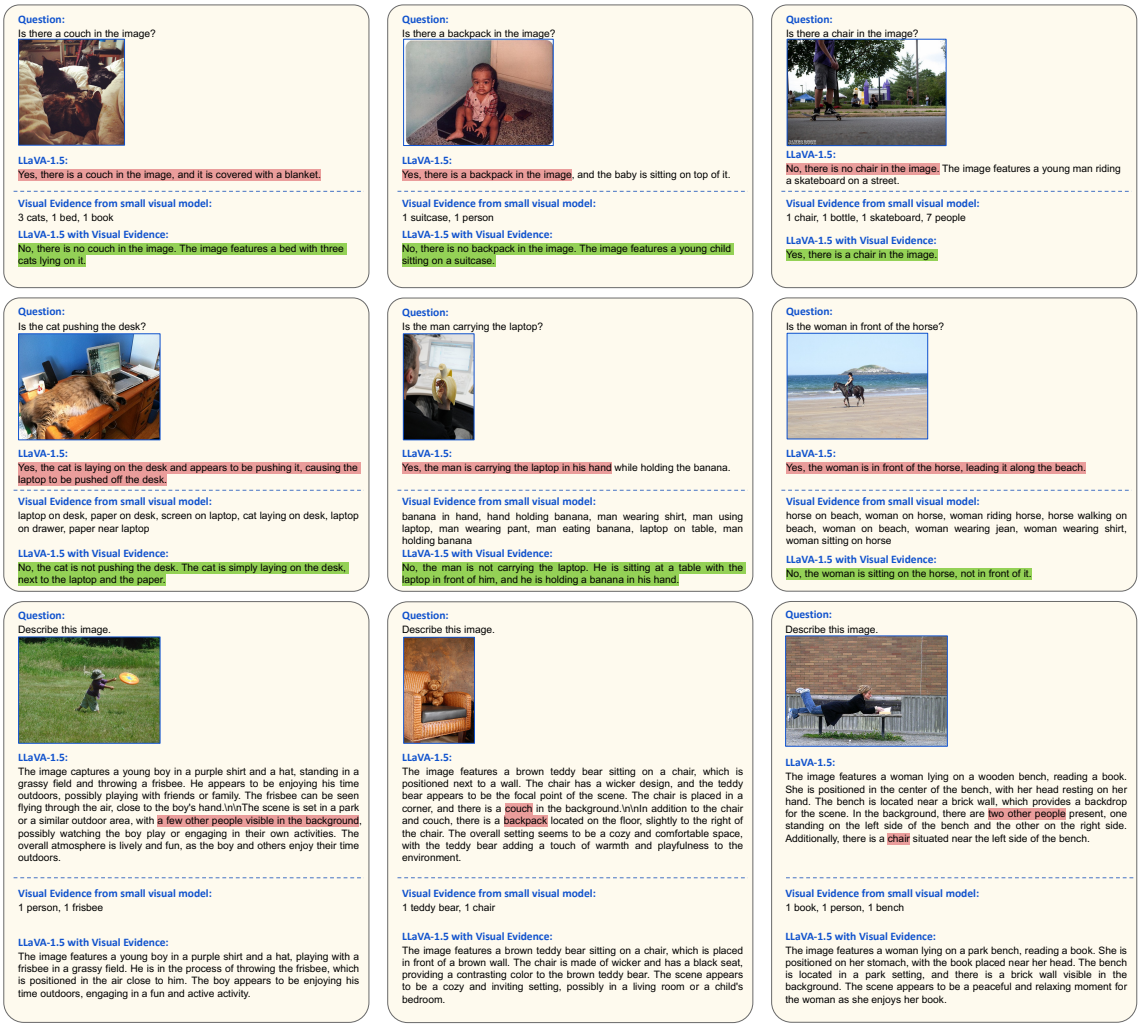


Figure 6: Some cases demonstrating that visual evidence prompting effectively reduce hallucination of LLaVA-1.5-7B.

