Attention Prompting on Image for Large Vision-Language Models

- Supplementary Material -

Contents

1 Examples 2

2 Notation Table 10

3 Observation and Discussion of API Method 13

3.1 CLS Token Similarity and Non-CLS Token Similarity . 13

3.2 Attribution Map Aggregation for CLIP Model 14

4 More Experimental Results and Implementation Details . . 14

4.1 Ensemble 14

4.2 Influence on Different VQA Abilities 15

4.3 Implementation Details 16

5 Limitation, Future Direction, and Potential Impact 18

1 Examples

Question

Question: *What fruit is in the left part of the fridge?*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

GPT-4V + Original Image:

*On the left side of the fridge, there is a clear container filled with **strawberries**. **Below that container is another one with blueberries**. Both strawberries and blueberries are types of fruit.*

GPT-4V + API-Generated Image:

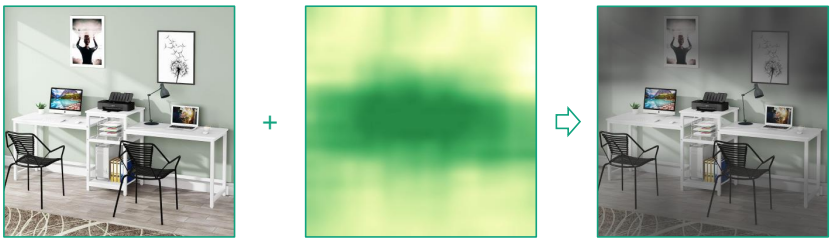
*In the left part of the fridge, there are **strawberries**. They appear to be stored in a clear, plastic **clamshell container**, which is quite common for berry packaging.*

Fig. 2: In complex images including multiple objects, our method accurately highlights the fruits and masks the other objects, thereby simplifying the scene and facilitating the LVLM's inference of spatial relationships.

Question

Question: *On the right desk, what is to the left of the laptop?*

Input Images



Original Image | Heatmap | API-Generated Image

Answers from LVLM

GPT-4V + Original Image:

To the left of the laptop on the right desk, there is ***a white cup or mug*** *that seems to be placed on* ***a light-colored coaster or a small dish****.*

GPT-4V + API-Generated Image:

To the left of the laptop on the right desk, there appears to be ***a desk lamp****.* ***The lamp is black****, and its design suggests it is adjustable, with a classic desk lamp shape and a directional shade that can be oriented to direct the light where needed.*

Fig. 3: Our method identifies regions related to the objects, thereby assisting the LVLM in spatial reasoning.

Question

Question: *What is the index of the step when we need to add all purpose flour?*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

LLaVA-1.5 + Original Image:

*The index of the step when we need to add all purpose flour is **step 5**.*

LLaVA-1.5 + API-Generated Image:

*The index of the step when we need to add all purpose flour is **step 3**.*

Fig. 4: Our method assists LVLM's recognition process by highlighting the corresponding steps in the flowchart.

Question

Question: *What are all the scene text in the image?*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

LLaVA-1.5 + Original Image:

*The scene text in the image reads "**S33020C**".*

LLaVA-1.5 + API-Generated Image:

*The scene text in the image reads "**S3302CD**".*

Fig. 5: In this example, our method enhances LVLM's OCR capability by masking background areas and highlighting the regions that require OCR.

Question

Question: *What is unusual about this image?*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

Gemini + Original Image:

*A man is ironing **clothes** while riding on the back of a moving vehicle.*

Gemini + API-Generated Image:

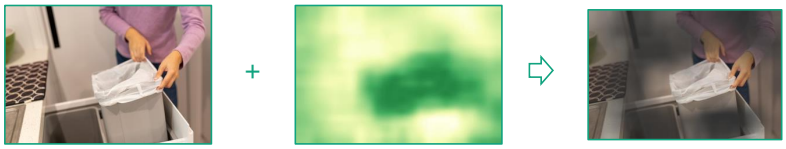
*A man is ironing a **shirt** while riding on the back of a moving vehicle.*

Fig. 6: In this example, our method highlights related regions and enables the LVLM to generate more detailed and accurate response.

