Token-Efficient VLM: High-Resolution Image Understanding via Dynamic Region Proposal

Supplementary Material

6. Ablation Studies

Importance of Dynamic Region Proposal in High-Resolution Bench. To highlight the significance of visual grounding and focused attention in high-resolution tasks, we compare TEVA with VILA-Qwen2, which refers to VILA [32] configured identically to TEVA in both training and testing, including the use of the same language model (Qwen2-7B [65]), MLP adaptor, and vision tower model (SigLIP [71]).

Different from VILA-Qwen2, TEVA leverages its Relevant Area Proposal (RAP) module and informative patch sampling strategy to selectively focus on critical regions of high-resolution images. As shown in Table 6, this concentrated attention results in significant performance improvements on tasks requiring precise spatial reasoning and fine-grained image understanding.

For example, TEVA outperforms VILA-Qwen2 by 28.7 points on the V^* attribute recognition (AR) benchmark [61], which focuses on specific small objects attribute. Additionally, TEVA achieves a 6.1 point improvement on the Remote Sensing task of the MME-RealWorld Perception dataset [79], which involves detailed analysis of related objects such as traffic signs, vehicles, and other high-resolution scene elements. These results demonstrate TEVA's superior capability in handling tasks that demand precise focus and visual grounding in high-resolution settings.

More evaluation on General Resolution VLM Tasks. We evaluate TEVA on 13 widely used general-resolution VLM benchmarks, including GQA [17], SQA [43], VQA^{T} (TextVQA) [54], POPE [29], SEED^{I} (SEED Image Bench) [26], LLaVA^{w} [37], MMMU^{t} (MMMU test split) [69], MMMU^{v} (MMMU validation split) [69], VQA^{v2} (VQA-V2) [14], MME [12], MMB^{EN} (MMbench English) [41], MMB^{CN} (MMbench Chinese) [41], and AI2D [22].

As summarized in Table 7, TEVA surpasses the baseline model, VILA-Qwen2, on 10 out of the 13 benchmarks, highlighting its robust generalization across a variety of general-resolution vision-language tasks. Notably, TEVA achieves a 13 point improvement on the multi-modality benchmark $\text{MMbench}^{\text{CN}}$ [41] and an impressive 88 point gain on the MME benchmark [12], demonstrating its strong performance on challenging multi-modal datasets.

7. More Qualitative Samples

Visualization Results at Different Resolutions. To highlight the importance of maintaining high image resolution for Vision-Language Models (VLMs) to accurately answer questions about fine-grained details in high-resolution images, we conducted experiments using QA tasks at varying image resolutions. As illustrated in Figure 6, reducing the image resolution significantly impacts the model's ability to discern critical details. For example, in Figure 6 (Left), when resizing an image from 2247 $\times$ 2500 pixels to 1024 $\times$ 1024 pixels or 672 $\times$ 672 pixels, the texture details of a distant car become blurred. This blurriness leads the VLM to either provide incorrect answers or fail to answer due to the loss of essential visual information.

This experiment underscores the necessity of preserving the original high-resolution during inference to ensure accurate visual understanding, which is a key motivation for developing TEVA.

Real-world Downstream Tasks. To further assess TEVA's performance in real-world downstream tasks that demand high-resolution details, we conducted experiments in industrial scenarios, autonomous driving, and remote sensing. As shown in Figure 7, TEVA's ability to focus on fine details allows it to effectively identify and interpret critical information in complex scenes. For example, in a remote sensing scenario (Figure 7 Right), TEVA can accurately focus on a specific vehicle and provide precise details about its status, demonstrating its capability to handle tasks requiring attention to intricate visual elements.

Handling Dense Information in Images. As discussed in Section 4.4, TEVA faces limitations when handling images densely packed with information across the entire frame. In such cases, the bounding box typically encompasses the entire image, making the usual region-focused strategy less effective. To address this, we implement a fallback mechanism that switches to uniform patch sampling for the entire image. For instance, as shown in Figure 8, when querying information from a text-dense document, the anchor prompt is set to "Text". Since the image predominantly contains text, dividing the limited patch budget between global and local patches would not be advantageous. Instead, reverting to uniform patch sampling ensures that the VLM maintains its performance by adequately capturing all relevant details.

