Echoes of Ownership: Adversarial-Guided Dual Injection for Copyright Protection in MLLMs

Supplementary Material

This supplementary material includes details of original MLLMs, fine-tuning and inference settings, and datasets used in fine-tuning. Furthermore, we provide the inference parameter analysis and utility performance of both original models and fine-tuned models. The contents are organized as follows:

A. Implementation Details

A.1. Details of original models

We choose LLaVA-1.5 [7] and Qwen2-VL [15] as the original MLLMs to obtain derivative MLLMs, and then construct triggers on the original MLLMs.

LLaVA-1.5. For LLaVA-1.5, we choose LLaVA-1.5-7B, a MLLM of end-to-end training, which consist of a frozen vision encoder CLIP ViT-14L [10], a visual language connector with two linear layers, and a large language model decoder LLaMA-2 with a total of 32 layers, and the 4096 hidden dimensions.

Qwen2-VL. For Qwen2-VL, we choose Qwen2-VL-2B-Instruct, which enables the model to dynamically process images of varying resolutions and also integrates multimodal rotary position embedding, facilitating the effective fusion of positional information. It is an end-to-end unified transformer architecture that integrates vision encoders and language models, and consists of a vision encoder with a resolution of 224×224, and a language model Qwen2 [14].

A.2. Details of downstream fine-tuning datasets

In the experiments, we chose five downstream task fine-tuning datasets to simulate various real-world scenarios. We provide a detailed description of those datasets in the following. Moreover, all train datasets were standardized into the ShareGPT format, specifically designed to simulate natural conversational flows.

Table 8. Fine-tuning setting in the experiments.

Hyperparameter	LoRA fine-tuning setting	Full fine-tuning setting
Optimizer	AdamW	AdamW
Learning rate	2e-4	1e-5
Batch size	8	4
LoRA rank	16	/
LoRA alpha	32	/
Training epochs	3	3
Gradient accumulation	1	2
Dtype	bfloat16	bfloat16
Lr scheduler	cosine	cosine
Warm-up epoch ratio	0.03	0.01

V7W. Visual7W (V7W) [19] is a dataset designed for comprehensive image content understanding, specifically tailored for VQA tasks. This dataset extends beyond raw images by incorporating region-specific question-answer annotations. It comprises 47,300 COCO-sourced images with 327,929 QA pairs, 1,311,756 human-generated multiple-choice questions, and 561,459 object groundings from 36,579 categories. The questions, structured exclusively as four-option multiple-choice items, are systematically organized around seven interrogative types (What, Where, How, When, Who, Why, Which), as shown in Figure 6.

ST-VQA. The ST-VQA [2] dataset comprises 23,038 images from six diverse sources (including scene-text benchmarks like COCO-Text and VizWiz, and general vision datasets such as ImageNet and Visual Genome) to mitigate inherent biases and enhance question variety. Each image contains two or more scene text instances, ensuring multiple answer options. The dataset provides 31,791 non-binary, unambiguous question-answer pairs requiring explicit reasoning about textual elements within visual contexts. It is split into training (19,027 images with 26,308 QA pairs) and evaluation subsets for standardized benchmarking, as shown in Figure 7.

TextVQA. TextVQA [13] is a standard benchmark for text-based visual reasoning, requiring models to read and reason about scene text within images to answer questions. The dataset comprises 28,408 images and 45,336 questions. It is split into training (21,953 images; 34,602 questions), validation (3,166 images; 5,000 questions), and test sets (3,289 images; 5,734 questions), as shown in Figure 8.

PaintingForm. The PaintingForm [1] dataset comprises 19,000 painting images paired with 50,000 expert analysis paragraphs focused exclusively on visual characteristics of

Table 9. Utility performance of LoRA fine-tuned LLaVA-1.5 models.

Datesets	V7W (ACC)	ST-VQA (ACC)	TextVQA (ACC)	PaintingF (BLEU / ROUGE)	MathV (ACC)
Before fine-tuning	0.6%	21.9%	15.6%	4.9% / 8.8%	27.1%
After fine-tuning	37.9%	65.2%	51.3%	13.3% / 15.9%	57.5%

Table 10. Utility performance of LoRA fine-tuned Qwen2-VL models.

Datesets	V7W (ACC)	ST-VQA (ACC)	TextVQA (ACC)	PaintingF (BLEU / ROUGE)	MathV (ACC)
Before fine-tuning	0.8%	29.8%	23.9%	7.6% / 11.9%	7.0%
After fine-tuning	37.9%	47.7%	78.7%	13.4% / 16.2%	66.6%

Table 11. Utility performance of full fine-tuned Qwen2-VL models.

