HoVLE: Unleashing the Power of Monolithic Vision-Language Models with Holistic Vision-Language Embedding

Supplementary Material

A. Implementation Details

Hyper-parameters. The hyper-parameters for three training stage of our HoVLE are listed in Table 7, Table 8, and Table 9, respectively.

Datasets. Table 10 and Table 11 list the detailed datasets used in the alignment and instruction tuning stages.

Hyper-parameters	Value
Resolution of Image tile	448×448
Amount of data	500M
Batch size	4096
Warmup steps	2000
Optimizer	AdamW
Peak learning rate	3×10^{-4}
Learning rate schedule	Constant
Weight decay	0.05
AdamW β	$(0.9, 0.999)$
AdamW ϵ	1×10^{-8}

Table 7. **Hyper-parameters for distillation stage**

Hyper-parameters	Value
Resolution of Image tile	448×448
Amount of data	50M
Batch size	4096
Warmup steps	100
Optimizer	AdamW
Peak learning rate	5×10^{-5}
Learning rate schedule	Cosine
Weight decay	0.01
AdamW β	$(0.9, 0.999)$
AdamW ϵ	1×10^{-8}

Table 8. **Hyper-parameters for alignment stage**

Hyper-parameters	Value
Resolution of Image tile	448×448
Amount of data	5M
Batch size	4096
Warmup ratio	0.03
Optimizer	AdamW
Peak learning rate	4×10^{-5}
Learning rate schedule	Cosine
Weight decay	0.01
AdamW β	$(0.9, 0.999)$
AdamW ϵ	1×10^{-8}

Table 9. **Hyper-parameters for instruction tuning stage**

task	dataset
Short Caption	Laion (en&zh) [61], COYO [4], COCO [41]
OCR	Wukong-OCR [16], LaionCOCO-OCR [62]
Detection	GRIT [59]
Conversation	All-Seeing (en&zh) [72]
Image-text instruction data (see Table 11)	

Table 10. **Summary of datasets used in the alignment stage.**

task	dataset
General QA	VQAv2 [15], GQA [25], OKVQA [53] VSR [44]
Science	AI2D [32], ScienceQA [49], Chemistry Data [38] TQA [33]
Medical	PMC-VQA [82], VQA-RAD [36], VQA-Med [2] Medical-Diff-VQA [22], PathVQA [19], SLAKE [42], PMC-CaseReport [76]
Chart	ChartQA [55], LRV-Instruction [45], PlotQA [58] Unichart [56], MMC-Inst [46], DVQA [28] TableMWP [50], FigureQA [29], MapQA [7] SciTSR [11], Fintabnet [84]
Mathematics	CLEVR [27], MetaMath [78], GeoQA+ [5] Geometry3k [48], GeoS [63], Unigeo [9] Super-CLEVR [40], MathQA [1]
Knowledge	Art500k [52], MovieNet [23], KonIQ-10k [20] KVQA [64], ViQuAE [37]
OCR	InfoVQA [57], TextVQA [67], ArT [12] CASIA [43], Chart-to-text [30], COCO [70] CTW [79], EATEN [17], ICDAR2019-LSVT [69] ICPR MTWI [18], NAF [14], ReCTS [81] TextOCR [68], LLaVAR [83], HME-100k [80] POIE [34], SROIE [24], ST-VQA [3] EST-VQA [73], IAM [54]
Document	DocVQA [13], DocReason25k [21]
Grounding	RefCOCO [31], RefCOCO+ [31], RefCOCOg [31] RD-BoxCoT [10]
Conversation	ALLaVA [8], LAION-GPT4V [35] MMDU [47], TextOCR-GPT4V [6]
Detection	Objects365 [66], V3Det [71]
Video	CLEVRER [77], EgoTaskQA [26], LSMDC [60] Mementos [74], STAR [75], NTU RGB+D [65] VideoChat2-IT* [39], VideoGPT+ [51],

Table 11. **Summary of datasets used in the instruction tuning stage.** *IT refers to the instruction tuning data in VideoChat2.

