Improving Large Vision and Language Models by Learning from a Panel of Peers

Supplementary Material

A. Implementation Detatils

A.1. Training Hyperparameters

In Table A.1, we list the detailed training dataset usage and hyperparameters. The training data are constructed based on the following datasets: `BLIP-LAION-CC-SBU` [23], which contains 558K image-text pairs from BLIP-captioned CC3M [19], SBU [31], and LAION400M [33] filtered by LLaVA; `LLaVA-Instruct-mix665k` [21], which contains 665k visual instruction-following data constructed to train the LLaVA family of models; and synthetic data created using images and questions from the `Cambrian-7M` dataset [36]. Unless otherwise specified, we randomly sample the indicated number of instances from each dataset during the training process. During training, we use Flash Attention [8], bfloat16, and PyTorch FSDP [43] to accelerate training efficiency.

A.2. Panel-of-Peers Models

Image Processing and Visual Representations We implement all image processing logic using the default image transforms provided by `torchvision` and the TIMM library [38]. We normalize pixel values using the default ImageNet normalization values. The default backbone employed by all visual representations that we evaluate in this work is a Vision Transformer [10] trained with the CLIP objective [32]; we extract patch features from the *penultimate* layer, following LLaVA [23].

Vision-Language Projector We use a simple 2-layer GELU MLP as the projector, which projects each patch independently into the embedding space of the language model.

Language Model We choose three models to create the Panel-of-Peers: Vicuna-7B [6], Mistral-7B [14], and -8B [2]. In order to combine the projected visual patch embeddings, we perform simple sequence-wise concatenation by placing the patch embeddings before the text embeddings.

A.3. Evaluation benchmarks

Systemic evaluations of the Panel-of-Peers regarding General VQA, knowledge, Chart&OCR, Hallucination, and Vision-Centric capabilities have been conducted. The benchmarks and datasets used are listed in Table A.2. During the evaluation, we use `VLMEvaKit` [11] as our primary evaluation toolkit.

A.4. Prompt Template

To evaluate model-generated responses within our Panel-of-Peers (PoP) framework, we designed a detailed prompt template to guide models in rating responses. This prompt was central to generating pseudo-rewards, which serve as feedback signals to enable self-improvement iterations. Each model evaluated the outputs of its peers based on a set of predefined criteria and aggregated their results using an ensemble strategy to achieve consensus. The prompt comprises three main components: *System Prompt*, *Evaluation Criteria*, and *Rating Guidelines*. It is structured as follows:

- **System Prompt:** The model is instructed to act as an expert evaluator tasked with assessing the quality of a response provided to a user's question. Both the question and its related image are provided for context.
- **Evaluation Criteria:** Responses are evaluated across five dimensions on an ordinal Likert scale:
  1. **Helpfulness:** Utility of the response in addressing the user's query (1 to 5 scale).
  2. **Correctness:** Accuracy and factuality of the response (1 to 5 scale).
  3. **Coherence:** Logical consistency and clarity of the response (1 to 5 scale).
  4. **Complexity:** Level of language sophistication, ranging from simple to expert-level (1 to 5 scale).
  5. **Verbosity:** Appropriateness of detail and conciseness (1 to 5 scale).
- **Rating Guidelines:** Models receive detailed explanations for scoring each dimension. For instance, a rating of 5 in Helpfulness indicates complete alignment with the user's intent, while a 1 represents a failure to address the query effectively. Similarly, Coherence is rated based on logical flow, with a 1 indicating substantial contradictions or redundancy.
- **Output Format:** To standardize results, models are instructed to provide evaluations in a strict JSON schema format, including scores for each criterion.

This prompt enabled consistent and systematic evaluation of the model-generated responses, ensuring that pseudo-rewards were aligned with the evaluation objectives outlined in our PoP framework.

