A APPENDIX: DISCUSSIONS

LIST OF REVISIONS

A.1 OUT-OF-DOMAIN DATA

Reviewer rr7K, Q2
Reviewer w1s8, Q1

We also conduct more evaluations on more out-of-domain datasets. Specifcally, we collect 2,540 additional samples from open-world datasets and scenarios to further evaluate the genaliation ability of our method. 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. The Object365 dataset is a collection of images that aims to provide a comprehensive representation of objects commonly found in indoor environments. It consists of over 365 object categories, with each category containing multiple images depicting different instances of the object. The OpenImage dataset consists of millions of images covering a wide range of categories, including objects, scenes, and events. It provides valuable annotations for each image, including object bounding boxes, class labels, and object relationships. Table 4 present the comparison with baseline results for the evaluation on out-of-domain datasets. The experimental results indicate that in out-of-domain open scenarios, incorporating visual evidence can still mitigate the hallucination of LVLMs significantly.

In 20.6% of the images, small model captures incorrect or partial correct object or relation information. With these visual evidences, only 8% of the false evidence confuse the LVLM and change the response from collect to wrong. Finally, we would like to point out that the contribution of our method lies in combining small and large models, utilizing the domain-specific knowledge of small models to complement the large models. In practical applications, it is possible to customize domain-specific small models to tailor different domain knowledge.

A.2 WHY NOT FINETUNING?

Reviewer w1s8, Q5

It is a common practice to fine-tune foundation models on specific tasks to enhance task performance or align the model's behavior with human expectations. It is well-known that the foundation models gain speciality to achieve exceptional performance on the fine-tuning task, but it can potentially lose its generality. This phenomenon is closely associated with the concept of catastrophic forgetting observed in deep neural networks.

Evaluation	Model	Accuracy	F1 Score	Yes (%)
Object Hallucination (Out-of-domain)	mPLUG-Owl	52.04	66.81	94.52
	+ Visual Evidence	**62.46**	**69.43**	72.67
	Qwen-VL-Chat	70.25	60.31	24.95
	+ Visual Evidence	**76.74**	**75.50**	45.01
Relation Hallucination (Out-of-domain)	mPLUG-Owl	58.52	69.06	84.07
	+ Visual Evidence	**72.41**	**75.29**	66.88
	Qwen-VL-Chat	73.93	71.54	41.71
	+ Visual Evidence	**75.98**	**72.84**	38.18

Table 4: Detailed object and relation hallucination evaluation results on out-of-domain datasets constructed from Object365 (2000 samples) and OpenImage (540 samples).

Previous work Zhai et al. (2023) has conducted fine-tuning experiments on LLaVA. As the fine-tuning progresses, LLaVA starts to hallucinate by disregarding the questions and exclusively generating text based on the examples in the fine-tuning datasets. As in the Table 3 in Zhai et al. (2023), after 1 epoch finetuning LLaVA-7b on MNIST, the accuracy on CIFAR-10 significantly drops from 56.71% to 9.27%. On the other hand, our prompt-based method does not modify the parameters of the model, and offer greater controllability, which is advantageous for preserving the model's original generalization capability.

A.3 Overlap between objects in evidence and questions

Reviewer YNjN, Q1

	correct → correct	correct → wrong	wrong → correct	wrong → wrong
Type A	139 (46.7%)	8 (2.7%)	110 (36.9%)	41 (13.8%)
Type B	415 (34.5%)	22 (1.8%)	563 (46.8%)	202 (16.8%)

Table 5: Robustness against the overlap between objects in questions and objects in evaluation datasets' questions. For example, "wrong → correct" denotes the samples that were initially answered incorrectly and answer correct after provided with visual evidence.

We calculate the current stats of the overlap between objects in questions and objects in evaluation datasets' questions. In the 3,000 visual evidence prompts, there are 298 prompts that contains object that are not in the question (Type A), and 1,202 prompts that contain objects that are not in the questions exclusively (Type B).

Following Figure 3, we calculate the stats of samples which were initally answered corrently/wrongly and answer correctly/wrongly after provided with Type A/B prompts (Table 5. In the 298 Type A prompts, 110 of which (36.9%) allievates the hallucination of LVLM with detr-resnet-101 on Qwen-VL-Chat. In the $1,202$ Type B prompts, 563 of which (46.8%) allievates the hallucination of LVLM.

Model	Setting	Accuracy
mPLUG-Owl	baseline	57.29%
	+ visual evidence	78.38%
	+ visual evidence to synonyms	71.54%
Qwen-VL-Chat	baseline	81.23%
	+ visual evidence	87.70%
	+ visual evidence to synonyms	86.53%

Table 6: Robustness against the object labels to synonyms.

