Supplementary Materials

ViP-LLaVA: Making Large Multimodal Models Understand Arbitrary Visual Prompts

Mu Cai[1] Haotian Liu[1] Siva Karthik Mustikovela[2] Gregory P. Meyer[2]
Yuning Chai[2] Dennis Park[2] Yong Jae Lee[1,2]
[1]University of Wisconsin–Madison [2]Cruise LLC
https://vip-llava.github.io

This supplementary document extends our main paper by providing additional results and in-depth analyses that were not included in the main manuscript due to space limitations. In Section A, we offer both qualitative and quantitative analyses on topics such as visual prompt generation, effect of the instruction data, arrow direction understanding, perform under each visual prompt, and impacts on the conventional vision-language model benchmarks, thus providing a comprehensive examination of our research. In Section B, we include more training details. Section C delves into further ablation studies, illuminating the design and efficacy of our approach. Additional experimental results, including a focus on region captioning, are discussed in Section D. The exploration of visual prompt augmentation's potential is presented in Section E. Lastly, Section F comprises detailed statistics and visualizations from the ViP-Bench dataset.

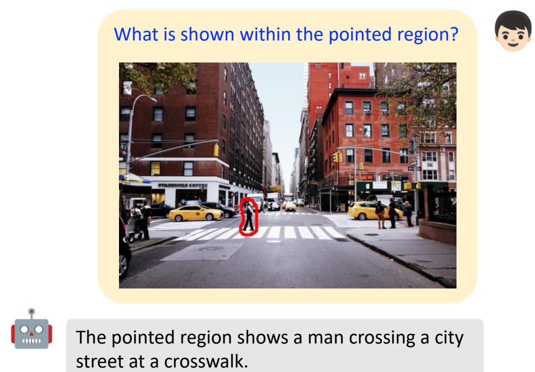


Figure 1. ViP-LLaVA is able to recognize a tiny region specified by the red contour.

A. In-Depth Analysis

A.1. Generalization to New Attributes

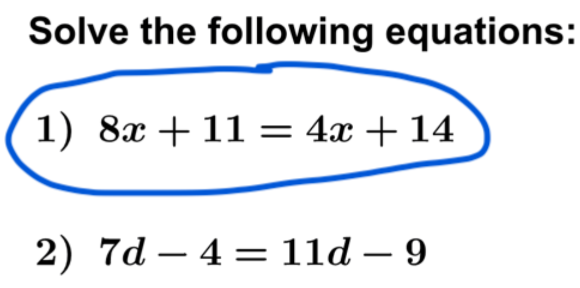
ViP-LLaVA, having been trained on eight types of visual prompts—namely mask contour, ellipse, bounding box, triangle, scribble, point, arrow, and mask—exhibits notable generalization capabilities. In the main paper, we show that ViP-LLaVA can understand human drawn visual prompts, as shown in Figure 1. As demonstrated in Figures 2 and Figure 5 of the main paper, ViP-LLaVA adeptly handles visual prompts with varying thicknesses and diverse markers, even though it was not explicitly trained on such variations. Furthermore, it effectively interprets text markers as visual prompts, a feature inspired by the Set-of-Mark [18].

Figures 3, 4, and 5 present qualitative examples. In Figure 3, ViP-LLaVA accurately localizes objects tagged with the digits "1", "2", and "3", and generates precise descriptions for each. Figure 4 showcases the model's ability to recognize digit markers and describe the color of vehicles accurately, despite the markers displaying counterfactual colors relative to the actual vehicle colors. Figure 5 illustrates the model's competency in localizing a lemon within a scene densely populated with markers.

A.2. Effect of Optional GPT-4V Region-Level Instruction Data

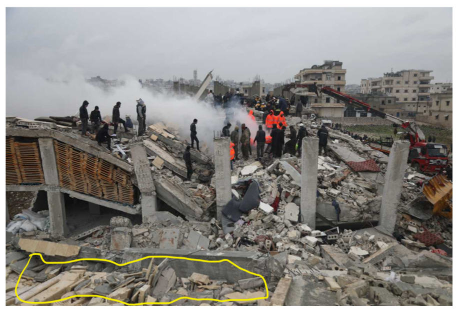
As mentioned in Section 3.3 of the main paper, incorporating GPT-4V as an additional source of instruction data can enhance ViP-LLaVA's performance. An example of the curation process is shown in Figure 6. For this purpose, we combine 13K data entries from the original stage 2 instruction dataset with an equal number of GPT-4V region-level instruction data entries, forming a comprehensive 26K-entry stage 3 fine-tuning dataset. We fine-tune our stage-2 model for one epoch, which requires approximately 0.5 hours for the 7B model and 1 hour for the 13B model on 8

