Visual Evidence Prompting Mitigates Hallucinations in Large Vision-Language Models

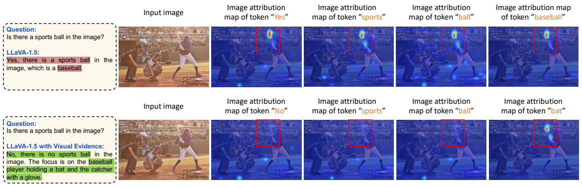
Wei Li[1], Zhen Huang[2], Houqiang Li[1], Le Lu[2], Yang Lu[3],
Xinmei Tian[1†], Xu Shen[2†], Jieping Ye[2]
[1]MoE Key Laboratory of Brain-inspired
Intelligent Perception and Cognition, University of Science and Technology of China
[2]Independent Researcher [3]Xiamen University
lwzkd@mail.ustc.edu.cn, shenxuustc@gmail.com, xinmei@ustc.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have shown impressive progress by integrating visual perception with linguistic understanding to produce contextually grounded outputs. Despite these advancements achieved, LVLMs still suffer from the hallucination problem, *e.g.*, they tend to produce content that does not exist in the input images. Our investigation suggests that such hallucinations often stem from the deficiencies in fine-grained comprehension on the visual aspect, particularly when visual scenes exhibit 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. While traditional visual models are not adept at interacting with humans, they excel at perceiving the fine-grained image contents. By 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 and attention on the images, reducing visual confusion by suppressing false activation while enhancing correct ones. Extensive experiments and in-depth analysis demonstrate the effectiveness of our method. 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.


Input image
Image attribution map of token "Yes"
Image attribution map of token "sports"
Image attribution map of token "ball"
Image attribution map of token "baseball"
Input image
Image attribution map of token "No"
Image attribution map of token "sports"
Image attribution map of token "ball"
Image attribution map of token "bat"

Question:
Is there a sports ball in the image?

LLaVA-1.5:
Yes, there is a sports ball in the image, which is a baseball.

Question:
Is there a sports ball in the image?

LLaVA-1.5 with Visual Evidence:
No, there is no sports ball in the image. The focus is on the baseball player holding a bat and the catcher with a glove.

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

Despite the success of LVLMs, they still grapple with a notable issue called *multimodal hallucination*. They tend to produce non-existent objects, relations, or attributes 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).

In this work, we begin by conducting both quantitative and qualitative analyses to investigate the underlying causes of hallucinations in LVLMs (see Sec. 2 for details). Our findings suggests that hallucinations primarily arise from the deficiencies in fine-grained comprehension on the visual aspect, which cause the model to confuse visually or semantically similar elements within the image.

Although various approaches have been proposed to mitigate hallucinations, existing methods are not explicitly 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

[†]Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4048–4080
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

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.

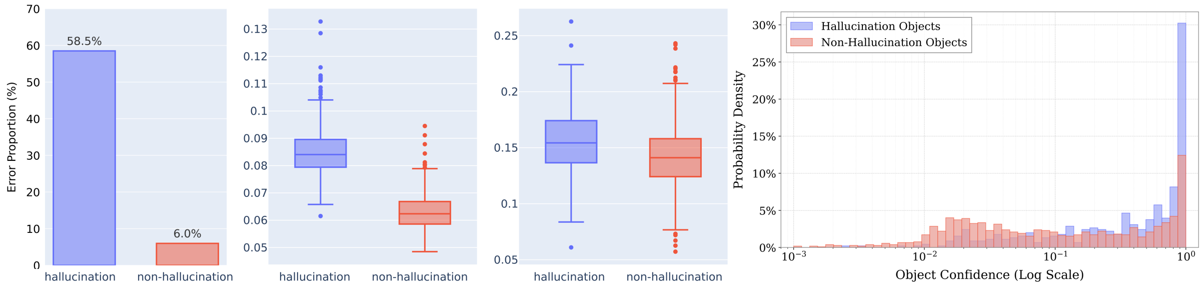
This work explores how the hallucinations of LVLMs can be mitigated by referring to visual evidence from small visual models. An example is shown in Fig. 3. The original LVLM produces inaccurate answer when queried about the presence of a chair within the image. However, small visual models, *i.e.*, 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 context, the model are able to generate correct responses. We refer to this training-free approach as *visual evidence prompting* (VEP).

Through detailed analysis, we find that visual evidence enables models to adjust and rectify the attribution and attention on the images, diminishing the degree of visual confusion by suppressing false activation while enhancing correct ones. Integrating visual evidence also corrects the prediction distribution for hallucination instances and elevates the confidence for non-hallucination samples. Extensive experiments on 11 LVLMs on 5 benchmarks show that visual evidence prompting mitigates the object, attribute, and relation hallucinations for both generative task and discriminative task, 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 LVLMs.

2 Preliminary Analysis

Qualitative Analysis. 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 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.



(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

(d) Internal object confidence within intermediate LVLM image representations

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. **(d)** Comparison of internal object confidence within intermediate vision tokens. These statistic analyses are conducted on POPE.

Quantitative Analysis. We observe that such phenomena are prevalent when hallucination happens, which motivates further statistical analyses to investigate their underlying causes.

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 definition of the attribution score is provided in in Appendix B. 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. In Fig. 2(d), we adopt the method proposed in Jiang et al. (2024) to compute the confidence scores with which the internal visual representations of the LVLM (*i.e.*, the vision tokens within the LLM) encode objects. We analyze the confidence distributions of vision tokens associated with hallucinated versus non-hallucinated objects. As shown in the figure, hallucinated objects consistently receive higher confidence scores than their non-hallucinated counterparts. This suggests that the model's internal visual representations tend to over-encode hallucinated content, highlighting a limitation in the LVLM's fine-grained visual perception capability.

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.

3 Visual Evidence Prompting

The goal of this work is to mitigate hallucinations in LVLMs by complementing them with fine-grained visual knowledge derived from small visual models. Generating answer A with input image I and question Q can be formulated within a probabilistic framework as estimating the conditional distribution $P(A|Q, I)$. The visual evidence prompting is formalized as $P(A|Q, I, VE)$, where VE is the extracted visual evidence 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.

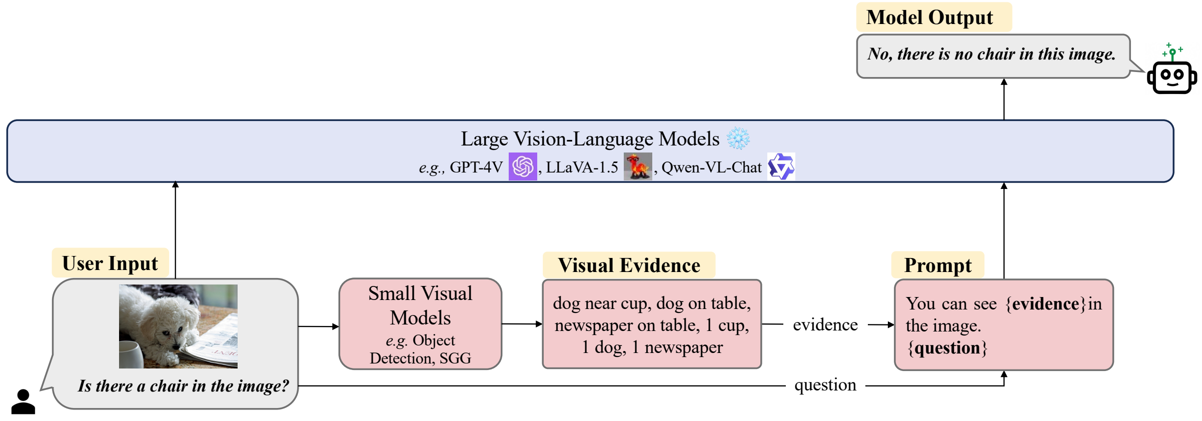

Model Output
No, there is no chair in this image.
Large Vision-Language Models
e.g., GPT-4V, LLaVA-1.5, Qwen-VL-Chat
User Input
Is there a chair in the image?
Small Visual Models
e.g. Object Detection, SGG
Visual Evidence
dog near cup, dog on table, newspaper on table, 1 cup, 1 dog, 1 newspaper
evidence
Prompt
You can see {evidence} in the image. {question}
question

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 context.

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.

The framework can be formulated as follows:

$$A = f_{LVLM}(I, Q, VE), VE = T[f_{SVM}(I)]. \qquad (1)$$

Here, SVM denotes small visual models, and T represents the process that transforms the structured output from small visual models into natural language.

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.

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" refers to approaches that do not rely 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. In this paper, visual models refer to the object detection and scene graph generation models. Other models such as 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 work.

4 Experiments

4.1 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.

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
Gemini 1.5 Pro (API)	82.43	-	8.70	-	72.70	-	69.06	-	14.36
+ Visual Evidence	**87.32**	+4.89	**7.63**	-1.07	**75.16**	+2.54	**71.13**	+2.07	13.82
Claude 3 (API)	75.40	-	5.34	-	75.91	-	69.57	-	17.36
+ Visual Evidence	**87.50**	+12.10	**5.0**	-0.34	**78.64**	+2.73	**70.57**	+1.00	14.90

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

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 C.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 4 close-source models as representatives, including MiniGPT-4 (Zhu et al., 2023), LLaVA (state-of-the-art open-source model), GPT4-V, Gemini 1.5 Pro and Claude 3. More details can be refer to Appendix C.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.

4.2 Results

Baselines performance. We firstly evaluate the hallucination performance on the datasets of POPE and AMBER of the 7 open-source models and 4 black-box APIs. The extensive evaluation results are presented in Tab. 2 and Tab. 15 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. 17 in Appendix. There are several in-

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

teresting 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 E.1.

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

Model	POPE		AMBER			
	Acc.	Δ	CHAIR$_{\downarrow}$	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.

4.4 Ablation of Visual Models

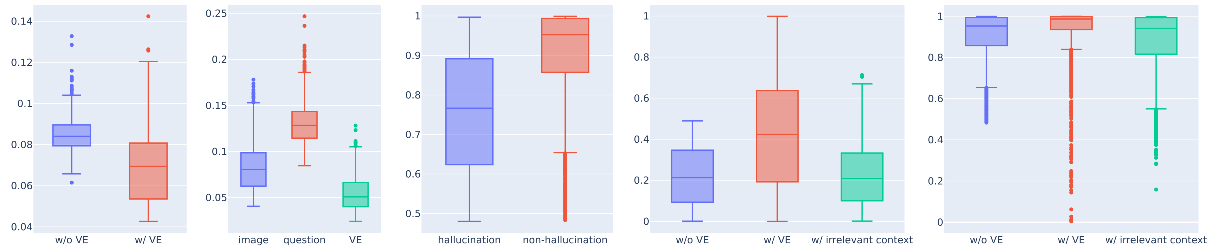
Given that the small visual models used in this work are contrasted with large vision-language models, presenting results using various versions of visual models will ensure a fair and comprehensive comparison. We employ 6 object detection models with different architectures (Fang et al., 2021; Carion 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. 21 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, model parameters, prompt templates are presented in the Appendix G.3, G.4, G.5 and G.6.

5 Analysis

5.1 How Does VEP Work?

5.1.1 Attribution and Confidence Analysis

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 dimin-



(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: Attribution and confidence analysis about how VEP works.

ishes 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 E.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)).