Model	baseline	+ object labels	+ relation labels
mPLUG-Owl	63.62%	71.41%	75.68%
Qwen-VL-Chat	62.58%	66.88%	68.46%

Table 7: Only use object labels as visual evidence for relation hallucination

Model	Setting	In-domain objects	Out-of-domain objects
mPLUG-Owl	Baseline	58.68%	48.45%
	+ Visual Evidence	65.38%	60.87%
Qwen-VL-Chat	Baseline	74.64%	67.87%
	+ Visual Evidence	79.77%	75.10%

Table 8: Performance on open-vocabulary objects.

With the help of ChatGPT, we also manually change the object appear in question to its synonyms respectively. The evaluation of object hallucination slightly decreases from 87.70% to 86.53% on Qwen-VL-Chat and from 78.38% to 71.54% on mPLUG-Owl, but there is still a non-trival improvement over the baseline especially on mPLUG-Owl, the results are shown in the Table 6.

A.4 Only object labels

Reviewer bJ8t, Q1

We conduct validation experiments using the detr-resnet-101 model to provide object labels as evidence for relation hallucination.

The results in Table 7 show that providing object labels as evidence also has some improvement although not as effective as relation label. We suppose it is because object labels themselves contain crucial object information from the image, which leads to mitigating relation hallucination. This result not only validates the necessity of relation labels but also further verifies that our approach is orthogonal to the specific task.

A.5 Open-vocabulary objects and few-shot relations

Reviewer bJ8t, Q2

We construct a new out-of-domain object hallucination dataset with 2000 samples using the test sets from Object365 (Shao et al., 2019) following the construction idea of POPE Li et al. (2023b). This dataset is divided into two parts. One part includes 80 objects that defined in COCO, while the other portion consists of objects that do not appear in COCO. The performance of these two parts are shown in the Table 8. It can be observed that there is a consistent improvement in performance for both in-domain and out-of-domain object categories.

We chose the bottom-10 tail relations as defined in (He et al., 2020) of VG to construct a medium-sized relation hallucination dataset with 1006 samples. We used OpenPSG as the SGG model and conducted experiments on Qwen-VL-Chat and mPLUG-Owl. The experiment results are shown in the table 9 below, and it can be seen that our framework still achieves significant improvements in few-shot relations.

B Appendix: Datasets

Object hallucination datasets. In POPE(Li et al., 2023b), 500 images are randomly selected from the validation set of COCO Vinyals et al. (2016), with more than three ground-truth objects in the annotations. For each image, 6 questions are constructed from annotations whose answers are "Yes". For questions with the answer "No", three strategies, *i.e.*, Random, Popular, Adversarial, are considered to sample their probing objects. The difficult of question increases from Random to

Model	Setting	Accuracy (%)
mPLUG-Owl	Baseline	61.23%
	+ Visual Evidence	67.89%
Qwen-VL-Chat	Baseline	55.67%
	+ Visual Evidence	68.63%

Table 9: Performance on few-shot relations.

Adversarial. For MSCOCO-Random, objects that do not exist in the image are randomly sampled. For MSCOCO-Popular, the top-3% most frequent objects in MSCOCO are selected. For MSCOCO-Adversarial, first rank all the objects according to their co-occurring frequencies with the ground-truth objects, and then select the top-k most frequent ones that do not exist in the image.

Relation hallucination datasets. Firstly we categorize the all 50 relationships into two groups, *i.e.*spatial and action relationships. The spatial relation categories include above, at, behind, in, in front of, near, on, on back of, over, underand laying on. The spatial relationship categories consist of carryingcovered incoveringeatingflying ingrowing onhanging from holdinglying onlooking atmounted onparked on, ridingsitting onstanding on, walking onwalking in, and watching. Subsequently, we proceed to select 7 spatial relationships, specifically above, at, behind, in, in front of, on, and under, as well as 9 action relationships, namely carrying, eating, holding, lying on, looking at, riding, sitting on, standing on, and walking on. 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 Table 10. 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 Table 10. Finally the dataset consists of 2400 triplets of image, question and answer, in which 1200 are "Yes" and 1200 are "No". In Figure 6, we show some cases in our dataset.

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 10: 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 6: Several cases in our proposed relation hallucination dataset are depicted in this figure, with the two on the left representing spatial relations and the two on the right illustrating action relations.