Question

Question: *Is the trash empty or full?*

Input Images



Original Image | Heatmap | API-Generated Image

Answers from LVLM

CogVLM + Original Image:

*The trash appears to be **full**, as evidenced by the overflowing white trash bag.*

CogVLM + API-Generated Image:

*The trash appears to be **empty** as there are no visible items inside.*

Fig. 7: In this example, where the question asks to determine whether the trash can is full, our method accurately highlights the area around the trash can's opening, thereby guiding the LVLM to make a correct judgment.

Question and Ground Truth Answer

Question: *What country is one book about? Answer the question using a single word or phrase.*

Ground Truth Answer: *China*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

LLaVA-1.5 + Original Image: ***France***

LLaVA-1.5 + API-Generated Image: ***China***

Fig. 8: In this example, where the question is related to books, our method accurately highlights the area where the books are located in the image.

Question and Ground Truth Answer

Question: *What is the largest measurement we can see on this ruler? Answer the question using a single word or phrase.*

Ground Truth Answer: *50*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

LLaVA-1.5 + Original Image: ***40***

LLaVA-1.5 + API-Generated Image: ***50***

Fig. 9: In this example, the largest measurement number 50 on the ruler is not fully displayed, leading to error in the baseline method. In contrast, as seen through the heatmap, our method emphasizes the bottom right corner of the image where the end of the ruler is located, thereby guiding the LVLM to provide the correct answer.

Question and Ground Truth Answer

Question: *Is the baby on the left of the dog in the image? Please answer yes or no.*

Ground Truth Answer: *No*

Input Images

Original Image | Heatmap | API-Generated Image

Answers from LVLM

GPT-4V + Original Image: ***Yes***

GPT-4V + API-Generated Image: ***No***

Fig. 10: Our method accurately emphasizes the baby and dog in the image, thereby facilitating the inference of their spatial relationship.

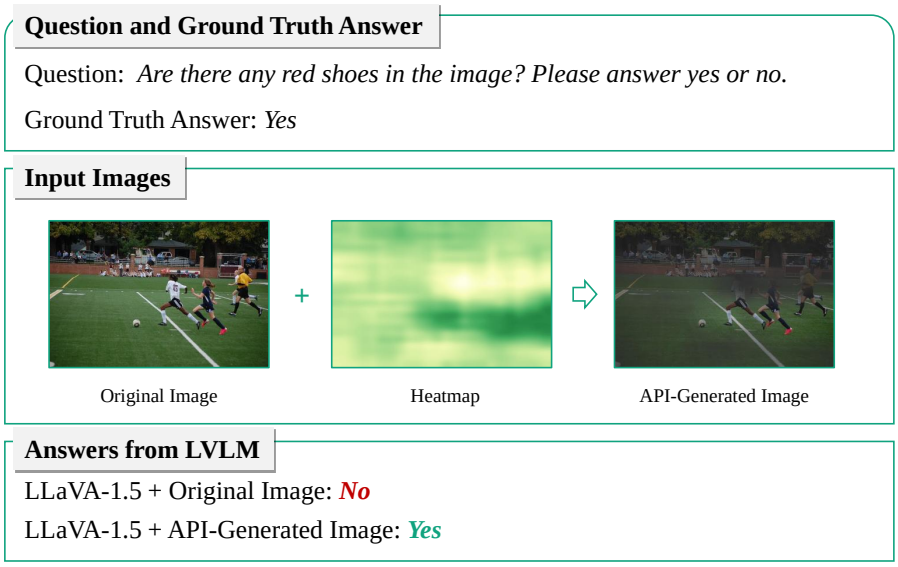


Fig. 11: In this example, the question is related to the shoes, which are small objects and are difficult to recognize for the model. Our method accurately located the shoes in the image, leading the LVLM to the correct answer.