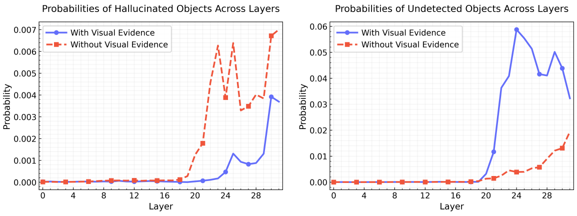
These intriguing phenomena suggest that the incorporation of visual evidence can guide the model in dynamically adjusting and rectifying the focus on image regions, diminishing the degree of confusion in visual information (suppressing erroneous activations while reinforcing correct activations). This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

5.1.2 Internal Interpretability Analysis

To analyze how visual evidence influences the model's internal behavior, we apply path patching (Wang et al., 2022) and logit lens (nostalgebraist, 2021) to conduct three rounds of backward tracing from the output answer. This interpretability analysis reveals the underlying information flow during object recognition, as detailed in the Appendix D. Our findings suggest that the model determines the presence of a queried object by leveraging object-related reference information encoded at the anchor token in the question, which is progressively transformed into the semantics of the correct answer (*e.g.*, from "visible" to "Yes") by some key attention heads. In the analysis process, we identify several special attention heads. For examples, head 14.24 attends to the relevant object region in the image, while head 13.28 primarily focuses on the object token in the question.

To analyze the impact of VEP, we analyze how hallucinated object information is encoded at the anchor token across different layers. We randomly sample 100 false positive samples and false negative samples respectively, and use logit lens to measure the encoding probabilities of these objects. As shown in Fig. 5, the introduction of VEP increases the probability of correctly encoding previously missed objects and decreases the probability of encoding spurious ones. This suggests that VEP helps refine the set of reference objects the model relies on for decision-making.

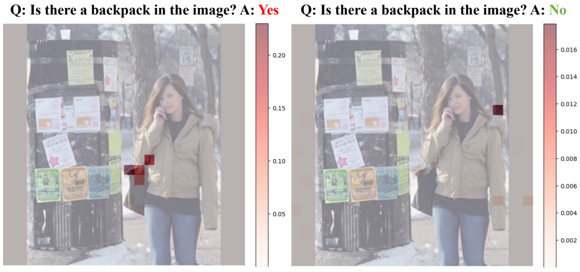
We further conduct qualitative case studies on the attention patterns of heads responsible for attending relevant object region. As shown in Fig. 6, in hallucinated cases, the head attends to visually similar but incorrect regions (*e.g.*, shoulder bag, even though the question asks about a backpack). After incorporating VE, such incorrect attention is reduced, resulting in correct prediction. These find-



(a) **(b)**

Figure 5: **(a)** Probabilities of hallucinated object information (false positives) encoded at the anchor token in the residual stream w/ and w/o VE. **(b)** Probabilities of undetected object information (false negatives) encoded at the anchor token in the residual stream w/ and w/o VE. Best viewed zoomed in.

ings indicate that VE mitigates hallucinations by rectifying internal attention patterns of some key heads, consistent with the observations discussed in Sec. 5.1.1.


Q: Is there a backpack in the image? A: Yes
0.20
0.15
0.10
0.05
Q: Is there a backpack in the image? A: No
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002

(a) w/o VE **(b) w/ VE**

Figure 6: Attention pattern visualization of key head 14.24 w/ and w/o VE.

5.2 Extra Hints or Complete Replacement for Visual Information?

To further investigate that whether providing visual evidence offers extra visual hints to the model, or dominates the image itself, we randomly replaced the original input image with a different image from the COCO2014 validation set. 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 5.1.1. We also observe differing behaviors between LLaVA-1.5 and LLaVA-1.6. We discuss this phenomenon in Appendix G.7.

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.

5.3 Mitigating Hallucinations beyond Objects

In LVLM, hallucinations not only include object hallucinations but also other types such as 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 model. The results are shown in the second last column of Tab. 2 and Tab. 15 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 and different kinds of hallucination are shown in Appendix F.2 and Appendix F.3 respectively. In Appendix E.7, we also demonstrate that other fine-grained tasks like object counting and OCR can also be enhanced effectively by our method. These results strongly validate the effectiveness and versatility of our framework.

5.4 Results on Open-world Scenario

An important concern is whether VEP is able to generalize to the open-vocabulary scenario. 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 E.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. Detailed error analysis in

LVLM	Visual model		Acc.
	Model name	mAP	(%)
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.

the Appendix E.4 further supports this argument.

5.5 Evaluation on General Multimodal Understanding Tasks

To assess how well our method performs on general multimodal understanding tasks, we evaluate baseline models and models 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 results suggest that our method can mitigate hallucinations without compromising the model's abilities across other dimensions.

6 Related Works

6.1 Hallucinations in LLMs

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 difficult to acquire grounded visual knowledge from a general knowledge base.

6.2 Hallucinations in LVLMs

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. In previous works, POPE (Li et al., 2023b) 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).

7 Conclusion

We have explored visual evidence prompting (VEP) as a simple and broadly applicable method for mitigating hallucinations in large vision-language models (LVLMs). Through comprehensive experiments on 11 models and various benchmarks, we demonstrate that VEP is an effective, robust, and general cure for LVLMs. We also conduct in-depth analysis to understand how VEP affects model behavior. We hope this work offers meaningful insights to advance the research on LVLMs.

8 Limitations

While our work sheds light on hallucination mitigation, there are several limitations to our work. **1) Limited Knowledge Integration.** Unlike fine-tuning, prompt-based strategies do not incorporate new knowledge into the model's parameters. Prior work (Zhai et al., 2023) has shown that excessive fine-tuning may cause models to hallucinate by overfitting to patterns in the training data while disregarding the input questions. In contrast, prompting preserves the model's original weights, offering greater controllability and maintaining generalization capabilities. However, this also limits its ability to embed domain-specific knowledge permanently. **2) Computational Overhead.** Introducing external visual models inevitably increases computational cost. Though our approach incurs significantly less overhead than instruction tuning. **3) Dependence on Visual Evidence Quality.** Although our experiments and analysis demonstrate the robustness of our method against imperfect visual evidence, its effectiveness still depends on the overall quality of the visual evidence. **4) Sensitivity to Prompt Design.** 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) Limited Task Coverage.** Our primary focus is on hallucinations involving objects, attributes, and relations. Although we present preliminary results on tasks like OCR and object counting in the Appendix, these are proof-of-concept. Rather than claiming a universal solution, we aim to show the potential of combining the generality of large models with the precision of small, specialized models. **6) Lack of Evaluation on Synthetic Domains.** We do not evaluate our method on datasets that do not have real-world scenes, such as MMMU, MathVision, and MathVista. Nevertheless, our approach has the potential to improve LVLM performance on these benchmarks. For example, small domain-specific models, such as image captioning models trained for chart or table understanding, could be developed using these datasets and directly incorporated into our framework to enhance LVLM capabilities. This is enabled by the model-agnostic and flexible nature of our framework, which allows for straightforward customization and incorporation of customized visual experts. We also discuss the complementarity of small and large models in Appendix A.

9 Acknowledgements

This work was supported in part by NSFC No. 62222117. We sincerely thank the meta-reviewer and the anonymous reviewers for their constructive and valuable feedback.

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 bimodal 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*.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*.

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.

nostalgebraist. 2021. Logit Lens on non-GPT2 models + extensions.

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.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

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.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

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 Complementarity of Small and Large Models

Our framework leverages the strengths of both specialized visual models and large vision language models (LVLMs), we discuss their complementary roles in supporting accurate and versatile visual understanding from two aspects.

(1) Open-ended dialogue: While small visual models excel in specialization and precision, they lack versatility and openness. In contrast, LVLM is highly versatile and capable of open-ended dialogue, but it suffers from lower precision. Our proposed framework effectively combines the strengths of both approaches. LVLM serves as the "brain" of the system, making it indispensable. If the query is a straightforward discriminative task, such as determining whether a specific object exists in an image (e.g., a yes/no question), small visual models can handle this through keyword search. Though, keyword search struggles with semantically equivalent but linguistically different queries. Moreover, for open-ended, generative questions like “Describe this image”, small visual models are insufficient. However, their outputs can provide critical and precise information to assist LVLM in generating more accurate answers.

(2) High recall of visual content: Small visual models are primarily designed for high-precision predictions rather than high-recall ones (though better models could offer both). On the other hand, LVLM provides comprehensive image understanding with high recall but lower precision, which can lead to hallucinations. The small visual model provides high-confidence supplementary information, ensuring that at least the presence of specific objects or relations in the image is reliably identified, thus assisting LVLM in making more accurate decisions. We have also analyzed this in Sections 5.1 and 5.2, demonstrating that the model synthesizes both textual and visual information to arrive at the final decision, rather than relying solely on visual evidence.

B 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.

$$\bar{\mathbf{E}}_\mathbf{i} = \sum_{j=1}^{h} (\nabla \mathbf{A}^\mathbf{i}_\mathbf{j} \odot \mathbf{A}^\mathbf{i}_\mathbf{j})^+, \qquad (2)$$

where $\odot$ is the Hadamard product, $\mathbf{A}^\mathbf{i}_\mathbf{j}$ denote the attention matrix of the head j in layer i, $\nabla \mathbf{A}^\mathbf{i}_\mathbf{j} := \frac{\partial logit\_current\_token}{\partial \mathbf{A}^\mathbf{i}_\mathbf{j}}$ 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.

$$\mathbf{R} \leftarrow \mathbf{R} + \bar{\mathbf{E}}_\mathbf{i} \cdot \mathbf{R}. \qquad (3)$$

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

C Further Details for Experimental Setup

C.1 Datasets

The experiments are mainly conducted on POPE, AMBER and RPE. We also conduct experiments using LLaVA-1.5 and LLaVA-1.6 on the MMHal-Bench (Sun et al., 2023). Please refer to Appendix G.2 for details.

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. 7, we show some cases in our dataset.

C.2 Details of Evaluating SOTA Models

We conduct experiments on GPT-4V (model version: gpt-4-1106-vision-preview), Gemini 1.5 Pro Flash, Claude 3 Haiku, and Qwen-VL-Max via API access. For open-source models, 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.

D Internal Interpretability Analysis

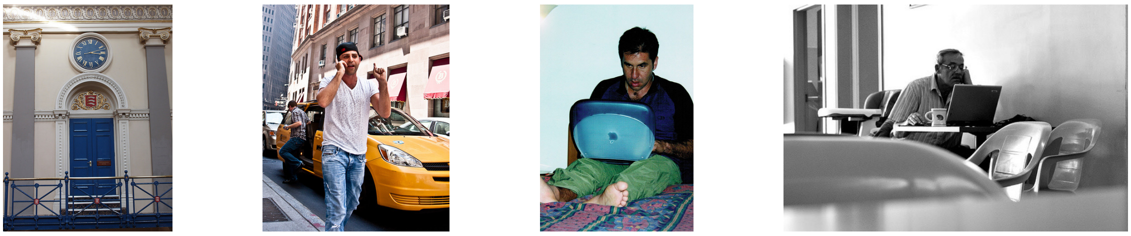
To better understand the internal pattern changes induced by visual evidence, we perform a tracing-

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}

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



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 7: 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.

based interpretability analysis on object recognition task. Starting from the model's final answer (*i.e.*, "Yes" or "No"), we trace the information flow involved in object recognition task. This analysis is conducted on the POPE dataset, where model responses are limited to binary choices ("Yes" or "No").

Our methodology proceeds as follows:

(1) At the token position generating the binary prediction (specifically, the colon ":" token), we apply path patching algorithm (Wang et al., 2022) to identify the key attention heads.

(2) We isolate the heads that encode the correct answer (*e.g.*, "Yes") and inspect their attention patterns to identify the most strongly attended tokens—denoted as token A.

(3) For each token A, we use logit lens (nostalgebraist, 2021) to compute the probability of the correct answer token across the residual streams of all layers, and determine the layer l where this probability is maximized.

(4) We then perturb each attention head prior to layer l at the position of token A using path patching, measuring the resulting change in the correct answer token's probability at layer l's residual stream. The head whose perturbation causes the largest probability drop is selected for further analysis.

(5) Finally, we examine the decoded content and attention patterns of these influential heads and recursively trace backward, in order to identify the specific visual and textual cues that guided the model's decision.