C Appendix: Detailed Object Hallucination Evluation

C.1 More results about object hallucination

The evaluation results on MSCOCO-Popular and MSCOCO-Random of Qwen-VL-Chat, mPLUG-Owl, MiniGPT-4 and LRV-instruction are presented in Table 11.

Object hallucination datasets	**Model**	Accuracy	F1 Score	Yes (%)
MSCOCO-Popular	mPLUG-Owl	57.96	68.96	85.39
	+ LRV-Instruction	71.81	74.48	60.26
	+ Visual Evidence	**77.63**	**77.86**	50.89
	MiniGPT-4	73.67	73.04	47.67
	+ LRV-Instruction	66.24	73.57	77.64
	+ Visual Evidence	**80.77**	**82.08**	57.30
	Qwen-VL-Chat	85.53	85.08	47.00
	+ Visual Evidence	**89.60**	**88.87**	43.40
MSCOCO-Random	mPLUG-Owl	62.86	72.11	81.55
	+ LRV-Instruction	79.41	80.90	55.64
	+ Visual Evidence	**81.18**	**81.22**	48.70
	MiniGPT-4	80.85	79.33	41.08
	+ LRV-Instruction	69.48	76.08	75.77
	+ Visual Evidence	**89.69**	**89.87**	50.24
	Qwen-VL-Chat	88.17	87.79	45.29
	+ Visual Evidence	**90.79**	**90.31**	43.51

Table 11: Detailed object hallucination evaluation results of LVLMs on MSCOCO-Popular and MSCOCO-Random using POPE evaluation pipeline.

D Appendix: Different Visual Models and LVLMs

D.1 More results about the performance of LVLMs incorporated with visual models of different capacities

In Figure 12, Figure 13 and Figure 14, we show more results on the MSCOCO-Popular and MSCOCO-Random about the performance of three LVLMs incorporated with visual models of different capacities. In Figure 15, we present the results on VG Relation Hallucination dataset 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 mPLUG-Owl and Qwen-VL-Chat. For example, RelTR achieves 5.92% and MOTIFS achieves 5.8% improvement on mPLUG-Owl. RelTR achieves 11.35% and MOTIFS achieves 12.55% improvement on Qwen-VL-Chat. The gains brought by different scene graph generation models to LVLM are within a stable range (saturated).

Reviewer bJ8t, Q4

D.2 More results on different LVLMs and larger detection models

Reviewer w1s8, Q4

We have also conducted experiments on LLaVA and LLaVA-1.5 to further validate the effectiveness of our method.

It is observed that the hallucination evaluation of LLaVA-1.5 is indeed state-of-the-art, with an accuracy of 84.47% for object hallucination. However, it still exhibits a significant amount of relation hallucination, with an accuracy of 70.38%. Besides LLaVA, visual evidence prompting further helps LLaVA-1.5 alleviate both object and relation hallucination capabilities 84.47% $\rightarrow$ 90.20%, 70.38%

Datasets	Visual model		Qwen-VL-Chat		
	Model name	mAP	Accuracy	F1 Score	Yes (%)
MSCOCO-Popular	-	-	85.53	85.08	47.00
	yolos-tiny	28.7	**85.90**	**84.27**	39.63
	yolos-small	36.1	**87.37**	**86.12**	41.03
	detr-resnet-50	42.0	**89.10**	**88.23**	42.63
	detr-resnet-101	43.5	**89.60**	**88.87**	43.40
MSCOCO-Random	-	-	88.17	87.79	45.29
	yolos-tiny	28.7	**86.74**	**85.43**	39.52
	yolos-small	36.1	**88.18**	**87.22**	40.96
	detr-resnet-50	42.0	**90.10**	**89.49**	42.61
	detr-resnet-101	43.5	**90.79**	**90.31**	43.51

Table 12: Object hallucination results of Qwen-VL-Chat incorporating visual evidence from different object detection models, *i.e*. yolos-tiny Fang et al. (2021), yolos-small Fang et al. (2021), detr-resnet-50 Carion et al. (2020) and detr-resnet-101 Carion et al. (2020). The mAP on COCO 2017 validation of different visual models is also reported.