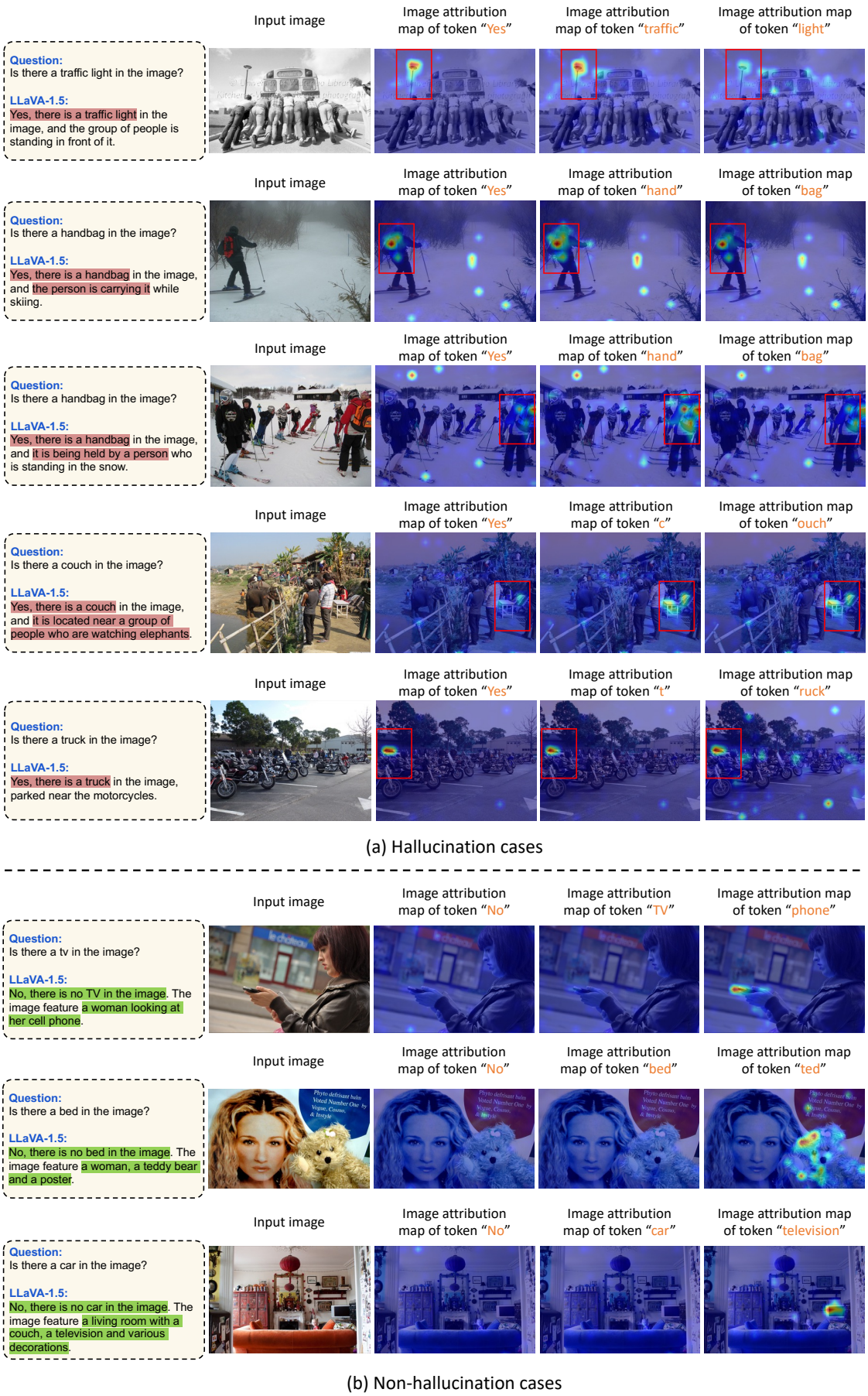


Figure 7: More qualitative comparison of image attribution maps between hallucination samples and non-hallucination ones.

Table 8: The negative relations candidate set used to contruct negative question are shown here. We also present the synonymous pairs used to ensure not select synonyms of the ground truth as probing relations

Relation type	Negative relations	Synonymous pairs
Spatial relation	**above, at, behind, in, in front of, on, under** *at the left of, at the right of*	above: {on} on: {above}
Action relation	**carrying, eating, holding, lying on, looking at, riding, sitting on, standing on, walking on** *walking in, watching, cutting, feeding, leaning on, jumping over, hugging, kissing, pushing, pulling, washing, kicking, draging*	walking on: {walking in, standing on} looking at: {watching}

