

---

## A APPENDIX

In the supplementary materials, we provide the following sections:

- (a) More implementation details in Section B.
- (b) Ablation study experiments in Section C.
- (c) Visualization result analysis in Section D.

## B IMPLEMENTATION DETAILS

For the initialization of the proposed prompt embedding network (PEN), we use Kaiming initialization technology [1]. The UAE-Large-V1<sup>1</sup> model is adopted as the pre-trained textual encoder to extract textual embeddings for the visual prompt.

## C ABLATION STUDY

Next, we conduct more ablation study experiments to provide deeper insight into the components of our proposed approach.

**Object Detector v.s. Segmentation Model.** To determine the effect of using an object detector or segmentation model to incorporate pixel-level semantics into the proposed visual prompt, we conduct an ablation study with the popular object detector GroundingDINO [2] and the segmentation model OpenSeed [3]. The results are shown in Table 1. We observe that both GroundingDINO and OpenSeed significantly boost performance across all benchmarks. However, utilizing OpenSeed achieves better performance gains due to its fine-grained mask regions. Thus, we adopt OpenSeed by default to generate object regions.

**The Effect of Fine-Tuning with the Visual Prompt.** As displayed in Table 2, the model fine-tuned with the proposed visual prompt (*i.e.*, the third row) achieves remarkably better performance than the one fine-tuned without our visual prompt (*i.e.*, the second row) across all benchmarks. Specifically, without using our visual prompt for fine-tuning, the model even shows performance degradation on Text-VQA benchmark [4] and has negligible gains on Science-QA [5], VQAv2 [6], MME-P [7], and MME-C [7] benchmarks. All these results demonstrate the superiority of the proposed method.

**Discussion.** We also compare our presented REVIP method with V\* [8], which employs an LLM-guided visual search mechanism to enhance MLLM’s contextual understating capacities. In Table 3, to ensure a fair comparison with V\* [8], we present our method’s results using the experimental settings from [8]. We also report the accuracy metrics (1st and 2nd rows) for the multimodal benchmarks as adopted by LLaVA [9] in its paper. As stated in V\* [8] and evident from the Table 3, there’s significant degraded performance on MM-Vet [10], LLaVA-Bench<sup>W</sup> [11] and MMBench [12]. In contrast, our method demonstrates consistent improvements across all these multimodal benchmarks. ViperGPT [13] also utilizes LLM to solve visual tasks, however, it focuses on code generation to solve complex tasks based on Codex, which isn’t related with our method.

## D VISUALIZATION RESULT ANALYSIS

We’ve provided more visualization results in Table 4, 5, 6, and 7. Compared to LLaVA-1.5 7B [9], our method generates more reasonable and accurate responses to the questions.

As shown in Table 4, our approach can generate accurate movie titles, such as “The Godfather”, and the two actors’ names, such as “Al Pacino” and “Robert De Niro”. Additionally, it provides a corresponding introduction, such as “The movie is a classic crime drama film directed by Francis Ford Coppola, known for its iconic characters, storytelling, and memorable scenes” for the left example. In the right example, our method generates the precise title “The Lord of the Rings: The Fellowship of

---

<sup>1</sup><https://huggingface.co/WhereIsAI/UAE-Large-V1>

Table 1: The ablation study of using an object detector or a panoptic segmentation model to extract object regions for pixel-level textual embeddings.

Method	Region Generator	VQAv2	GQA	VisWiz	SQA <sup>1</sup>	VQA <sup>T</sup>	MME-P	MME-C	MMB	MM-Vet	POPE	MMMU
Mipha-3B	-	81.3	63.9	45.7	70.9	56.6	1488.9	295.0	69.7	32.1	86.7	32.5
Mipha-3B <sup>+</sup>	GroundingDINO	82.0 <sup>↑</sup>	64.9 <sup>↑</sup>	46.4 <sup>↑</sup>	71.4 <sup>↑</sup>	57.2 <sup>↑</sup>	1491.7 <sup>↑</sup>	350.2 <sup>↑</sup>	71.0 <sup>↑</sup>	34.5 <sup>↑</sup>	88.4 <sup>↑</sup>	32.9 <sup>↑</sup>
<b>Mipha-3B<sup>+</sup> (Ours)</b>	<b>OpenSeed</b>	<b>82.4<sup>↑</sup></b>	<b>65.3<sup>↑</sup></b>	<b>47.0<sup>↑</sup></b>	<b>71.8<sup>↑</sup></b>	<b>57.8<sup>↑</sup></b>	<b>1501.2<sup>↑</sup></b>	<b>369.1<sup>↑</sup></b>	<b>71.5<sup>↑</sup></b>	<b>35.1<sup>↑</sup></b>	<b>88.7<sup>↑</sup></b>	<b>33.5<sup>↑</sup></b>

Table 2: The ablation study of fine-tuning with and without the proposed visual prompt. The first (Mipha-3B), second (Mipha-3B<sup>\*</sup>) and third (Mipha-3B<sup>+</sup>) rows mean Mipha-3B baseline, fine-tuning on Mipha-3B without and with the proposed visual prompt using LoRA [14].

