BlueLM-V-3B: Algorithm and System Co-Design for Multimodal Large Language Models on Mobile Devices

Supplementary Material

Algorithm 1 Relaxed Aspect Ratio Matching

function RELAXED_ASPECT_RATIO_MATCHING(original_size: $(W_{\text{orig}}, H_{\text{orig}})$, possible_ratios: List of (m, n))
    Initialize: best_fit ← None, $R_{e,\text{max}} \leftarrow 0$, $R_{w,\text{min}} \leftarrow \infty$
    $(W_{\text{orig}}, H_{\text{orig}}) \leftarrow$ original_size
    for each (m, n) in possible_ratios **do**
        $(W, H) \leftarrow (384 \times m, 384 \times n)$
        scale $\leftarrow \min\left(\frac{W}{W_{\text{orig}}}, \frac{H}{H_{\text{orig}}}\right)$
        $\delta W \leftarrow \text{int}(W_{\text{orig}} \cdot \text{scale})$
        $\delta H \leftarrow \text{int}(H_{\text{orig}} \cdot \text{scale})$
        $R_e \leftarrow \min(\delta W \cdot \delta H, W_{\text{orig}} \cdot H_{\text{orig}})$
        $R_w \leftarrow W \cdot H - R_e$
        if $(R_e - R_{e,\text{max}}) > \alpha \cdot R_{e,\text{max}}$ **or** $((R_{e,\text{max}} - R_e) < \alpha \cdot R_{e,\text{max}}$ and $R_w < R_{w,\text{min}})$ **then**
            $R_{e,\text{max}} \leftarrow R_e$
            $R_{w,\text{min}} \leftarrow R_w$
            best_fit $\leftarrow (m, n)$
        end if
    end for
    return best_fit
end function

1. Relaxed Aspect Ratio Matching

To further expand the content of the main text, here we provide more about the relaxed aspect ratio matching method in this section.

Pseudocode: We present the pseudocode for our proposed relaxed aspect ratio matching method, as shown in Alg. 1. To be specific, we change the updating logic of LLaVA-NeXT by adding a parameter α such that when:

$$R_e - R_{e,\text{max}} > \alpha \cdot R_{e,\text{max}}, \tag{1}$$

or

$$(R_{e,\text{max}} - R_e) < \alpha \cdot R_{e,\text{max}} \text{ and } R_w < R_{w,\text{min}}, \tag{2}$$

we then update

$$R_{e,\text{max}} \leftarrow R_e, \quad R_{w,\text{min}} \leftarrow R_w, \tag{3}$$

and record the according aspect ratio. This increases the likelihood of selecting aspect ratios with smaller R_e but also smaller R_w.

Case Study: Here we present real cases where LLaVA-NeXT [45] and InternVL 1.5 [12] result in significant image enlargement, as illustrated in Fig. 1. In Fig. 1A (from COYO-300M Dataset [3]), LLaVA-NeXT selects a resolution of 384×768 for an image originally sized 380×393. Similarly, Fig. 1B (from CogVLM-SFT dataset [82]) shows InternVL 1.5 selecting a resolution of 1920×384 for an image initially sized 500×102. In contrast, our proposed relaxed aspect ratio matching selects 384×384 for the 380×393 image and 768×384 for the 500×102 image.

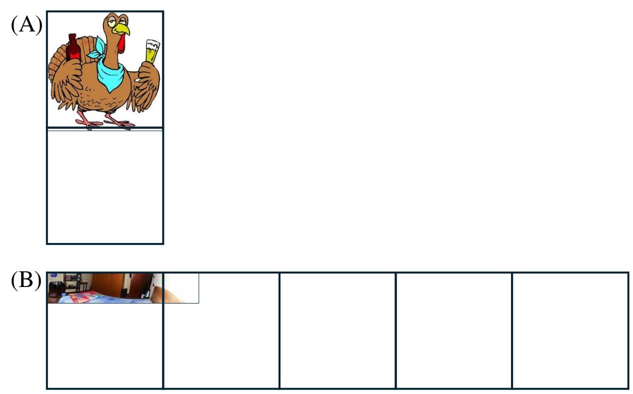


Figure 1. **Case study.** (A) LLaVA-NeXT chooses resolution 384×768 for an image with the original size of 380×393. (B) InternVL 1.5 chooses resolution 1920×384 for an image with the original size of 500×102.

