

Don't Show Pixels, Show Cues: Unlocking Visual Tool Reasoning in Language Models via Perception Programs

Supplementary Material

Methods	HardBLINK			
	3	4	5	Avg.
🔥 Aurora [1]	66.90	60.50	54.80	60.73
🔥 ReVPT-7B [35]	68.55	55.65	60.48	61.56
❄️ Visual Sketchpad [10] (GPT-5 Mini)	71.77	62.90	56.45	63.71
❄️ GPT-5 Mini	62.10	53.23	41.49	52.42
❄️ GPT-5 Mini + P ²	82.26	87.90	74.19	81.45
❄️ Gemini 2.5 Pro	66.13	55.65	45.16	55.65
❄️ Gemini 2.5 Pro + P ²	77.42	83.87	71.77	77.68
❄️ InternVL3.5-2B	39.52	28.23	20.97	29.57
❄️ InternVL3.5-2B+P ²	35.48	29.03	32.26	32.25
❄️ InternVL3.5-4B	55.65	36.29	29.84	40.59
❄️ InternVL3.5-4B+P ²	72.58	70.97	66.13	69.89
❄️ Qwen3VL-4B	55.65	46.77	38.71	47.04
❄️ Qwen3VL-4B+P ²	73.39	61.29	48.39	61.02

Table 4. **HardBLINK Breakdown.** We report accuracy (%) on different sub-tasks in HardBLINK [1]: 3 point, 4 point and 5 point. Each setting is increasingly difficult from the prior, wherein more candidate points are presented to the MLLM.

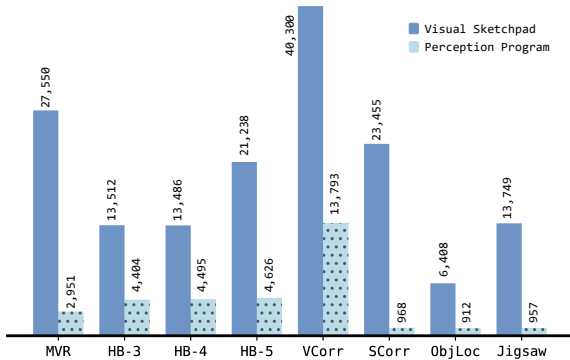


Figure 7. **Average Tokens/Sample.** Comparison of Visual Sketchpad (with GPT-5 Mini as LLM) and GPT-5 Mini with P² on average token per sample across all six sub-tasks. P² incurs significantly lower token cost.

7. Perception Program Details

In this section, we discuss additional details about Perception Programs. Mainly, we provide samples of prompts for both frontier and open-source MLLMs. We also detail in-context (ICL) example that we use to query the open-source MLLMs. Recall that frontier models, GPT-5 Mini and Gemini 2.5 Pro, work as is and do not require any ICL examples. However, for both Qwen3VL and InternVL3.5, we provide a single in-context example as part of the system prompt, see Fig. 8

for an illustrative example for one question in multi-view reasoning task.

Note that the P² rationale as part of in-context sample is taken from GPT-5 Thinking [15]. We prompt GPT-5 Thinking with the question and its P² and ask it to output a short rationale on how it uses P² to compute the answer. We include this obtained rationale as an in-context example for open-source MLLMs. This procedure is similar for all the tasks from the BLINK benchmark we consider in this work, along with HardBLINK. Note that we do the same for raw tool setting and provide exhaustive descriptions of the tool in terms of how to use it to get to the answer, see Fig. 9.

An important distinction to note is for InternVL3.5 (both 2B and 4B variants), we also include additional instructions. In all problems, we direct it to not copy the in-context example as is, along with some problem-specific orientations and clarification. For multi-view reasoning, we mention that clockwise and left are used interchangeably. For relative depth, we explain that the comparison of which point is closest is based on depth range and not coordinates. For semantic correspondence, we reiterate to use similarity scores and not the coordinates for comparison. In visual correspondence we emphasize not to directly use euclidean distance between coordinates in different images to conclude which point corresponds to REF. However, these clarifications are not necessary for Qwen3VL.

For closed-source frontier LLMs, such as GPT-5 Mini and Gemini 2.5 Pro, we just query with the question along with the tool or P² and do not provide any example. The frontier LLMs are able to understand how to use P² on their own.