D.1 Tracing Back from Answer Token

We initiate our analysis by tracing backward from the answer token. Aiming to identify the key attention heads for object recognition, we randomly select 20 samples from the POPE dataset where the model successfully performs this task (*i.e.*, generate correct answers "Yes"). Notably, expanding the sample size of path patching results in a similar distribution of key heads, a phenomenon consistent with the findings of Wang et al. (2022). We firstly construct counterfactual question that are designed to suppress object recognition—aiming to avoid activating this capability of the model as much as possible. Examples of both original and counterfactual question are shown in Tab. 9. We then employ path patching (Wang et al., 2022) to localize the attention heads that contribute most to the generation of the answer token.

Fig. 8(a) illustrates the distribution of the identified key heads. The key heads appear to be sparsely distributed. From this set, we select the top five heads—head 16.0, head 15.31, head 14.20, head

Data	Question
X_r	`Is there a {object} in the image?`

X_c	`Is this image from an outdoor setting?`
	`Is this image taken in daylight?`
	`Was this image taken in a park?`
	`Is this image geometrically complex?`
	`Does the image contain a solid color background?`

Table 9: Example of probing data X_r (original question) and counterfactual data X_c in first round path patching in Sec. D.1.

13.16, and head 16.15—for further analysis. Using Logit Lens (nostalgebraist, 2021), we decode the representations at these heads by projecting them into the vocabulary space. Notably, head 16.0 yields top-ranked tokens that are semantically aligned with the correct answer, whereas the other heads produce less meaningful outputs. Some examples are show in Tab. 10.

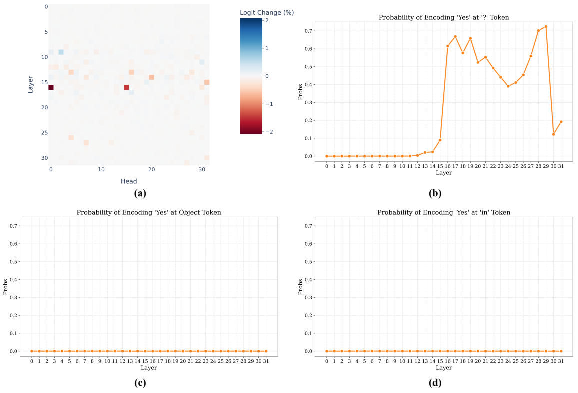


Figure 8: **(a)** Distribution of key heads in first round path patching that mostly influence final answer generation. **(b)** Probability of the correct answer "Yes" being encoded at the final question token ("?") on the residual streams across each layer of the LLM. **(c)** Probability of "Yes" being encoded at the object token in the question across residual streams. **(d)** Probability of "Yes" being encoded at the preposition "in" token in the question across residual streams. The probabilities are obtained by averaging across 100 random samples. Best viewed with zoom for clarity.

To further trace the source of the "Yes"-related signal, we visualize the attention patterns of the key heads, some examples are shown in Fig. 9. In nearly all cases, head 16.0 consistently attends to the final token in the question—the "?" token—while the other heads primarily attend to the object token or the preposition token "in". In contrast, randomly selected heads often attend to the <bos> token, potentially due to the attention sink effect (Xiao et al., 2023). We further apply Logit Lens to compute the probability of encoding correct answer "Yes" token from the residual stream at these attended token positions. The results in Fig. 8 show that the final question token "?" carries a high predictive probability (up to 0.7–0.8), while the object token and "in" tokens contribute minimally, with probabilities near to zero.

These observations suggest that head 16.0 plays a pivotal role in *transfering* answer-relevant information, specifically the answer token "Yes", from *the last token in question* to generate the model's final decision.

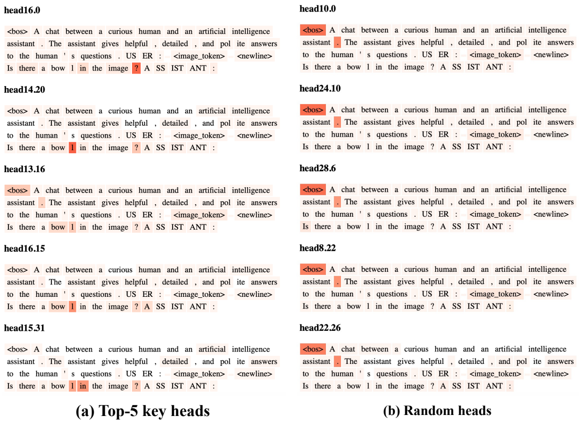

head16.0
head14.20
head13.16
head16.15
head15.31
head10.0
head24.10
head28.6
head8.22
head22.26
<bos> A chat between a curious human and an artificial intelligence assistant . The assistant gives helpful , detailed , and pol ite answers to the human ' s questions . US ER : <image_token> <newline> Is there a bow l in the image ? A SS IST ANT :
(a) Top-5 key heads
(b) Random heads

Figure 9: **(a)** Attention pattern of the top heads identified as having the greatest impact on answer generation in Sec. D.1. **(b)** Attention pattern of randomly selected heads.

Head	Top tokens in projection
16.0	`yes, Yes, yes, Yes, YES`
15.31	`there, presence, \u5426, Yes, achi`
13.16	`ferrer, \ufffd, \u2010, \u00e9n, hell`
14.20	`Mel, ritz, operator, ner, \uc7ac`
16.15	`ouv, EXISTS, SHA, abol, igny`

Table 10: Decoded content of the identified top key heads in first round path patching in Sec. D.1, projected into the vocabulary space using the Logit Lens.

D.2 Tracing Back from Last Token of Question

In Sec. D.1, we conjecture that head 16.0 is responsible for transferring answer-relevant information from the final token of the question to generate the model's decision. Therefore, we conduct a further backward analysis starting from the last token of the question. As shown in Fig. 8(b), we observe a sharp increase in the probability of encoding the

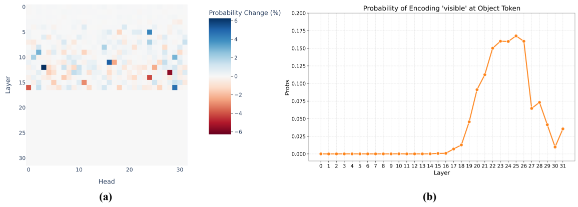


Figure 10: **(a)** Distribution of key heads in second round path patching that mostly influence the process that encoding the correct answer at the final question token. **(b)** Probability of "visible" token being encoded at the object token in the question on the residual streams across each layer of the LLM. The probabilities are obtained by averaging across 100 random samples. Best viewed with zoom for clarity.

correct answer at the final question token at layer 16.

Following this observation, we perform path patching by perturbing all attention heads preceding layer 16. We then track changes in the probability of encoding the correct answer at the final question token within the residual stream of layer 16. This allows us to identify the key heads that most influence this encoding. In this second round path patching experiment, the counterfactual question X_c is "Is this image from an outdoor setting?".

Fig. 10 presents the distribution of key heads identified through path patching at the final token position of the question. As shown, only a small number of heads have a substantial impact. Following the approach illustrated in Sec. D.1, we apply the Logit Lens to decode the top key heads and analysis their attention patterns. Among these, three heads stand out. Head 13.4, attends to both the image region corresponding to the object and the object token in the question. Its decoded output prominently features semantically related tokens such as "detected", "presence" and "visible". Head 13.28, primarily attends to the object token in the question. Decoding its output reveals the presence of the token "Yes". Another head 14.24 focuses mainly on the image region corresponding to the object queried in the question. Examples of decoding results and attention patterns are provided in Tab. 11 and Fig. 11.

We further utilize the Logit Lens to decode the information encoded at the object token position in the residual stream. The decoded output is semantically concentrated on tokens like "visible" and "present", with "visible" emerging as the most probable token. The distribution of "visible" token probabilities across layers in the residual stream is shown in Fig. 10(b).

The above analysis and findings suggest that the encoded information about the correct answer "Yes" at the final token of the question is derived from higher-level concepts, such as "visible" and "present", which are encoded in *the object token*. These concepts are formed by several key attention heads, including head 13.4**, head** 14.24**, and head** 13.28**, which examine and combine information from the object region in the image and the object token in the question.**

Head	Top tokens in projection
13.4	**available**, **visible**, **available**, **disponible**, **observable**, Bedeut, u0150, onymes, **ailable**, **existsdetection**, **detected**, \u8a71, \u25c4, subset, \u4ef6, gres, Unterscheidung, alert, **captured**, **visible**
13.28	dispon, sj\ŏ0f6, disponible, available, available, loyd, assa, rvm, j\ŏ0fa, rizzak, Shaw, **yes**, **Yes**, \u043d\u0432\u0430, aland, **Yes**, azon, pilot, **yes**, \u0161\u010d
14.24	alom, tera, bers, (, Terr, **Visible**, heim, dup, ld, ershell, \u00fcn

Table 11: Decoded content of the identified top key heads in second round path patching in Sec. D.2, projected into the vocabulary space using the Logit Lens.

D.3 Tracing Back from Object Token in the Question

To further investigate how the object token in the question comes to encode semantic information related to "visible", we perform a third round of path patching at the object token position. Specifically, we trace this process by measuring the probability of encoding the token "visible" at the object token in the residual stream at layer 22. This layer is chosen based on the observation in Fig. 10(b), where the probability begins to level off and then shows signs of decline.

Following the methodology described in Sec. D.2, we construct counterfactual question X_c and perturb all attention heads preceding layer 22 to identify the contributing key heads. The resulting distribution is shown in Fig. 12. Among the top ten heads identified, five are found to predominantly attend to the tokens "a" and "there". Decoded outputs of these heads exhibit semantic content closely aligned with concepts like "visible" and "available". Examples are provided in Tab. 12 and Fig. 13.

We further decode the residual stream at the "a" and "there" token positions. In mid-to-late lay-

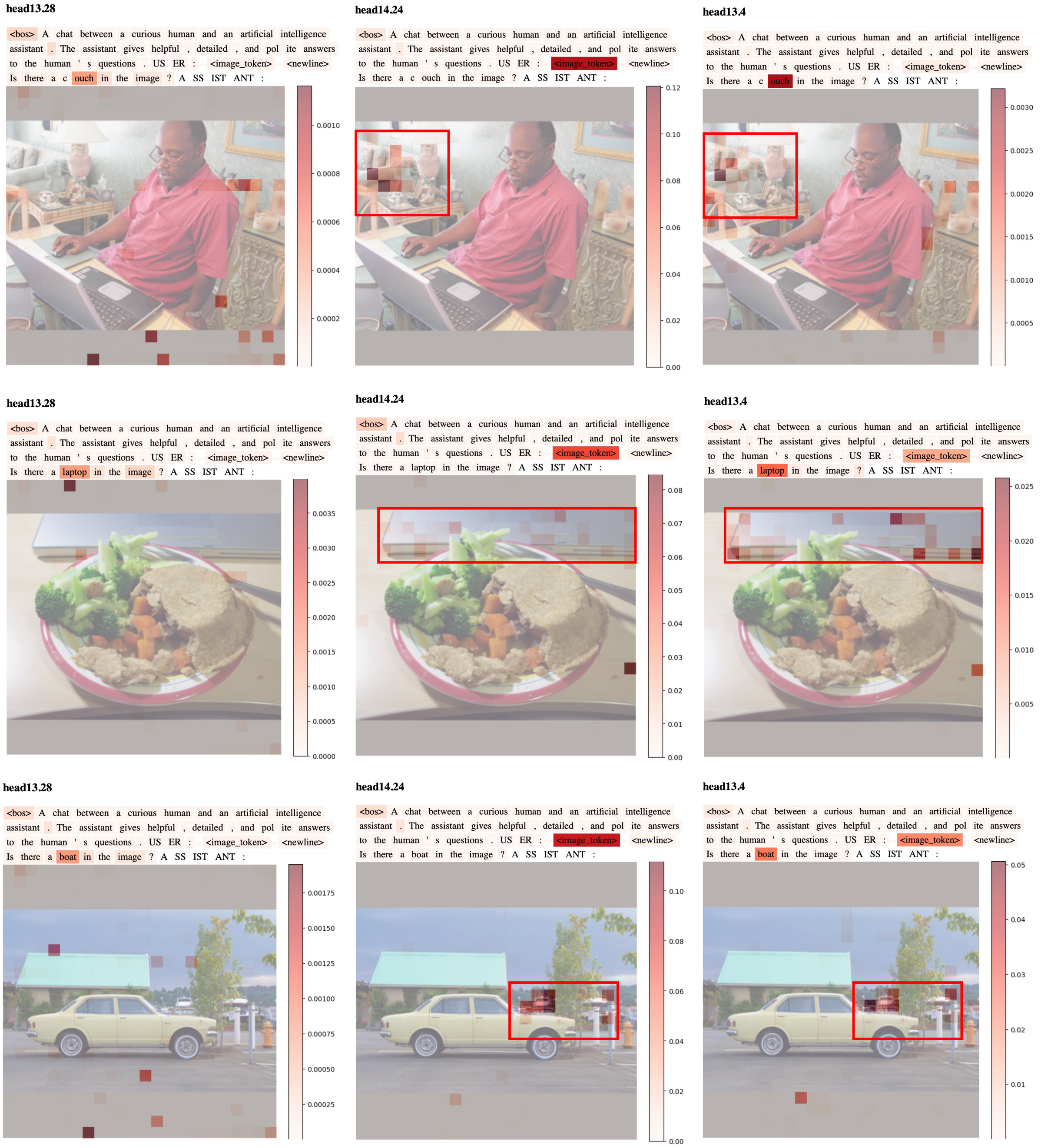


Figure 11: Attention pattern visualization of key heads identified in second round path patching in Sec. D.2.