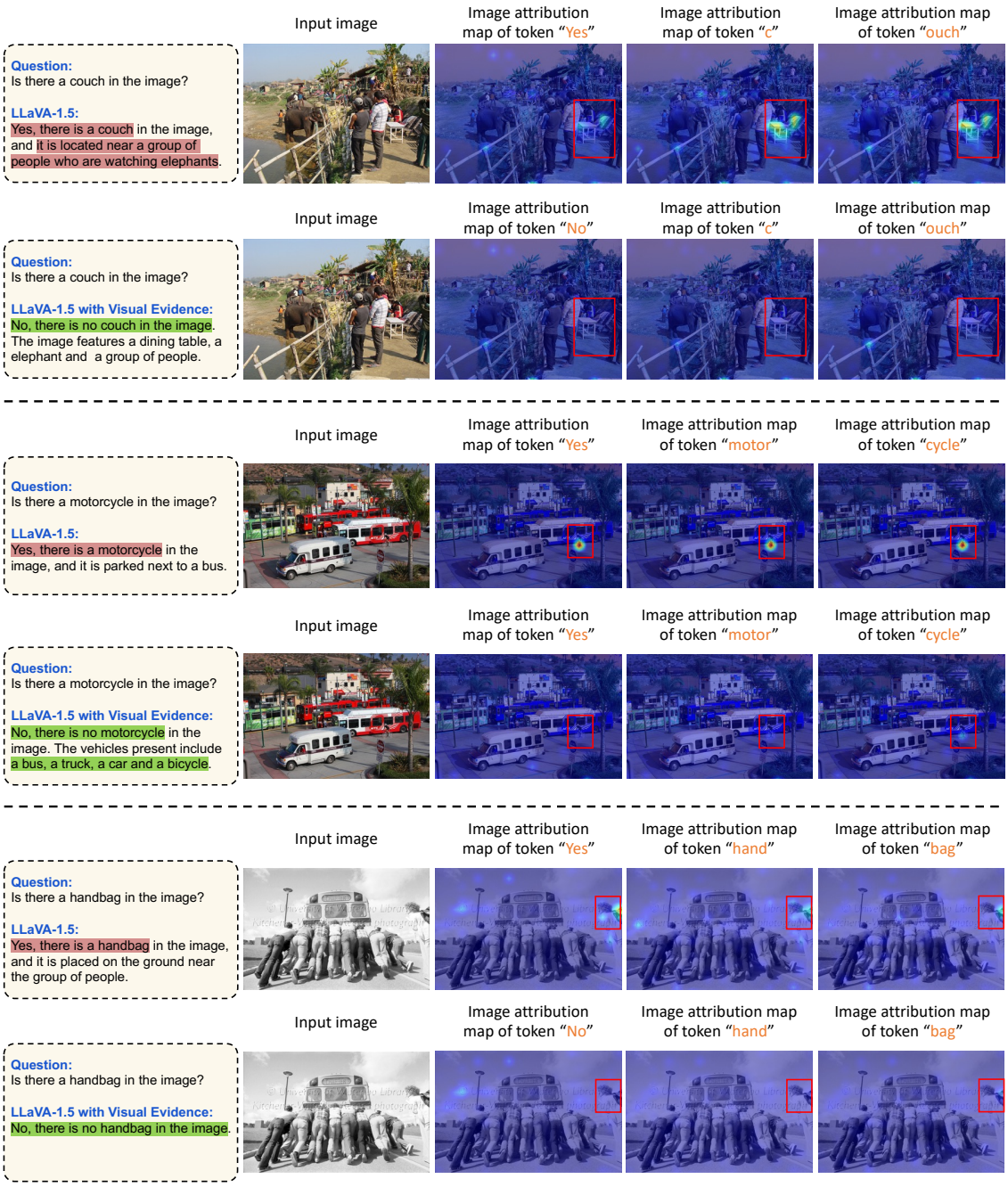


Figure 8: More qualitative comparison of image attribution maps before and after the integration of visual evidence.

of detecting inconspicuous objects from small visual models. Semantically similar word confusion (1.8%) and spurious detection of nonexistent objects (3.3%) are less, indicating that the inaccurate information introduced by small visual models are not common and the inaccurate information in the visual evidence is negligible.

C.5 Error Examples

In Fig. 9, we show some error examples of each error type. For example, in the case of the error type "annotation noise", an image with no car present is incorrectly labeled as "yes" (the correct answer is "no", as the question asked, "Is there a car in the image?"). In the case of the error type "inconspicuous object undetected", the small visual model fail to detect a cellphone held by a passenger on a bus, leading to an incorrect answer from the model. In the case of the error type "context ignored by LVLM", the small visual model's output is correct, but the model ignore the context and persisted with its original incorrect answer. In the case of the error type "spurious detection", the small visual model incorrectly detect a car that is not actually present in the image. This may have been caused by a blurry object in the background that slightly resembled the shape of a car. In the case of the error type "semantically similar word confusion", the small visual model correctly identify an oven, but the question is about a toaster, causing confusion to the model.

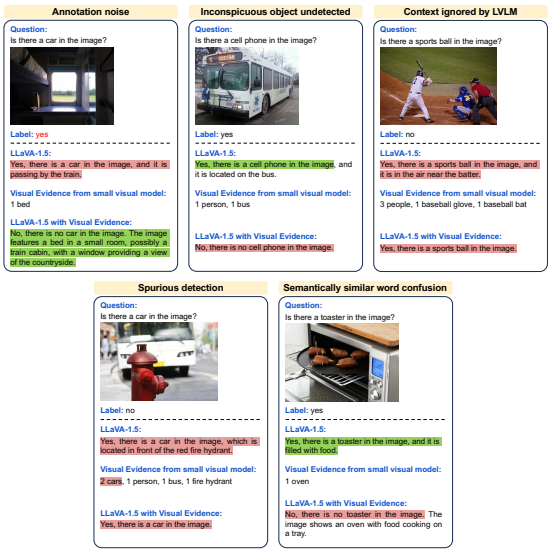


Figure 9: Examples of each error type.