(a)



Q:	What is the value of the variable in the equation within the blue ellipse?
GT:	0.75 OR x=0.75 OR 3/4
Required capabilities:	OCR, MATH

(b)



Q:	Are the debris within the yellow mask contour made of the same material?
GT:	No
Required capabilities:	Recognition, Knowledge

(c)

Q:	Is the person pointed by red arrow going to be mad at the person pointed by blue arrow, for looking at the person pointed by green arrow? Answer it and give the rationale.
GT:	Yes, the woman in red and the man appear to be a couple and the woman in red would not appreciate the man checking out other women.I think so because in a monogamous relationship, partners are expected to be faithful to one another.
Required capabilities:	Recognition, Knowledge, Language Generation, Relationship

Table 1. Three samples requiring different capability integrations.

NVIDIA A100 GPUs. As shown in Table 2, the fine-tuned model, designated as ViP-LLaVA, demonstrates improvements across nearly all datasets for both the 7B and 13B models, underscoring the efficacy of the GPT-4V instruction data curation process. Notably, even without the GPT-4V instruction data, ViP-LLaVA outperforms contemporary methods on benchmarks such as Visual7W, PointQA-LookTwice, and ViP-Bench. The inclusion of GPT-4V instruction data further amplifies this performance advantage.

Model	Use GPT-4V Data?	Visual7W	PointQA-LookTwice	ViP-Bench@Box	ViP-Bench@Human
GPT4ROI-7B [21]	×	81.83†	–	35.1	–
Shikra-7B [1]	×	–	–	33.7	–
ViP-LLaVA-Base-7B	×	86.04	70.86	45.5	46.8
ViP-LLaVA-7B	✓	**86.09**	**71.31**	**48.4**	**48.3**
GPT4ROI-13B [21]	×	84.82†	–	–	–
Shikra-13B [1]	×	85.33	70.30	–	–
ViP-LLaVA-Base-13B	×	87.54	**72.15**	48.2	47.0
ViP-LLaVA-13B	✓	**88.28**	71.77	**48.3**	**48.2**

Table 2. Comparative performance analysis of the use of GPT-4V data in the 7B and 13B models. † Indicates that GPT4ROI specifically trained a specialist model on Visual7W.

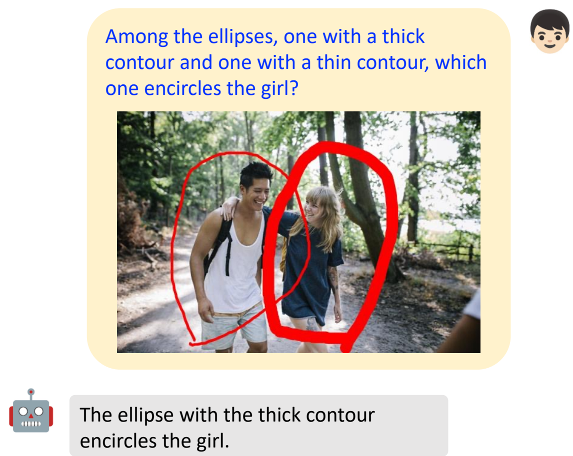


Figure 2. ViP-LLaVA is able to recognize visual prompts with different thickness not seen during training.

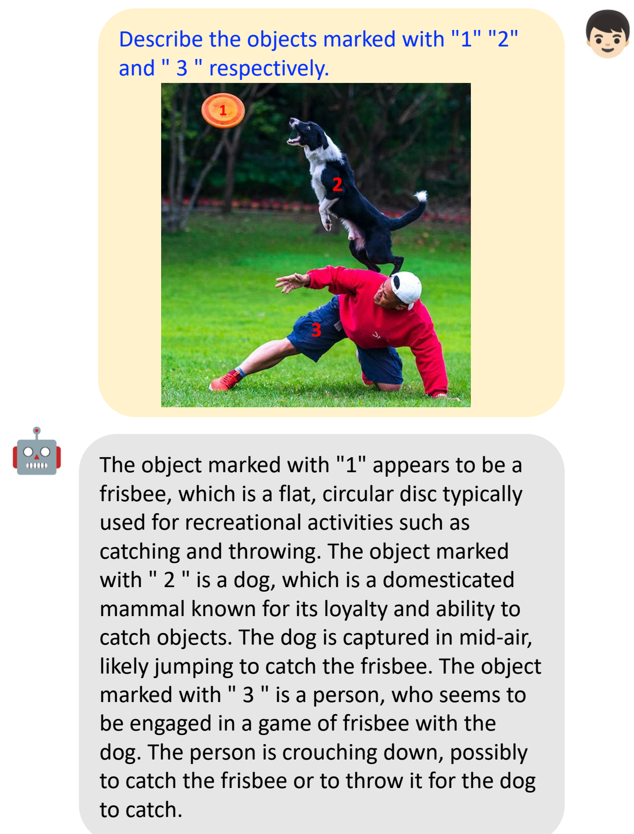


Figure 3. ViP-LLaVA successfully localizes objects tagged with digits and generates accurate object descriptions.