2. Open-source Training Dataset

Here we provide the open-source dataset used to train BlueLM-V-3B in the fine-tuning stage in Tab. 1.

Task	Dataset
Text-only	ALLaVA [6], ScienceQA [50], Orca-Math [57], OpenOrca [39], MetaMathQA [90], WizardLM [84], MathInstruct [78]
Caption	TextCaps [69], Screen2Words [80], VizWiz [19], Laion [64], COCO [10], LLaVA [46], ALLaVA [6], SVIT [94], SA1B [30], VSR [41], Chart2Text [28], MultiMath [60], ArXivCap [36], COYO [3]
OCR	Wukong [16], HierText [47], TextOCR [72], WildReceipt [74], DocILE [70], SVRD [91], DocLayNet [61], XFUND [85], COCO-Text [79], SROIE [22], FUNSD [24], CORD [58], Paper2Fig100k [63], Docmatix [32], LAION-2B-OCR [40], SynthDoG [29], WebSight [33], DeepForm [75], Kleister [73], TabFact [9]
VQA	LVIS-Instruct4V [81], CLEVR [25], TallyQA [1], LNQA [62], Geo170K [67], ALLaVA [6], DocVQA [53], ChartQA [52], ArxivQA [36], GEOS [65], PMC-VQA [92], KVQA [66], Geometry3K [48], MapQA [5], PlotQA [55], ViQuAE [34], VQA-RAD [31], ST-VQA [2], TextVQA [71], LLaVAR [93], SIBR [87], MMC-Inst [43], IconQA [49], GQA [23], SciGraphQA [37], LRV-Instruction [42], DVQA [26], InfographicVQA [54], FigureQA [27], WikiTableQuestions [59], TAT-DQA [95], VisualMRC [76], ScienceQA [50], OCR-VQA [56], WebSRC [11], PathVQA [20], UniGeo [7], ScreenQA [21], VizWiz [18], SVIT [94], CogVLM [82], FM-IQA [14], VQAv2 [15], OK-VQA [51], EST-VQA [83], VisDial [13], Shikra [8], Super-CLEVR [38], LLaVA [44], IDK [4], AlfWorld [68], M-HalDetect [17], Cambrian7M [77], LLaVA-OneVision [35], mPLUG-DocOwl [88], UReader [89]

Table 1. **Training data.** This table presents the open-source datasets used in the fine-tuning stage, corresponding with the categories and data volume in Tab. 1 of the main text.

Please note that some datasets may belong to more than one category, and there may be overlapping data among these datasets.

3. Hyper-parameters for Training

We list the hyper-parameters for the pre-training stage (stage 1) and fine-tuning stage (stage 2) in Tab. 2 and Tab. 3 respectively.

Configuration	Stage 1
LLM Sequence Length	4096
Dynamic Resolution	None (384×384)
Optimizer	AdamW
Optimizer Hyperparams	$\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$
Peak LR	10^{-3}
LR Schedule	Cosine Decay
Weight Decay	0.05
Training Steps	3.434k
Warm-up Steps	34
Global Batch Size	720
Gradient Accumulation	1
Numerical Precision	`bfloat16`

Table 2. **Hyper-parameters.** Hyper-parameters for the pre-training stage (stage 1).

Configuration	Stage 2
LLM Sequence Length	4096
Dynamic Resolution	Up to 16 patches (1536×1536)
Optimizer	AdamW
Optimizer Hyperparams	$\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$
Peak LR	10^{-4}
LR Schedule	Cosine Decay
Weight Decay	0.05
ViT Layer-wise LR Decay	0.9
Training Steps	131k
Warm-up Steps	1310
Global Batch Size	5760
Gradient Accumulation	8
Numerical Precision	`bfloat16`

Table 3. **Hyper-parameters.** Hyper-parameters for the fine-tuning stage (stage 2).

Please note that due to the upsampling of certain datasets with smaller data volumes, the product of Training Steps and Global Batch Size may exceed the total data volume.

4. Model Accuracy after Quantization

INT4 LLM quantization will result in decreased accuracy on mobile phone NPUs. Both MediaTek and Qualcomm currently do not support group-wise quantization on NPUs, which is crucial for maintaining model accuracy after quantization on traditional GPUs and CPUs. In this case, we do not claim perfect accuracy after quantization in the paper. In our implementation, we do not specifically design the quantization algorithm. The PTQ algorithm for the MediaTek chip is based on per-channel quantization (utilizing MediaTek's NeuroPilot), similar to Qualcomm's QNN SDK. We evaluate the accuracy across 4 tasks after performing quantized inference on the NPU of the MediaTek 9300 chip, achieving an acceptable average accuracy retention at 88.7%. For reference, please refer to PowerInfer-2 [86].