C.6 More Cases on out-of-domain images

Following the idea of CLIP (Radford et al., 2021), we selected 10 out-of-domain datasets from the 27 datasets used to test the zero-shot generalization performance of CLIP. These 10 datasets are Caltech-101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), Birdsnap (Berg et al., 2014), Flowers102 (Nilsback and Zisserman, 2008), CLEVRCounts (Johnson et al., 2017), Country211 (Radford et al., 2021), Food101 (Bossard et al., 2014), SUN397 (Xiao et al., 2010), HatefulMemes (Kiela et al., 2020), and STL10 (Coates et al., 2011). Then, we randomly selected two images from each dataset, one for evaluating object hallucination and the other for evaluating relation hallucination. As shown in Fig. 10 and Fig. 11, we can see that even when providing incorrect visual evidence to the model, it still maintains its original correct answer, which further verifies the model's robustness and adaptivity to incorrect evidence in open scenarios.

C.7 Some Cases on Object Counting and OCR

In Fig. 12 and Fig. 13, we also demonstrate some cases of GPT-4V that other tasks like object counting and OCR can also be enhanced effectively by our method. For instance, consider the rightmost image in the second row of Fig. 12, which unambiguously depicts five dogs and four people. When queried regarding the number of people and dogs present in the image, GPT-4V erroneously stated, "There are five people and six dogs." However, the object detection small visual model can provided accurate visual evidence of "5 dogs, 4 people". With this visual evidence, GPT-4V amended its statement to "There are four people and five dogs in the image." Similarly, in the leftmost image of the first row in Fig. 13, a vehicle's license plate reads "OZL7H33". GPT-4V's initial response about the license plate number inaccurately reported the sequence as "OZL733", neglecting the "H". Yet, upon utilizing the OCR small visual model, the visual evidence "OZL7H33" was yielded accurately. After incorporating the visual evidence, GPT-4V was able to answer correctly ("The license plate number of this car is OZL7H33.")