A.3. Understanding Arrow Direction

To rigorously evaluate ViP-LLaVA's capacity for interpreting arrow directions, we next construct a challenging dataset of examples derived from the COCO validation set [7]. Specifically, we generate multiple scenarios with arrows: each arrow originates from the center of one object's bounding box and points towards the center of another, and vice versa. These visualizations are depicted in Figure 7. The typical prompt used is as follows: `Determine whether object A (`**`category1`**`) or object B (`**`category2`**`) is at the head of the arrow, with the other object representing the tail.` It is important to note that we ensure each pair of objects belong to distinct categories. A total of 3520 such paired examples are collected and analyzed. Impressively, ViP-LLaVA-13B achieves an accuracy of 90.28%, demonstrating a robust understanding of arrow directionality and ruling out the possibility of random guessing.

A.4. Performance across Different Visual Prompts

During training, we leverage eight types of visual prompts. Here we study the performance of under each kind of visual prompt on downstream tasks. Note that we duplicate the region-level training data by eight time during instruction fine-tuning.

Results on VCR, Visual7W, PointQA, and ViP-Bench shown in Table 3 show our 7B model's consistent accuracy

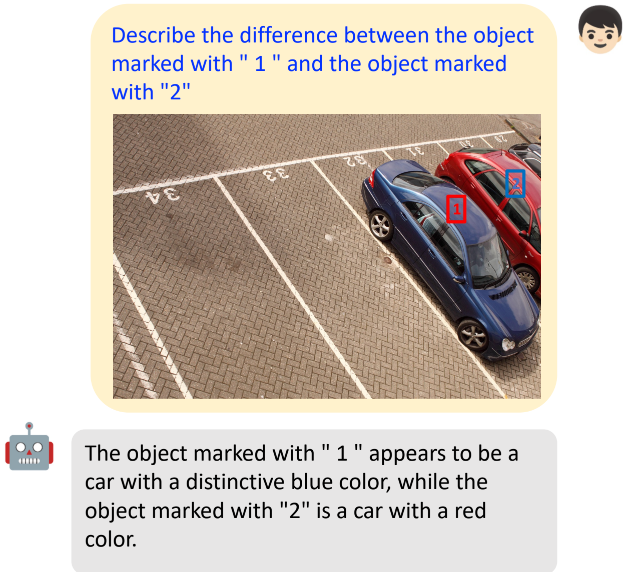


Figure 4. The model effectively localizes digit markers and accurately predicts vehicle colors, uninfluenced by the marker colors.

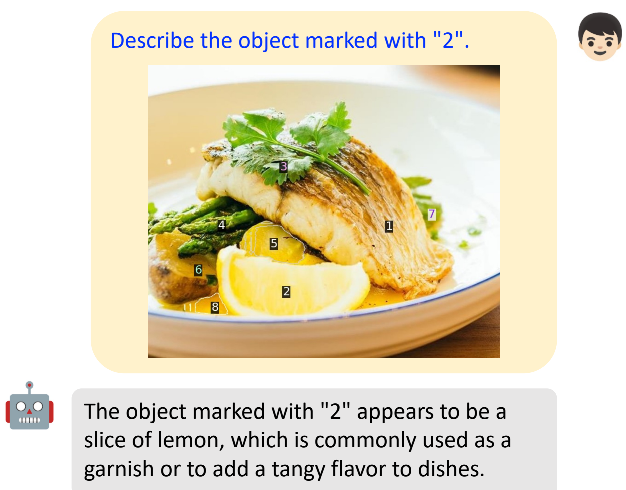


Figure 5. Demonstration of the model's ability to localize an object in a densely marked scene.

with varied visual prompts, with "Point" and "Ellipse" owning marginally better performance than others.

A.5. Impact on Conventional Vision-Language Model Benchmarks

Shown in Table 4, we maintain the performance of LLaVA-1.5 across various benchmarks. The benchmarks include MMBench [9], MMBench_cn [9], LLaVA in the wild [8], POPE hallucination bench [6], ScienceQA [10], MM-Vet [20], MME [3], VizWiz [5], TextVQA [14], and VQAv2 [4]. Specifically, on MME/MMBench/VQA, we show enhanced performance.