B. Additional Experiments

B.1. Comparison with State of the Art

We compare against the top 49 models on the OpenVLM leaderboard, highlighting the performance of our models

Prompt Template for Generating Responses from the Panel of Peers

[System Prompt]
You are an expert evaluation model. You are asked to evaluate the AI assistant's response to a user's question based on an image. You will see the user's question, the related image, and the AI's response.

[Evaluation Criteria]
Please rate the response using a 5-point Likert scale across the following dimensions: Helpfulness, Correctness, Coherence, Complexity, and Verbosity.

[Rating Guidelines]
- **Helpfulness:** Rate from 1 (not useful at all) to 5 (extremely helpful).
- **Correctness:** Score from 1 (completely incorrect) to 5 (fully correct and accurate).
- **Coherence:** Evaluate from 1 (completely incoherent) to 5 (perfectly coherent and clear).
- **Complexity:** Assess from 1 (basic, understandable by children) to 5 (expert level, specialized vocabulary).
- **Verbosity:** Judge from 1 (very concise) to 5 (highly detailed and verbose).

[Brief Definitions]
- **Helpfulness** relates to the utility of the response in addressing the user's need.
- **Correctness** ensures the response is factual and free from errors.
- **Coherence** checks for logical flow and consistency in the response.
- **Complexity** reflects the sophistication of language and concepts used.
- **Verbosity** measures the brevity or expansiveness of the response.

Here is the question and the assistant response:
[Question]
{question}
[Assistant Response]
{response}

[JSON Output]
Your answer should look like this. Only output result in the following JSON schema format:
{"Helpfulness": (int), "Correctness": (int), "Coherence": (int), "Complexity": (int), "Verbosity": (int) }

Figure A.1. **Evaluating Synthetic Responses.** We use the following prompt template, which is used to evaluate responses from the Panel-of-Peers.

	`Stage I`	`Stage II`	`Stage III`
Config	Alignment	SFT	PoP
Training Hyper-Parameters			
Optimizer	AdamW	AdamW	AdamW
Learning Rate	2e-3	2e-5	6e-5
Weight Decay	0.0	0.0	0.0
Training Epochs	1	1	2
Warmup Ratio	0.003	0.003	0.003
Learning Rate Scheduler	Cosine	Cosine	Cosine
Batch Size Per GPU	16	8	8
Maximum Token Length	2048	2048	2048
Unfreeze LLM	✗	✓	✓
Training Data			
Dataset	`BLIP-LAION-CC-SBU`	`LLaVA-Instruct-mix665k`	Sampled from `Cambrian-7M`
Data Size	558K	665K	3 ×300K
Data Type	Pair	Instruction	Synthetic
Training Cost			
GPU Device	8×NVIDIA A100-80GB	8×NVIDIA A100-80GB	8×NVIDIA A100-80GB
Training Time	∼6h	∼10h	∼90h

Table A.1. **Training recipes** for PoP. The three training stages are introduced in Section 3. `Stage I`: Alignment training, `Stage II`: Instruction Tuning, `Stage III`: Panel-of-Peers Learning.

using PoP. Our models include PoP-Vicuna, PoP-Mistral, PoP-LLaMA3, and their single-try counterparts, which are evaluated in 15 benchmarks against a broad spectrum of state-of-the-art methods.

Capability	Dataset	Task description	Eval Split	Metric
General VQA	MM-Vet [41]	Multi-disciplinary QA	-	GPT-4 Eval [41]
	MMBench [24]	Multi-disciplinary QA	dev	GPT-3.5 Eval [24]
	SEED-Bench [17]	Multi-disciplinary QA	-	Multi-choice Acc
Knowledge	AI2D [15]	Science Diagrams	test	Multi-choice Acc
	MMMU [42]	College-level Multi-disciplinary	val	Multi-choice Acc
	MMStar [4]	Misc Multi-disciplinary	-	Multi-choice Acc
	MathVista [27]	General Math Understanding	min	GPT-4 Eval
	ScienceQA [26]	High-school Science	val	Multi-choice Acc
Chart&OCR	ChartQA [28]	Chart Understanding	test	Relaxed Accuracy
	TextVQA [34]	OCR; Reasoning	val	VQAScore
	OCR-Bench [25]	OCR; Multi-disciplinary	-	Acc
	OCRVQA [29]	Document OCR	TESTCORE	Acc
Hallucination	POPE [20]	Yes/No Hallucinations	-	Acc, F1-score
	HallusionBench [12]	Visual Hallucination	-	Acc, F1-score
Vision Centric	RWQA [39]	Real-world QA	dev	Multi-choice Acc