ers, these tokens are found to encode information related to the presence of objects in the image. Notably, starting from layer 17, the "a" token consistently yields object-related content in its decoded output. The "there" token also contributes to object-level semantics, though to a lesser extent. Additional decoding examples are shown in Tab. 13.

These findings suggest that the object token acquires object-related information from the "a" and "there" tokens, which act as anchors. We therefore denote them as *anchor tokens*. This information is subsequently transformed into higher-level semantics, such as "visible", by several key attention heads. Since each head operates on inputs drawn from the residual stream and the residual stream at anchor token already encodes object-level features. We conjecture that these heads likely serve as the function for semantic abstraction and propagation.

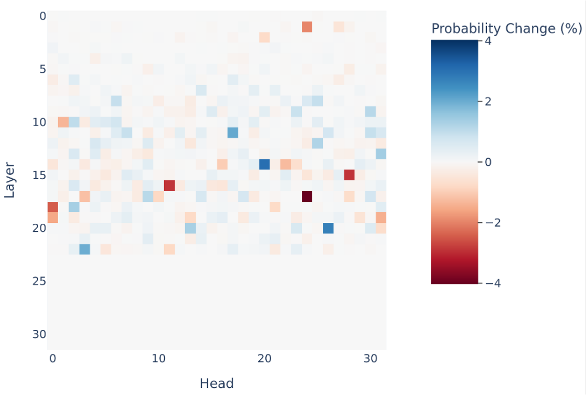


Figure 12: Distribution of key heads in third round path patching that mostly influence the process that encoding the "visible" related semantic at the object token in question.

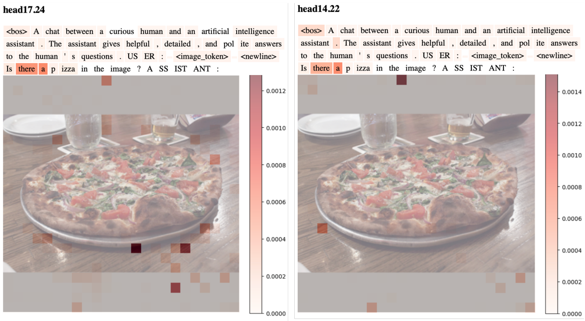


Figure 13: Attention pattern visualization of key heads identified in third round path patching in Sec. D.3.

Head	Top tokens in projection
17.24	**available**, **visible**, **available**, **disponible**, **observable**, Bedeut, \u0150, onymes, **ailable**, **exists**
14.22	**dispon**, sj\u00f6, **disponible**, **available**, **available**, loyd, assa, rvm, j\u00fa, riz

Table 12: Decoded content of the identified top key heads identified in third round path patching in Sec. D.3, projected into the vocabulary space using the Logit Lens.

D.4 Conclusion about the Tracing Process

Starting from the answer token, we conduct three rounds of backward tracing and uncover several interesting patterns. We summarize the internal information flow within the model during object recognition as follows:

(1) The anchor tokens "a" and "there" encode information related to objects in the image. The object token receives this object-related information from anchor token "a" and "there", which is subsequently transformed by some attention heads into higher-level semantics such as "visible".

(2) The final token of the question (*i.e.*, "?" token) extracts this "visible"-related semantic information from the object token. Combined with visual features from the corresponding object region in the image, this contributes to the generation of the "Yes" signal.

(3) The final input token to the model (typically ":" token) retrieves the "Yes" signal from the final token of the question (*i.e.*, token "?") and produces the output answer token "Yes".

This overall information flow is illustrated in Fig. 14.

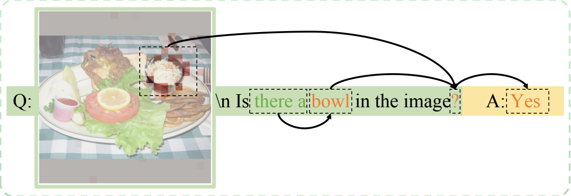


Figure 14: Overall information flow when model perform object recognition task.

Since our tracing leads back to the anchor token, which encodes substantial object-level information, we are further able to compare how this token's representation changes under two settings: with and without visual evidence. This comparison enables us to understand how the incorporation of VE

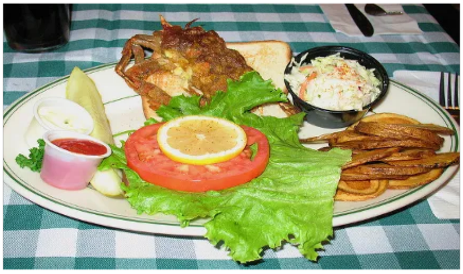
Image	Layer	Top tokens in projection
	24	fork, plate, sand, table, plate, tables, \u6c99, sal, variety, glass, bow, place, serving, Sand, cole, tom, usammen, plates, Hamb, difference
	25	fork, plate, sand, tom, tom, table, bow, \u6c99, plate, tables, usammen, variety, Tom, glass, Sand, cole, Bowl, Hamb, Tom, nap

Table 13: Decoded content of the residual stream at anchor token, projected into the vocabulary space using the Logit Lens.

affects the internal object information the model refer to during decision-making.

Although this interpretability analysis is not the main contribution of our work and is limited in scope, as it focuses on simple yes-or-no object recognition tasks, it nevertheless highlights a promising direction for future research. The internal mechanisms of LVLMs remain largely opaque, and we hope our findings can provide some meaningful insights for the broader research community.

E Qualitative Results

E.1 Effect of Visual Evidence

Fig. 15 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.

E.2 Comparison of Image Attribution Map between Hallucination and Non-hallucination Samples

In Fig. 16, 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.

E.3 Comparison of Image Attribution Map before and after the Incorporation of VE

In Fig. 17, 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.

E.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 E.5 for error examples of each type):

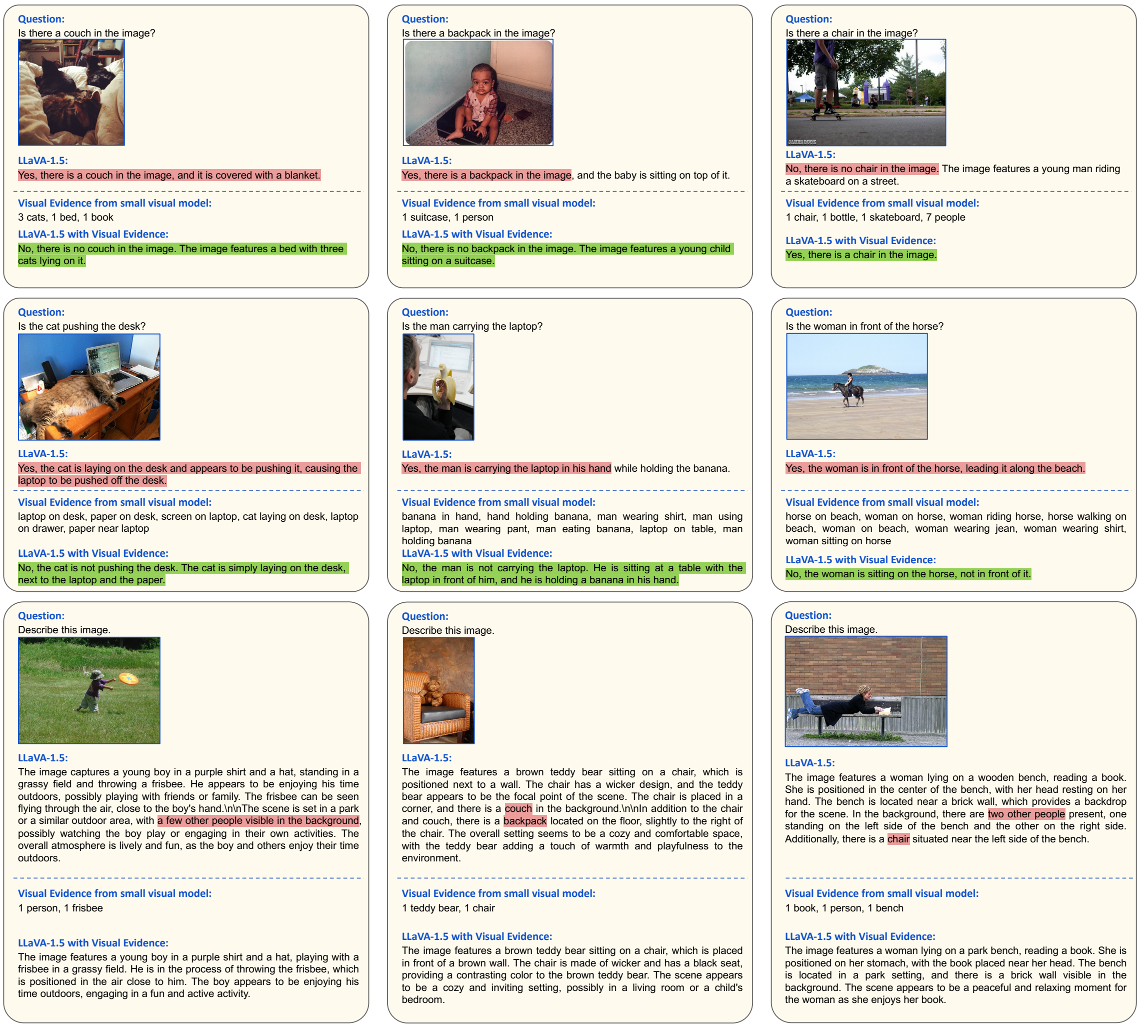
(1) *Annotation noise (*29.2%*)*: Objects being annotated as present in the image when they are not, or vice versa.

(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",



Question:
Is there a couch in the image?

LLaVA-1.5:
Yes, there is a couch in the image, and it is covered with a blanket.

Visual Evidence from small visual model:
3 cats, 1 bed, 1 book

LLaVA-1.5 with Visual Evidence:
No, there is no couch in the image. The image features a bed with three cats lying on it.

Question:
Is there a backpack in the image?

LLaVA-1.5:
Yes, there is a backpack in the image, and the baby is sitting on top of it.

Visual Evidence from small visual model:
1 suitcase, 1 person

LLaVA-1.5 with Visual Evidence:
No, there is no backpack in the image. The image features a young child sitting on a suitcase.