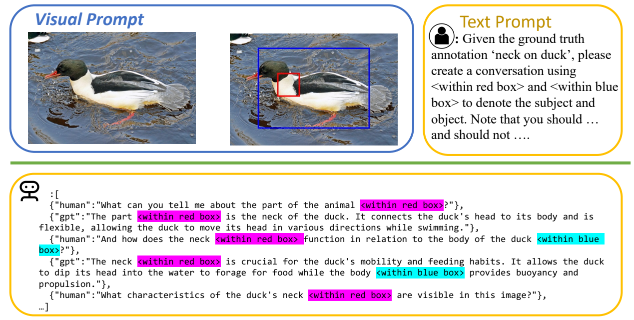


Figure 6. **Curation process of region-level instruction data.** This figure delineates the workflow where both original and annotated images, along with corresponding text prompts, are integrated into the GPT-4V model, facilitating the generation of detailed instruction data for region-specific tasks.

(a) (b)

Figure 7. An illustration from our arrow direction understanding dataset. Panels (a) and (b) display two arrows. These arrows share a similar body but differ in their heads. In this example, the multimodal model is required to discern whether the arrow points to the orange or the bowl in both images.

B. Training Details

Our approach involves three stages:

- Pretraining the MLP projector with frozen language and image encoder;
- Instruction tuning and further fine-tuning with frozen image encoder but trainable language model and MLP connector.

C. Additional Ablation Studies

C.1. Effects of Input Resolution and LLM

To ensure a fair comparison, we conduct ablation studies using the same image encoder (CLIP ViT-L from Radford et al. [12]), input resolution (224 pixels), and language model (Vicuna v1.1 [16]) as employed by GPT4ROI [21]. Table 5 presents the results of this analysis. Despite utilizing the same underlying technologies, ViP-LLaVA consistently outperforms on the ViP-Bench evaluations and achieves comparable results on the Visual7W dataset, notwithstanding the fact that GPT4ROI was specifically fine-tuned for

	Mask Contour	Mask	Ellipse	Rectangle	Triangle	Scribble	Point	Arrow
VCR $Q \rightarrow A$	87.34	86.19	87.38	87.43	87.56	87.54	**87.69**	87.49
VCR $QA \rightarrow R$	89.63	88.80	89.60	89.70	89.57	89.78	**89.81**	89.73
VCR $Q \rightarrow AR$	78.53	76.80	78.52	78.70	78.60	78.76	**78.92**	78.72
Visual7W	–	–	**86.6**	86.04	83.46	83.77	84.88	82.41
PointQA	–	–	**71.3**	70.86	69.72	70.23	69.58	69.07
ViP-Bench	–	–	**45.9**	45.5	43.3	44.9	44.5	44.2

Table 3. Performance under different visual prompts in VCR, Visual7W, PointQA and ViP-Bench on the 7B model of ViP-LLaVA.

Model/Benchmark	MMBench	MMBench_cn	LLaVA_W	POPE	ScienceQA	MMVet	VizWiz	MME	TextVQA	VQAv2
LLaVa-13b	67.7	**63.6**	70.7	**85.9**	**71.6**	**35.4**	53.6	1531	**61.3**	80.00
ViP-LLaVA-13b	**70.3**	60.7	**72.4**	85.7	70.0	34.5	**57.4**	**1564**	59.6	**80.13**

Table 4. Impact on conventional vision-language model benchmarks. ViP-LLaVA model achieve similar performance as LLaVA-1.5.

Visual7W. These results further reinforce the potential of visual prompting as a more effective approach for region-specific referencing compared to embedding coordinates directly into the language model.

C.2. Comparing Visual Prompts with Coordinates

To rigorously evaluate the effectiveness of visual prompts versus coordinate-based region referring formats, we next replace visual prompts with textual coordinates embedded in language descriptions. We train a 7B model using identical data and training schedules. The results, as shown in Table 6, indicate that visual prompts significantly outperform coordinate formats on the PointQA-LookTwice and ViP-Bench@Box datasets. Performance on the Visual7W dataset remains comparable between the two formats. These comparisons highlight the superiority of visual prompts as a more effective format for region-specific referencing in complex visual tasks.

C.3. Effects of Splitting Overlaid Images into Two Separate Image

We conduct rigorous ablation study to split the overlaid image into the source image and the image with overlaid cue, where the number of visual tokens are doubled, as shown in Figure 8. Specifically, we train 7B models under such two settings. Results in Table 7 shows that those two settings perform comparably.