2 Notation Table

Although the definitions of all symbols are included within the main text, we provide a comprehensive notation table in Tabs. 7 and 8 to facilitate easy reference and a macro-level understanding of the concepts involved in each part of the method.

Table 7: The notations used in the manuscript.

Symbol	Definition	Mainly used in
f	LVLM used for inference	Entire Sec. 3
g	Auxiliary LVLM used for attribution map extraction	Entire Sec. 3
$\mathcal{A}$	Annotation function, which is the proposed method	Entire Sec. 3
I	Original image	Entire Sec. 3
I^a	Image with annotations, which is obained by visual prompting method	Entire Sec. 3
Ψ	Attribution map in the token space, which is extracted from the auxiliary LVLM and is used to generate the heatmap	Entire Sec. 3
Φ	Heatmap in the pixel space, which will be overlied on the original image	Entire Sec. 3
T^i	Input text query	Entire Sec. 3
T^o	Output text response	Entire Sec. 3
$A^{(l,h)}$	Attention map in the l-th transformer layer corresponding to the h-th head	Entire Sec. 3
g_{clip}	CLIP model	Sec. 3.1
$\hat{I}$	Image feature generated by CLIP, which is able to calculate the similarity	Sec. 3.1
$\hat{T}$	Text feature generated by CLIP, which is able to calculate the similarity	Sec. 3.1
L	Number of transformer layers within the CLIP vision encoder	Sec. 3.1
MSA	Multihead Self-Attention structure	Sec. 3.1
MLP	Multi-Layer Perceptron structure	Sec. 3.1
Z^l	Input token sequence for the l-th transformer layer	Sec. 3.1
$[Z]_{\text{cls}}$	Value of the cls token within the token sequence Z.	Sec. 3.1

Table 8: The notations used in the manuscript.

Symbol	Definition	Mainly used in
$\mathcal{L}$	Linear transformation in the CLIP model, which is performed after the transformer structure, before calculating the similarity score	Sec. 3.1
L'	In the similarity decomposition of the CLIP model, only the MSA output of last $L - L'$ layers are considered. L' is the starting layer index.	Sec. 3.1
$V^{(l,h)}$	Value matrix in the l-th layer corresponding to the h-th head	Sec. 3.1
$W^{(l,h)}$	Weight matrix in the l-th layer used to merge the multiple attention heads and corresponds to the h-th head. For each head, after the the multiplication between the attention map and the value matrix, we have a matrix with the size of $T \times D'$. To aggregate the matrices from all heads, a weight matrix with the size of $(H \times D') \times D$ is used. $W^{(l,h)}$ is obtained from splitting this large weight matrix.	Sec. 3.1
$B^{(l)}$	Bias matrix in the l-th layer used to merge the multiple attention heads	Sec. 3.1
$A^{(l,h)}_{\text{cls},t}$	Attention value of the class token towards the t-th token in $A^{(l,h)}$	Sec. 3.1
$V^{(l,h)}_{t,:}$	t-th row of $V^{(l,h)}$	Sec. 3.1
H	Number of attention heads	Sec. 3.1
T	Number of tokens	Sec. 3.1
η^l_t	MSA output of the l-th layer corresponding to the t-th patch(token)}	Sec. 3.1
ψ_t	η^l_t summing over the layer index	Sec. 3.1
Ψ^{cls}	Attribution map generated from the CLS token	Sec. 3.1
Ψ^{comp}	Complementary attribution map generated using the non-CLS token	Sec. 3.1
Z^{text}	N tokens corresponding to the text query	Sec. 3.2
Z^{img}	$P \times P$ tokens corresponding to the image patches	Sec. 3.2
Z^{out}	M tokens generated by the LLaVA model	Sec. 3.2
$A^{(\bar{L},h)}_{m,t}$	Attention value in $A^{(\bar{L},h)}$ from the m-th token to the t-th token	Sec. 3.2
$\hat{\Phi}$	Raw heatmap, which is generated by resizing the attribution map	Sec. 3.3