As shown in Table 5, TEVA without RAP and dynamic patch sampling (fallback for every image) is comparable to baseline VILA and higher with RAP and dynamic patch sampling. We want to emphasize that TEVA is designed to prioritize general high-resolution image understanding tasks that require focusing on identifying informative candidates within the image, rather than processing text-intensive tasks that require global information across the entire image without a specific candidate focus. This design is especially valuable in applications like autonomous driving and remote sensing. We recognize the importance of text-intensive OCR tasks and plan to address them in future work while continuing to develop TEVA's capabilities for general high-resolution tasks.

Result on Image Captioning with Subject-oriented Guidance. Leveraging its ability to focus on fine details in high-resolution images, TEVA can generate captions emphasizing specific areas of an image with remarkable precision. As demonstrated in Figure 10, when generating a caption that prioritizes a person in the image, TEVA accurately identifies the individual as the primary subject and provides detailed, contextually correct information. In contrast, even the current state-of-the-art VLM, e.g., ChatGPT-4o [50], often hallucinates details and generates incorrect information. This highlights TEVA's capability to produce subject-oriented, detail-specific captions, a feature that has the potential for broader applications in fields requiring precise visual understanding.

Results on Reasoning case. TEVA excels in complex reasoning, as shown in Figure 9, accurately under stands and captures spatial relationships between regions (extracted with anchors "girl" and "woman"). Unlike TEVA w/o RAP, which hallucinates, TEVA w/ RAP succeeds, highlighting RAP's role, combined with dynamic patch sampling, in processing detailed visual information.

8. Details of Subject-Oriented Data Curation

Dataset Curation Pipeline. As illustrated in Figure 11, we employ structured prompts to generate multiple subject-oriented QA pairs from detailed image captions and record the "Asked Object" (e.g., butterfly) for each question. This approach effectively produces a large-scale dataset of QA pairs. However, the generated pairs may contain noise, such as nonexistent objects caused by misdescriptions in the captions or inaccuracies introduced by the QA generation process of Llama3.1 [59].

To enhance dataset quality, we use the "Asked Object" as an anchor prompt to validate its presence in the image. Specifically, we extract the bounding box of the corresponding object using an open-vocabulary detection method, Grounding DINO [39], in our setup. If a bounding box is detected with high confidence for the "Asked Object", we retain the QA pair. Otherwise, we discard it, as the absence of a bounding box suggests that the "Asked Object" might not exist in the image.

This pipeline ensures both efficiency, by leveraging text captions for initial generation, and quality, through a robust data cleansing process. Additionally, we provide more examples of the Subject-Oriented Dataset in Figure 12.

Method	ChartQA	DocVQA	InfoQA
VILA-Qwen2	60.7	50.6	31.6
TEVA w/o RAP	62.3	50.0	31.4
TEVA w/ RAP	**63.8**	**51.3**	**31.9**

Table 5. Comparison between TEVA, TEVA without RAP and the baseline VILA-Qwen2 on Text-Dense Dataset ChartQA [45], DocVQA [46] and InfoQA [47].

9. Implementation Details

In Table 8, we provide the details of the training hyperparameters used across the four stages of training TEVA. All training processes were conducted using 64 NVIDIA A100 GPUs. For the Supervised Fine-Tuning (SFT) stage, we build on VILA [32] SFT blend. Specific training dataset details are outlined in Table 9.

Figure 6. Visualization of inference at different resolutions.

Method	LLM	V^*		MME-RealWorld^{PRC}				MME-RealWorld^{RSN}		
		AR	SRR	OCR	RS	MO	AD	OCR	MO	AD
VILA-Qwen2	Qwen2-7B	47.8	59.2	54.6	32.1	34.8	31.9	42.2	28.5	33.2
TEVA-3B	Sheared-llama-2.7B	60.0	63.2	43.1	32.3	32.6	33.0	41.8	33.5	29.9
TEVA-7B	Qwen2-7B	**76.5**	**77.6**	**60.7**	**38.2**	**36.9**	**35.2**	**49.4**	**34.7**	**35.5**

Table 6. Comparison between TEVA and VILA-Qwen2 based on the same training configuration on high-resolution benchmarks.