8. Additional Related Work

In this section, we give a non-comprehensive summary of methods from the related work, expanding on some that were briefly mentioned while also introducing additional ones. We additionally note that several prior state-of-the-art BLINK results were obtained by methods that do not rely on tools, which we also include here.

8.1. Tool-Use through Program Synthesis

Methods in this category leverage program generation, typically Python, to structure the model’s reasoning process. Instead of reasoning purely in natural language, these approaches produce executable code that coordinates external vision modules to enable compositional and interpretable visual reasoning.

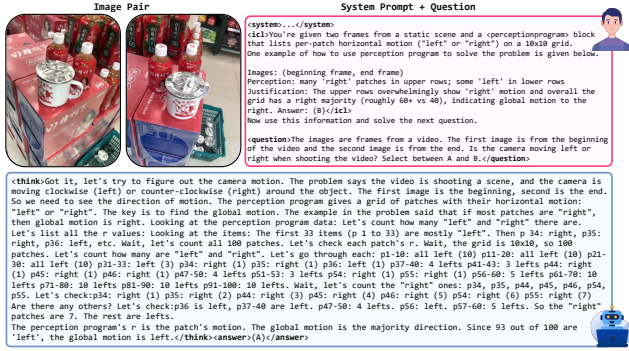


Figure 8. **Open-Source Prompt with P^2 ICL.** We present a sample prompt for open-source MLLMs (e.g., Qwen3VL and InternVL3.5). We include a single in-context example describing the use of P^2 . Both Qwen3VL and InternVL3.5 reason with the given P^2 to compute the **correct answer (A)** to the question.

- **VisProg [9]:** A neuro-symbolic approach in which the model uses in-context learning to generate modular Python programs that call vision models and image-processing tools. Demonstrates strong performance in compositional VQA, reasoning over image pairs, object tagging, and language-guided image editing.
- **ViperGPT [19]:** Reduces the burden on large MLLMs by equipping GPT-based models with an API of callable vision-related subroutines. The model generates Python programs executed on images or video, improving visual grounding and compositional question answering, including cases requiring external knowledge.
- **Thyme [33]:** Enhances logical reasoning by enabling models to perform image-level manipulations such as cropping, rotation, and contrast adjustment. Trained via a two-stage pipeline combining supervised fine-tuning and GRPO with adaptive temperature sampling.
- **MMFactory [4]:** Addresses deployment challenges such as performance constraints and computational limits. Proposes a model that composes programmatic solutions from a tool repository while also suggesting metrics and benchmarks, taking user illiteracy into account to improve real-world usability.

8.2. Tool-Use through Chain-of-Thought

These methods integrate tool usage directly into the reasoning trajectory of the model, often allowing the system to call vision specialists or perform visual operations as part of its intermediate reasoning steps rather than through offline program synthesis.

- **LATTE [14]:** Introduces 8B vision-language models trained to incorporate outputs from multiple vision specialists as part of a think-act-observe reasoning loop. Supports tasks including object recognition, depth estimation, text extraction, and mathematical operations.
- **VisualSketchpad [10]:** A hybrid method that integrates

code generation and tool usage into chain-of-thought reasoning. Inspired by sketch-based human problem solving and implemented on a GPT backbone.

- **VigoRL [17]:** Highlights the gap between the success of RL in math/coding and its limited impact on visually grounded tasks. Proposes grounding reasoning steps in image regions and enabling zoom operations to focus on visually relevant details.
- **ReVPT [35]:** Uses a GRPO-based RL framework to train multimodal LLMs to reason with visual tools such as detection, zooming, edge analysis, and depth estimation. Shows notable improvements on perception-heavy benchmarks (e.g., KV-Bench, BLINK, MMVP, MMStar), with particularly strong gains for 2B-scale models.
- **Visual Perception Token [30]:** Trains models to emit tool-calling tokens that enable selective invocation of external visual modules. Provides tools for region selection-then-zooming and for supplying enriched vision tokens from an auxiliary vision tower.

8.3. Other Prominent Methods

This group includes methods that achieve strong results in multimodal reasoning without relying primarily on explicit tool calls. Many provide alternative training paradigms or new datasets that improve perception or reasoning capabilities.