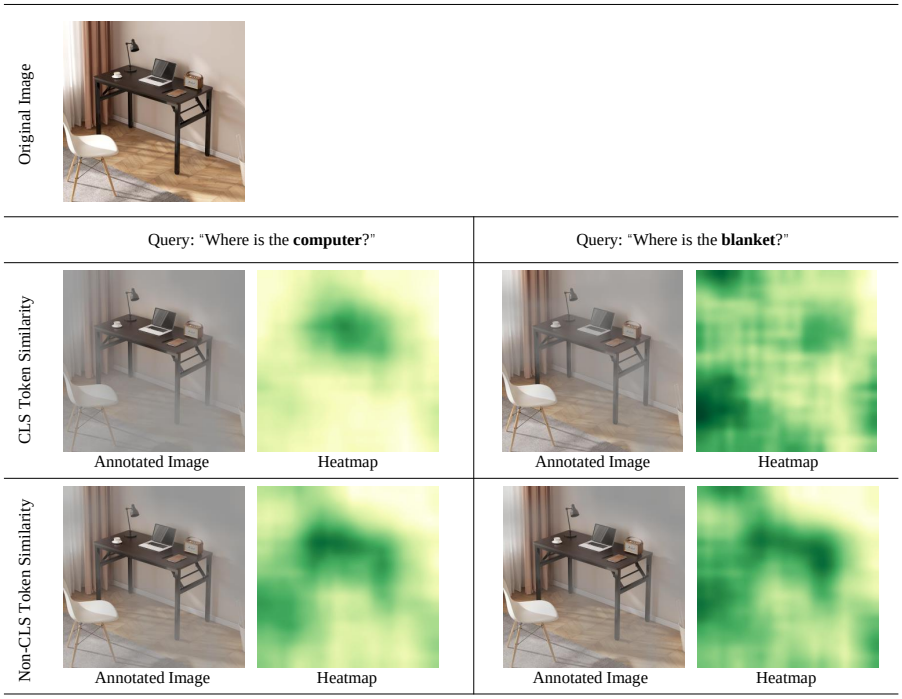


Fig. 12: Comparison between the functionality of CLS token similarity and the Non-CLS token similarity.

3 Observation and Discussion of API Method

3.1 CLS Token Similarity and Non-CLS Token Similarity

To extract heatmaps from the CLIP model, we designed two complementary types of attribution maps: one based on the decomposition of similarity between the feature of the CLS token and text feature, and the other measuring the similarity between the feature of the Non-CLS tokens and text feature. Fig. 12 compares the differences in functionality between these two types of attribution maps. The third row in the image shows the heatmap generated solely based on Ψ^{cls} and its resulting annotated image. The fourth row shows the heatmap obtained solely from Ψ^{comp}. Firstly, we can observe that when the query changes, Ψ^{cls} can highlight different parts of the image corresponding to different queries. It selects the areas where the blanket and computer are located based on the query. However, Ψ^{comp} does not show significant differences in response patterns to different queries. On the other hand, Ψ^{comp} can filter out the background of the image, leaving the objects, which potentially can be used in the process of VQA. For instance, when the query explicitly mentions "computer", Ψ^{cls} completely ignores the chair and blanket in the lower left corner, but Ψ^{comp} still assigns

high values to these areas. Therefore, we combine Ψ^{cls} and Ψ^{comp} to form a complete attribution map.