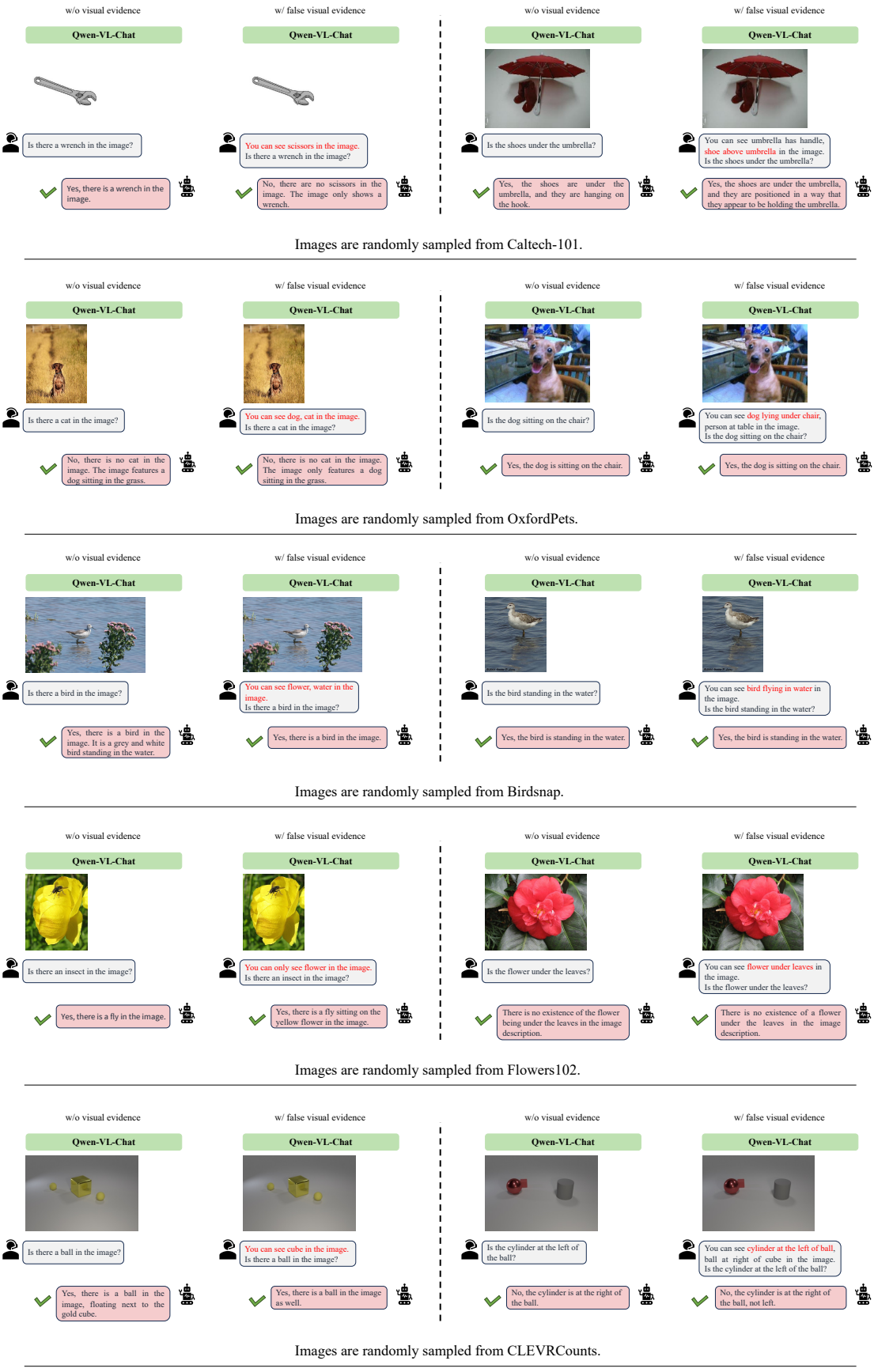


Figure 10: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.

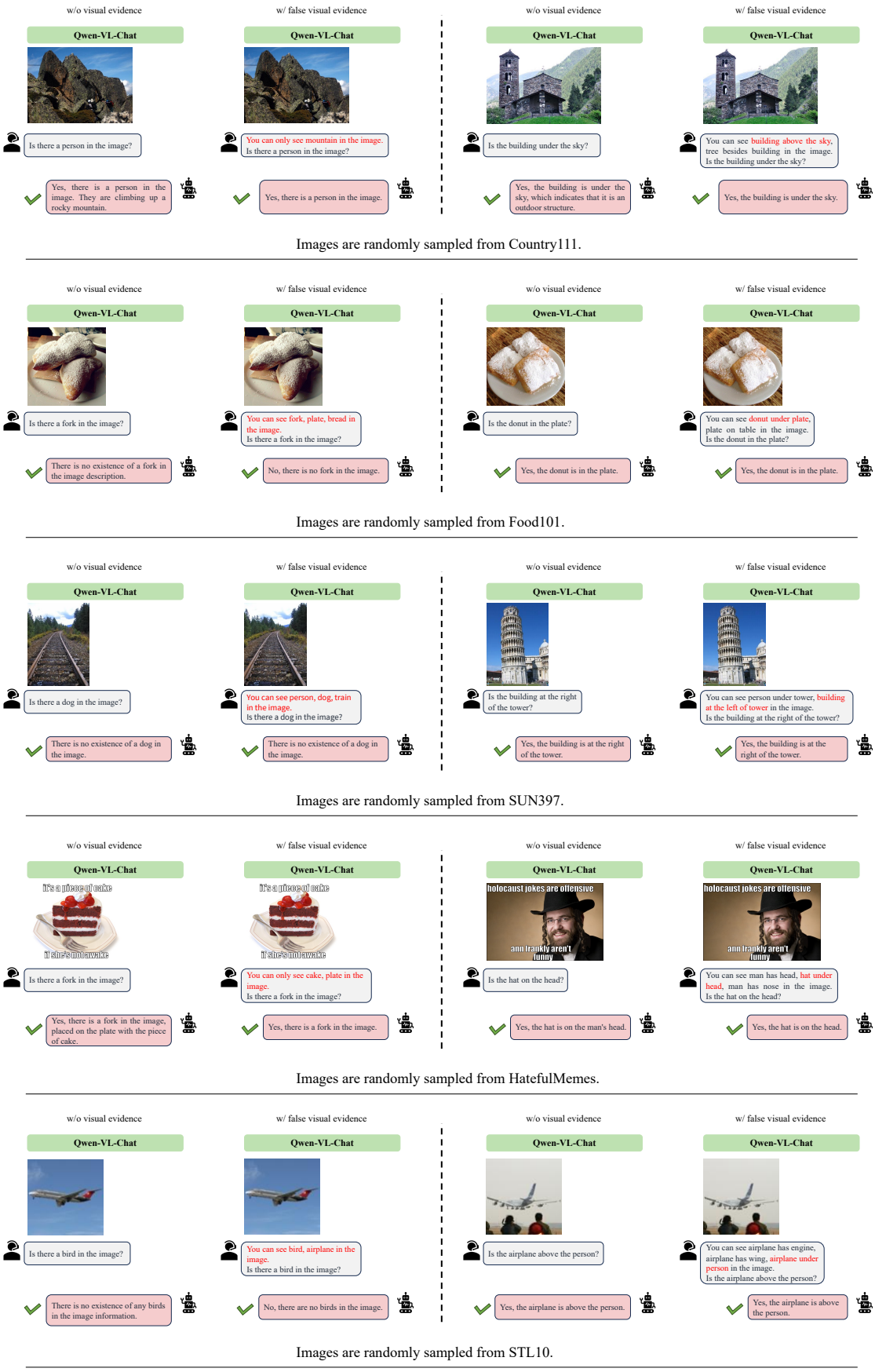


Figure 11: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.