- **TULIP [21]:** Based on Llama-3.2-11B, this method addresses limitations of CLIP/SigLIP in detailed visual interpretation. Introduces contrastive and reconstruction objectives and provides a drop-in replacement vision tower, yielding improved BLINK performance.
- **PerceptionLM [2]:** Promotes transparency by avoiding reliance on closed-source vision model annotations. Constructs a fully open perception-language model (8B LLM + vision tower) that achieves strong BLINK performance without using tools.
- **Zebra-CoT [11]:** Tackles the scarcity of high-quality sketch/diagram reasoning data by releasing a new interleaved image-text dataset and training the Anole-7B model on it. Instead of using external tools, the model directly generates auxiliary images.
- **OVR [27]:** Argues that prior RL approaches under-scale cognitive behavior training. Proposes a large two-stage RL paradigm achieving strong gains in mathematical reasoning and BLINK tasks.

All of the methods discussed above combine to form a wide-range of baselines we compare our proposed P^2 to in Tables 1 and 2; also refer to Sec. 4.

9. Additional Experimental Details

In Sec. 5.1 we discussed the quality of visual interpretation of current MLLMs. We expand the discussion on on vi-

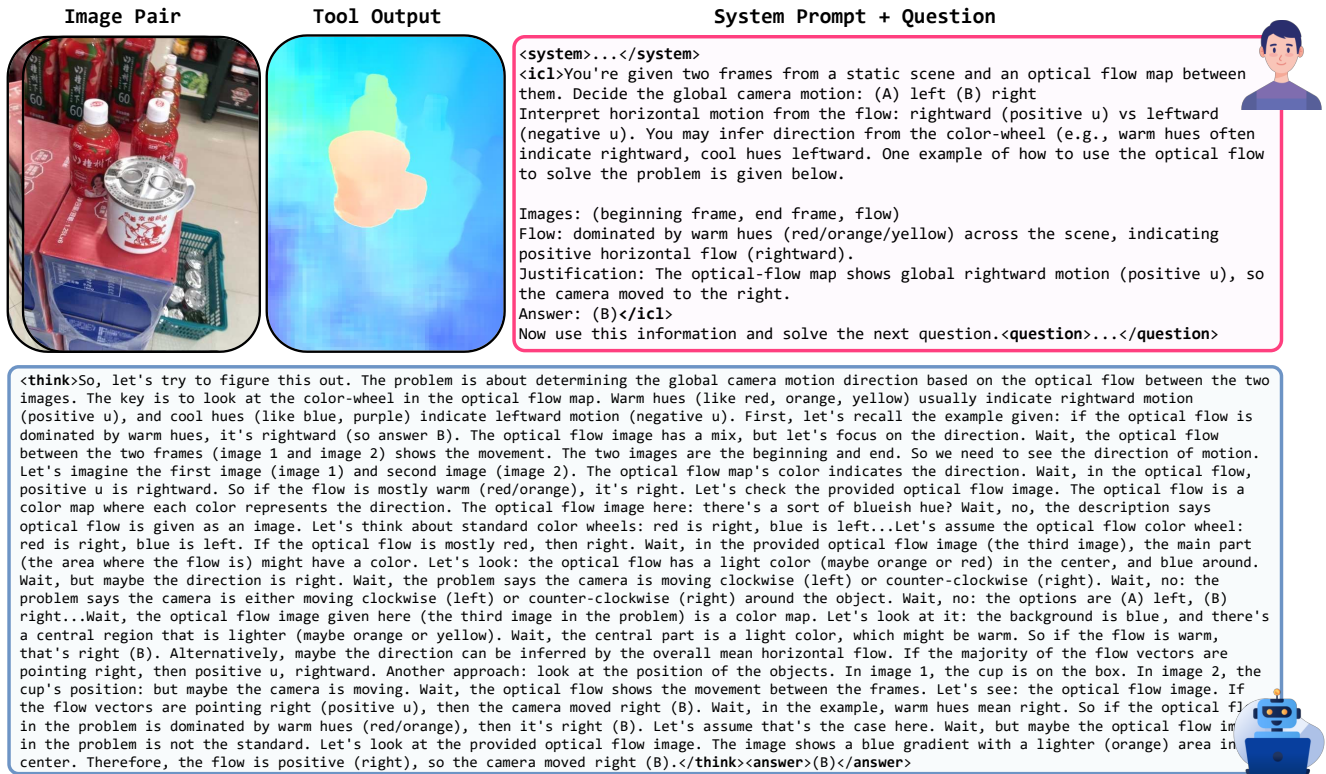


Figure 9. **Open-Source Prompt with Tool ICL.** We present a sample prompt for open-source MLLMs (e.g., Qwen3VL and InternVL3.5). We include a single in-context example describing the use of optical flow as tool output. Note how the example clearly illustrates that blue hues indicate left while warm hues indicate right motion, the MLLM (Qwen3VL in this example) concludes the same that flow is dominated by blue hues, yet gives the **wrong answer (B)**. Note that MLLM also uses a lot more tokens than its P² counterpart (exhausting almost the entire 8192 token budget), we use ... for brevity purposes in this illustrative figure.

sual correspondence task and describe the two baselines we included, namely Naive and Oracle, also see Fig. 6.

9.1. Additional Details on Visual Correspondence

When evaluating the reconstructed P² results, we considered the naive Euclidean and oracle baselines. We describe their setup as follows.

Naive. It receives ground-truth coordinates corresponding to the BLINK alternatives: REF in the reference image and A, B, C, D, E in the target image. It simply gives the answer as the point whose coordinate is closest to REF in the normalized coordinate space. This method completely disregards the visual content of the image and is therefore unsuitable to solve the task of visual correspondence. Its performance of 85% indicates that most pairs of images indeed have low camera movement, which we later confirm in Sec. 9.2.