Table A.2. **Overall descriptions of the evaluation benchmarks** for evaluating capabilities, including GeneralVQA, Knowledge, Chart&OCR, Hallucination and Vision Centric Benchmarks.

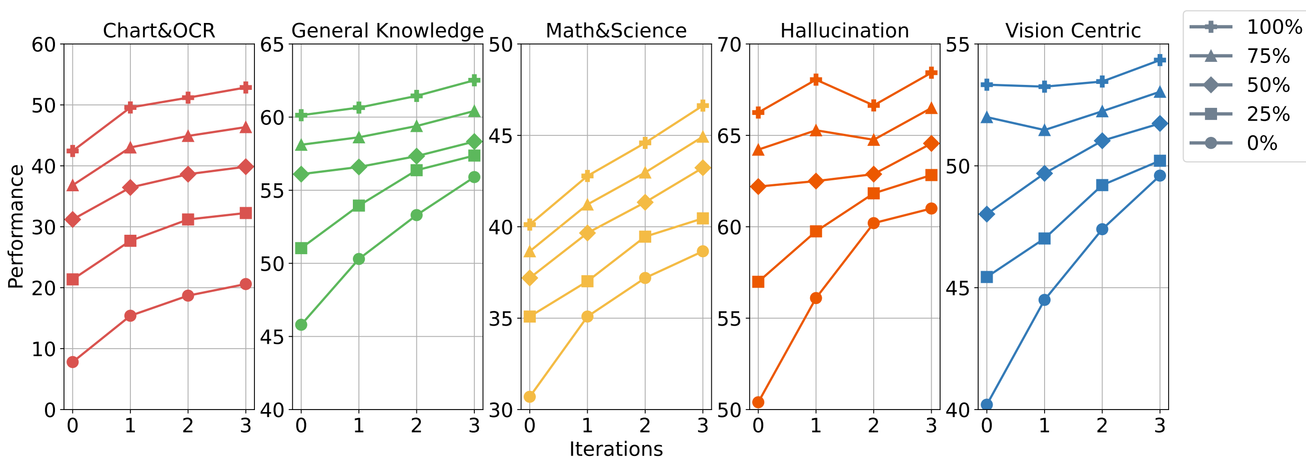

Figure A.2. **Learning a New Skill from Peers (OCR).** We start with a model with very limited OCR knowledge ($\approx 0\%$) and use PoP to iteratively teach OCR skills. The performance is evaluated across multiple categories, including Chart & OCR, General Knowledge, Math & Science, Hallucination, and Vision-Centric tasks.

Our best-performing models, PoP-LLaMA3 and mt-PoP-LLaMA3, achieve an average score of 56.3% and 59.7%, starting from a score of 48.9%. Compared to proprietary models like GPT4-o [30] and Gemini-1.5 [35], our models lag behind approximately 20 percentage points in performance. A similar gap is observed when compared with open-source state-of-the-art models, such as Qwen2-VL-72B [37], InternVL2-Llama3-76B [5], and NVLM-D-72B [7]. Compared to models of the same size category but trained on significantly more data and higher-resolution inputs, our best-performing models lag behind the recently released Qwen2-VL-7B [37], the LLaVA-OneVision family [18], and the Molmo family [9] by approximately 10 percentage points. Compared to models of the same size category trained on similar budgets, our best-performing model surpasses all the LLaVA-NeXT family [22] except for models larger than 30B by approximately 5 percentage points. We remark that our models use 224x224 pixels as the input resolution compared