Question:
Is there a couch in the image?

GPT-4V:
There are nine bottles in the image.

Visual Evidence from small visual model:
8 bottles

GPT-4V with Visual Evidence:
There are 8 bottles in the image.

Question:
How many people are there in the image?

GPT-4V:
There are nine people in the image.

Visual Evidence from small visual model:
10 people, 2 benches, 1 chair

GPT-4V with Visual Evidence:
There are 10 people in the image.

Question:
How many apples are there in the image?

GPT-4V:
There are six apples in the image.

Visual Evidence from small visual model:
8 apples, 1 bowl

GPT-4V with Visual Evidence:
There are 8 apples in the image.

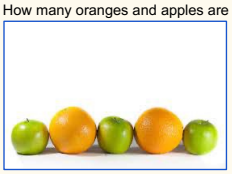
Question:
How many cats and dogs are there in the image?

GPT-4V:
In the image, there are three cats and four dogs.

Visual Evidence from small visual model:
2 cats, 4 dogs

GPT-4V with Visual Evidence:
In the image, there are 4 dogs and 2 cats.

Question:
How many oranges and apples are there in the image?



GPT-4V:
There are three oranges and two apples in the image.

Visual Evidence from small visual model:
3 apples, 2 oranges

GPT-4V with Visual Evidence:
There are a total of 5 fruits in the image, with 3 apples and 2 oranges.

Question:
How many people and dogs are there in the image?

GPT-4V:
In the image, there are five people and six dogs.

Visual Evidence from small visual model:
5 dogs, 4 people

GPT-4V with Visual Evidence:
There are four people and five dogs in the image.

Figure 12: Some cases in object counting task when applying VEP to GPT-4V.

Question:
What is the license plate number of this car?

GPT-4V:
OZL733

Visual Evidence from small visual model:
OZL7H33

GPT-4V with Visual Evidence:
The license plate number of this car is OZL7H33.

Question:
What is the license plate number of this car?

GPT-4V:
01A26FG

Visual Evidence from small visual model:
018266FG

GPT-4V with Visual Evidence:
The license plate number of this car is 018266FG.

Question:
What is the license plate number of this car?

GPT-4V:
431 U 10857

Visual Evidence from small visual model:
431K11856

GPT-4V with Visual Evidence:
431K11856

Question:
What is the license plate number of this car?

GPT-4V:
4URT725

Visual Evidence from small visual model:
8UBK725

GPT-4V with Visual Evidence:
The license plate number of this car is "8UBK725."

Question:
What is the license plate number of this car?

GPT-4V:
MK62 ACZ

Visual Evidence from small visual model:
WK62ACZ

GPT-4V with Visual Evidence:
The license plate number of the car is "WK62 ACZ."

Question:
What is the license plate number of this car?

GPT-4V:
SH461A

Visual Evidence from small visual model:
SH4161A

GPT-4V with Visual Evidence:
The license plate number of this car is SH4161A.

Figure 13: Some cases in OCR task when applying VEP to GPT-4V.

D Detailed Results and Analysis

D.1 Main Results of Other Five Models

In Tab. 9, we present the experiment results of five additional models beyond the main text on the POPE, AMBER, and RPE benchmarks. The results clearly illustrate that, across both discriminative and generative tasks, the integration of our method significantly enhances performance for models of varying sizes and architectures. This highlights the plug-and-play attribute and effectiveness of our approach.

D.2 Fine-grained Results on RPE

In this section, we comprehensively demonstrate and analyze the model's performance across diverse relationship categories. In Fig. 14, the performance of LLaVA-1.5 with and without corresponding visual evidence is presented for each relationship category in RPE, where spatial relationships are depicted on the left and action relationships on the right. Based on the depicted results, it is evident that LLaVA-1.5 exhibits varying degrees of improvement across different relationship categories with the integration of visual evidence. Notably, a significant enhancement is observed in the action relationship category. Overall, the model outperforms the spatial relationship in the context of action relationships. This discrepancy could be attributed to the finer-grained nature of spatial relationships within images, which demand a higher level of comprehension capability.

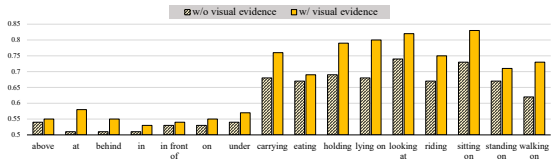


Figure 14: The effect of incorporating visual evidence on the performance of LLaVA-1.5-7B across different relation categories in RPE has been presented in this figure.