Question:
Is there a chair in the image?

LLaVA-1.5:
No, there is no chair in the image. The image features a young man riding a skateboard on a street.

Visual Evidence from small visual model:
1 chair, 1 bottle, 1 skateboard, 7 people

LLaVA-1.5 with Visual Evidence:
Yes, there is a chair in the image.

Question:
Is the cat pushing the desk?

LLaVA-1.5:
Yes, the cat is laying on the desk and appears to be pushing it, causing the laptop to be pushed off the desk.

Visual Evidence from small visual model:
laptop on desk, paper on desk, screen on laptop, cat laying on desk, laptop on drawer, paper near laptop

LLaVA-1.5 with Visual Evidence:
No, the cat is not pushing the desk. The cat is simply laying on the desk, next to the laptop and the paper.

Question:
Is the man carrying the laptop?

LLaVA-1.5:
Yes, the man is carrying the laptop in his hand while holding the banana.

Visual Evidence from small visual model:
banana in hand, hand holding banana, man wearing shirt, man using laptop, man wearing pant, man eating banana, laptop on table, man holding banana

LLaVA-1.5 with Visual Evidence:
No, the man is not carrying the laptop. He is sitting at a table with the laptop in front of him, and he is holding a banana in his hand.

Question:
Is the woman in front of the horse?

LLaVA-1.5:
Yes, the woman is in front of the horse, leading it along the beach.

Visual Evidence from small visual model:
horse on beach, woman on horse, woman riding horse, horse walking on beach, woman on beach, woman wearing jean, woman wearing shirt, woman sitting on horse

LLaVA-1.5 with Visual Evidence:
No, the woman is sitting on the horse, not in front of it.

Question:
Describe this image.

LLaVA-1.5:
The image captures a young boy in a purple shirt and a hat, standing in a grassy field and throwing a frisbee. He appears to be enjoying his time outdoors, possibly playing with friends or family. The frisbee can be seen flying through the air, close to the boy's hand.\n\nThe scene is set in a park or a similar outdoor area, with a few other people visible in the background, possibly watching the boy play or engaging in their own activities. The overall atmosphere is lively and fun, as the boy and others enjoy their time outdoors.

Visual Evidence from small visual model:
1 person, 1 frisbee

LLaVA-1.5 with Visual Evidence:
The image features a young boy in a purple shirt and a hat, playing with a frisbee in a grassy field. He is in the process of throwing the frisbee, which is positioned in the air close to him. The boy appears to be enjoying his time outdoors, engaging in a fun and active activity.

Question:
Describe this image.

LLaVA-1.5:
The image features a brown teddy bear sitting on a chair, which is positioned next to a wall. The chair has a wicker design, and the teddy bear appears to be the focal point of the scene. The chair is placed in a corner, and there is a couch in the background.\n\nIn addition to the chair and couch, there is a backpack located on the floor, slightly to the right of the chair. The overall setting seems to be a cozy and comfortable space, with the teddy bear adding a touch of warmth and playfulness to the environment.

Visual Evidence from small visual model:
1 teddy bear, 1 chair

LLaVA-1.5 with Visual Evidence:
The image features a brown teddy bear sitting on a chair, which is placed in front of a brown wall. The chair is made of wicker and has a black seat, providing a contrasting color to the brown teddy bear. The scene appears to be a cozy and inviting setting, possibly in a living room or a child's bedroom.

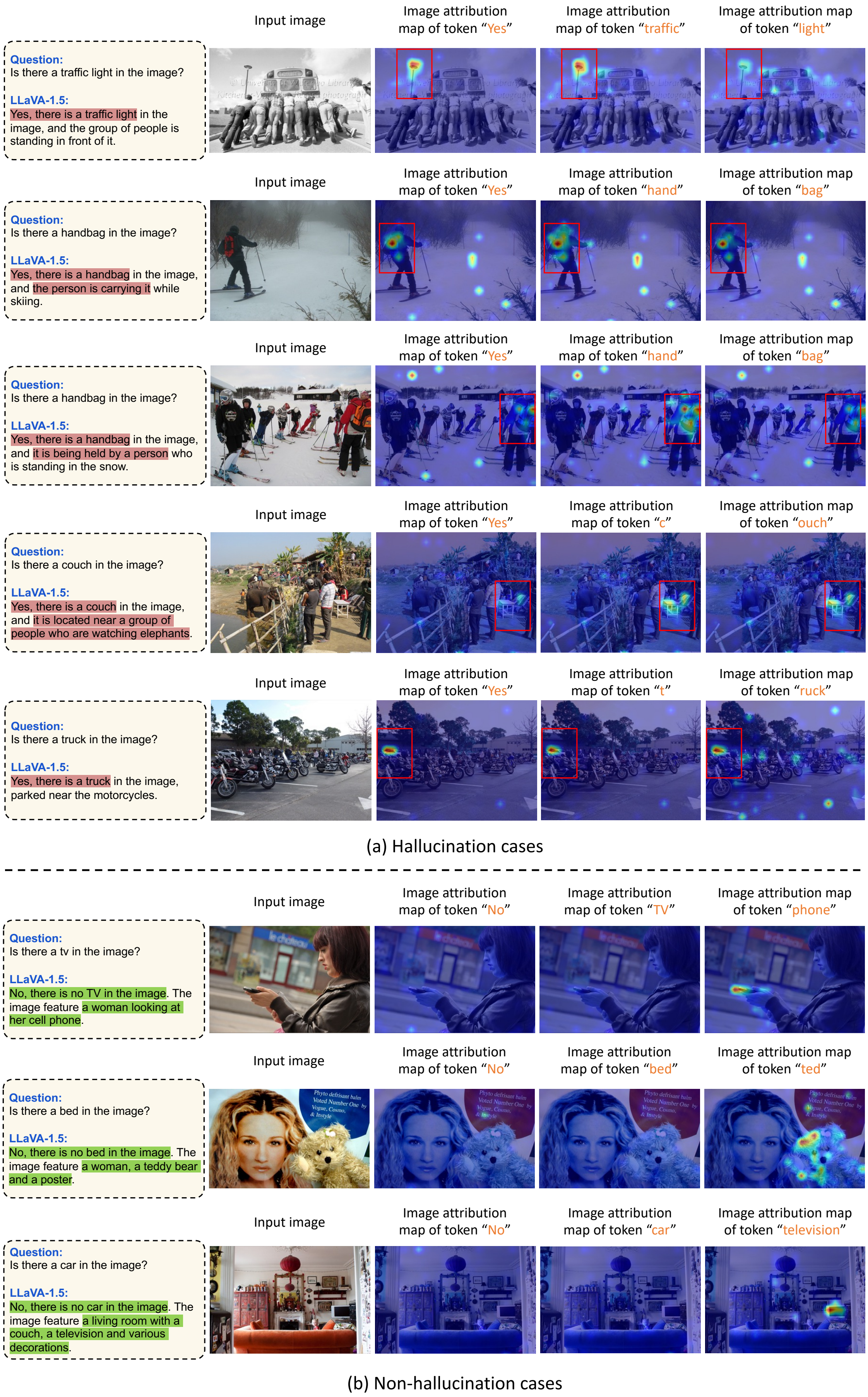
Question:
Describe this image.

LLaVA-1.5:
The image features a woman lying on a wooden bench, reading a book. She is positioned in the center of the bench, with her head resting on her hand. The bench is located near a brick wall, which provides a backdrop for the scene. In the background, there are two other people present, one standing on the left side of the bench and the other on the right side. Additionally, there is a chair situated near the left side of the bench.

Visual Evidence from small visual model:
1 book, 1 person, 1 bench

LLaVA-1.5 with Visual Evidence:
The image features a woman lying on a park bench, reading a book. She is positioned on her stomach, with the book placed near her head. The bench is located in a park setting, and there is a brick wall visible in the background. The scene appears to be a peaceful and relaxing moment for the woman as she enjoys her book.

Figure 15: Some cases demonstrating that visual evidence prompting effectively reduce hallucination of LLaVA-1.5-7B.


Input image
Image attribution map of token “Yes”
Image attribution map of token “traffic”
Image attribution map of token “light”
Image attribution map of token “Yes”
Image attribution map of token “hand”
Image attribution map of token “bag”
Image attribution map of token “Yes”
Image attribution map of token “hand”
Image attribution map of token “bag”
Image attribution map of token “Yes”
Image attribution map of token “c”
Image attribution map of token “ouch”
Image attribution map of token “Yes”
Image attribution map of token “t”
Image attribution map of token “ruck”
Image attribution map of token “No”
Image attribution map of token “TV”
Image attribution map of token “phone”
Image attribution map of token “No”
Image attribution map of token “bed”
Image attribution map of token “ted”
Image attribution map of token “No”
Image attribution map of token “car”
Image attribution map of token “television”

Question:
Is there a traffic light in the image?

LLaVA-1.5:
Yes, there is a traffic light in the image, and the group of people is standing in front of it.

Question:
Is there a handbag in the image?

LLaVA-1.5:
Yes, there is a handbag in the image, and the person is carrying it while skiing.

Question:
Is there a handbag in the image?

LLaVA-1.5:
Yes, there is a handbag in the image, and it is being held by a person who is standing in the snow.

Question:
Is there a couch in the image?

LLaVA-1.5:
Yes, there is a couch in the image, and it is located near a group of people who are watching elephants.

Question:
Is there a truck in the image?

LLaVA-1.5:
Yes, there is a truck in the image, parked near the motorcycles.

(a) Hallucination cases

Question:
Is there a tv in the image?

LLaVA-1.5:
No, there is no TV in the image. The image feature a woman looking at her cell phone.

Question:
Is there a bed in the image?

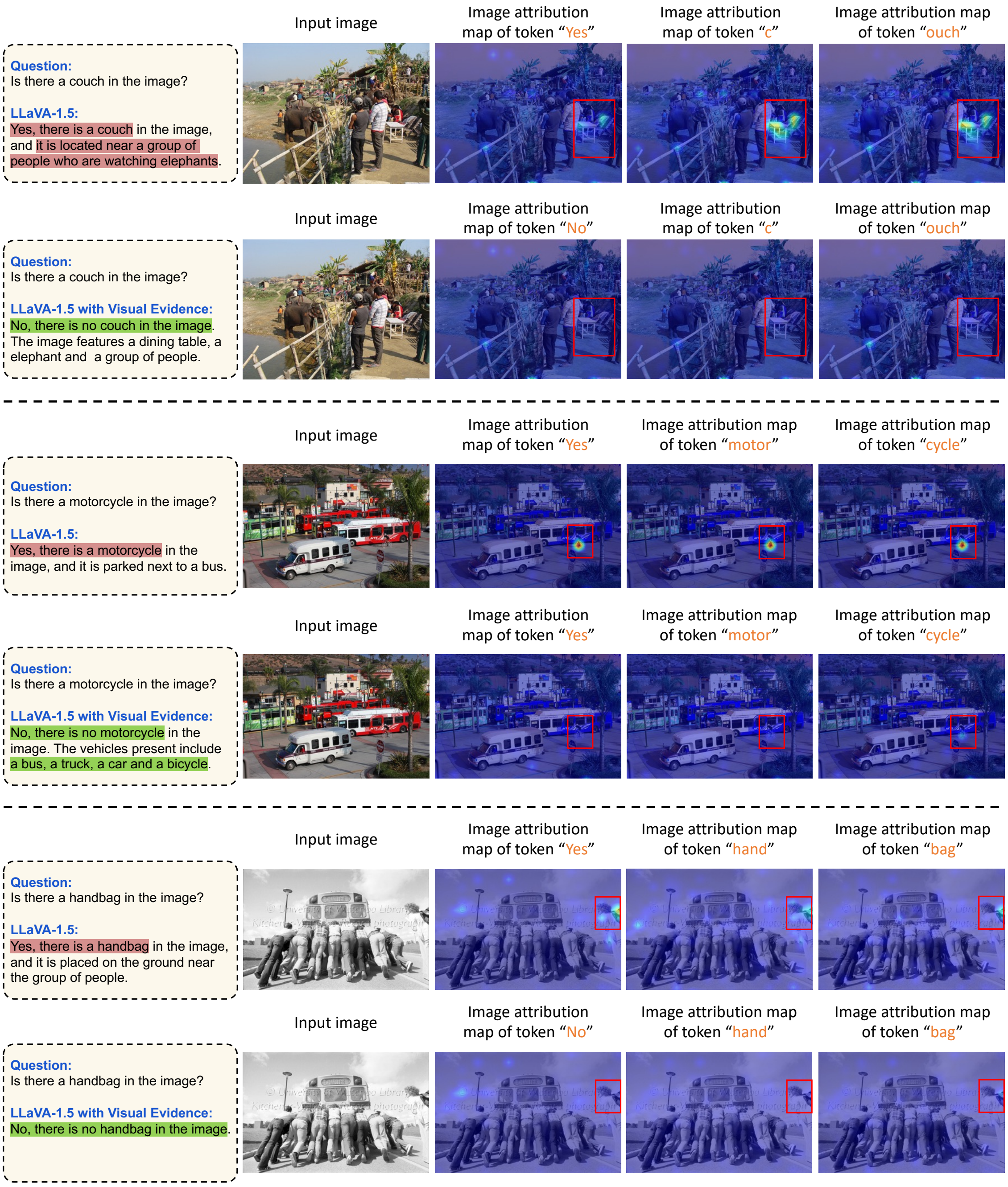
LLaVA-1.5:
No, there is no bed in the image. The image feature a woman, a teddy bear and a poster.