	OCRBench	DocVQA	MMVet	ScienceQA	AVG
BF16 (A100)	829	86.6	61.8	94.0	81.3
MediaTek 9300	746	75.4	52.2	86.3	72.1
Retained (%)	90.0	87.1	84.5	91.8	88.7

Table 4. **Model accuracy after quantization.** We evaluate the accuracy across 4 tasks after performing quantized inference on the NPU of the MediaTek 9300 chip, achieving an acceptable average accuracy retention at 88.7%

The accuracy retention of LLMs on Qualcomm chips using QNN is also around 80%-90%. We can adopt QAT, Q-LoRA, or advanced PTQ methods to improve accuracy in the future.

5. Energy Consumption

The energy consumption of BlueLM-V-3B is tested on MediaTek 9300 using the balanced performance mode to align with real-world scenarios. We achieve a token generation speed of 19.4 token/s, which is sufficient for daily use. We use the metric of peak power and average Joules per token as in PowerInfer-2. For reference, we also provide Table 8 of PowerInfer-2 (the shaded entries), which deploys LLMs on Qualcomm chips. We can achieve lower energy consumption in two aspects: one is our smaller model size (3B), and the other is the MediaTek 9300's inherently better energy efficiency performance compared to chips like the Qualcomm 8Gen3.

Framework	Ours	PowerInfer-2	QNN	llama.cpp
Peak Power (W)	4.584	5.095	5.133	4.065
Energy (J/token)	0.215	0.257	0.373	0.672

Table 5. **Energy consumption.** Energy consumption of BlueLM-V-3B on MediaTek 9300 with the balanced performance mode.

6. Other Deployment Platform

We provide the peak mode deployment statistics on the NPU of the Qualcomm Snapdragon 8Gen3 SoC using QNN SDK. We now support the 1-patch ViT inference. Deploying on the Qualcomm 8Gen3 SoC achieves a slightly faster speed than MediaTek 9300. We will provide complete support for Qualcomm chips in future works.

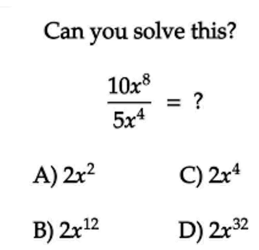
SoC	ViT (s)	Input (token/s)	Output (token/s)
MediaTek 9300	0.30	758	24.4
Qualcomm 8Gen3	0.24	800	27.2

Table 6. **Deployment efficiency.** Deployment statistics on the Qualcomm 8Gen3 SoC.

7. Visualization on Practical Cases

We present some demos of BlueLM-V-3B on mobile devices.

Math Ability



Human: Solve this problem.

Response: To solve this problem, we need to simplify the given expression and compare it to the options provided.

The expression is:

\[\frac{10x^8}{5x^4} = ? \]

First, we simplify the expression by dividing the coefficients and subtracting the exponents of the same base:

\[\frac{10}{5} \cdot \frac{x^8}{x^4} = 2 \cdot x^{8-4} = 2 \cdot x^4 \]

So the simplified expression is:

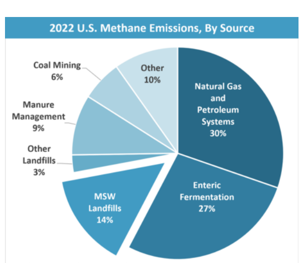
\[2x^4 \]

Now, we compare this to the options:

A) \(2x^2 \)
B) \(2x^{12} \)
C) \(2x^4 \)
D) \(2x^{32} \)

The correct answer is C) \(2x^4 \), as it matches the simplified expression.

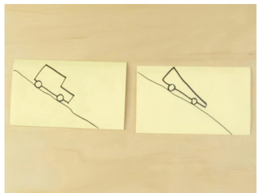
Chart Understanding



Human: What is the percentage difference between Natural Gas and Petroleum Systems and MSW Landfills?

Response: The pie chart shows that Natural Gas and Petroleum Systems make up 30% of methane emissions, while MSW Landfills account for 14%. The percentage difference between Natural Gas and Petroleum Systems and MSW Landfills is 30% - 14% = 16%.