Datasets	Visual model		mPLUG-Owl		
	Model name	mAP	Accuracy	F1 Score	Yes (%)
MSCOCO-Popular	-	-	57.96	68.96	85.39
	yolos-tiny	28.7	**70.15**	**70.74**	51.87
	yolos-small	36.1	**73.92**	**73.74**	49.33
	detr-resnet-50	42.0	**76.24**	**76.67**	51.62
	detr-resnet-101	43.5	**77.63**	**77.86**	50.89
MSCOCO-Random	-	-	62.86	72.11	81.55
	yolos-tiny	28.7	**73.06**	**73.03**	48.38
	yolos-small	36.1	**76.61**	**76.73**	49.00
	detr-resnet-50	42.0	**80.65**	**80.78**	49.01
	detr-resnet-101	43.5	**81.18**	**81.22**	48.70

Table 13: Object hallucination results of mPLUG-Owl incorporating visual evidence from different object detection models, *i.e*. yolos-tiny Fang et al. (2021), yolos-small Fang et al. (2021), detr-resnet-50 Carion et al. (2020) and detr-resnet-101 Carion et al. (2020). The mAP on COCO 2017 validation of different visual models is also reported.

Datasets	Visual model		MiniGPT-4		
	Model name	mAP	Accuracy	F1 Score	Yes (%)
MSCOCO-Popular	-	-	73.67	73.04	47.67
	detr-resnet-50	42.0	**80.70**	**81.99**	57.17
	detr-resnet-101	43.5	**80.77**	**82.08**	50.89
MSCOCO-Random	-	-	80.85	79.33	41.08
	detr-resnet-50	42.0	**89.55**	**88.20**	49.83
	detr-resnet-101	43.5	**89.69**	**89.87**	50.24

Table 14: Object hallucination results of MiniGPT-4 incorporating visual evidence from different object detection models, detr-resnet-50 Carion et al. (2020) and detr-resnet-101 Carion et al. (2020). The mAP on COCO 2017 validation of different visual models is also reported.

Model	Visual model		Performance of LVLMs		
	Model name	mAP	Accuracy	F1 Score	Yes (%)
mPLUG-Owl	-	-	62.58	71.18	79.83
	RelTR	18.9	**68.50**	**73.06**	66.92
	MOTIFS	20.0	**68.38**	**73.17**	67.88
	OpenPSG	28.4	**68.25**	**72.82**	66.83
Qwen-VL-Chat	-	-	63.62	46.99	18.62
	RelTR	18.9	**74.97**	**75.62**	52.70
	MOTIFS	20.0	**75.80**	**76.44**	52.77
	OpenPSG	28.4	**76.17**	**77.00**	53.71

Table 15: Relation hallucination results of Qwen-VL-Chat and mPLUG-Owl 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.

$\rightarrow 75.08\%$, thereby providing further evidence of the effectiveness of our method. This indicates that not only can different small models help alleviate hallucinations in large models, but a single small model can consistently alleviate hallucinations in large models of different sizes and trained on different datasets. This result further confirms the complementarity between large and small models and the necessity of our framework.

We also conduct experiments with larger open-source detection models, DINO (Zhang et al., 2022), which is the top-tier model with 58.0 mAP in COCO 2017 val (detr-resnet-101 has 43.5 mAP). The results are shown in Table 17, it can be observed that as the mAP increases, the small model consistently provides a boost to the large model, though it gradually saturates. Reviewer w1s8, Q3

Evaluation	Model	Accuracy
Object Hallucination	LLaVA	60.23
	+ Visual Evidence	**77.43**
	LLaVA-1.5	84.47
	+ Visual Evidence	**90.20**
Relation Hallucination	LLaVA	64.49
	+ Visual Evidence	**70.54**
	LLaVA-1.5	70.38
	+ Visual Evidence	**75.08**

Table 16: Object and relation hallucination evaluation results on LLaVA and LLaVA-1.5.

E Appendix: More Ablations

E.1 Visual results of erroneous evidence

We show some cases where the model insists on the correct answer when wrong visual evidence is provided and some cases where the model was misled by the wrong evidence.

E.2 More results about the ablation of question templates in object and relation hallucination evaluation

To verify the stability of our method against different question prompt templates,, As shown in Table 18 and Table 19, under different question templates, Qwen-VL-Chat shows consistent performance gain with low standard deviations in both object hallucination and relation hallucination datasets. Such results further validate the robustness of our method. We also present the results of robustness study of visual evidence prompting against different input prompts on relation evaluation.

Model	Visual model Model name	mAP	Performance of LVLMs Accuracy
mPLUG-Owl	-	-	57.29
	yolos-tiny	28.7	70.23 (+12.94)
	yolos-small	36.1	73.44 (+3.21)
	detr-resnet-50	42.0	76.55 (+3.11)
	detr-resnet-101	43.5	78.38 (+1.83)
	DINO-4scale-swin	58.0	**79.44 (+1.06)**
Qwen-VL-Chat	-	-	81.23
	yolos-tiny	28.7	**83.73**
	yolos-small	36.1	**85.47**
	detr-resnet-50	42.0	**87.10**
	detr-resnet-101	43.5	**87.70**
	DINO-4scale-swin	58.0	**89.17 (+1.46)**

Table 17: Object hallucination results of Qwen-VL-Chat and mPLUG-Owl incorporating visual evidence from larger version object detection model (Liu et al., 2022). The values in parentheses indicate the performance improvement compared to the previous row's setting in the table.