Question:
Is there a car in the image?

LLaVA-1.5:
No, there is no car in the image. The image feature a living room with a couch, a television and various decorations.

(b) Non-hallucination cases

Figure 16: More qualitative comparison of image attribution maps between hallucination samples and non-hallucination ones.


Input image
Image attribution map of token "Yes"
Image attribution map of token "c"
Image attribution map of token "ouch"
Input image
Image attribution map of token "No"
Image attribution map of token "c"
Image attribution map of token "ouch"
Input image
Image attribution map of token "Yes"
Image attribution map of token "motor"
Image attribution map of token "cycle"
Input image
Image attribution map of token "Yes"
Image attribution map of token "motor"
Image attribution map of token "cycle"
Input image
Image attribution map of token "Yes"
Image attribution map of token "hand"
Image attribution map of token "bag"
Input image
Image attribution map of token "No"
Image attribution map of token "hand"
Image attribution map of token "bag"

Question:
Is there a couch in the image?

LLaVA-1.5:
Yes, there is a couch in the image, and it is located near a group of people who are watching elephants.

Question:
Is there a couch in the image?

LLaVA-1.5 with Visual Evidence:
No, there is no couch in the image. The image features a dining table, a elephant and a group of people.

Question:
Is there a motorcycle in the image?

LLaVA-1.5:
Yes, there is a motorcycle in the image, and it is parked next to a bus.

Question:
Is there a motorcycle in the image?

LLaVA-1.5 with Visual Evidence:
No, there is no motorcycle in the image. The vehicles present include a bus, a truck, a car and a bicycle.

Question:
Is there a handbag in the image?

LLaVA-1.5:
Yes, there is a handbag in the image, and it is placed on the ground near the group of people.

Question:
Is there a handbag in the image?

LLaVA-1.5 with Visual Evidence:
No, there is no handbag in the image.

Figure 17: More qualitative comparison of image attribution maps before and after the integration of visual evidence.

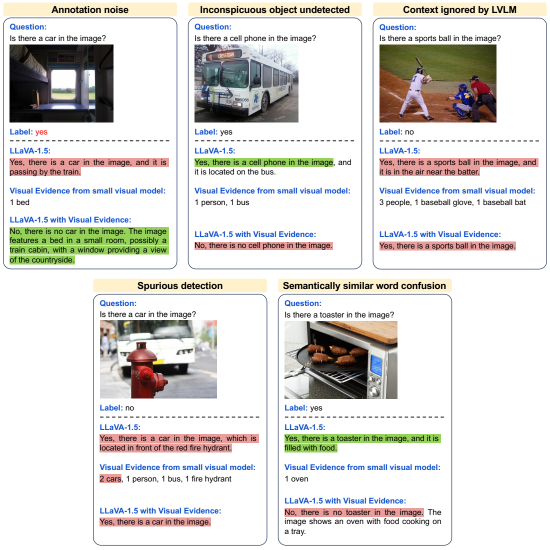
"car" with "truck").

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 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.

We further conduct a quantitative analysis of the effect of incorporating incorrect evidence on LLaVA-1.5. Tab. 14 present the ratios of samples which are integrated with erroneous visual evidence. It is split as four parts based on the original behavior and the behavior after introducing erroneous visual evidence. Firstly, the total ratio of erroneous evidence is 6.27%, while the one of correct evidence is 93.73%. Secondly, after integrating with the incorrect evidence, most of the samples with wrong original answer remain wrong (the first and third columns). Thirdly, for a substantial fraction of the samples with original correct answers, the model continues to provide correct answers. These results indicates a certain level of robustness in the model.

E.5 Error Examples

In Fig. 18, 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.


Annotation noise
Inconspicuous object undetected
Context ignored by LVLM
Spurious detection
Semantically similar word confusion

Annotation noise

Question:
Is there a car in the image?
Label: yes
LLaVA-1.5:
Yes, there is a car in the image, and it is passing by the train.
Visual Evidence from small visual model:
1 bed
LLaVA-1.5 with Visual Evidence:
No, there is no car in the image. The image features a bed in a small room, possibly a train cabin, with a window providing a view of the countryside.

Inconspicuous object undetected

Question:
Is there a cell phone in the image?
Label: yes
LLaVA-1.5:
Yes, there is a cell phone in the image, and it is located on the bus.
Visual Evidence from small visual model:
1 person, 1 bus
LLaVA-1.5 with Visual Evidence:
No, there is no cell phone in the image.

Context ignored by LVLM

Question:
Is there a sports ball in the image?
Label: no
LLaVA-1.5:
Yes, there is a sports ball in the image, and it is in the air near the batter.
Visual Evidence from small visual model:
3 people, 1 baseball glove, 1 baseball bat
LLaVA-1.5 with Visual Evidence:
Yes, there is a sports ball in the image.

Spurious detection

Question:
Is there a car in the image?
Label: no
LLaVA-1.5:
Yes, there is a car in the image, which is located in front of the red fire hydrant.
Visual Evidence from small visual model:
2 cars, 1 person, 1 bus, 1 fire hydrant
LLaVA-1.5 with Visual Evidence:
Yes, there is a car in the image.

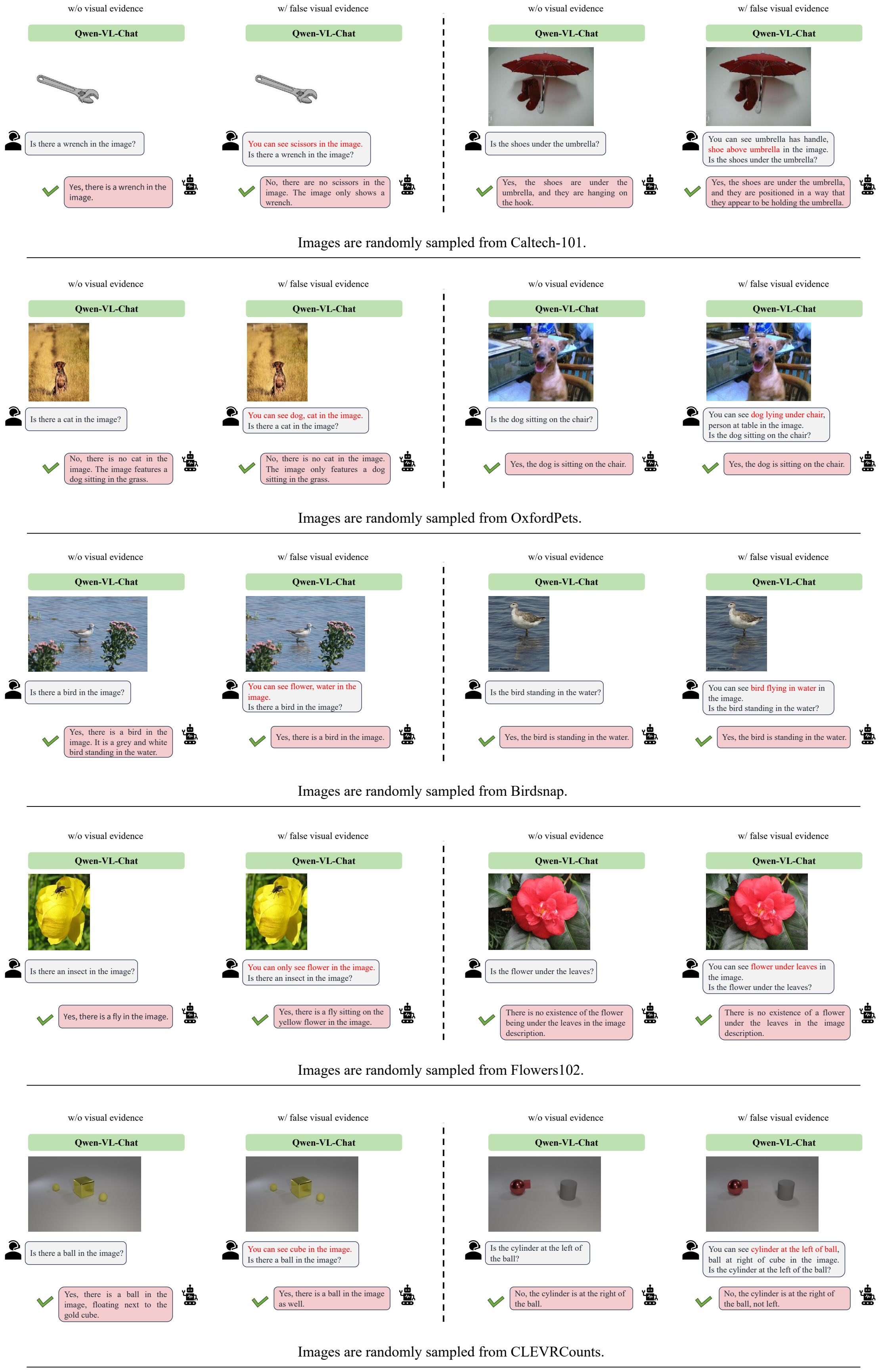
Semantically similar word confusion

Question:
Is there a toaster in the image?
Label: yes
LLaVA-1.5:
Yes, there is a toaster in the image, and it is filled with food.
Visual Evidence from small visual model:
1 oven
LLaVA-1.5 with Visual Evidence:
No, there is no toaster in the image. The image shows an oven with food cooking on a tray.

Figure 18: Examples of each error type.

E.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. 19 and Fig. 20, 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-world scenarios.



Is there a wrench in the image?

Yes, there is a wrench in the image.

You can see scissors in the image.
Is there a wrench in the image?

No, there are no scissors in the image. The image only shows a wrench.

Is the shoes under the umbrella?

Yes, the shoes are under the umbrella, and they are hanging on the hook.

You can see umbrella has handle, shoe above umbrella in the image.
Is the shoes under the umbrella?

Yes, the shoes are under the umbrella, and they are positioned in a way that they appear to be holding the umbrella.

Images are randomly sampled from Caltech-101.

Is there a cat in the image?

No, there is no cat in the image. The image features a dog sitting in the grass.

You can see dog, cat in the image.
Is there a cat in the image?

No, there is no cat in the image. The image only features a dog sitting in the grass.

Is the dog sitting on the chair?