E More Experiment Results

E.1 More Results on POPE

More quantitative results on POPE COCO-Popular and COCO-Random are shown in Tab. 10. The results also indicate that the incorporation of visual evidence leads to significant enhancements in all models across both two datasets, providing further validation of the efficacy of our method.

E.2 Ablation of Visual Models

E.2.1 Performance Correlation of LVLM and Small Models

In Fig. 15, we demonstrate the performance changes of LLaVA-1.5 after incorporating visual evidence from small visual models with varying capabilities. Overall, as the performance of the small visual model increases, so does the performance of LLaVA-1.5, showing a positive correlation. However, when the performance of the small visual model exceeds 40.0, the gains in LLaVA-1.5's performance become marginal and a saturation trend begins to appear. This implies that there is a limit to the effectiveness of a certain type of evidence. At this point, it may be more valuable to explore new types of visual models to gather new kinds of evidence, *e.g.*, the ones with relations, attributes or other detailed descriptions. These findings suggest that a highly proficient capable visual model is not a prerequisite for our method and small visual model with decent performance can greatly mitigate hallucinations of LVLM. As the visual evidence produced by small visual model serves as complementary information for LVLM, the precision of the compact visual model takes precedence over its recall due to the robustness and its capacity of large models to synergistically and adaptively integrate parametric knowledge with contextual external knowledge (Neeman et al., 2023).

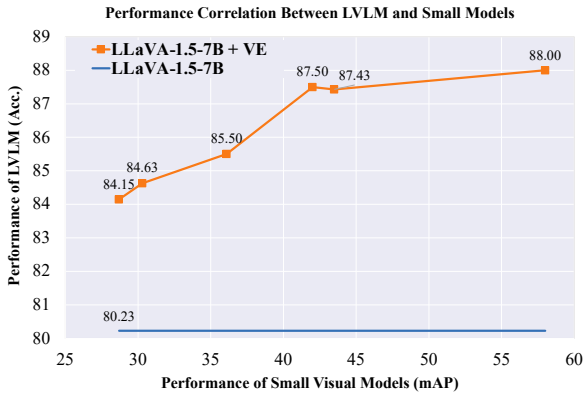


Figure 15: The performance correlation between LVLM and small visual models. Specifically, the metric used to evaluate the smaller visual models' performance is the mAP on the COCO 2017 validation set. The performance of LVLM is the accuracy on POPE.

E.2.2 Comparison between Open-set and Close-set Visual Model

In Tab. 6 of the main text, we employ an open-vocabulary detection model for ablation studies.

Model	POPE		AMBER				RPE		Latency token/sec.
	Acc.	Δ	CHAIR↓	Δ	Acc.	Δ	Acc.	Δ	
MiniGPT-4	71.16	-	14.13	-	64.28	-	65.11	-	24.33
+ Visual Evidence	**80.47**	+9.31	**13.63**	-0.50	**69.68**	+5.40	**72.03**	+6.92	22.64
Qwen-VL-Chat	82.33	-	7.30	-	80.80	-	64.08	-	20.47
+ Visual Evidence	**88.17**	+5.84	**6.03**	-1.27	**80.26**	-0.54	**75.99**	+11.91	17.74
LLaVA-1.5-13B	78.70	-	7.02	-	70.72	-	61.33	-	17.50
+ Visual Evidence	**86.80**	+8.10	**6.18**	-0.84	**72.82**	+2.10	**71.25**	+9.92	15.62
LLaVA-1.6-13B	85.60	-	8.49	-	74.17	-	73.61	-	13.68
+ Visual Evidence	**90.43**	+4.83	**7.45**	-1.04	**79.55**	+5.38	**73.25**	-0.36	12.60
Qwen-VL-Max (API)	87.90	-	6.11	-	83.79	-	63.39	-	11.13
+ Visual Evidence	**90.66**	+2.76	**6.79**	+0.68	**83.97**	+0.18	**75.81**	+12.42	9.88

Table 9: The main results of other five models on POPE, AMBER and RPE dataset.

Model	POPE COCO-Popular		POPE COCO-Random	
	Acc.	Δ	Acc.	Δ
MiniGPT-4	74.10	-	81.27	-
+ Visual Evidence	**81.00**	+6.90	**89.93**	+8.66
Qwen-VL-Chat	86.80	-	89.52	-
+ Visual Evidence	**90.13**	+3.33	**91.20**	+1.68
LLaVA-1.5-7B	85.83	-	90.41	-
+ Visual Evidence	**90.80**	+4.97	**93.02**	+2.61
LLaVA-1.5-13B	84.53	-	88.28	-
+ Visual Evidence	**90.40**	+5.87	**93.33**	+5.05
LLaVA-1.6-7B	88.96	-	91.58	-
+ Visual Evidence	**92.03**	+3.07	**93.57**	+1.99
LLaVA-1.6-13B	89.23	-	91.99	-
+ Visual Evidence	**91.67**	+2.44	**93.09**	+1.10
MiniGPT-4-v2	81.27	-	89.24	-
+ Visual Evidence	**87.53**	+6.26	**92.89**	+3.65

Table 10: More results on POPE COCO-Popular and COCO-Random.