Datesets	V7W (ACC)	ST-VQA (ACC)	TextVQA (ACC)	PaintingF (BLEU / ROUGE)	MathV (ACC)
Before fine-tuning	0.8%	29.8%	23.9%	7.6% / 11.9%	7.0%
After fine-tuning	36.5%	88.8%	78.6%	13.5% / 15.9%	86.0%

artwork. Designed to advance multimodal AI for deep understanding of artistic elements, as shown in Figure 9.

MathV360k. MathV360K [11] is a comprehensive multimodal benchmark synthesized from 24 open-source datasets to advance mathematical visual reasoning. Curated through a rigorous selection process, the dataset originates from 40K high-quality images filtered by visual clarity and cognitive complexity. Each image is enriched with 360K diverse instruction-tuning pairs targeting five high-level reasoning domains: Figure Question Answering (FQA), Geometry Problem Solving (GPS), Math Word Problems (MWP), Textbook Question Answering (TQA), and Visual Question Answering (VQA). This multi-domain architecture addresses critical gaps in existing resources by enhancing image comprehension and mathematical reasoning capabilities, as shown in Figure 10.

A.3. Fine-tuning setting

For both full and LoRA [5] fine-tuning settings, the detailed training configurations are summarized in Table 8. All LoRA fine-tuned models are evaluated in their merged form. For Qwen2-VL, multimodal downstream task fine-tuning is conducted based on the LlamaFactory [18] project.

A.4. Inference setting

We use "generate" function for LLaVA-1.5's inference in all experiments. We set inference parameters such as temperature at 0.5, Top-p at 0.5, num-beams at 1, and max-new-tokens at 128. For Qwen2-VL, we use default inference setting with max-new-tokens at 128.

B. Additional Experiments

B.1. Utility of the fine-tuning models

We test the utility performance of our two original MLLMs before and after LoRA fine-tuning on downstream task datasets. For full fine-tuning, we report the utility performance of Qwen2-VL before and after fine-tuning on downstream task datasets. As detailed in Tables 9, 10, and 11 fine-tuning substantially enhanced MLLM performance on target tasks. This indicates that there has been a significant change through fine-tuning the model parameters, demonstrating that fine-tuned models effectively simulate real-world application scenarios. Evaluation was conducted on 5,000 randomly sampled test/validation instances from each dataset. For dataset V7W, ST-VQA, TextVQA, and MathV360k, we computed accuracy (ACC) based on answer correspondence. For the dataset PaintingForm, we evaluated BLEU [9] and ROUGE [6] scores.

B.2. Impact on inference parameter settings

In practical model deployment, the performance of trigger tracing may vary due to the randomization effects caused by the configuration of sampling parameters such as temperature and top-P during downstream model inference. We test two fine-tuned variants of LLaVA-1.5 and Qwen2VL across temperature and Top-p values ranging from 0.1 to 1.0 in 0.1 increments. As shown in Figure 11, the results show that the ASR fluctuation stayed within ±1 %, confirming that AGDI is resilient to inference randomization, such as sampling parameter variation.

Q: Why does the second computer have fish on the screen?
A: Its screen saver is running.

Q: Where is the image taken?
A: Near to house.

Q:What kind of bus is this?
A: Double decker bus.

Figure 6. Examples of Visual7W dataset.

Q: What does Espressa sell?
A: Pizza.

Q: What does the writing on the bottom of the picture read?
A: Big Kites.

Q: Who is the sponsor of this tournament
A: Blackrock.

Figure 7. Examples of ST-VQA dataset.

Q: What letters are embellished on the parachute?
A: Raf.

Q: What man is mentioned on the back of the truck?
A: Ice cream man

Q: What is the last number of the plane?
A: 3.

Figure 8. Examples of TextVQA dataset.

B.3. More multimodal datasets fine-tuned models

Tracking the copyright of models fine-tuned for diverse downstream tasks is essential for strong MLLM copyright protection. To further evaluate our method's robustness, we conduct validation across five prevalent task domains: remote sensing, medical, scientific, autonomous driving, and finance. Specifically, we select RSVQA [8], DriveLM [12], FinVis [17], PathVQA [4] and SciVerse [3]. The details of these five datasets are as follows:

- **RSVQA** is the first remote sensing VQA dataset constructed by automatically extracting information from OpenStreetMap. It comprises two versions based on low-resolution Sentinel-2 satellite imagery and high-resolution aerial imagery, respectively, and covers five types of

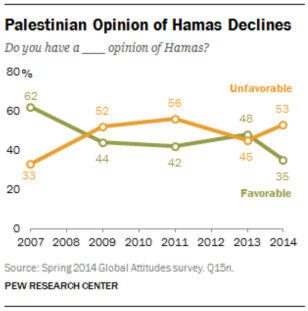
Q: Craft a well-structured paragraph that delves into the formal analysis of visual elements, ensuring a comprehensive exploration of their intrinsic qualities.