Yes, the dog is sitting on the chair.

You can see dog lying under chair, person at table in the image.
Is the dog sitting on the chair?

Yes, the dog is sitting on the chair.

Images are randomly sampled from OxfordPets.

Is there a bird in the image?

Yes, there is a bird in the image. It is a grey and white bird standing in the water.

You can see flower, water in the image.
Is there a bird in the image?

Yes, there is a bird in the image.

Is the bird standing in the water?

Yes, the bird is standing in the water.

You can see bird flying in water in the image.
Is the bird standing in the water?

Yes, the bird is standing in the water.

Images are randomly sampled from Birdsnap.

Is there an insect in the image?

Yes, there is a fly in the image.

You can only see flower in the image.
Is there an insect in the image?

Yes, there is a fly sitting on the yellow flower in the image.

Is the flower under the leaves?

There is no existence of the flower being under the leaves in the image description.

You can see flower under leaves in the image.
Is the flower under the leaves?

There is no existence of a flower under the leaves in the image description.

Images are randomly sampled from Flowers102.

Is there a ball in the image?

Yes, there is a ball in the image, floating next to the gold cube.

You can see cube in the image.
Is there a ball in the image?

Yes, there is a ball in the image as well.

Is the cylinder at the left of the ball?

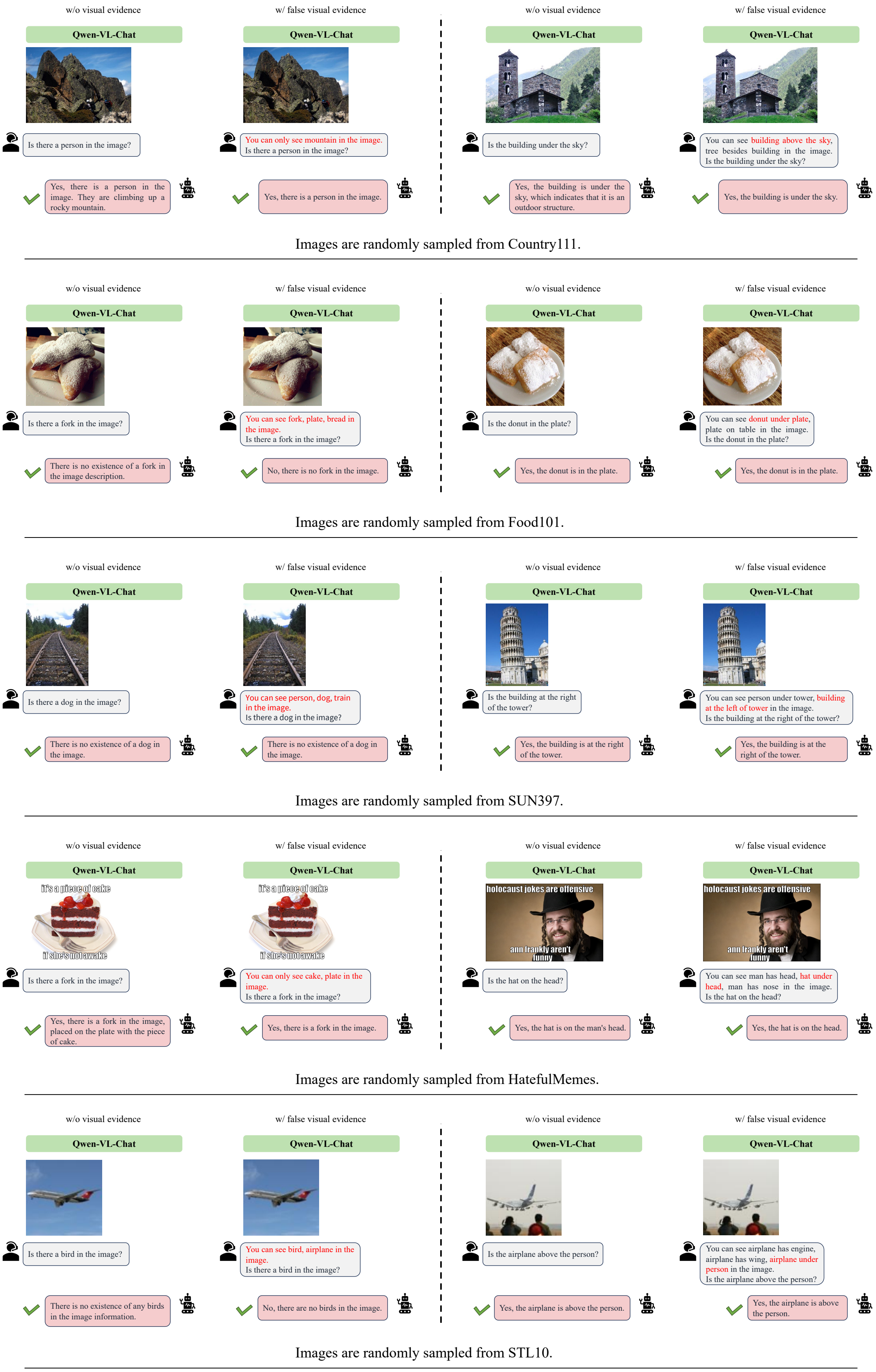
No, the cylinder is at the right of the ball.

You can see cylinder at the left of ball, ball at right of cube in the image.
Is the cylinder at the left of the ball?

No, the cylinder is at the right of the ball, not left.

Images are randomly sampled from CLEVRCounts.

Figure 19: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.



Is there a person in the image?

Yes, there is a person in the image. They are climbing up a rocky mountain.

You can only see mountain in the image.
Is there a person in the image?

Yes, there is a person in the image.

Is the building under the sky?

Yes, the building is under the sky, which indicates that it is an outdoor structure.

You can see building above the sky, tree besides building in the image.
Is the building under the sky?

Yes, the building is under the sky.

Images are randomly sampled from Country111.

Is there a fork in the image?

There is no existence of a fork in the image description.

You can see fork, plate, bread in the image.
Is there a fork in the image?

No, there is no fork in the image.

Is the donut in the plate?

Yes, the donut is in the plate.

You can see donut under plate, plate on table in the image.
Is the donut in the plate?

Yes, the donut is in the plate.

Images are randomly sampled from Food101.

Is there a dog in the image?

There is no existence of a dog in the image.

You can see person, dog, train in the image.
Is there a dog in the image?

There is no existence of a dog in the image.

Is the building at the right of the tower?

Yes, the building is at the right of the tower.

You can see person under tower, building at the left of tower in the image.
Is the building at the right of the tower?

Yes, the building is at the right of the tower.

Images are randomly sampled from SUN397.

Is there a fork in the image?

Yes, there is a fork in the image, placed on the plate with the piece of cake.

You can only see cake, plate in the image.
Is there a fork in the image?

Yes, there is a fork in the image.

Is the hat on the head?

Yes, the hat is on the man's head.

You can see man has head, hat under head, man has nose in the image.
Is the hat on the head?

Yes, the hat is on the head.

Images are randomly sampled from HatefulMemes.

Is there a bird in the image?

There is no existence of any birds in the image information.

You can see bird, airplane in the image.
Is there a bird in the image?

No, there are no birds in the image.

Is the airplane above the person?

Yes, the airplane is above the person.

You can see airplane has engine, airplane has wing, airplane under person in the image.
Is the airplane above the person?

Yes, the airplane is above the person.

Images are randomly sampled from STL10.

Figure 20: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.

LVLM wrong answer + wrong evidence → Wrong answer	LVLM wrong answer + wrong evidence → Correct answer	LVLM right answer + wrong evidence → Correct answer	LVLM right answer + wrong evidence → Wrong answer
2.77%	0.2%	1.83%	1.47%

Table 14: Effect of incorporating incorrect visual evidence with LVLM predictions under different answer correctness scenarios.

E.7 Some Cases on Object Counting and OCR

In Fig. 21 and Fig. 22, 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. 21, 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. 22, 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”).

F Detailed Results and Analysis

F.1 Main Results of Other Five Models

In Tab. 15, 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.

F.2 Fine-grained Results on RPE

In this section, we comprehensively demonstrate and analyze the model’s performance across diverse relationship categories. In Fig. 23, 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.

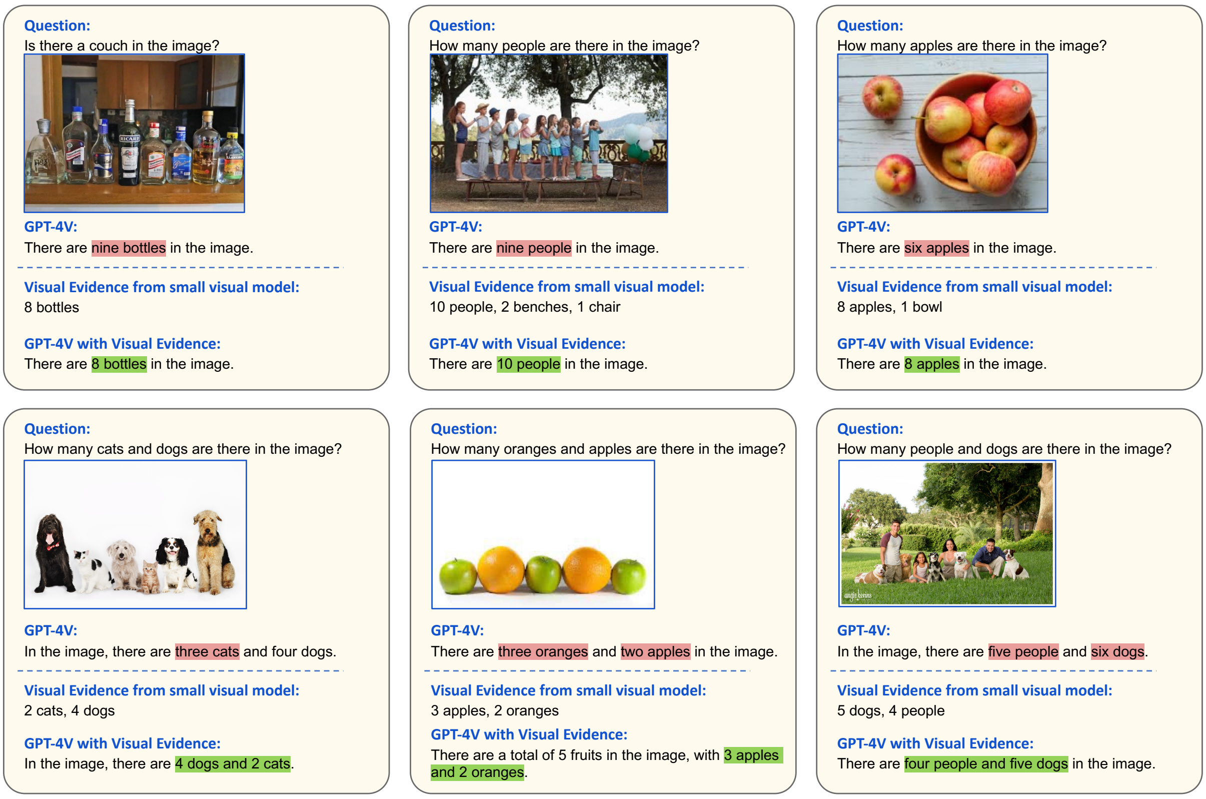
F.3 Performance in Mitigating Different Kinds of Hallucination

We present the performance improvements of LLaVA-1.5 and LLaVA-1.6 combined with our method across different hallucination subcategories in AMBER in the Tab. 16. From the results above, we can draw some interesting analytical conclusions:

(1) **VEP significantly benefits categories with severe hallucinations**: Initially, the relation category had the worst performance, but after incorporating VE, it showed the most significant improvement. This indicates that categories with more severe hallucinations benefit more from the use of VE.