Figure 7: Some cases wherein the initial response was correct, and the model persistently produced correct response despite provided with erroneous visual evidence. The incorrect evidence is underlined in red font. The annotated object categories is presented below each case.

Figure 8: Some cases wherein the initial response was correct, and the model produce wrong response when provided with erroneous visual evidence. The incorrect evidence is underlined in red font. The annotated object categories is presented below each case.

Question Prompt Templates	Accuracy		F1 Score	
	Baseline	Baseline + VE	Baseline	Baseline + VE
Is there a <object> in the image?	80.93	**87.73**	81.10	**86.63**
Does the image contains a <object>?	83.32	**87.37**	82.46	**86.01**
Is there any <object> present in the image?	80.53	**87.03**	80.80	**85.76**
Can you see a <object> in the image?	80.85	**87.17**	80.46	**85.81**
Avg±Std.	81.41±1.11	**87.32**±0.26	81.20±0.76	**86.05**±0.35

Table 18: The evaluation results of Qwen-VL-Chat on MSCOCO-Adversarial before and after incorporating visual evidence across diverse question prompt templates are presented in this table.

Question Prompt Templates	Accuracy		F1 Score	
	Baseline	Baseline + VE	Baseline	Baseline + VE
Is the <subject> <relation> the <object>?	63.62	**74.97**	46.99	**75.66**
Can you see the <subject> <relation> the <object>?	55.58	**73.14**	21.62	**73.35**
Is the <subject> <relation> the <object> in the image?	64.96	**74.99**	51.36	**74.86**
Can you see the <subject> <relation> the <object> in the image?	57.33	**73.15**	27.99	**72.95**
Avg±Std.	60.37±3.99	**74.06**±0.92	36.99±12.48	**74.21**±1.10

Table 19: The evaluation results of Qwen-VL-Chat on VG Relation Hallucination dataset before and after incorporating visual evidence across diverse question prompt templates are presented in this table.

Visual Evidence Prompt Templates	Accuracy	F1 Score
{question}	63.62	46.99
Evidence: There are {evidence} in the image.\n Let's refer to the evidence and then answer the following question.\n{question}	74.48	75.34
Evidence: You can see {evidence} in the image.\n Let's consider the evidence and then answer the following question.\n{question}	74.97	75.66
Evidence: You can see {evidence} in the image.\n {question} According to the image and evidence, the answer is	75.27	73.48
You can see {evidence} in the image.\n Then answer the question based on what you see: {question}	**75.50**	**78.52**
It's a beautiful day.\n{question}	53.33	13.85
There is nothing in the image.\n{question}	55.56	20.68

Table 20: Robustness study of Qwen-VL-Chat against template measured on the VG Relation Hallucination dataset.

F Appendix: Case Study

F.1 More cases on out-of-domain images

More out-of-domain cases are shown here. 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 & 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 Figure 11 and Figure 12, 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 to incorrect evidence in open scenarios.

Reviewer rr7K, Q2

Reviewer w1s8, Q1

F.2 Some cases on object counting and OCR

In Figure 9 and Figure 10, we also show some cases that other tasks like object counting and OCR can also be enhanced by our framework.