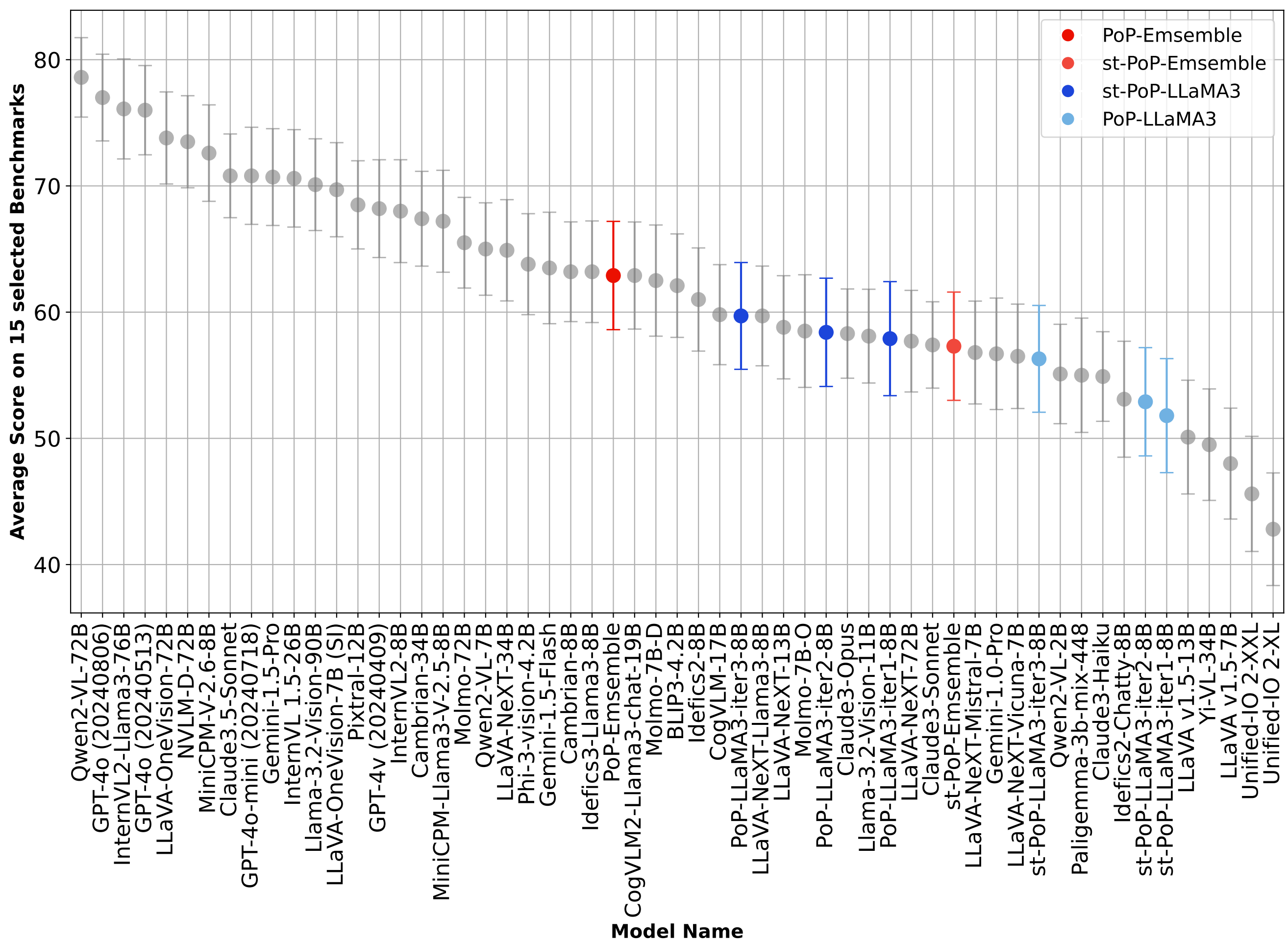

Figure B.1. **Evaluation results of our approach on 15 selected benchmarks in the OpenVLM Leaderboard.** The figure displays 49 selected LVLMs (until 2024.10.30) in descending order of average score. When calculating the average score, the scores of each benchmark are normalized to the range of 0 to 100.

to 768x768 pixels of the NeXT family.