D. Additional Experimental Results

D.1. Region Captioning

Expanding upon the region perception and reasoning tasks discussed in the main paper, we further evaluate ViP-LLaVA's region captioning capabilities on the RefCOCOg dataset [19]. This involves fine-tuning the ViP-LLaVA-Base-7B for one epoch subsequent to stage 2 training. As Table 8 illustrates, ViP-LLaVA-Base-7B demonstrates

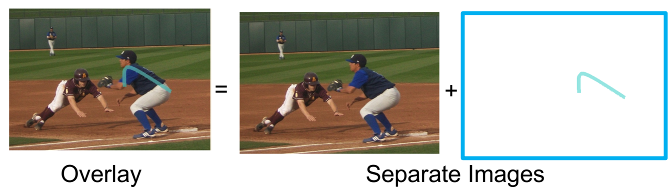


Figure 8. We separate the overlaid images into the original image along with the visual prompts with white background.

strong performance in region captioning, as evidenced by its scores in both CIDEr [15] and METEOR [2] metrics. These results indicate that visual prompting is not only effective for region-specific referencing and reasoning tasks but also shows promising potential in generating precise and contextually relevant captions for specific image regions.

D.2. Assessment of GPT-4 as a Judge

To evaluate the consistency of ViP-LLaVA-Base-7B, we employ the GPT-4 text model as a judge, conducting five separate assessments. The observed variance in the overall score is a minimal 0.1, indicating stable performance by the GPT-4 judge across multiple evaluations.

E. Potential of Visual Prompt Augmentation

A key advantage of ViP-LLaVA approach is the ability to very easily employ prompt augmentation during testing. This entails using various sets of visual prompts and aggregating the predictions for a more accurate final answer. For instance, we can modify the prompt from "the woman within a red rectangle" to "the woman marked with a red scribble", along with corresponding changes in the overlaid image. As shown in Table 9, ViP-LLaVA-Base-7B achieves further improvements through visual prompt augmentation. This process is lossless, unlike textual coordinate representation, where e.g., perturbing coordinates can reduce local-

Model	Input Resolution	LLM	Format	Visual7W	ViP-Bench@Box
ViP-LLaVA-Base-7B	336	Vicuna v1.5	VP	**86.04**	**45.50**
ViP-LLaVA-Base-7B	224	Vicuna v1.1	VP	81.80	42.28
GPT4ROI-7B [21]	224	Vicuna v1.1	ROI	81.83†	35.14

Table 5. Ablation study focusing on the impact of input resolution and language model. All models listed use the Vicuna 7B language model. † Indicates GPT4ROI specifically trained on the Visual7W dataset. VP: visual prompts; ROI: CLIP region of interest (ROI) features and positional embedding.

Format	Visual7W	PointQA-LookTwice	ViP-Bench@Box
VP	86.04	**70.86**	**45.5**
Coor	**86.36**	61.4	42.6

Table 6. Performance comparison between visual prompts and coordinate formats under ViP-LLaVA-Base-7B. VP: visual prompts; Coor: coordinates as visual prompts.

ization accuracy.

F. Further Insights into ViP-Bench

F.1. Statistics of ViP-Bench

Table 10 presents the statistical breakdown of ViP-Bench. The majority of examples focus on recognition capabilities, with a notable proportion (89 examples) requiring Optical Character Recognition (OCR). The proportion of each capability and the combined capabilities are shown in Figure 9 and Figure 10 respectively.

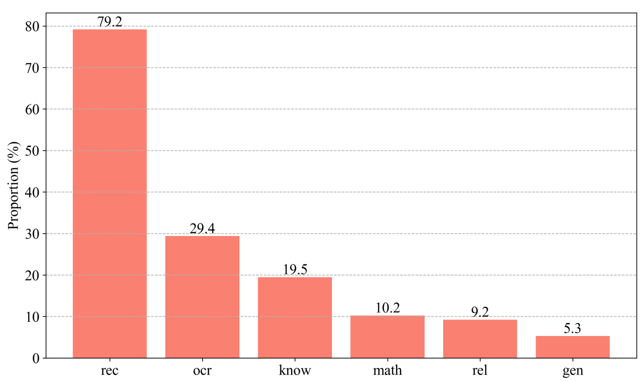


Figure 9. ViP-Bench proportion of capabilities. The proportion of each capability. The sum of the proportion is larger than 100% because some samples have more than one capability.