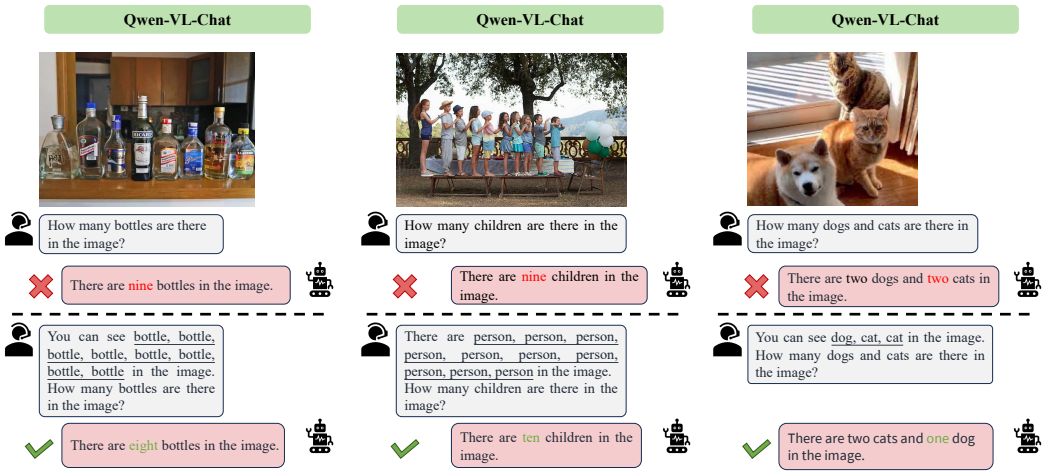

Figure 9: Some cases of Qwen-VL-Chat on object counting task.

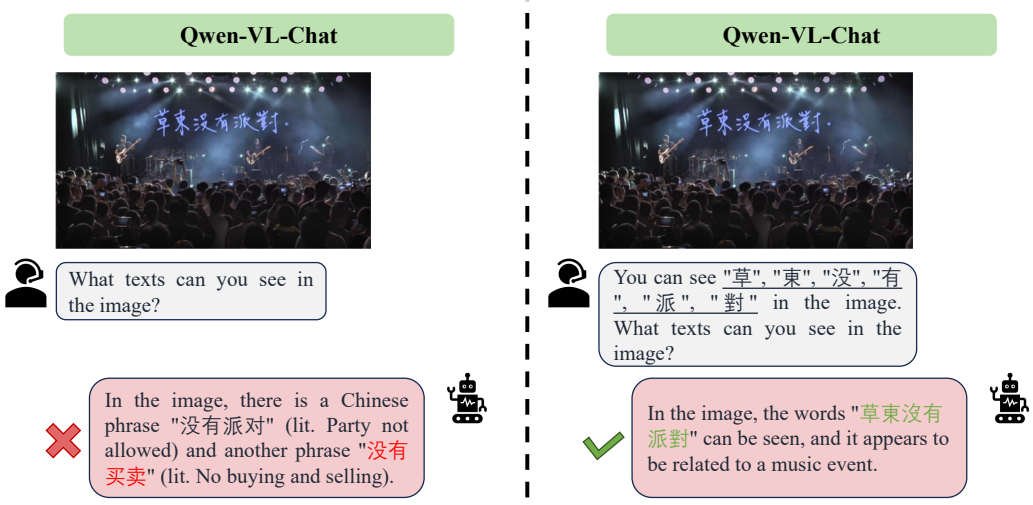

Figure 10: A case of Qwen-VL-Chat on OCR task.