These results demonstrate the efficacy of our approach in using peer evaluations to improve model performance, effectively increasing the average score by approximately 12% compared to the original `LLaVA-1.5-7b` model. Figure B.1 illustrates a comparative analysis of the top 49 models on the OpenVLM leaderboard [11], highlighting the performance of our models using peer-to-peer learning.

B.2. Extra Results on Learning and Ability from Scratch

In addition to the ablation study presented in the main manuscript, where we evaluated the ability of the Panel-of-Peers (PoP) framework to teach a model OCR capabilities, we expanded the analysis to include the performance of the *OCR-Dumb* model across other benchmark categories. Figure A.2 provides a comprehensive view of the model's iterative performance improvement across five categories: *Chart&OCR*, *General Knowledge*, *Math and Science*, *Hallucination*, and *Vision-Centric* tasks.

The experiment began with an OCR-Dumb model trained with varying proportions of OCR knowledge (0%, 25%, 50%, 75%, and 100%). Interestingly, the results demonstrate that as OCR knowledge increases, the model's performance steadily improves not only in OCR-related tasks but also in other categories. Notable observations include:

- **Chart and OCR:** Performance rises sharply with increased OCR knowledge, validating the importance of reading capabilities for interpreting structured visual data.
- **General Knowledge:** Gains in this category suggest that improved text recognition contributes to better multimodal understanding and reasoning.
- **Math and Science:** Enhanced OCR capabilities positively impact tasks involving numerical and scientific reasoning,

Capability	Benchmark	Iteration 0			Iteration 1			Iteration 2			Iteration 3		
GeneralVQA	MMBench [24]	62.4	66.5	65.6	65.3	65.5	68.7	65.7	66.1	69.9	67.0	67.4	71.3
	MM-Vet [41]	21.1	32.9	26.2	24.5	31.9	29.5	29.5	31.6	31.9	30.5	32.5	33.0
	SEED-Bench [17]	64.6	65.8	61.6	68.7	61.6	65.5	65.1	66.2	66.9	66.8	67.9	68.6
Knowledge	†AI2D [15]	62.0	55.5	61.1	66.0	59.1	65.1	65.8	60.2	66.1	64.6	62.9	71.4
	MMMU [42]	32.7	35.7	33.6	35.5	38.8	36.6	39.1	36.4	36.9	39.9	37.1	37.6
	MMStar [4]	36.4	33.1	38.6	37.4	34.0	39.6	40.8	39.6	40.2	41.5	40.9	45.3
	MathVista [27]	30.3	25.6	30.3	33.1	31.2	33.1	34.9	33.8	35.5	35.0	34.9	37.7
	†ScienceQA [26]	58.0	66.8	71.2	62.4	67.1	73.1	66.1	71.9	75.4	68.0	74.0	77.6
Chart&OCR	†ChartQA [28]	39.6	31.9	40.4	42.4	42.7	43.3	46.3	45.1	45.7	48.4	47.1	47.8
	†TextVQA [34]	44.9	45.5	44.9	48.4	49.0	48.4	50.2	50.3	49.3	52.2	52.3	51.2
	OCR-Bench [25]	33.6	31.8	33.9	34.7	33.8	35.0	39.5	38.7	38.3	41.3	41.6	44.5
	OCRVQA [29]	59.7	60.6	57.7	62.7	63.6	60.6	60.9	62.4	61.7	61.4	62.9	62.2
Hallucination	POPE [20]	87.0	86.1	84.8	85.1	86.8	83.0	86.2	86.4	84.1	86.1	86.3	85.0
	HallusionBench [12]	30.4	27.6	32.4	34.7	32.6	37.1	30.7	31.7	30.7	28.2	31.8	36.5
Vision Centric	RWQA [39]	53.1	54.8	48.9	54.6	53.2	50.3	53.0	49.6	52.9	53.4	50.0	53.3
	Average	47.7	48.0	48.7	50.4	50.1	51.2	51.6	51.3	52.4	51.2	51.6	53.7