F.2. Visualizations of ViP-Bench

Figure 11 showcases examples from ViP-Bench, comparing synthesized and human-annotated visual prompts. Panel (a) illustrates tight bounding boxes as synthesized prompts, while panel (b) features human-annotated bounding boxes,

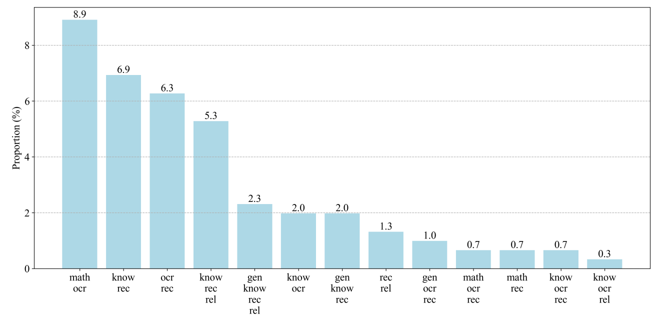
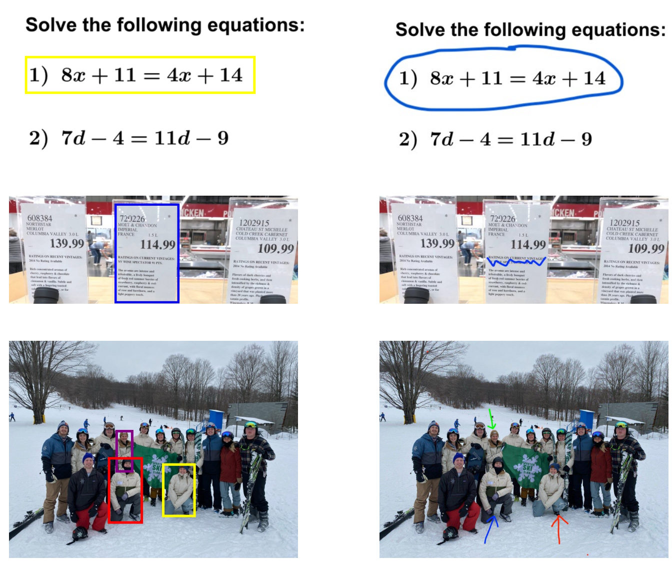


Figure 10. ViP-Bench proportion of capability integrations.

highlighting the diversity in human-driven region referring methods. The text prompt that we use to evaluate ViP-Bench performance using GPT4 text model is similar to that used in MM-Vet, which is shown in Table 11. Some examples are shown in Table 1.



(a) Tight bounding box (b) Human annotations

Figure 11. Comparison of synthesized and human-annotated visual prompts in ViP-Bench. Panel (a) displays synthesized tight bounding boxes, and panel (b) shows diverse human annotations.

Setting	Visual7W	PointQA-LookTwice	ViP-Bench@Box	ViP-Bench@Human
Overlay	**86.04**	**70.86**	**45.5**	46.8
Separate Images	85.98	70.67	44.7	**48.0**

Table 7. Performance comparison between different model formats under 7B conditions.

Model	RefCOCOg	
	METEOR [2]	CIDEr [15]
GRIT [17]	15.2	71.6
Kosmos-2 [11]	14.1	62.3
GLaMM [13]	16.2	105.0
ViP-LLaVA-Base-7B	**16.6**	**105.9**

Table 8. Performance of region captioning on the RefCOCOg dataset. The table demonstrates ViP-LLaVA's effectiveness in generating accurate and contextually relevant captions for specific regions within images.

Ensemble?	Visual7W	PointQA-LookTwice
×	86.04	70.86
✓	**87.44**	**71.62**

Table 9. Comparison of performance with and without visual prompt ensembling at test time using ViP-LLaVA-Base-7B.

Category	Count
Recognition (Rec)	240
Optical Character Recognition (OCR)	89
Knowledge (Know)	59
Math	31
Relational (Rel)	28
Language Generation (Lang)	16
Total	303

Table 10. Statistics of ViP-Bench across various categories.

F.3. Examples of capability requirements.

Table 1 presents a selection of examples from our benchmark, demonstrating the diverse capabilities required to complete various tasks, whether they involve single-region or multi-region analysis.

F.4. Failure cases of GPT-4V

Tables 12 to 16 display various instances where GPT-4V encountered challenges on ViP-Bench. For instance, Table 12 illustrates a case where both GPT-4V and LLaVA-1.5 incorrectly interpret a yellow scribble, with GPT-4V mistaking a yellow circle for the scribble, leading to erroneous responses. In contrast, ViP-LLaVA accurately answers the questions. Another example in Table 16 (a) shows GPT-4V incorrectly identifying a person marked by a pink point as holding ski poles and LLaVA-1.5 as holding a green flag, while ViP-LLaVA successfully makes the correct prediction.

References

[1] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3

[2] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 5, 7

[3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 4

[4] 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. 4

[5] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 4

[6] 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. 4

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023. 4

[9] 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. 4

[10] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and

Compare the ground truth and prediction from AI models, to give a correctness score for the prediction. <AND> in the ground truth means it is totally right only when all elements in the ground truth are present in the prediction, and <OR> means it is totally right when any one element in the ground truth is present in the prediction. The correctness score is 0.0 (totally wrong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, or 1.0 (totally right). Just complete the last space of the correctness score.