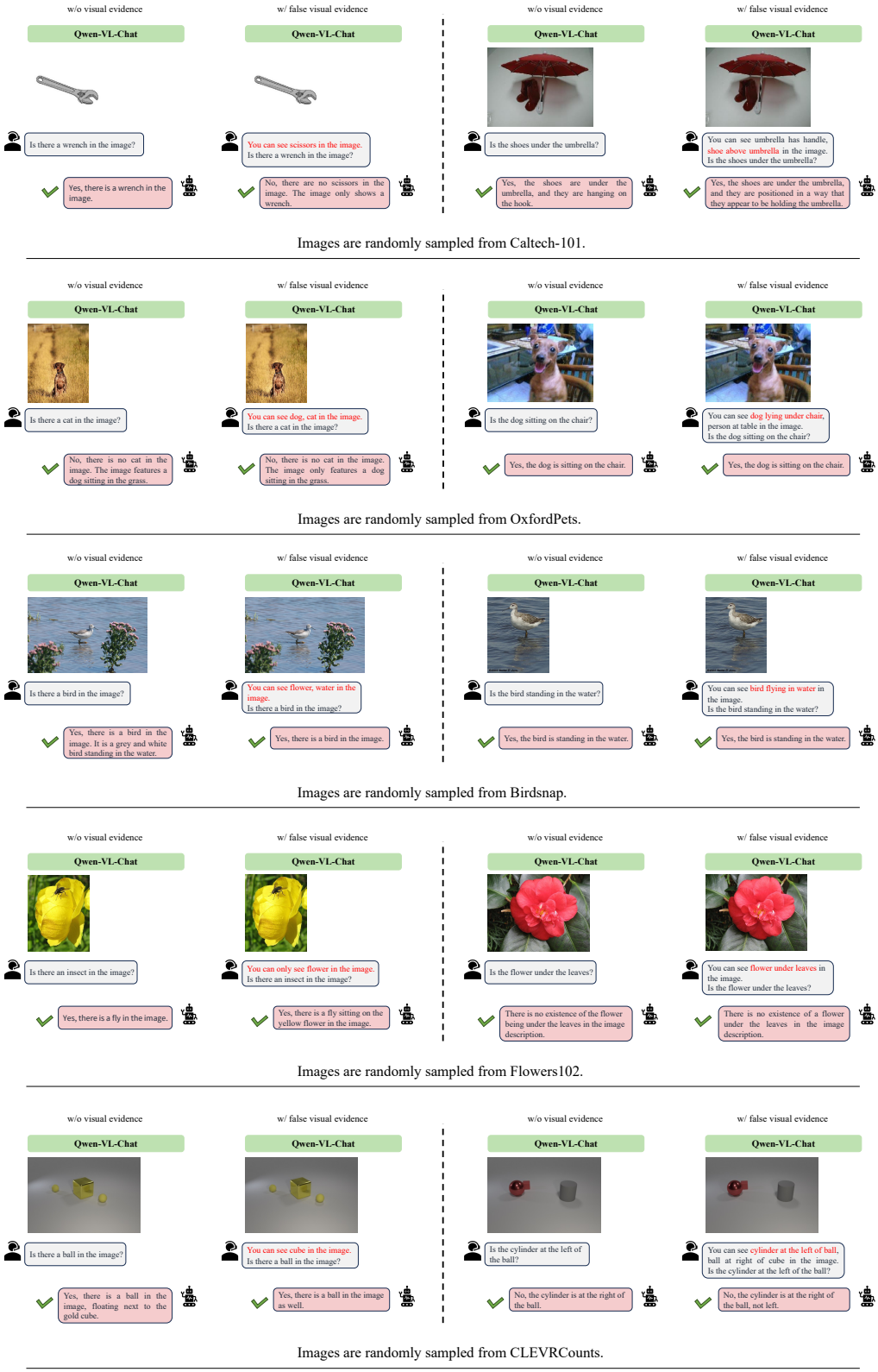

Figure 11: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.