Scientific Understanding



Human: Which car in the picture is more aerodynamic?

Response: The car on the right appears more aerodynamic due to its streamlined shape with a pointed front end and a sloped rear end, which reduces air resistance and could potentially allow it to move more efficiently through the air.

Document Understanding

Human: What is the image about?

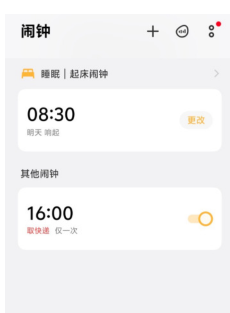
Response: The image depicts "The Nobel Prize in Chemistry" for the year 2024, featuring three scientists: David Baker, Demis Hassabis, and John M. Jumper. Each is recognized for their contributions to computational protein design or structure prediction in science. The poster highlights their significant achievements in the field of chemistry and celebrates their awards.

Multilingual Ability

Human: What are all the scene text in the image, and what do they mean?

Response: The scene text in the image is "Welcome bienvenida Bienvenue Willkommen benvenuto". It is a multilingual sign that says "Welcome" in English, Spanish, French, German and Italian respectively.

Screen Understanding



Human: 我有几个闹钟?分别提醒我做什么?

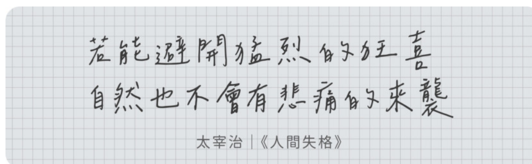
Response: 有两个闹钟。一个是明天早上8点30分起床的闹铃，另一个是今天下午16点00分的快递取件。

Image Description

Human: 详细描述照片中的内容。

Response: 这张图片展示了一个卡通风格的男孩雕塑。他闭着眼睛，嘴巴张开，似乎在笑或者打哈欠。他的头发是黑色的，有点凌乱。他穿着一件红色的上衣，上面有黄色的纽扣和装饰性的图案，还有一条蓝色的裤子，裤脚处有白色的毛边装饰。男孩的脚上穿着棕色的鞋子，鞋带也是黑色的。这个雕塑放置在一个圆形的底座上，底座看起来像是用石头或类似材料制成，颜色为浅灰色。背景是模糊的，没有太多细节，主要是白色和灰色的色调，这使得雕塑更加突出。整体来看，这个雕塑给人一种轻松愉快的感觉。

Chinese OCR



Human: 告诉我图里面写了什么。

Response: 若能避開猛烈的狂喜 自然也不會有悲痛的來襲 太宰治|《人間失格》

Common Sense

Human: 我能吃这些蘑菇么?

Response: 不，这些红盖白点的蘑菇可能有毒，不建议食用。

Knowledge

Human: Introduce me the author of this painting.

Response: This painting is by Claude Monet, a renowned French painter who was a founding member of the Impressionist movement. Monet is celebrated for his innovative use of light and color to capture the fleeting moments of life and landscapes.

Multi-turn Conversation

Human: What are the departure point and destination of this train, and where is the next stop?

Response: The departure point is Praha hl.n. and the destination is Wien Hbf. The next stop is Brno-Židenice.

Human: What is the train speed now?

Response: The train speed is 160 km/h.

References

[1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8076–8084, 2019. 2

[2] 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. 2

[3] 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, 2

[4] Sungguk Cha, Jusung Lee, Younghyun Lee, and Cheoljong Yang. Visually dehallucinative instruction generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5510–5514. IEEE, 2024. 2

[5] 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. 2

[6] 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. 2

[7] 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. 2

[8] 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. 2

[9] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. 2019. 2

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[11] Xingyu Chen, Zihan Zhao, Lu Chen, Danyang Zhang, Jiabao Ji, Ao Luo, Yuxuan Xiong, and Kai Yu. Websrc: A dataset for web-based structural reading comprehension. *arXiv preprint arXiv:2101.09465*, 2021. 2

[12] 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. 1

[13] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2

[14] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. 2

[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. 2

[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. *NeurIPS*, 35: 26418–26431, 2022. 2

[17] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 2

[18] 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 *CVPR*, pages 3608–3617, 2018. 2

[19] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer, 2020. 2

[20] 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. 2

[21] Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Victor Carbune, Jason Lin, Maria Wang, Srinivas Sunkara, Yun Zhu, and Jindong Chen. Screenqa: Large-scale question-answer pairs over mobile app screenshots. *arXiv preprint arXiv:2209.08199*, 2022. 2