Method	GQA	SQA	VQA^{T}	POPE	SEED^{I}	LLaVA^{w}	MMMU^{t}	MMMU^{v}	VQA^{v2}	MME	MMB^{EN}	MMB^{CN}	AI2D
VILA-Qwen2	**64.3**	**94.3**	64.1	87.7	**74.4**	91.2	40.7	44.3	82.7	1529.1	72.4	66.8	75.3
TEVA-7B	64.1	92.5	**72.5**	**87.9**	74.1	**94.3**	**41.9**	**48.4**	**82.8**	**1617.1**	**80.7**	**79.8**	**76.3**

Table 7. Comparison between TEVA and the baseline VILA-Qwen2 under identical training configurations across 13 general-resolution VLM benchmarks, including GQA [17], SQA [43], VQA^{T} (TextVQA) [54], POPE [29], SEED^{I} (SEED Image Bench) [26], LLaVA^{w} [37], MMMU^{t} (MMMU test split) [69], MMMU^{v} (MMMU validation split) [69], VQA^{v2} (VQA-V2) [14], MME [12], MMB^{EN} (MMbench English) [41], MMB (MMbench Chinese)^{CN} [41] and AI2D [22].

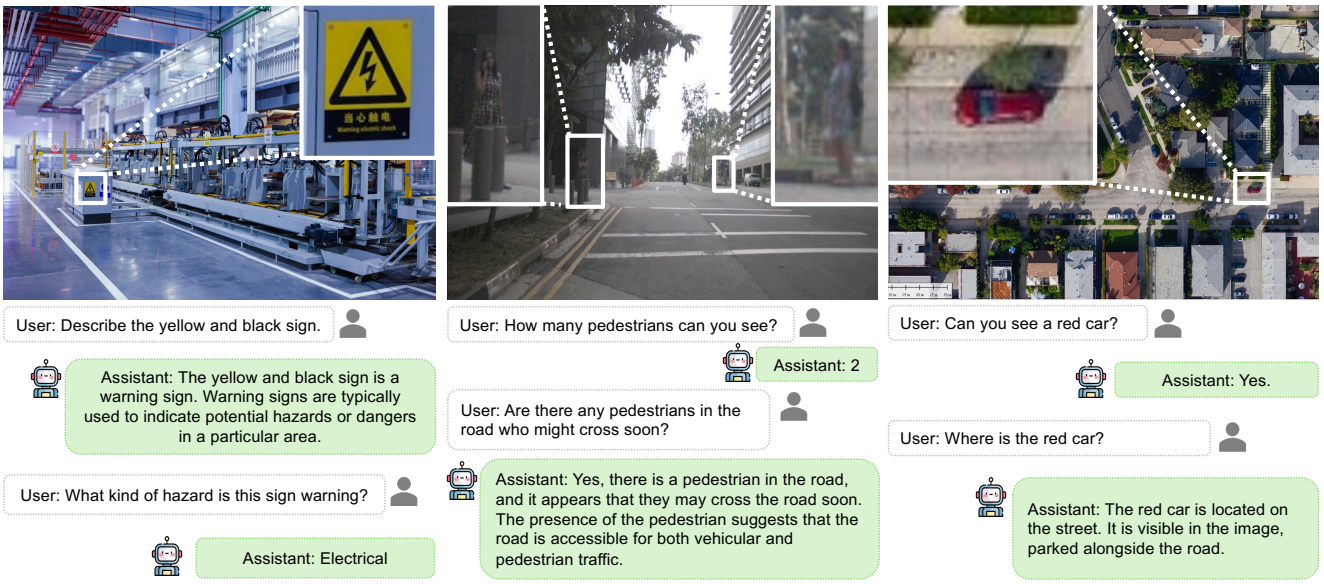

Figure 7. Visualization of Real-world downstream tasks. Left: Industrial Scenario. Middle: Autonomous Driving. Right: Remote Sensing.