A: This painting showcases a vibrant and colorful composition, with a focus on two round fruits placed prominently in the center of the canvas. The artist's use of bold, expressive brushstrokes creates a sense of movement and energy within the piece, particularly evident in the swirling patterns of light and shadow that dance across the surface of the fruits. The rich, warm tones of reds, yellows, and greens add a sense of depth and dimension to the objects, making them appear almost three-dimensional. The background of the painting is kept relatively simple, with soft, blurred forms that serve to highlight the detailed rendering of the fruits themselves. The play of light and shadow on the textured surfaces of the fruits creates a dynamic interplay of highlights and shadows, adding to the overall sense of realism in the piece. The composition is balanced and harmonious, with the two fruits positioned in such a way as to draw the viewer's eye towards the center of the painting. Overall, this work demonstrates the artist's skill in capturing the beauty and vitality of everyday objects through his masterful use of color, light, and texture.

Q: Craft a well-structured paragraph that delves into the formal analysis of visual elements, ensuring a comprehensive exploration of their intrinsic qualities.

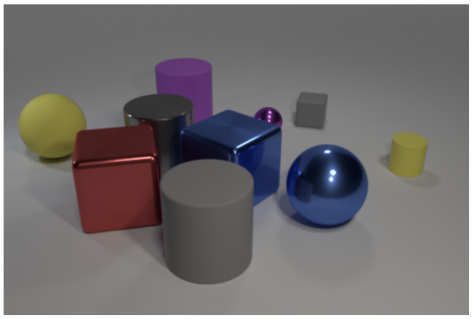
A: The seascape depicts a tumultuous ocean under a dramatic sky. The waves crash against the rocky shore, creating a sense of movement and energy. The sky is filled with dark clouds, which contrast with the bright blue of the water. The use of light and shadow creates a sense of depth and atmosphere. The painting is executed in a realistic style, with attention to detail. The brushstrokes are fluid and expressive, capturing the movement of the waves and the texture of the rocks. The overall effect is one of drama and beauty, as the viewer is drawn into the power and majesty of the sea.

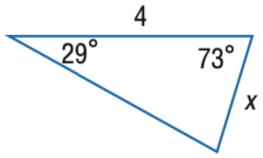
Figure 9. Examples of PaintingForm dataset.



Q: Hint: Please answer the question and provide the final answer at the end. Question: In the year 2014, which opinion is dominant?

A: The answer is Unfavorable.





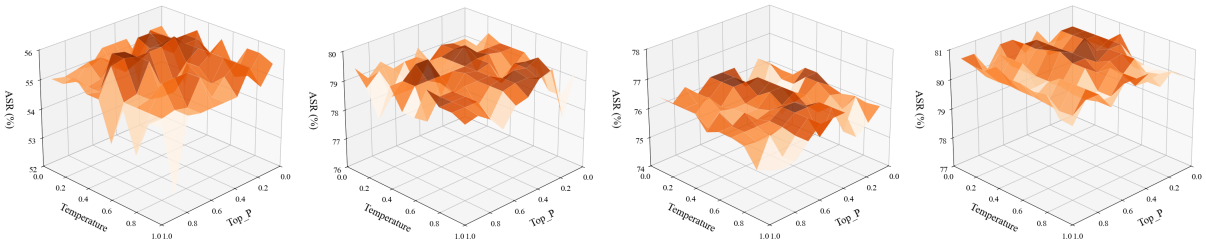
Q: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end. Question: Find x. Round side measure to the nearest tenth. Choices: (A) 2.0 (B) 4.1\n(C) 7.9 (D) 8.1

A: The answer is A.

Q: Hint: Please answer the question and provide the final answer at the end. Question: Subtract all rubber blocks. Subtract all cubes. How many objects are left?

A: The answer is 6.

Figure 10. Examples of MathV360k dataset.



(a) **LLaVA-1.5 (ST-VQA).** (b) **LLaVA-1.5 (PaintingF).** (c) **Qwen2-VL (ST-VQA).** (d) **Qwen2-VL (PaintingF).**

Figure 11. Hyperparameter analysis results of inference settings (Top-p and temperature). (a)(b) ASR on ST-VQA and PaintingF for LLaVA-1.5; (c)(d) ASR for Qwen2-VL.