Trainable Modules	MMB	MME	SEED	TextVQA	InfoVQA	DocVQA
None	66.1	1698	66.7	55.0	43.0	71.8
Full Model	72.0	1862	70.9	62.0	51.4	82.1
Holistic Embedding	69.9	1806	69.5	60.4	48.2	78.2

Table 12. **Ablation on instruction tuning strategy.**

B. More Ablation Studies

Instruction Tuning Strategy. We also tried to only fine-tune the holistic embedding and freeze the LLM during instruction tuning, as shown in Table 12. It's shown that instruction tuning only the holistic embedding can also

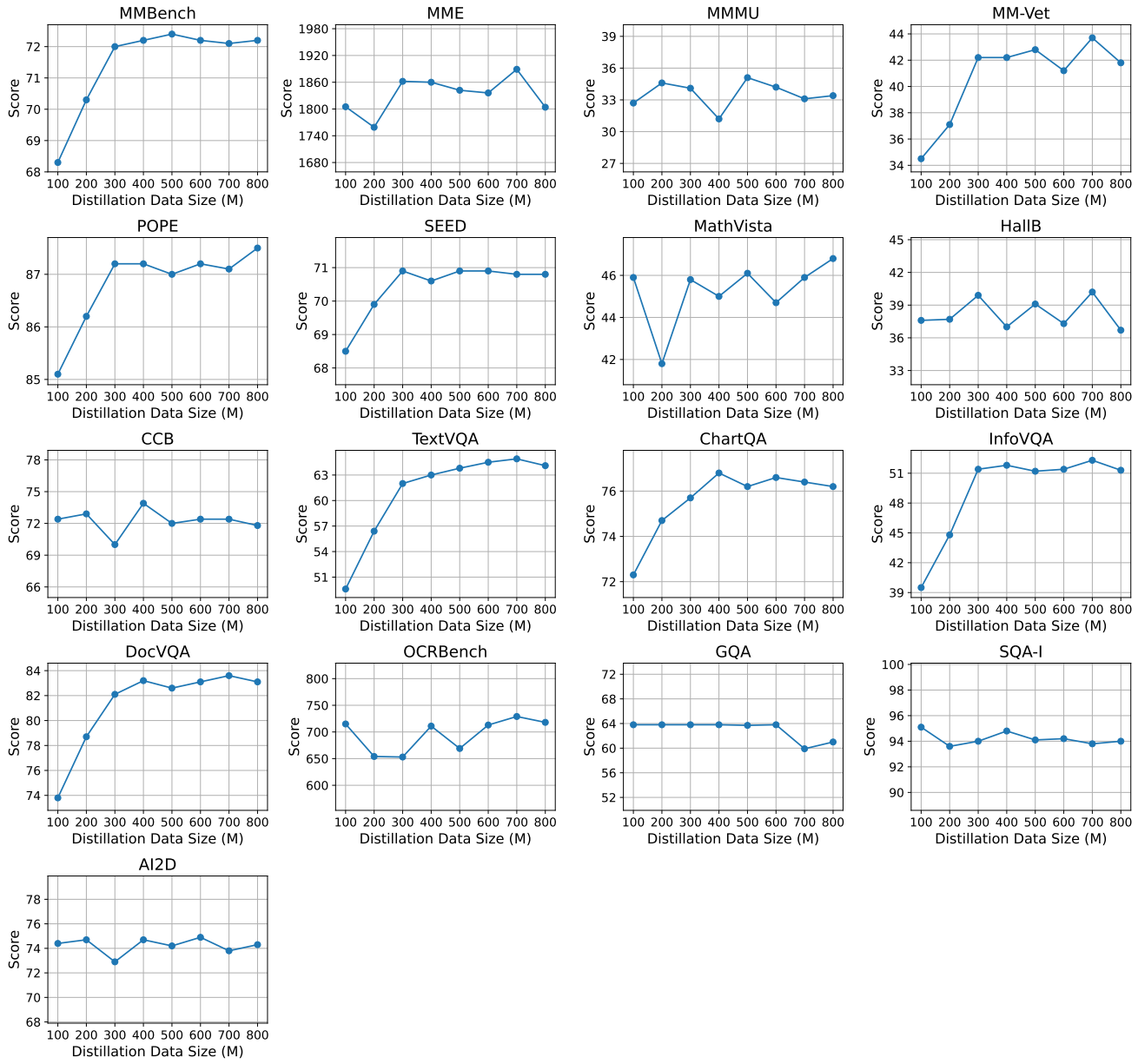

Figure 6. **Distillation data scaling performance on 17 benchmarks.**

Model	Tile Resolution	MMB	MME	MMMU	MM-Vet	POPE	SEED	MathVista	HallB	CCB
InternVL2	448×448	73.2	1877	34.3	44.6	88.3	71.6	46.4	37.9	74.7
HoVLE (HD)	336×336	72.9	1869	33.6	40.4	87.8	70.2	46.1	39.4	71.6
HoVLE (HD)	392×392	72.3	1838	34.2	43.9	87.4	70.8	45.3	38.7	70.6
HoVLE (HD)	448×448	73.3	1862	32.2	43.8	87.4	70.9	49.2	38.4	74.3

Table 13. **Inference speed-performance trade-off of HoVLE (HD) on general VLM benchmarks.**

achieve decent results close to full modeling tuning. This may imply that HoVLE has advantage in retaining the full capabilities of LLMs, unlike previous compositional and monolithic VLMs that still require tuning the LLMs.

Distillation Data Scaling. We provide model performance change as data scales up on 17 benchmarks in Figure 6. It's shown that the model performance continues to improve on 9 benchmarks, while it oscillates on other 8 benchmarks.

Model	Tile Resolution	TextVQA	ChartQA	InfoVQA	DocVQA	OCRBench	GQA	SQA-I	AI2D
InternVL2	448×448	73.4	76.2	57.7	85.9	784	61.0	84.9	74.1
HoVLE (HD)	336×336	68.1	76.7	55.6	84.2	739	63.7	94.6	73.1
HoVLE (HD)	392×392	70.0	78.3	55.4	85.4	737	64.9	94.4	73.3
HoVLE (HD)	448×448	70.9	78.6	55.7	86.1	740	64.9	94.8	73.0

Table 14. **Inference speed-performance trade-off of HoVLE (HD) on visual question answering benchmarks.**