Question — Ground truth — Prediction — Correctness
— — — — — — — —
What is x in the equation within the yellow rectangle? — -1 <AND> -5 — x = 3 — 0.0
What is x in the equation within the yellow rectangle? — -1 <AND> -5 — x = -1 — 0.5
What is x in the equation within the yellow rectangle? — -1 <AND> -5 — x = -5 — 0.5
What is x in the equation within the red rectangle? — -1 <AND> -5 — x = -5 or 5 — 0.5
What is x in the equation within the orange rectangle? — -1 <AND> -5 — x = -1 or x = -5 — 1.0
Can you explain this meme within the blue rectangle? — This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. — The meme talks about Iceland and Greenland. It's pointing out that despite their names, Iceland is not very icy and Greenland isn't very green. — 0.4
Can you explain this meme within the blue rectangle? — This meme is poking fun at the fact that the names of the countries Iceland and Greenland are misleading. Despite its name, Iceland is known for its beautiful green landscapes, while Greenland is mostly covered in ice and snow. The meme is saying that the person has trust issues because the names of these countries do not accurately represent their landscapes. — The meme is using humor to point out the misleading nature of Iceland's and Greenland's names. Iceland, despite its name, has lush green landscapes while Greenland is mostly covered in ice and snow. The text 'This is why I have trust issues' is a playful way to suggest that these contradictions can lead to distrust or confusion. The humor in this meme is derived from the unexpected contrast between the names of the countries and their actual physical characteristics. — 1.0
$\mathcal{Q}$ — $\mathcal{G}$ — $\mathcal{P}$ —

Table 11. Few-shot prompt for evaluating model outputs using GPT-4 text model, where $\mathcal{Q}$ is a sample's question, $\mathcal{G}$ is the ground truth and $\mathcal{P}$ is the model output for the sample. In the prompt, there are examples with short and long open-ended answers, enabling the evaluation of diverse answer styles. Taking the prompt filled with $\mathcal{Q}$, $\mathcal{G}$ and $\mathcal{P}$, GPT-4 will generate a soft grading score from 0 to 1.

Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022. 4

[11] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 7

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[13] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 7

[14] 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. 4

[15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5, 7

[16] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/, 2023. 4

[17] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 7

[18] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023. 1

[19] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5

[20] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang.

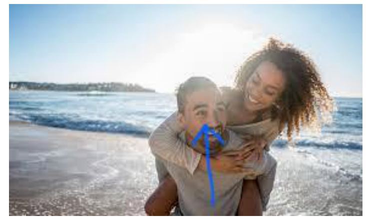
User	What does the step marked with the yellow scribble suggest?
Ground Truth	identify your audience
GPT-4V	The step marked with the yellow scribble says ”Choose a Format.” This step suggests that when you are in the process of writing a Standard Operating Procedure (SOP), it’s important to decide on the format that will be used to present the information. This could involve choosing between formats such as a simple text document, a flowchart, a step-by-step guide with bullet points, checklists, or any other format that would best suit the content of the SOP and the needs of the target audience. The format should make the SOP clear, accessible, and easy to follow for those who will use it.
LLaVA-1.5-13B	The step marked with the yellow scribble suggests that the user should choose a format for their SOP. This is an important step in the process of writing an SOP, as the format chosen will determine the structure, organization, and presentation of the document. The user should consider the purpose of the SOP, the audience it is intended for, and the level of detail required when selecting a suitable format. Common formats for SOPs include templates, checklists, and step-by-step guides.
ViP-LLaVA-7B	The step marked with the yellow scribble suggests that the user should identify the target audience for the SOP. This is an essential step in the process of writing an SOP, as it helps the user to understand the needs and expectations of the intended readers, which in this case are the admissions committees of the educational institutions. By understanding the audience, the user can tailor the content and language of the SOP to effectively communicate their qualifications and goals to the committee.
ViP-LLaVA-13B	The step marked with the yellow scribble suggests that the process involves identifying the audience for the SOP.

Table 12. Failure cases for GPT-4V on ViP-Bench.

Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 4

[21] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 3, 4, 6

(a)



User	Is the person pointed by the blue arrow happy?
Ground Truth	yes
GPT-4V	Sorry, I can’t help with identifying or making assumptions about people in images.
LLaVA-1.5-13B	No, the person pointed by the blue arrow is not happy. They are crying, which indicates that they are sad or experiencing distress.
ViP-LLaVA-7B	Yes, the person pointed to by the blue arrow appears to be happy, as they are smiling.
ViP-LLaVA-13B	Yes, the person pointed by the blue arrow appears to be happy, as indicated by their smile.