Table 12. Copyright tracking performance on more multimodal downstream models.

Method	Qwen2-VL					InternVL3.5				
	Pathvqa	DriveVQA	FinVis	RSVQA	SciVerse	Pathvqa	DriveVQA	FinVis	RSVQA	SciVerse
Ordinary	41%	52%	50%	55%	69%	30%	27%	26%	36%	46%
RNA	39%	43%	43%	46%	54%	29%	19%	17%	30%	35%
PLA	55%	76%	72%	76%	86%	40%	46%	52%	64%	73%
AGDI	**58%**	**80%**	**77%**	**79%**	**88%**	**45%**	**50%**	**57%**	**70%**	**76%**

question-answer pairs: counting, existence, area estimation, comparison, and urban-rural classification.

- **DriveLM** is a graph-structured VQA dataset for autonomous driving, available in both real-world and simulated versions. By integrating semi-regularized with fully automated annotation, it achieves superior scale, coverage, and logical complexity over existing benchmarks, offering a generalizable platform for training and evaluating vision-language models in autonomous driving domain.
- **FinVis** presents the first two-stage multimodal instruction dataset designed for financial chart analysis. Featuring a pretraining stage for vision-language alignment on historical charts and an innovative instruction-tuning stage that incorporates future data for forecasting, the dataset is structured as image, instruction, answer triplets. It provides comprehensive support for professional tasks including chart description, financial question answering, and trend prediction.
- **PathVQA** is a medical visual question answering dataset derived via a semi-automated pipeline from textbooks and digital libraries. It is designed to simulate the American Board of Pathology examinations, with open-ended clinical questions constituting 50.2% of its content.
- **SciVerse** introduces a multimodal scientific assessment dataset spanning physics, chemistry, and biology. It analyzes LLM capabilities in knowledge, vision, and reasoning by varying the knowledge and visual complexity of the problems. This is paired with a novel scientific CoT evaluation strategy to progressively pinpoint knowledge and logic errors, providing deep diagnostic insights into models' problem-solving gaps.

In the fine-tuning setup, we employ the full training set for PathVQA, whereas a 30k-sample subset of the training set is used for each remaining dataset. As shown in Table 12, we report the ASR of LoRA fine-tuned variants of Qwen2-VL-2B-Instruct and InternVL3.5-2B-HF. The results demonstrate that AGDI achieves superior copyright tracing performance compared with the baselines across a wide range of multimodal and fine tuning scenarios.

B.4. Experiments on more MLLMs

We also evaluate the copyright tracking performance on the advanced MLLM such as InternVL3.5 [16]. InternVL3.5 is an open-source multimodal model series featuring a "ViT-MLP-LLM" architecture that significantly enhances reasoning capabilities via cascade reinforcement learning and innovatively introduces a Visual Resolution Router (ViR) for dynamic visual token compression together with Decoupled Vision-Language Deployment. We report the ASR on the InternVL3.5-HF 2B and 8B parameter scale models. As shown in Tables 13 and 14, consistent performance confirms that our method generalizes effectively across different architectures and increasing parameter scales. These results demonstrate that the effectiveness of our method is not limited by model size or architecture, exhibiting strong generalization to fine-tuned models.

Table 13. Copyright tracking performance on InternVL3.5 8B fine-tuned models.

MLLM	Method	V7W	ST-VQA	TextVQA	PaintingF	MathV
InternVL3.5-8B	Ordinary	14%	21%	15%	9%	16%
	RNA	16%	22%	15%	11%	20%
	PLA	44%	55%	46%	31%	49%
	AGDI	**58%**	**63%**	**62%**	**45%**	**58%**

B.5. Tracking results under input transformations.