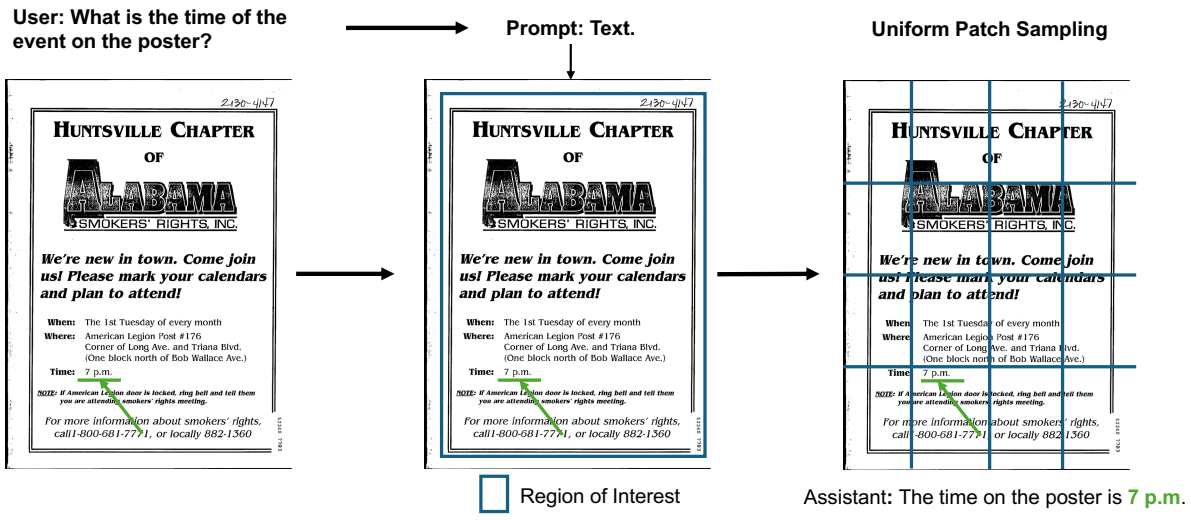

Figure 8. For images containing dense information, where the focused region encompasses a significant portion of the entire image (greater than a threshold ratio), the system falls back to a uniform patch sampling strategy. For example, when answering a query about the time based on a poster with dense text information, the selective area corresponding to the anchor prompt "Text" covers nearly the entire image. In such cases, the uniform patch sampling strategy is applied to ensure comprehensive coverage of the image and to accurately capture all relevant details.

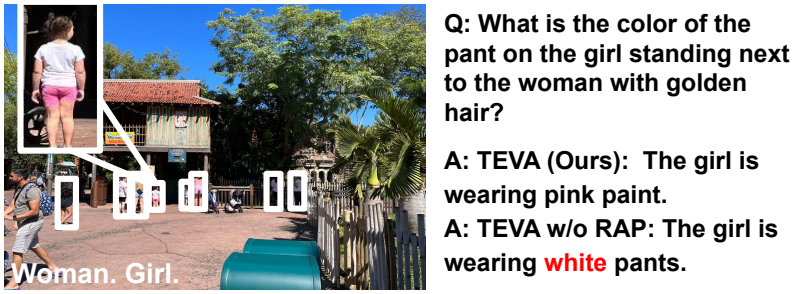

Figure 9. Visualization of Reasoning Task.

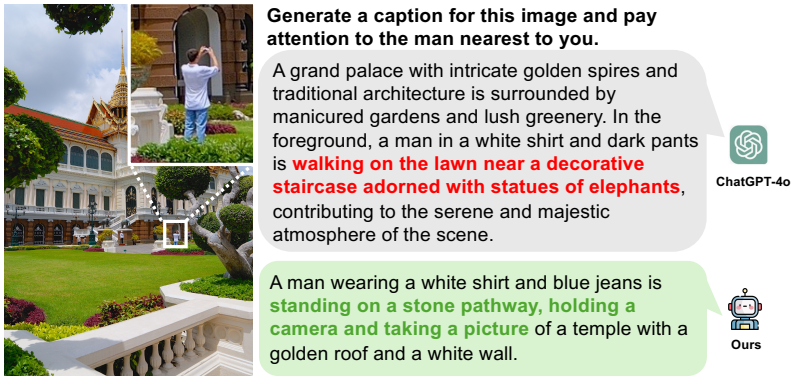

Figure 10. Visualization of image caption generation with subject-oriented guidance