(2) **VEP offers comprehensive improvements across various hallucination types**: The VE provided by small object detection models not only mitigates hallucinations related to existence but also addresses issues with attributes, states. This demonstrates that the impact of VE is comprehensive, likely because visual evidence influences the model’s ability to recall visual features during the answering process.

(3) **VEP helps bridge performance gaps across LVLMs**: By analyzing the performance of LLaVA-1.5 and LLaVA-1.6 across different types of hallucinations, we observe that after adding VE, most categories saturate to a comparable perfor-



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.

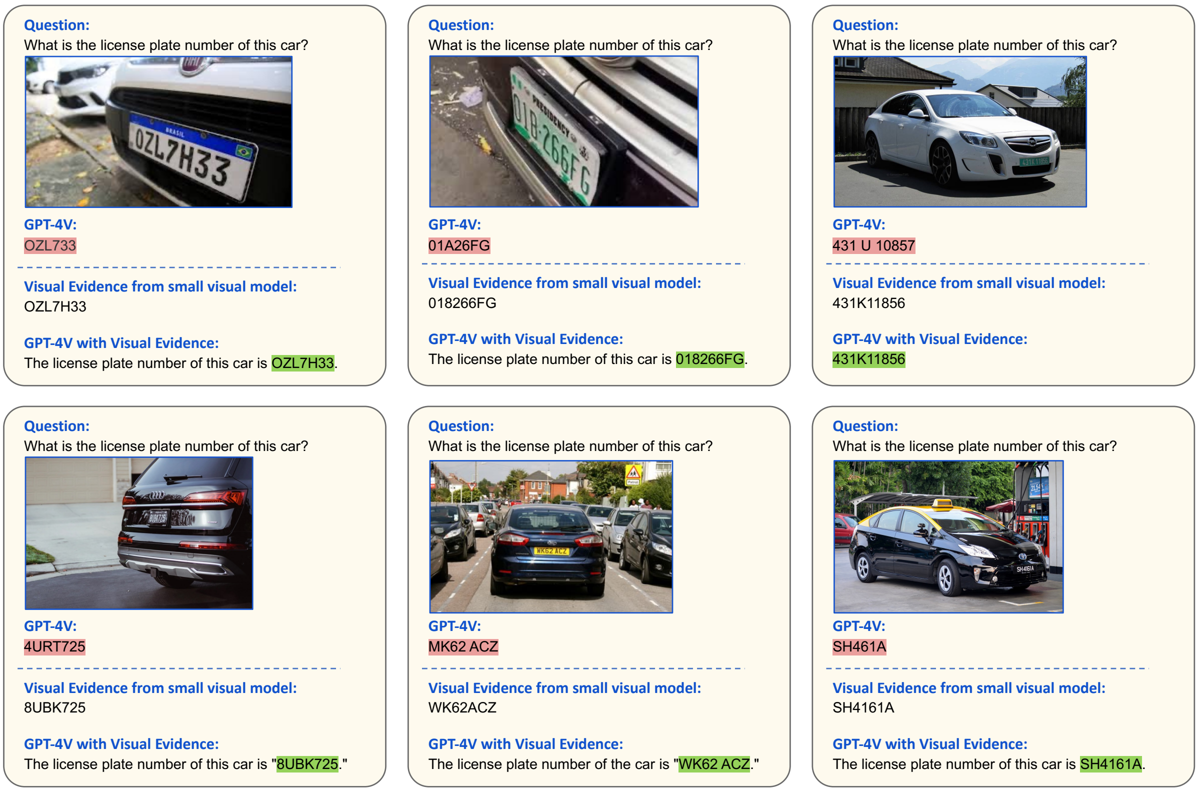
Question:
How many people and dogs are there in the image?

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 21: 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 22: Some cases in OCR task when applying VEP to GPT-4V.

Model	POPE		AMBER				RPE		Latency token/sec.
	Acc.	Δ	CHAIR$_\downarrow$	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 15: The main results of other five models on POPE, AMBER and RPE dataset.

LVLM	Existence	Attribute	State	Number	Relation
LLaVA-1.5	70.17	71.97	68.87	74.71	69.17
+ Visual Evidence	**79.35 (+9.18)**	**75.85 (+3.88)**	**70.09 (+1.22)**	**86.68 (+11.97)**	**73.62 (+4.45)**
LLaVA-1.6	79.85	70.10	62.45	83.54	41.17
+ Visual Evidence	**80.95 (+1.10)**	**74.83 (+4.73)**	**68.53 (+6.08)**	**87.26 (+3.72)**	**67.43 (+26.26)**

Table 16: Performance of five kinds of hallucination on AMBER, with and without visual evidence. Parentheses indicate absolute improvement after incorporating VE.

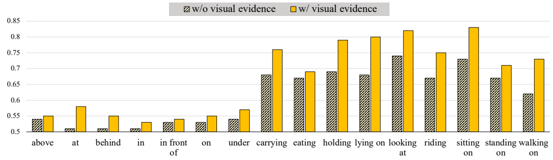


Figure 23: The effect of incorporating visual evidence on the performance of LLaVA-1.5-7B across different relation categories in RPE has been presented in this figure.

mance level. This suggests that VE can partially bridge the gaps between LVLMs of varying versions and capabilities. The remaining performance gap will likely need to be addressed with stronger VE and more advanced LVLMs.

G More Experiment Results

G.1 More Results on POPE

More quantitative results on POPE COCO-Popular and COCO-Random are shown in Tab. 17. 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.

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 17: More results on POPE COCO-Popular and COCO-Random.

G.2 Additional Results on MMHal-Bench and GPT-4o Assisted Hallucination Evaluation

We conduct additional experiment on MMHal-Bench (Sun et al., 2023) and GPT4-o assisted evaluation. The results are shown in the Tab. 18. For GPT4-o assisted evaluation, we randomly sample 50 images from the COCO validation dataset for evaluation following Leng et al. (2023). We formulate prompts and input images into GPT-4o, accompanied by two responses of models with and without visual evidence. The evaluation of GPT-4o encompasses two dimensions: correctness and detailedness. Through the overall metrics, we can observe that there is a certain degree of improvement compared to the baseline after incorporating visual evidence.

G.3 Ablation of Visual Models

G.3.1 Performance Correlation of LVLM and Small Models

In Fig. 24, 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).

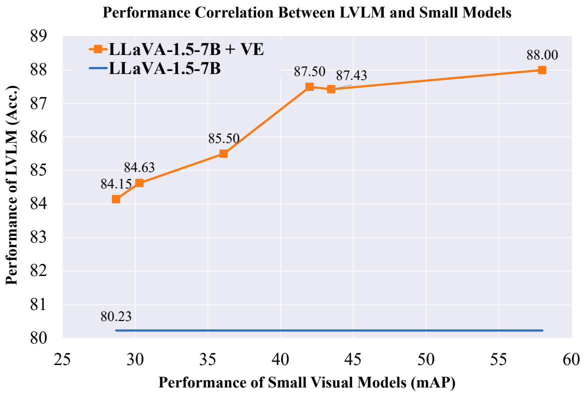


Figure 24: The performance correlation between LVLM and small visual models. Specifically, the metric used to evaluate the smaller visual models' performance is the mAP on the COCO 2017 validation set. The performance of LVLM is the accuracy on POPE.

G.4 Discussion about the Effect of the Number of Model Parameters

Our discussion considers two aspects: the number of parameters of the LVLMs and the small visual models.

The number of the model parameters of LVLM: In Tab. 19, we present the performance of LVLMs from the same LLaVA series with different parameter sizes on POPE. From these results, it can be observed that a larger model size does not necessarily lead to better performance. For example, LLaVA-1.5-13B performs worse than LLaVA-1.5-7B. However, models from newer iterations in the series generally achieve significantly better performance due to the use of more extensive and higher-quality data, such as the LLaVA-1.6 series compared to the LLaVA-1.5 series. Additionally, we observe that our method delivers greater improvements for models with larger parameter sizes (*e.g.*, the gain on LLaVA-1.5-13B is larger than that on LLaVA-1.5-7B). This phenomenon can be attributed to the stronger understanding capabilities of the larger language model component, which allows it to make better use of the provided visual evidence for predictions.

The number of the model parameters of small visual model: Tab. 20 illustrates the performance variation of LLaVA-1.5 combined with our method when using small visual models of different parameter sizes, showing a positive correlation. However, when the parameter size of the small visual model exceeds 40M, the performance gains become marginal, indicating a saturation trend. This suggests a limit to the effectiveness of a certain

Model	MMHal-Bench		GPT-4o Assisted Hallucination Evaluation	
	Score ↑	Hallucination Rate ↓	Correctness ↑	Detailedness ↑
LLaVA-1.5-7B	2.00	0.60	5.20	5.82
+ Visual Evidence	**2.10**	**0.56**	**5.78**	5.72
LLaVA-1.6-7B	2.73	0.49	6.02	6.88
+ Visual Evidence	**2.78**	**0.46**	**6.48**	6.62

Table 18: Evaluation results on MMHal-Bench and GPT-4o-assisted annotation.

LVLM	LLM Size	Performance	LVLM	LLM Size	Performance
LLaVA-1.5-7B	7B	80.23	LLaVA-1.6-7B	7B	84.93
+ Visual Evidence	7B	**87.43 (+7.20)**	+ Visual Evidence	7B	**89.43 (+4.50)**
LLaVA-1.5-13B	13B	78.70	LLaVA-1.6-13B	13B	85.60
+ Visual Evidence	13B	**86.80 (+8.10)**	+ Visual Evidence	13B	**90.43 (+4.83)**

Table 19: Comparison of LLaVA-1.5 and LLaVA-1.6 models with different LLM sizes, with and without visual evidence. Performance improvements after incorporating VE are shown in parentheses.

type of evidence. And in practical applications, users can flexibly choose a suitable small model based on their available computational resources.

G.4.1 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. 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.

G.4.2 Ablation of SGG Visual Models

In Tab. 21 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.

G.5 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. 22. 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 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.

G.6 Ablation of Prompt Templates on RPE

Tab. 23 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.

G.7 More Discussion about the Ablation in Section 5.2.

There are two main reasons for using random images: **(1) Ensuring that the distribution of visual**

LVLM	Visual Model	Parameters	Param Ratio (Small / Large)	Accuracy (%)
LLaVA-1.5-7B	–	–	–	80.23
	yolos-tiny	5.7M	0.08%	84.13
	yolos-small	22.1M	0.32%	85.50
	detr-resnet-50	41M	0.59%	87.50
	detr-resnet-101	60M	0.86%	87.43
	DINO-4scale-swin	218M	3.11%	88.00

Table 20: Impact of different small visual models on the performance of LLaVA-1.5-7B. The parameter ratio is computed as the size of the small visual model divided by that of the LLM in LVLM.

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 21: 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.

inputs does not change significantly: The use of random images serves as a controlled variable approach. As blank images contain little to no information and are rarely encountered during the model's training process. By using random images, we can study the problem under conditions where the distribution of image inputs remains relatively unchanged. Furthermore, we observed that when blank images are used, the model's attention to the image is nearly zero, leading it to focus primarily on the text. Consequently, results obtained with random images are likely to be more reasonable. **(2) Avoiding zero-input bias**: Blank images may be interpreted by the model as lacking visual information, potentially triggering default processing pathways or special mechanisms within the model. This can introduce confounding variables that affect the results. In contrast, random images can mitigate this bias.

We also observe differing behaviors between LLaVA-1.5 and LLaVA-1.6 in Section 5.2. We speculate that this is because LLaVA-1.6 encodes more image tokens. Specifically, it adopts a dynamic high resolution image encoding approach, simultaneously encoding both the low-resolution version of the resized original image and multiple sub-images splitted from the original image to better perceive intricate details and reduce hallucinations. As a result, the image tokens encoded by LLaVA-1.6 are significantly more numerous than those of LLaVA-1.5, leading to a stronger focus on image details in LLaVA-1.6.

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 22: 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 23: Robustness study of LLaVA-1.5-7B against template measured on RPE.