Table B.1. **Evaluation on 15 vision-language benchmarks.** We compare the performance of the single-try Panel-of-Peers (st-PoP). We have separated the benchmarks into five categories. Columns show three training iterations for = PoP-Mistral, = PoP-Vicuna, and = PoP-LLaMA3. † indicates that the training set has been observed in our data mixture.

where understanding text is critical.

- **Hallucination:** Improvements here indicate that OCR knowledge helps reduce misalignments and inconsistencies in model outputs at the beginning. However, this improvement plateaus if the model starts with more OCR knowledge.
- **Vision-Centric:** Even tasks not directly reliant on OCR knowledge show gradual improvement, though to a lesser extent, with more OCR knowledge. This emphasizes the holistic impact of PoP training.

These results show the applicability of Peer-to-Peer Learning, demonstrating its ability to transfer knowledge, including OCR, while simultaneously increasing performance in various multimodal tasks. This highlights the effectiveness of PoP as a self-improvement mechanism, enabling models to iteratively learn new capabilities and address their initial weaknesses.

B.3. Extra Details on the Panel-of-Peers Ensemble as a Zero-Shot Evaluator

We present more details on the experiments in Section 5.2. For models with more than 3B parameters, we included Phi-3-Vision [1], BLIP3 [40], and Paligemma [3]. In the more than 7B range, we selected LLaVA-NeXT-Llama3, LLaVA-NeXT-Mistral, LLaVA-NeXT-Vicuna [22], and Idefics2 [16]. For models exceeding 10B parameters, we picked CogVLM2-Chat [13], LLaVA-NeXT-Vicuna-13B [22], and Llama-3.2-Vision [2]. For models with more than 30B parameters, we incorporated InternVL2-26B, InternVL 1.5-26B [5], Cambrian-34B [36], and LLaVA-NeXT-Yi [22]. Each panel performed response regeneration and evaluations. However, this is an evaluation-only method, enabling the creation of an ensemble using their consensus.

B.4. Extra Results of Our Trained Models

We present the specific scores of each of the members of the panel of peers, outlined in Table 2 of the main manuscript, where we presented the average scores of the whole panel.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 5

[2] AI@Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1, 5

[3] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele

Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 5

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 3, 5

[5] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3, 5

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 1

[7] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint*, 2024. 3

[8] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 1

[9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 3

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[11] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024. 1, 4

[12] T Guan, F Liu, X Wu, R Xian, Z Li, X Liu, X Wang, L Chen, F Huang, Y Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. arxiv. 10.48550. *arXiv preprint arXiv.2310.14566*, 2023. 3, 5

[13] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 5

[14] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*, 2023. 1

[15] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 3, 5

[16] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 5

[17] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 3, 5

[18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 3, 5

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1

[22] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3, 5

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[24] 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. 3, 5

[25] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 3, 5

[26] 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. 3, 5

[27] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 5

[28] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022. 3, 5

[29] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3, 5

[30] OpenAI. Gpt-4o system card, 2024. Accessed: 2024-09-30. 3

[31] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 1

[32] 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. 1

[33] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[34] 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 (CVPR)*, 2019. 3, 5

[35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[36] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1, 5

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[38] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 1

[39] x.ai Team. Grok-1.5 vision preview, 2024. 3, 5

[40] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024. 5

[41] 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. In *Forty-first International Conference on Machine Learning*, 2024. 3, 5

[42] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 3, 5

[43] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023. 1