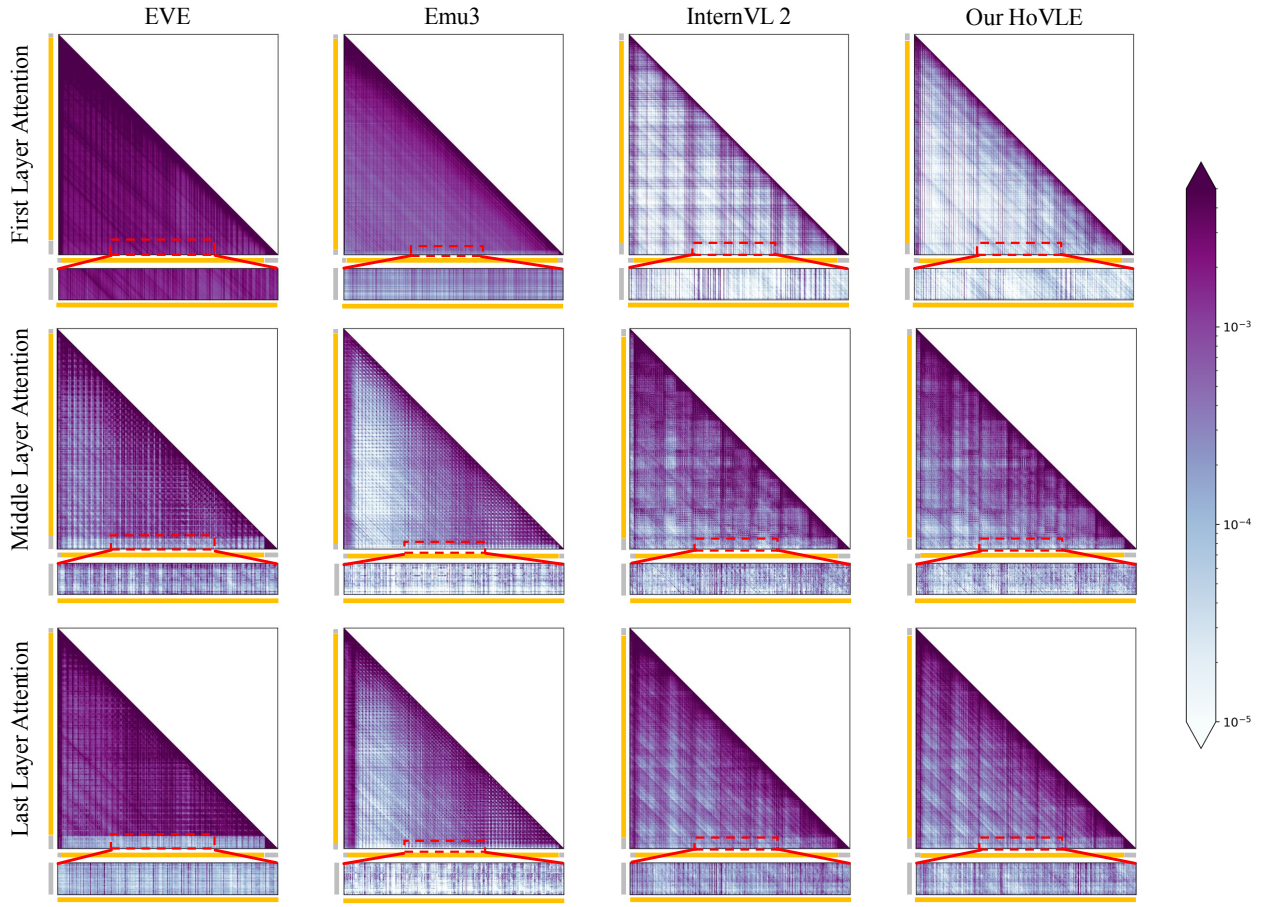

Figure 7. **Attention Maps for EVE, Emu3, InternVL2 and our HoVLE at the first, middle and last layers of LLM backbones**. Y-axis represents query tokens, and X-axis represents key tokens, with text modality tokens in gray and image modality tokens in yellow. All four models share the same input, but the sequence lengths of input tokens are different due to different image pre-processing. We highlight text-to-image attention below each full attention map. Our HoVLE, like the compositional InternVL2, has sparse attention across all network layers, while other monolithic models Emu3 and EVE have denser attention in shallow layers.

We hypothesize that the performance of these 8 benchmarks is bottle-necked by the LLM, not the holistic embedding. In contrast, the other benchmarks continues to benefit from additional distillation data, demonstrating the effectiveness of the distillation stage.

Speed-Performance Trade-off. We provide the detailed performance of HoVLE (HD) with different tile resolutions in Table 13 and Table 14. With tile resolution 336, HoVLE (HD) can already achieve comparable results with InternVL2. As the tile resolution increases, the performance of HoVLE (HD) steadily improves, especially on visual question answering benchmarks.

C. Attention Map Visualization

Figure 7 presents the visualization of attention map in the first, middle and last layers of EVE, Emu3, InternVL2 and our HoVLE. It's shown that the text-to-image attention of previous monolithic models, like EVE and Emu3, displays

dense pattern at the first layer, and gradually becomes sparse in deeper layers. On the contrary, compositional VLMs, like InternVL2, and our HoVLE possess sparse text-to-image attention throughout all LLM layers.