Oracle. An implementation of correct usage of the P² to navigate through correspondences is what we term as an oracle. Concretely, we first store the reference (REF) point coordinates. We then scan the correspondence P² to find all the candidate points and read their 'c' coordinates. For

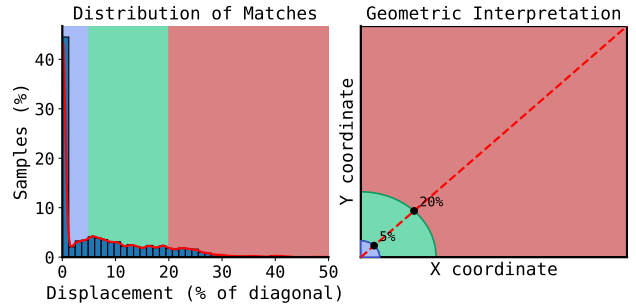


Figure 10. **Correspondence Distribution.** Illustration of distribution of correspondence markers in the visual correspondence task from BLINK validation set.

each candidate, we compute its Euclidean distance to the reference point and select the neighbor with the smallest distance. We then take this neighbor's 'r' coordinate as the mapped position in the second image. Finally, we compare this mapped location with the coordinates of alternatives A, B, C, D, E, and choose the point whose coordinates are closest in Euclidean distance as the correspondence of the original reference point.

9.2. Distribution of Displacements

We here plot the distribution of visual correspondences in the BLINK validation set. The left pane illustrates the histogram and density of LoFTR [18] displacements across the whole dataset as a percentage of the diagonal. To give the reader a rough visual reference of the displacement ranges, we illustrate regions closer than 5% (blue) and 20% (green) of the diagonal, considering the normalized coordinate space. Colors between panes correspond to matching regions. We can see that the majority of displacements are closer than 5% of image diagonal, further evidencing the fact that this dataset is biased towards low displacement between images.

9.3. Breakdown on HardBLINK

In Tab. 4, we present results on each sub-task, 3-, 4-, and 5-point, in HardBLINK benchmark introduced in Bigverdi et al. [1] to complement the results presented in Tab. 1 (the HardBLINK performance reported there is the average of these sub-tasks).

Across all three HardBLINK settings, we observe a consistent trend: performance drops as the number of candidate points increases, but P² substantially narrows this gap. Specialized baselines such as Aurora [1], ReVPT [35], and Visual Sketchpad (GPT-5 Mini) [10] achieve average accuracies between 60.73% and 63.71%, with modest degradation from 3-point to 5-point tasks. In contrast, raw MLLMs struggle more severely as difficulty increases, i.e., GPT-5 Mini falls from 62.10% on 3-point to 41.49% on 5-point sub-task. Both GPT-5 Mini and Gemini 2.5 Pro immensely benefit from P², and even the smaller InternVL3.5-4B and Qwen3VL-4B observe +29.30% and +13.98% increase.