Compared to the closed-set detection model, the visual evidence provided by the open-vocabulary model result in a smaller gain (+4.4% vs. +7.77%). However, the advantage of the open-vocabulary visual model lies in its open-ended object categories, which allows it to handle and recognize a wider and more comprehensive range of objects, such as rare or uncommon objects, making it more suitable for open-world scenarios. How to leverage higher-quality open-vocabulary visual models to further reduce hallucinations of LVLMs in open-world settings is a direction in our future work.

E.2.3 Ablation of SGG Visual Models

In Tab. 11 we present the results on RPE of Qwen-VL-Chat incorporated with different scene graph generation models. This results demonstrate that different scene graph generation models (RelTR, MOTIFS and OpenPSG) have comparable improvements on Qwen-VL-Chat. For example, RelTR achieves 11.77% and OpenPSG achieves 12.92% improvement on Qwen-VL-Chat. The gains brought by different scene graph generation models to LVLM are within a stable range.

E.3 Ablation of Prompt Templates on POPE

To validate the robustness of visual evidence prompting against input prompts, we evaluate LLaVA-1.5-7B with various templates on object evaluations in Tab. 12. The difference in accuracy is significant depending on the sentence. In this experiment, the one with more reasoning style achieves the best results. For example, the 5_{th} template adds a new prompt that tells the LVLMs

LVLM	Visual model		Acc.
	Model name	mAP	(%)
Qwen-VL-Chat	-	-	64.08
	RelTR	18.9	**75.85**
	MOTIFS	20.0	**76.89**
	OpenPSG	28.4	**77.00**

Table 11: Relation hallucination results of Qwen-VL-Chat incorporating visual evidence from different scene graph generation models, *i.e.* RelTR (Cong et al., 2023), MOTIFS (Zellers et al., 2018) and OpenPSG (Yang et al., 2022). The Recall@20 on PSG benchmark of different visual models is also reported.

that the evidence may be irrelevant to the query. In contrast, when we use misleading or irrelevant templates, the performance does not improve. The results indicate that the performance is improved if the text is written in a way that encourages referring to the evidence. We also explore the enhancement of the model through chain-of-thought (CoT) prompt. It is observed that incorporating the CoT prompt led to a measurable improvement in performance. Moreover, the combination of our method with the CoT prompt yielded a performance that is superior by 2% compared to that reported in the main experiment. Given that the purpose of this paper is to verify the effectiveness of the method, we select the simplest template.

E.4 Ablation of Prompt Templates on RPE

Tab. 13 shows the results of robustness study against prompt template on relation hallucinations. Similar to the ablation results on object hallucination, the results on RPE also indicate that the performance is improved if the prompt is written in a way that encourages referring to the evidence. If we use misleading or irrelevant prompt templates, the performance fluctuates around the baseline.

Visual Evidence Prompt Templates	Acc. (%)
{question}	80.23
You can see {evidence} in the image.\n{question}	87.43
There are {evidence} in the image.\n{question}	86.33
{evidence} are existing in the image.\n{question}	86.03
The following object are existing in the image: {evidence}.\n{question}	**88.67**
There are {evidence} in the image.\n Question: {question}\n Please answer the question use the image information and context information. If there is no relevant information in the provided context, try to ignore the context and answer yourself.	**89.97**
{question} Let's think step by step.	83.05
You can see evidence in the image.\n{question} Let's think step by step.	**89.57**
It's a beautiful day.\n{question}	78.20
This is a full black image.\n{question}	79.93

Table 12: Robustness study of LLaVA-1.5-7B against prompt template measured on POPE.

Visual Evidence Prompt Templates	Acc. (%)
{question}	61.92
Evidence: There are {evidence} in the image.\n Let's refer to the evidence and then answer the following question.\n{question}	66.42
Evidence: You can see {evidence} in the image.\n Let's considering the evidence and then answer the following question.\n{question}	68.00
Evidence: You can see {evidence} in the image.\n{question} According to the image and evidence, the answer is	66.62
You can see {evidence} in the image.\n Then answer the question based on what you see: {question}	**71.00**
It's a beautiful day.\n{question}	62.08
This is a full black image.\n{question}	62.42

Table 13: Robustness study of LLaVA-1.5-7B against template measured on RPE.