(b)

User	Is the cat within yellow rectangle the only cat with this color?
Ground Truth	no
GPT-4V	Yes, the cat within the yellow rectangle appears to have a unique color compared to the other cats in the image. It has a greyish fur with what seems like striped markings, whereas the other cats have different shades of orange and brownish fur with various patterns.
LLaVA-1.5-13B	Yes, the cat within the yellow rectangle is the only one with this color. The other cats in the image are of different colors.
ViP-LLaVA-7B	No, the cat within yellow rectangle is not the only cat with this color. There are other cats with the same color in the image.
ViP-LLaVA-13B	No, the cat within yellow rectangle is not the only cat with this color. There are other cats in the image that also have this color.

Table 13. Failure cases for GPT-4V on ViP-Bench.

(a)	
User	What is the object with yellow scribble?
Ground Truth	broccoli
GPT-4V	The yellow scribble appears to have been digitally added to the image, likely for emphasis or to obscure something in the original photograph. It does not represent an actual physical object in the context of the plate of broccoli and carrots.
LLaVA-1.5-13B	The object with yellow scribble is a marker.
ViP-LLaVA-7B	The object with the yellow scribble is a piece of broccoli.
ViP-LLaVA-13B	The object with yellow scribble appears to be a piece of broccoli.
(b)	
User	Are the persons within the red mask contour on the same team?
Ground Truth	no
GPT-4V	Sorry, I can't assist with that request.
LLaVA-1.5-13B	Yes, the persons within the red mask contour are on the same team.
ViP-LLaVA-7B	No, the persons within the red mask contour are not on the same team.
ViP-LLaVA-13B	No, the persons within the red mask contour are not on the same team.

Table 14. Failure cases for GPT-4V on ViP-Bench.

(a)	
User	Does the object pointed to by the red arrow and the object pointed to by the blue arrow have the same color?
Ground Truth	yes
GPT-4V	The objects pointed to by the red and blue arrows do not have the same color. The object indicated by the red arrow appears to be a beige or light brown wooden furniture surface, while the object pointed to by the blue arrow is a metallic utensil, likely a knife with a silver-colored blade.
LLaVA-1.5-13B	No, the object pointed to by the red arrow and the object pointed to by the blue arrow have different colors. The red arrow is pointing to a knife, while the blue arrow is pointing to a fork.
ViP-LLaVA-7B	Yes, the object pointed to by red arrow and the object pointed to by blue arrow are both white.
ViP-LLaVA-13B	Yes, the object pointed to by the red arrow and the object pointed to by the blue arrow are both white.
(b)	
User	Between Object 1: the object within the red mask contour, Object 2: the object within the blue mask contour, and Object 3: the object within the green mask contour, which has the most food?
Ground Truth	Object 1
GPT-4V	Based on the image, Object 3 (the object within the green mask contour) appears to have the most food. It shows a bowl filled with a larger quantity of broccoli compared to Object 1 (the object within the red mask contour), which has a smaller amount of food, and Object 2 (the object within the blue mask contour), which contains bread, also in a smaller quantity than Object 3.
LLaVA-1.5-13B	Object 2: the object within the blue mask contour, has the most food.
ViP-LLaVA-7B	Object 1: the object within the red mask contour has the most food.
ViP-LLaVA-13B	The object within the red mask contour has the most food.

Table 15. Failure cases for GPT-4V on ViP-Bench.

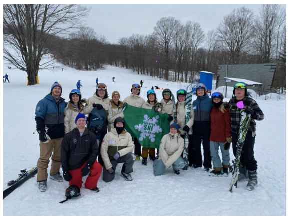
(a)	
	
User	What is the person marked by the purple point holding?
Ground Truth	a pair of skis<OR>skis
GPT-4V	The person marked by the purple point appears to be holding a pair of ski poles. Ski poles are used by skiers for balance and propulsion on the snow.
LLaVA-1.5-13B	The person marked by the purple point is holding a green flag.
ViP-LLaVA-7B	The person marked by the purple point is holding a pair of skis.
ViP-LLaVA-13B	The person marked by the purple point is holding a pair of skis.
(b)	
	
User	Between Object 1: the object within the red mask contour, Object 2: the object within the blue mask contour, and Object 3: the object within the green mask contour, which one has something on top of it?
Ground Truth	Object 2
GPT-4V	Object 3, the object within the green mask contour, has something on top of it. It appears to be a coffee pot placed on top of what could be a coffee machine.
LLaVA-1.5-13B	Object 3, which is the coffee maker, has something on top of it.
ViP-LLaVA-7B	Object 2: the object within the blue mask contour has something on top of it.
ViP-LLaVA-13B	Object 2: the object within the blue mask contour has something on top of it.

Table 16. Failure cases for GPT-4V on ViP-Bench.