References

[1] Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*, 2019. 1

[2] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019. 1

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 1

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1

[5] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, 2022. 1

[6] Jimmy Carter. Textocr-gpt4v. https://huggingface.co/datasets/jimmycarter/textocr-gpt4v, 2024. 1

[7] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022. 1

[8] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024. 1

[9] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 1

[10] 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. 1

[11] Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019. 1

[12] Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1571–1576. IEEE, 2019. 1

[13] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*, 2017. 1

[14] Brian Davis, Bryan Morse, Scott Cohen, Brian Price, and Chris Tensmeyer. Deep visual template-free form parsing. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 134–141. IEEE, 2019. 1

[15] 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. 1

[16] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022. 1

[17] He Guo, Xiameng Qin, Jiaming Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Eaten: Entity-aware attention for single shot visual text extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 254–259. IEEE, 2019. 1

[18] Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. Icpr2018 contest on robust reading for multi-type web images. In *2018 24th international conference on pattern recognition (ICPR)*, pages 7–12. IEEE, 2018. 1

[19] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1

[20] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1

[21] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024. 1

[22] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images, 2023. 1

[23] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 1

[24] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 1

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5, 1

[26] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35: 3343–3360, 2022. 1

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1

[28] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 1

[29] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 1

[30] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022. 1

[31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1

[32] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. pages 235–251, 2016. 5, 1

[33] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007, 2017. 1

[34] Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. Visual information extraction in the wild: practical dataset and end-to-end solution. In *International Conference on Document Analysis and Recognition*, pages 36–53. Springer, 2023. 1

[35] LAION. Laion-gpt4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023. 1

[36] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1

[37] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022. 1

[38] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *arXiv preprint arXiv:2408.07246*, 2024. 1

[39] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. pages 22195–22206, 2024. 1

[40] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023. 1

[41] 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. 1

[42] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 1

[43] Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. Casia online and offline chinese handwriting databases. In *2011 international conference on document analysis and recognition*, pages 37–41. IEEE, 2011. 1

[44] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1

[45] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[46] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774*, 2023. 1

[47] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024. 1

[48] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*, 2021. 1

[49] 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. 35:2507–2521, 2022. 5, 1

[50] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin

Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 1

[51] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 1

[52] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191, 2017. 1

[53] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

[54] U-V Marti and Horst Bunke. The iam-database: an english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5: 39–46, 2002. 1

[55] 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 *ACL*, pages 2263–2279, 2022. 5, 1

[56] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 1

[57] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. pages 1697–1706, 2022. 5, 1

[58] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 1

[59] 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. 1

[60] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 1

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. 35:25278–25294, 2022. 5, 6, 1

[62] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en. https://laion.ai/blog/laion-coco/, 2022. 1

[63] Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1466–1476, 2015. 1

[64] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8876–8884, 2019. 1

[65] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1

[66] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[67] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. pages 8317–8326, 2019. 5, 1

[68] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8802–8812, 2021. 1

[69] Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1557–1562. IEEE, 2019. 1

[70] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 1

[71] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854, 2023. 1

[72] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 1

[73] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020. 1

[74] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024. 1

[75] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reason-

ing in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 1

[76] Chaoyi Wu. Pmc-casereport. https://huggingface.co/datasets/chaoyi-wu/PMC-CaseReport, 2023. 1

[77] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 1

[78] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023. 1

[79] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shi-Min Hu. A large chinese text dataset in the wild. *Journal of Computer Science and Technology*, 34(3): 509–521, 2019. 1

[80] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. *arXiv preprint arXiv:2203.01601*, 2022. 1

[81] Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1577–1581. IEEE, 2019. 1

[82] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023. 1

[83] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 1

[84] Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706, 2021. 1