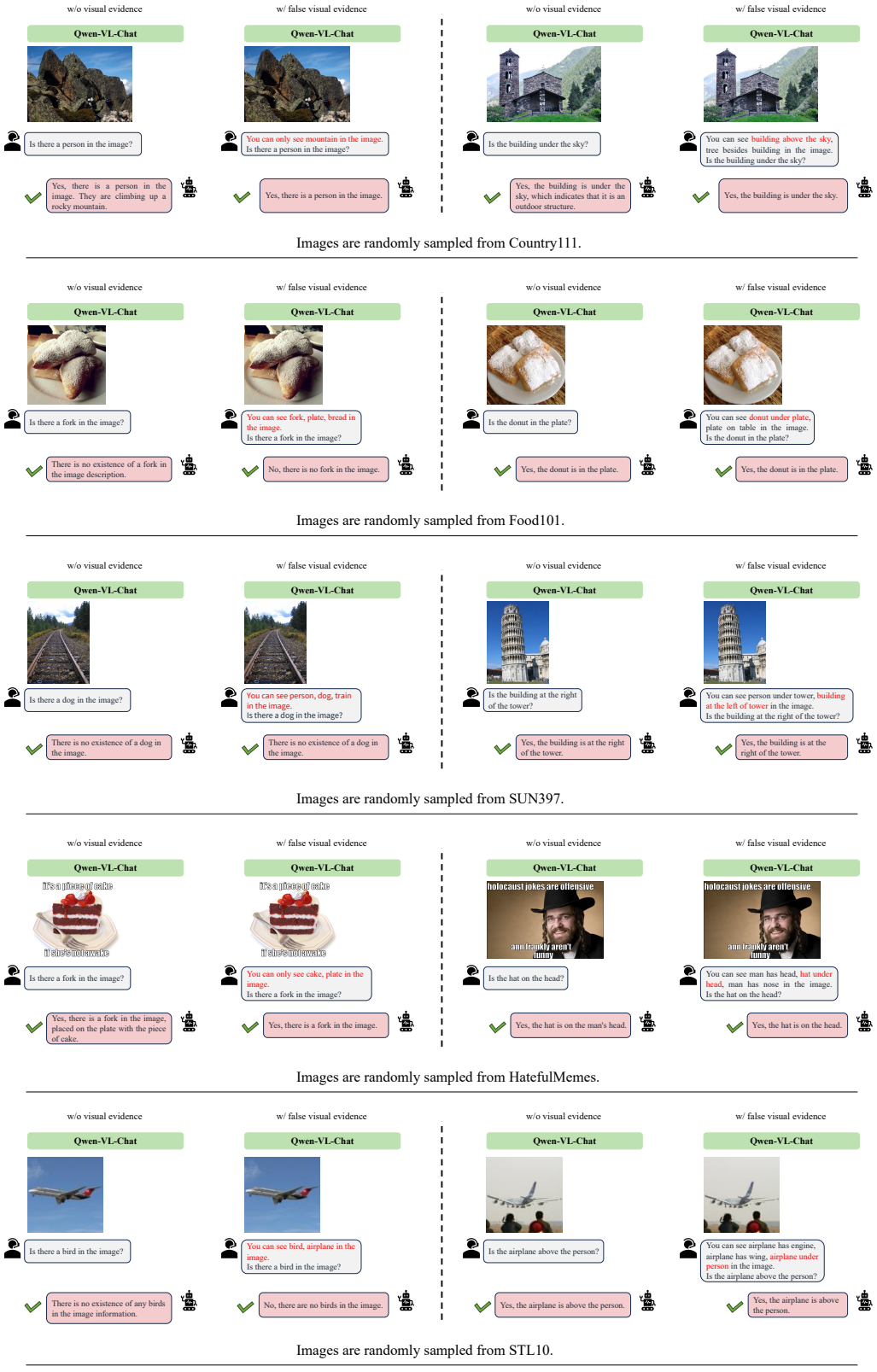

Figure 12: Some open-scenario cases from different out-of-domain datasets when LVLM are provided with false visual evidence.

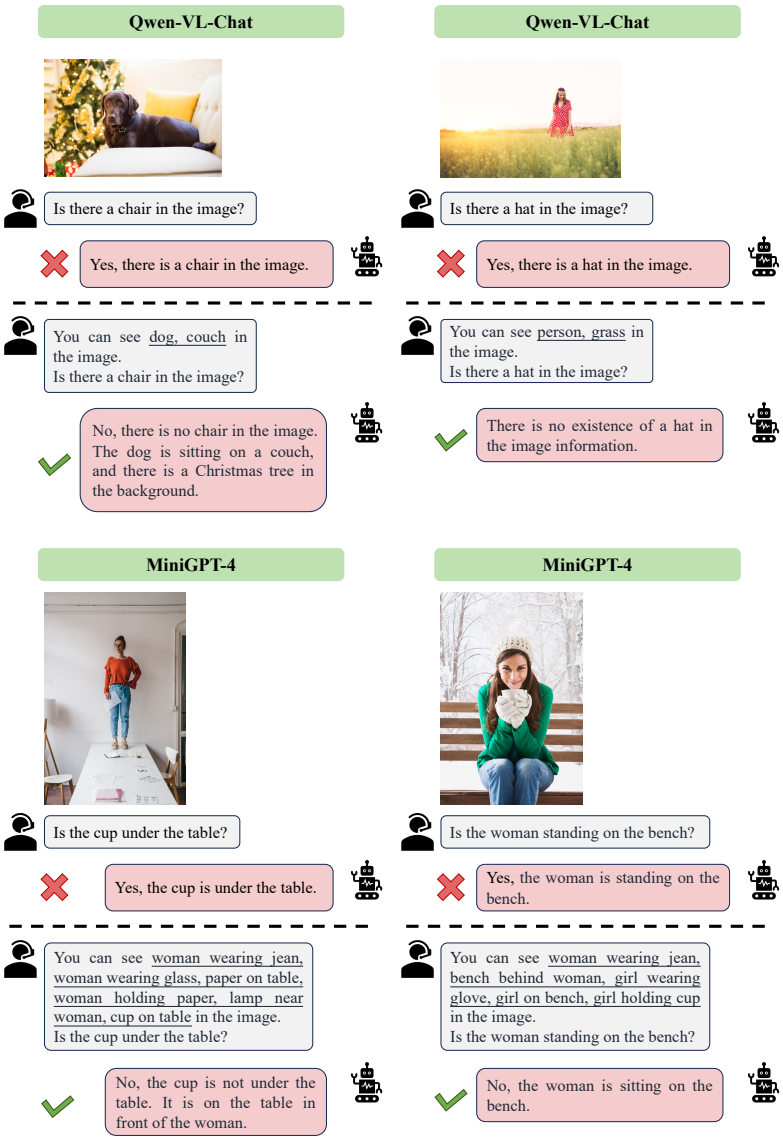

Figure 13: More out-of-domain cases are shown in this figure, the images are from winoground (Thrush et al., 2022).