[22] 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. 2

[23] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 2

[24] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 1–6. IEEE, 2019. 2

[25] 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 *CVPR*, pages 2901–2910, 2017. 2

[26] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656, 2018. 2

[27] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Fig-

ureQA: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 2

[28] 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. 2

[29] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[31] 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. 2

[32] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions., 2024. 2

[33] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024. 2

[34] 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. 2

[35] 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. 2

[36] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 2

[37] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023. 2

[38] 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. 2

[39] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". OpenOrca: An open dataset of GPT augmented FLAN reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`, 2023. 2

[40] Yiqi Lin, Conghui He, Alex Jinpeng Wang, Bin Wang, Weijia Li, and Mike Zheng Shou. Parrot captions teach clip to spot text. In *European Conference on Computer Vision*, pages 368–385. Springer, 2025. 2

[41] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 2

[42] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[43] 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. 2

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[45] 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. 1

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 2

[47] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2022. 2

[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. 2

[49] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 2

[50] 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. 2

[51] 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. 2

[52] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2

[53] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for VQA on document images. In *WACV*, pages 2200–2209, 2021. 2

[54] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. InfographicVQA. In *WACV*, pages 1697–1706, 2022. 2

[55] 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. 2

[56] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual question answering by reading text in images. In *ICDAR*, pages 947–952, 2019. 2

[57] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024. 2

[58] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019. 2

[59] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015. 2

[60] Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*, 2024. 2

[61] B Pfitzmann, C Auer, M Dolfi, AS Nassar, and PWJ Staar. Doclaynet: A large humanannotated dataset for document-layout analysis (2022). *URL: https://arxiv. org/abs/2206*, 1062. 2

[62] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020. 2

[63] Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3689–3698, 2023. 2

[64] 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. *NeurIPS*, 35:25278–25294, 2022. 2

[65] 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. 2

[66] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. KVQA: Knowledge-aware visual question answering. In *AAAI*, pages 8876–8884, 2019. 2

[67] Aditya Sharma, Aman Dalmia, Mehran Kazemi, Amal Zouaq, and Christopher J Pal. Geocoder: Solving geometry problems by generating modular code through vision-language models. *arXiv preprint arXiv:2410.13510*, 2024. 2

[68] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[69] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. 2

[70] Štěpán Šimsa, Milan Šulc, Michal Uřičář, Yash Patel, Ahmed Hamdi, Matěj Kocián, Matyáš Skalický, Jiří Matas, Antoine Doucet, Mickaël Coustaty, et al. Docile benchmark for document information localization and extraction. In *International Conference on Document Analysis and Recognition*, pages 147–166. Springer, 2023. 2

[71] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, pages 8317–8326, 2019. 2

[72] 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. 2

[73] Tomasz Stanisławek, Filip Graliński, Anna Wróblewska, Dawid Lipiński, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemysław Biecek. Kleister: Key information extraction datasets involving long documents with complex layouts. In *ICDAR*, pages 564–579. Springer, 2021. 2

[74] Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*, 2021. 2

[75] S Svetlichnaya. Deepform: Understand structured documents at scale. 2020. 2

[76] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. VisualMRC: Machine reading comprehension on document images. In *AAAI*, pages 13878–13888, 2021. 2

[77] 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. 2

[78] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*, 2024. 2

[79] 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. 2

[80] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021. 2

[81] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023. 2

[82] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023. 1, 2

[83] 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. 2

[84] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023. 2

[85] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. Xfund: a benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, 2022. 2

[86] Zhenliang Xue, Yixin Song, Zeyu Mi, Le Chen, Yubin Xia, and Haibo Chen. Powerinfer-2: Fast large language model inference on a smartphone. *arXiv preprint arXiv:2406.06282*, 2024. 3

[87] Zhibo Yang, Rujiao Long, Pengfei Wang, Sibo Song, Humen Zhong, Wenqing Cheng, Xiang Bai, and Cong Yao. Modeling entities as semantic points for visual information extraction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15358–15367, 2023. 2

[88] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*, 2023. 2

[89] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023. 2

[90] 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. 2

[91] Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *International Conference on Document Analysis and Recognition*, pages 536–552. Springer, 2023. 2

[92] 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. 2

[93] 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. 2

[94] Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 2

[95] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022. 2