3.2 Attribution Map Aggregation for CLIP Model

First, Eq. (7) in the maintext can be rewritten as $1-(1-\Psi^{cls})(1-\Psi^{comp})$, where since Ψ^{cls} and Ψ^{comp} are cosine similarities, both $(1-\Psi^{cls})$ and $(1-\Psi^{comp})$ range between 0 and 1. Thus, the final mask is related to the product of the two parts, $(1-\Psi^{cls})$ and $(1-\Psi^{comp})$. If Ψ^{cls} and Ψ^{comp} are considered binary, then $(1-\Psi^{cls})(1-\Psi^{comp})$ can be approximated as an OR operation between $(1-\Psi^{cls})$ and $(1-\Psi^{comp})$. That is, when either $(1-\Psi^{cls})$ or $(1-\Psi^{comp})$ is 0, the equation will be 1, and only when both are 1, the equation will be 0. This means that for patch i, as long as either attribution map Ψ^{cls} or Ψ^{comp} highlights this patch, the final attribution map Ψ will also highlight this patch. Only when both Ψ^{cls} and Ψ^{comp} consider patch i unimportant, the final attribution map will ignore this patch.

Experimental findings, as shown in Fig. 12, indicate that, on one hand, Ψ^{comp} can indiscriminately choose all entities, whereas Ψ^{cls} selects entities explicitly mentioned in the query. The highlighted area in Ψ^{cls} can be understood as a subset of the highlighted area in Ψ^{comp}. On the other hand, both Ψ^{cls} and Ψ^{comp} will ignore non-informative parts of the image. Therefore, in actual non-binary cases, the computation of Eq. (7) can be described as an algorithm: first, apply a mask to non-informative areas (*i.e.*, instruct the LVLM to ignore these patches) because these patches will not be selected by either Ψ^{cls} or Ψ^{comp}. For the remaining areas, which are patches with objects directly mentioned in the query or other entities potentially related to the query, a multiplication of Ψ^{cls} and Ψ^{comp} further highlights the patches with objects appearing in the query because they have greater weight in Ψ^{cls}.

4 More Experimental Results and Implementation Details

4.1 Ensemble

Table 9: Ensemble of visual prompts generated from different LVLM.

	LLaVA-Bench
w/o prompt	102.00
Ours (CLIP)	103.30 (+1.30)
Ours (LLaVA)	103.60 (+1.60)
Ours (CLIP+LLaVA)	104.80 (+2.80)

When the auxiliary LVLM and the LVLM used for inference are different, our approach can be seen as ensembling the knowledge of the auxiliary LVLM into the LVLM used for inference through visual prompts. Under this definition, baseline methods like FGVP and SoM can also be considered a form of ensemble, not between LVLMs but between a vision model (segmentation model) and an LVLM. From the experimental results, our method is the first effective ensemble method that is based on visual prompting in a VQA context.

In traditional ensemble methods that are based on output aggregation, the number of models to be ensembled can be more than 2. However, in our method, we ensemble only two models, namely, an auxiliary LVLM and an LVLM for inference. To achieve an ensemble of more than two models, we conduct the following experiment. We use GPT-4V as the inference model and experiment on the LLaVA-Bench (in-the-wild) dataset, Instead of using a single annotated image. We input the annotated images generated by both $\mathcal{API}$ +CLIP and CLIP+LLaVA simultaneously into GPT-4V, while keep using the original question without additional prompts as the textual query. The experimental results in Tab. 9, show that the ensemble of $\mathcal{API}$ +CLIP and CLIP+LLaVA can further improve performance.

4.2 Influence on Different VQA Abilities

To thoroughly understand the impact of our method on various capabilities of LVLMs, we report the performance changes across different specific abilities on the MM-Vet dataset using the CogVLM model as the inference model and CLIP as the mask model. The results are shown in Tab. 10. It is observed that our method enhances all categories of capabilities in the MM-Vet dataset. Notably, our method is particularly beneficial for OCR and Math abilities. The significant improvement in OCR capability is attributed to our method's highlighting of relevant areas, allowing the model to focus only on regions related to answering the question. This narrows down the scope of the OCR task, thereby enhancing OCR performance. Consequently, the improvement in mathematical ability is closely linked to the enhancement in OCR capability. Since addressing math-related questions in images first requires performing OCR tasks, the improvement in OCR also contributes to the enhancement of mathematical abilities.

Table 10: The influence of our method on various categories of LVLM capabilities.