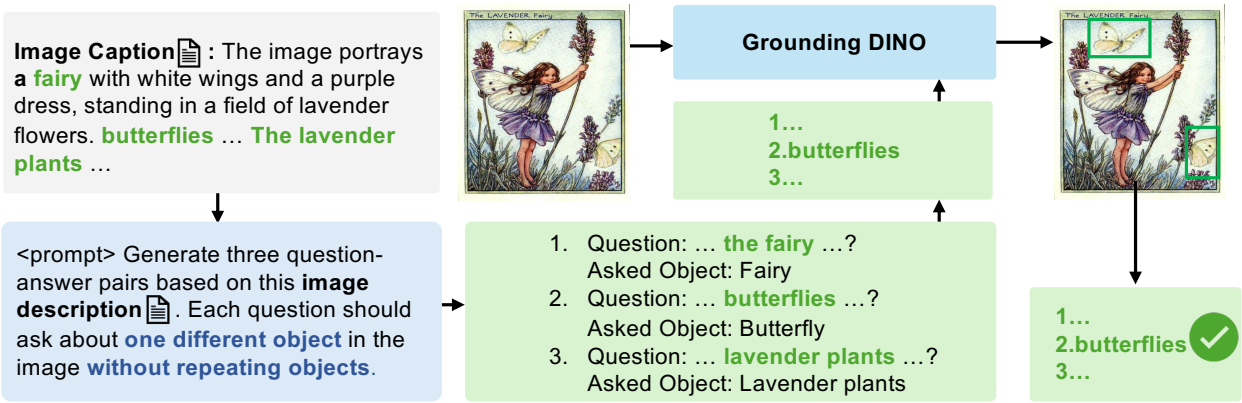

Figure 11. Pipeline of Subject-Oriented Data Curation. We first utilize structured prompts to create multiple subject-oriented QA pairs based on detailed image captions. Each QA pair records the "Asked Object" (e.g., butterfly) for validation. To ensure high-quality data, an open-vocabulary object detection model, Grounding DINO [39], is employed to verify the presence of the "Asked Object" in the image. Objects are retained only if a high-confidence bounding box is detected, discarding noisy pairs that involve nonexistent objects.

Image Caption: The image is a vibrant advertisement for PETA's 40th anniversary online auction. Dominating the center of the image is a sleek, silver car, its polished surface reflecting the surrounding elements. To the left of the car, a **woman** is seen cradling a dog in her arms, their bond evident in their close proximity. On the right side of the **car**, a pair of roller skates and a coffee cup are placed, suggesting a leisurely day out. The background is a rich blue, providing a stark contrast to the objects in the foreground. The text "Check Out PETA's 40th Anniversary Online Auction!" is prominently displayed, inviting viewers to participate in the event. The **logos** of Dunkin' Donuts and Baskin Robbins are also present, indicating their involvement or sponsorship of the event. The objects, people, and text are arranged in a balanced manner, creating a harmonious composition that effectively conveys the message of the advertisement.

1. Question: What is the **woman** in the image doing?
   Answer: The woman is cradling a dog in her arms.
2. Question: What is the color of the **car** in the image?
   Answer: The car is silver.
3. Question: What **logos** are visible in the image?
   Answer: The logos of Dunkin' Donuts and Baskin Robbins are present.

Image Caption: The image is an advertisement for the Xbox game "Conker: Live & Reloaded". The main focus of the image is a cartoon **squirrel**, who is dressed in a military uniform and is holding a gun. The squirrel is standing in front of a group of other cartoon animals, who are also armed and appear to be ready for battle. The background of the image is a dark, cloudy sky, which adds to the intense atmosphere of the scene. **The text** "Conker: Live & Reloaded" is prominently displayed in the center of the image, indicating the name of the game. The **overall** color scheme of the **image** is dark and muted, with the exception of the bright orange text, which stands out against the darker background. The relative positions of the objects suggest a sense of depth and perspective, with the squirrel in the foreground and the other animals in the background. The image does not contain any other discernible text or objects."

1. Question: What is written in the center of the image? (**text**)
   Answer: "Conker: Live & Reloaded."
2. Question: What is the **cartoon squirrel** holding in the image?
   Answer: A gun.
3. Question: How would you describe the color scheme of the **image**?
   Answer: The overall color scheme is dark and muted, with the exception of the bright orange text.

Figure 12. Example QA pairs of the Subject-Oriented Data Samples.

Table 8. Training hyperparameters for TEVA.

Configuration	Alignment Stage	Vision Tower Adapting	Integrated Interpretation	Instruction Fine-tuning
Unfreeze ViT	×	✓	✓	✓
Unfreeze LLM	×	×	✓	✓
Unfreeze VL Adapter	✓	✓	✓	✓
Image resolution	384	1536	1536	1536
Optimizer			AdamW	
Global learning rate	$1e^{-3}$	$1e^{-4}$	$5e^{-5}$	$5e^{-5}$
ViT learning rate	$1e^{-3}$	$1e^{-3}$	$1e^{-4}$	$5e^{-5}$
Warming up ratio	0.03	0.03	0.03	0.03
Learning rate schedule	cosine decay	cosine decay	cosine decay	cosine decay
Global batch size	256	1024	1024	1024
Numerical precision	bfloat16	bfloat16	bfloat16	bfloat16

Table 9. Training Dataset for TEVA Across Stages 1 to 4. Note that we have one more stage "Vision Tower Adaptation" than the common VLMs [27, 32] used to adapt the vision tower for flexibly ordered tokens.