Method	Visual Prompt	VQAv2	GQA	VisWiz	SQA <sup>1</sup>	VQA <sup>T</sup>	MME-P	MME-C	MMB	MM-Vet	POPE	MMMU
Mipha-3B	-	81.3	63.9	45.7	70.9	56.6	1488.9	295.0	69.7	32.1	86.7	32.5
Mipha-3B <sup>*</sup>	$\times$	81.4 <sup>↑</sup>	64.3 <sup>↑</sup>	45.9 <sup>↑</sup>	71.0 <sup>↑</sup>	56.5 <sup>↓</sup>	1489.2 <sup>↑</sup>	303.2 <sup>↑</sup>	70.4 <sup>↑</sup>	33.5 <sup>↑</sup>	87.4 <sup>↑</sup>	32.6
<b>Mipha-3B<sup>+</sup> (Ours)</b>	<b>✓</b>	<b>82.4<sup>↑</sup></b>	<b>65.3<sup>↑</sup></b>	<b>47.0<sup>↑</sup></b>	<b>71.8<sup>↑</sup></b>	<b>57.8<sup>↑</sup></b>	<b>1501.2<sup>↑</sup></b>	<b>369.1<sup>↑</sup></b>	<b>71.5<sup>↑</sup></b>	<b>35.1<sup>↑</sup></b>	<b>88.7<sup>↑</sup></b>	<b>33.5<sup>↑</sup></b>

Table 3: The results compared with V<sup>\*</sup> [8] has been reported, we adopt the setting and benchmarks from V<sup>\*</sup> [8], the compared benchmarks including MME [7], MMBench [12], SEED-Bench(Img) [15], MM-Vet [10], LLaVA<sup>W</sup>: LLaVA-Bench<sup>W</sup> [16] and POPE [17]. LLaVA-7B<sup>+</sup> denotes LLaVA (7B)<sup>+</sup> using presented REVIP method.

Method	MME	MMBench	SEED-Bench(Img)	MM-Vet	LLaVA <sup>W</sup>	POPE
LLaVA (7B) [8]	1051.2	34.4	41.8	30.4	62.6	76.5
V <sup>*</sup> (7B) [8]	1128.9 <sup>↑</sup>	33.1 <sup>↓</sup>	41.7 <sup>↓</sup>	27.7 <sup>↓</sup>	59.1 <sup>↓</sup>	82.4 <sup>↑</sup>
<b>LLaVA (7B)<sup>+</sup> (Ours)</b>	<b>1156.4<sup>↑</sup></b>	<b>37.5<sup>↑</sup></b>	<b>43.2<sup>↑</sup></b>	<b>34.5<sup>↑</sup></b>	<b>64.5<sup>↑</sup></b>	<b>82.8<sup>↑</sup></b>

the Ring” and provides an accurate introduction, such as “The movie is an epic fantasy adventure that follows the journey of a fellowship of characters”. In contrast, LLaVA’s responses are very general without fully understanding the global contexts within the images.

In Table 5, our method not only recognizes the text “Me: I’ll do it at 8. Time: 8.05. Looks like I gotta wait till 9 now”, but also understands its implication, such as “the puppet is aware of the time and is intentionally delaying any work or task it might be assigned”, in the left image. For the right image, our method generates more potential options, such as a “private school”, “a religious school”, etc. In contrast, LLaVA’s answers are either inappropriate or too limited.

Finally, as shown in Table 6 and 7, our method understands spatial relationships and accurately recognizes the text within the images. For example, in the left example of Table 6, our approach accurately names the person on the left as “Keira Knightley”. It also recognizes the plate number “S3302CD” in the left example of Table 7, while LLaVA’s answers are all incorrect. This indicates the strong fine-grained multimodal understanding capacity of our proposed method.

108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161

---

**Visual input example, context-related questions:**

---



Can you give a short introduction to this movie?



Can you give a short introduction to this movie?

LLaVA-7B

The image features a scene from a movie where two men are standing close to each other, with one man placing his hand on the other's shoulder. They are both wearing tuxedos, which suggests that they are attending a formal event or a special occasion. The men appear to be engaged in a conversation or interacting with each other in some way, possibly discussing the event or sharing a moment of camaraderie.

The image features a young man wearing a brown shirt and holding a ring in his hand. He appears to be in a dark setting, possibly a cave or a dungeon. The man is looking at the camera, possibly engaging in a conversation or sharing a story. The ring he is holding suggests that the scene might be related to a fantasy or adventure movie, where the ring could represent a magical or powerful object.