	Capability					
	Recognition	OCR	Knowledge	Generation	Spatial Relationship	Math
w/o prompt	54.9	42	43.9	42.6	50.1	3.5
Ours	55.3	48.3	45.6	46	51.2	14.6

4.3 Implementation Details

Pre-trained weight and $\mathcal{API}$. During the mask generation phase, we used the CLIP-ViT-L-336 model [8] released by OpenAI and the LLaVA-1.5-13B model [6]. In the inference process, we utilized the released weight of LLaVA-1.5-13B model [6] and cogvlm-chat-v1.1 model [11]. We use the "gpt-4-1106-vision-preview" and "gemini-pro-vision" models for GPT-4V [14] and Gemini [10] $\mathcal{API}$, respectively. All local experiments were deployed on a single A100 GPU.

Query GPT-4V and Gemini. For GPT-4V and Gemini, we used python APIs for batch querying. When encountering errors due to server or network issues, we paused for a while and retried the query once. If the error persisted, we recorded the response as an empty string. If a query was detected against security policy, such as person identification, we did not retry and directly recorded the responses from GPT-4V and Gemini as empty strings.

Baselines. The "w/o prompt" baseline is implemented by directly querying the LVLM with the question together with the original image. Following [5], the "Step-by-Step" baseline is implemented by inputting the original image and query in the format of

> [Question] Let's think step by step.

For the experiments with FGVP [13] and SoM [12], we query the LVLM with the corresponding annotated image and the original question, which is also the same when we implement our method. The only difference among the experiments with FGVP, SoM and our method is the annotated image. For the FGVP method, the annotation process is aligned with the default of the released code. For the SoM method, we choose SAM [4] as the segmentation model and keep all other parameters aligned with the default setting in the released code.

Implementation on each dataset. Our implementation on various datasets adopts the approach from LLaVA [7]. The evaluation process of each dataset adheres to its official usage protocols or its official template, when it is accessible. (1) **LLaVA-Bench (in-the-Wild)** [7] is a dataset comprising real-world scenes, drawings, memes, and other types of images, along with open-ended questions. It focuses on testing LVLMs' capabilities in QA, detailed description, and complex reasoning. In our implementation, the textual prompt is directly the question from the dataset. We record the LVLM's complete answer and use the GPT-based evaluation tool officially released by LLaVA-Bench (in-the-Wild) to score the answers. (2) **MM-Vet** [15] is a comprehensive dataset containing various types of images, including real-world scenes, artworks, statistical graphs, memes, etc., along with open-ended questions. Each question involves multiple aspects of visual and language abilities, such as recognition + spatial awareness or OCR + Math. In our implementation, the textual prompt is directly the question from the dataset. We record the LVLM's complete answer and use the GPT-based evaluation tool officially released by MM-VET to score the answers. (3) **MME** [3] is a dataset that includes images of real-world scenes, artworks, logos, etc., along with True-False questions. This dataset involves abilities

in commonsense reasoning, numerical calculation, and text translation, among others. Given its binary response format (yes or no), we add "Please answer yes or no" as an additional textual prompt to the original question. We evaluate the performance by the matching accuracy between LVLM's answers and the ground truth. (4) The **MMMU** [16] dataset encompasses multi-discipline questions requiring college-level expertise for responses. The questions are either multiple-choice or can be answered with simple data or phrases. For multiple-choice questions, we guide the LVLM to directly answer the corresponding option by adding "Answer with the option's letter from the given choices directly" after the original question and options. For other questions, we add "Answer the question using a single word or phrase." to the original question. Our experiment is conducted using the validation set of MMMU. Evaluation is based on the matching accuracy between LVLM's answers and the ground truth. (5) The **TextVQA** [9] dataset contains real-world images with text, where the questions can be answered with simple words or phrases, mainly testing the LVLM's OCR and reasoning abilities. We add "Answer the question using a single word or phrase" after the original question to guide the LVLM to directly respond to the query without providing additional explanations. Our experiment is conducted using the validation set of TextVQA. The evaluation score is the matching accuracy between LVLM's answers and the ground truth. (6) The **VisWiz** [1] dataset is collected from questions about real-world images asked by blind people and manually annotated answers. The questions can be answered with simple words or phrases. However, since the questions are from blind individuals, some questions are unanswerable based on the image alone and thus are marked as unanswerable. To address this, we concatenate the following prompt after the original question: "When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase" Our experiment is conducted using the validation set of VisWiz. Evaluation is based on the matching accuracy between LVLM's answers and the ground truth.