Stage	**Dataset**	**Samples**
Alignment Stage	LLaVA-Pretrain [35]	595k
Vision Tower Adapting	Proposed Subject-Oriented Dataset	4M
Integrated Interpretation	Proposed Subject-Oriented Dataset ShareGPT4V Caption Dataset	4M 1M
Instruction Finetuning	**Image Caption**: TextCaps [53], ShareGPT4V-100K [4], LLaVAR [78], Image Paragraph Captioning [24] **VQA**: ScienceQA-train [43], LLaVA-SFT [35], SHERLOCK [16], GQA-train [17], Geo170K [13], VQAv2-train [14], DocVQA [21], OCRVQA [48], OKVQA [44], ViQuAE [25], CLEVR [20], SynthDoG-en [23], WIT(Subset) [55], TextVQA-train [54], MMC-Instruction [34], AI2D [22], ChartQA [45], LLaVA-OneVision(Subset) [27] **Text-only**: FLAN-1M [60], MathInstruct [68]	4.17M

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. 2022. 1

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 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. 4, 5

[4] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. 5, 7

[5] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1

[6] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2025. 2, 5, 6, 7

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 7

[8] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*, 2024. 2

[9] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *arXiv preprint arXiv:2404.06512*, 2024. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. $vila^2$: Vila augmented vila. *arXiv preprint arXiv:2407.17453*, 2024. 5

[12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 1, 3

[13] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 7

[14] 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. 5, 7, 1, 3

[15] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13806, 2024. 2, 4

[16] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022. 7

[17] 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. 1, 3, 7

[18] IDEFICS. Introducing idefics: An open reproduction of state-of-the-art visual language model. 2023. 7

[19] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015. 4

[20] 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. 7

[21] 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. 7

[22] 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. 5, 7, 1, 3

[23] 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*, pages 498–517. Springer, 2022. 7

[24] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017. 7

[25] 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. 7

[26] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 3

[27] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 7

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 5, 7, 8, 1, 3

[30] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 2, 7

[31] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024. 2, 6, 7, 8

[32] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 6, 7, 8, 1, 2

[33] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 1, 2

[34] 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. 7

[35] 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, 2, 7

[36] 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, 7

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NIPS*, 2024. 1, 5, 7, 3

[38] Haogeng Liu, Quanzeng You, Xiaotian Han, Yiqi Wang, Bohan Zhai, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Infimm-hd: A leap forward in high-resolution multimodal understanding. *arXiv preprint arXiv:2403.01487*, 2024. 1, 2

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 2, 5

[40] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 7

[41] 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? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 5, 7, 1, 3

[42] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024. 1

[43] 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. 5, 7, 1, 3

[44] 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. 7

[45] 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, 7

[46] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 2

[47] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 2

[48] 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. 7

[49] Ravina Mithe, Supriya Indalkar, and Nilam Divekar. Optical character recognition. *International journal of recent technology and engineering (IJRTE)*, 2(1):72–75, 2013. 1

[50] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023. 1, 6, 7, 8, 2

[51] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4

[52] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? In *European Conference on Computer Vision*, pages 444–462. Springer, 2025. 1, 2, 7

[53] 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, 2020. 7

[54] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 5, 7, 8, 1, 3

[55] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 7

[56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 7

[57] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024. 7

[58] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 3, 5, 2

[60] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 7

[61] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 2, 4, 5, 6, 7, 8, 1

[62] Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023. 6

[63] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022. 2

[64] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 2

[65] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[66] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2, 4

[67] Xiaoyu Yue, Shuyang Sun, Zhanghui Kuang, Meng Wei, Philip HS Torr, Wayne Zhang, and Dahua Lin. Vision transformer with progressive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 387–396, 2021. 2

[68] Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023. 7

[69] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024. 5, 7, 1, 3

[70] Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, and Manling Li. Halle-switch: Controlling object hallucination in large vision language models. *arXiv e-prints*, 2023. 2

[71] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2, 1

[72] Yufei Zhan, Hongyin Zhao, Yousong Zhu, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon-g: Bridging vision-language and vision-centric tasks via large multimodal models. *arXiv preprint arXiv:2410.16163*, 2024. 2, 4, 5, 6, 7

[73] Yufei Zhan, Yousong Zhu, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Griffon v2: Advancing multimodal

perception with high-resolution scaling and visual-language co-referring. *arXiv preprint arXiv:2403.09333*, 2024. 1, 2, 4

[74] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 2

[75] Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *Proceedings of International Conference on Learning Representations (ICLR)*, 2025. 7

[76] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 2

[77] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[78] 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, 7

[79] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024. 5, 6, 7, 8, 1

[80] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1