Overall, P² not only lifts all base models, but is particularly effective in the more challenging 4- and 5-point regimes, where raw MLLMs otherwise collapse.

10. LLM Usage Statement

In this manuscript, we used several MLLMs as part of our experimental setup and we have described the necessary details in Secs. 4 and 7. Other than that, we also used LLMs (ChatGPT) to help with refining the manuscript in terms of fixing grammatical errors in writing and with plotting codes for various figures. The authors did not use any LLM in any part of ideation, experimental design, analysis of results, and implementation of core methodology.

References

- [1] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845, 2025. 1, 2, 3, 5, 7, 4
- [2] Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 2, 5, 6
- [3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasapat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5, 6
- [4] Wan-Cyuan Fan, Tanzila Rahman, and Leonid Sigal. Mmfactory: A universal solution search engine for vision-language tasks. *arXiv preprint arXiv:2412.18072*, 2024. 3, 5, 6, 7, 2
- [5] Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025. 3
- [6] Stephanie Fu, Tyler Bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: Vlms overlook their visual representations, 2025. 1, 2, 3
- [7] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llm-det: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14987–14997, 2025. 5
- [8] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024. 1, 2, 5, 6, 8
- [9] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14953–14962, 2023. 2, 3
- [10] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *Advances in Neural Information Processing Systems*, 37:139348–139379, 2024. 3, 5, 6, 7, 8, 1, 2, 4
- [11] Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, et al. Zebra-cot: A dataset for interleaved vision language reasoning. *arXiv preprint arXiv:2507.16746*, 2025. 5, 6, 2
- [12] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer, 2024. 7
- [13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

- In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 8
- [14] Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, Ranjay Krishna, and Silvio Savarese. Latte: Learning to think with vision specialists. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2025. 2, 3, 5, 6, 8
- [15] OpenAI. Gpt-5 system card, 2025. 5, 6, 1
- [16] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 1
- [17] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025. 2, 5, 7
- [18] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 5, 8, 4
- [19] 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. 2, 3
- [20] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 5
- [21] Zineng Tang, Long Lian, Seun Eisape, Xudong Wang, Roei Herzig, Adam Yala, Alane Suhr, Trevor Darrell, and David M Chan. Tulip: Contrastive image-text learning with richer vision understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4267–4277, 2025. 2, 5, 6
- [22] Qwen Team. Qwen3 technical report, 2025. 5, 6
- [23] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5
- [24] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 3
- [25] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5, 6
- [26] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5
- [27] Yana Wei, Liang Zhao, Jianjian Sun, Kangheng Lin, Jisheng Yin, Jingcheng Hu, Yinmin Zhang, En Yu, Haoran Lv, Zejia Weng, et al. Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. *arXiv preprint arXiv:2507.05255*, 2025. 5, 6, 2
- [28] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 5, 8
- [29] Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. *arXiv preprint arXiv:2506.17218*, 2025. 1, 2, 3, 5, 6, 7
- [30] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multimodal large language model. *arXiv preprint arXiv:2502.17425*, 2025. 2
- [31] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 1
- [32] Bob Zhang, Haoran Li, Tao Zhang, Cilin Yan, Jiayin Cai, and Yanbin Hao. Improving the reasoning of multi-image grounding in mllms via reinforcement learning. *arXiv preprint arXiv:2507.00748*, 2025. 7
- [33] Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025. 2, 3, 5, 6
- [34] Zhehao Zhang, Ryan Rossi, Tong Yu, Franck Dernoncourt, Ruiyi Zhang, Jiuxiang Gu, Sungchul Kim, Xiang Chen, Zichao Wang, and Nedim Lipka. Vipact: Visual-perception enhancement via specialized vlm agent collaboration and tool-use. *arXiv preprint arXiv:2410.16400*, 2024. 3
- [35] Zetong Zhou, Dongping Chen, Zixian Ma, Zhihan Hu, Mingyang Fu, Sinan Wang, Yao Wan, Zhou Zhao, and Ranjay Krishna. Reinforced visual perception with tools. *arXiv preprint arXiv:2509.01656*, 2025. 2, 5, 6, 1, 4