**Ours**

The image features a scene from the movie "The Godfather", where two men played by actors Al Pacino and Robert De Niro, are standing close to each other. One man has his hand on the other's shoulder, and they appear to be engaged in a conversation or a confrontation. The movie is a classic crime drama film directed by Francis Ford Coppola, known for its iconic characters, storytelling, and memorable scenes.

The image features a scene from the movie "The Lord of the Rings: The Fellowship of the Ring", where the character Gandalf is holding out his hand to the hobbit Frodo, offering him the One Ring. This iconic moment captures the essence of the story, where Frodo is entrusted with the task of destroying the ring to prevent its power from falling into the wrong hands. The movie is an epic fantasy adventure that follows the journey of a fellowship of characters as they embark on a perilous quest to rid Middle-earth of the dark lord Sauron and his army of orcs.

---

Table 4: The challenging examples on LLaVA-1.5, our approach can provide reasonable reponses for the tricky questions which need to fully understand the contexts of the images.

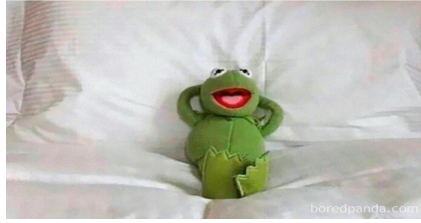
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215

---

**Visual input example, context-related questions:**

---

**Me:** I'll do it at 8  
**Time:** 8.05  
**Me:** looks like I gotta wait till 9 now



Can you explain this meme?



What kind of school does the boy go to? Answer it and give the rationale.

**LLaVA-7B** The meme features a green Kermit the Frog puppet lying on a bed, with a caption that reads, "Me looks like I gotta wait till 9" The puppet appears to be relaxing on the bed, possibly waiting for someone or something. The image is a playful and lighthearted representation of a common situation where people might have to wait for a certain time before they can proceed with their plans.

The boy likely goes to a private school, as he is wearing a suit and tie, which is a common attire for students in private schools. The presence of a tie suggests that the school has a formal dress code, and the boy is dressed appropriately for the occasion.

**Ours** The image features a green frog, which is a popular character from the television show "The Muppet Show", the frog is sitting on a bed, appearing to be in a relaxed and comfortable position. The puppet is holding its belly and seems to be enjoying its time. The bed has a white sheet, and the scene appears to be set in a cozy bedroom. The meme is captioned with a humorous message which reads "Me: I'll do it at 8. Time: 8.05. Looks like I gotta wait till 9 now", this implies that the puppet is aware of the time and is intentionally delaying any work or task it might be assigned. The meme is likely meant to be amusing and entertaining, as it combines the idea of a puppet with a human-like sense of humor and awareness of time.

The boy in the image is wearing a tie, which suggests that he might be attending a school with a formal dress code or a school that requires students to wear ties as part of their uniform. This could be a private school, a religious school, or a school with a strong emphasis on discipline and professionalism. The presence of a woman, who is wearing a dress, in the image, possibly his mother, indicates that the boy might be going to school with her support and guidance.

Table 5: The challenging examples on LLaVA-1.5, our approach can provide reasonable responses for the tricky questions which need to fully understand the contexts of the images.

216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269

Visual input example, relationship-aware questions:		
		
	Who is the person on the left?	Who is the person in the bottom right?
LLaVA-7B	The person on the left is Reese Witherspoon.	The person in the bottom right is a man named Tom Hiddleston.
Ours	The person on the left is Keira Knightley.	The person in the bottom rights is Jamie Dor-nan.


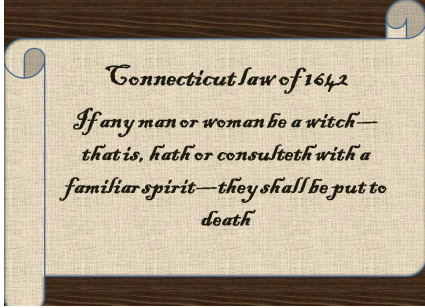
Table 6: The challenging examples on LLaVA-1.5, our approach can provide reasonable reponses for the tricky questions which need to fully understand the relationships among objects within images.



---

**Visual input example, text-related questions:**

---

		
	What are all the scene text in the image?	What are all the scene text in the image?
LLaVA-7B	The scene text in the image is “S3302CD”.	The scene text in the image is “Connecticut Law of 1652”.
<b>Ours</b>	The scene text in the image includes the license plate number “S3302CD” which is written in white on a black background.	The scene text in the image includes “Connecticut Law of 1642”, which states “If any man or woman be a witch - that is hath or consulteth with a familiar spirit - they shall be put to death”.

---

Table 7: The challenging examples on LLaVA-1.5, our approach can provide reasonable responses for the tricky questions which need to accurately recognize the texts within the images.

## REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.
- [2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.
- [3] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023.
- [4] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.
- [5] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.
- [7] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- [8] Penghao Wu and Saining Xie. V\*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.
- [9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

- 
- [10] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.
- [11] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.
- [12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.
- [13] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898, 2023.
- [14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [15] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.