Prompts used in the Self-Reflection experiment. For the textual self-reflection experiment, we use a two-round chat. In the first round, we directly ask the LVLM to answer the query and record the answer. In the second round, we use a prompt in the format of

> For the Question "[Question]", Your previous answer is "[Answer in the Round 1]". Evaluate the quality of the answer and provide a new answer.

We record the response of the second round and extract the answer by manually delete the sentences related to the quality evaluation of previous answer. The extracted answer is stored as the final answer. For the "$\mathcal{API}$ + reflection via re-emphasize" setup, we input the annotated image together with the prompt in the format of

> [Question] (Hint: The answer is related to the unmasked visible regions).

For the "$\mathcal{API}$ + reflection via evaluation" setup, we input the annotated image together with the prompt in the format of

> For this image, the question is "[Question]". Evaluate whether the unmasked visible regions of the image alone can provide an answer to the question. If they suffice to answer the question, respond with letter "T". If they do not support an answer to the question, reply with the letter "F".

If the LVLM responses with "F", we query it again using the original image and the question, and then use the response as the final answer. If the LVLM responses with "T", we query it again using the annotated image and the question, and then use the response as the final answer.

5 Limitation, Future Direction, and Potential Impact

Limitation and future direction. An essential component of this work is the extraction of attribution maps based on an auxiliary LVLM. The introduction of an auxiliary LVLM enhances the performance of visual prompting methods but also introduces some limitations and new research opportunities. First, generating visual prompts based on an LVLM incurs additional computational costs, either from an extra execution of the same LVLM or a forward pass through another LVLM. Note that this is a limitation, exploring ways to reduce this additional overhead, such as using lightweight LVLMs to generate visual prompts to achieve a weak-to-strong effect [2, 17], is a worthwhile research direction. Secondly, our current selection of auxiliary LVLMs is not adaptive; we cannot automatically choose a more suitable auxiliary LVLM for different image-query pairs. This is another limitation of our method and a potential research direction with promise.

Potential impact. The potential social impacts of this work mainly include two aspects. The first aspect is the potential accumulation of bias and unfairness due to the introduction of an extra LVLM. The bias and unfairness of the auxiliary LVLM may accumulate through our visual prompts into the final inference process. The other aspect is the creation of a new possibility for attacks, namely, by attacking the auxiliary LVLM to generate harmful visual prompts, thereby attacking the LVLM. Because the attack is based on the visual prompts in the pixel space, such attacks might be more covert and difficult to detect.

References

1. Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., Yeh, T.: Vizwiz: nearly real-time answers to visual questions. In: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology. p. 333–342 (2010)
2. Burns, C., Izmailov, P., Kirchner, J.H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., Wu, J.: Weak-to-strong generalization: Eliciting strong capabilities with weak supervision (2023)
3. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: MME: A comprehensive evaluation benchmark for multimodal large language models. CoRR **abs/2306.13394** (2023)
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: arXiv (2023)
5. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
6. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
7. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
9. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR). pp. 8317–8326 (2019)
10. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
11. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., Tang, J.: Cogvlm: Visual expert for pretrained language models (2023)
12. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. CoRR (2023)
13. Yang, L., Wang, Y., Li, X., Wang, X., Yang, J.: Fine-grained visual prompting. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)
14. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision) (2023)
15. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2023)
16. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502 (2023)
17. Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y.X., Wang, W.Y.: Weak-to-strong jailbreaking on large language models (2024)