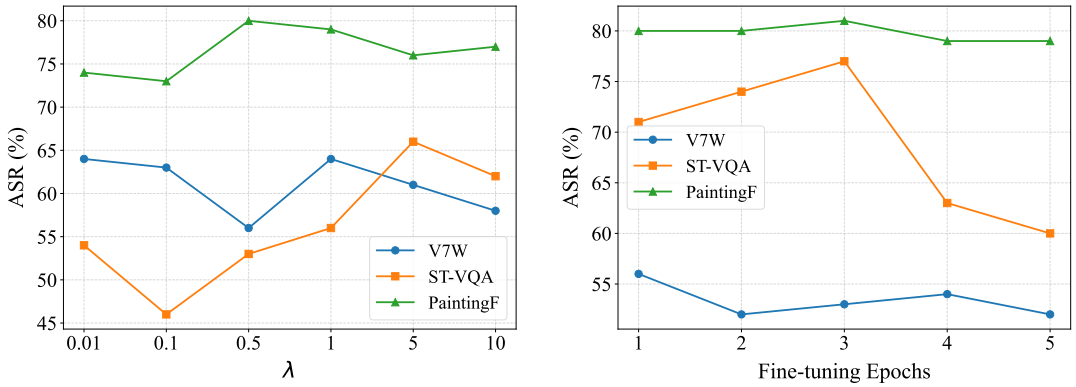
We report the copyright tracing results of our method under several common input level perturbations, including JPEG compression, Gaussian noise, and image resizing, to evaluate the robustness of the generated trigger images. The maximum magnitude of the Gaussian noise is set to 5, and the resized image resolution is fixed at 256. In Table 15, results on LLaVA-1.5 variants demonstrate that our method exhibits robustness against input transformations.

B.6. Ablation study of more parameters

We add the ablation study for parameter λ and fine-tuning epochs in Figure 12. The results in Figure 12a show consistent and stable performance across various λ values, demonstrating the robustness of our method to this hyperparameter. The results in Figure 12b shows that ASR stabilizes with more finetuning, proving increased scale cannot bypass our protection. Our setup aligns with practical scenarios relying on lightweight methods. Furthermore, given the trade-off between cost and utility, aggressive operations like full retraining are impractical because they degrade the model's performance and commercial value.

Table 14. Copyright tracking performance on InternVL3.5 2B fine-tuned models.

Method	LoRA Fine-tuning					Full Fine-tuning				
	V7W	ST-VQA	TextVQA	PaintingF	MathV	V7W	ST-VQA	TextVQA	PaintingF	MathV
Ordinary	25%	38%	32%	12%	23%	35%	38%	30%	21%	30%
RNA	23%	40%	30%	10%	20%	33%	40%	31%	18%	26%
PLA	39%	53%	45%	23%	34%	46%	53%	42%	38%	45%
AGDI	**42%**	**55%**	**47%**	**30%**	**41%**	**49%**	**56%**	**44%**	**49%**	**51%**



(a) **Parameter** λ (b) **Fine-tuning epochs**

Figure 12. Ablation study for λ and fine-tuning epochs: (a) The impact of loss function parameter λ on tracking performance; (b) The impact of downstream fine-tuning epochs on tracking performance.

Table 15. Copyright tracking under input transformations.

	V7W	ST-VQA	TextVQA	PaintingF	MathV
Original	64%	56%	36%	79%	30%
Resizing	42%	40%	26%	53%	19%
Gaussian	60%	53%	31%	73%	27%
JPEG	40%	39%	24%	42%	16%

B.7. Sensitivity analysis of trigger selection

We provide a sensitivity analysis of trigger selection for copyright tracking performance. Table 16 shows that tracking results depend more on the specific model than the trigger pairs. Furthermore, we use five diverse QA pairs and 200 random images per pair to prevent bias and ensure consistent results on the models.

Table 16. ASR on 5 different QA pairs.

	QA1	QA2	QA3	QA4	QA5
V7W	63%	75%	70%	68%	37%
ST-VQA	58%	62%	44%	61%	48%
PaintingF	81%	86%	69%	93%	62%

B.8. Robustness of system prompt variations

In real-world scenarios, malicious users or downstream developers usually modify the system prompt. We test the ASR on the fine-tuned models under system prompt variations. We design two system prompt as follows.

- As a clinical psychologist, use an empathetic tone and prioritize asking questions to guide emotional expression before offering advice.
- As a technical interviewer from a top tech firm, evaluate the candidate's programming basics and provide feedback after each answer.

Table 17 shows consistent ASR across different models before and after system prompt modifications. These results confirm system prompt robustness of AGDI.

Table 17. System prompt experiments on LLaVA-1.5 LoRA variants.

	V7W	ST-VQA	TextVQA	PaintingF	MathV
Original	64%	56%	36%	79%	30%
Sys prompt1	61%	53%	32%	73%	26%
Sys prompt2	63%	54%	32%	74%	25%

B.9. CLIP-like module stability evidence

We measure the similarity drift between 200 triggers and target texts across various fine-tuned models. As shown in

Table 18. Average similarity drift(%) on the fine-tuned models.

Similarity drift	Pair1	Pair2	Pair3	Pair4	Pair5
Base→ V7W	4.4%	9.3%	1.0%	2.1%	6.1%
Base→ MathV	6.9%	7.1%	3.3%	0.5%	1.2%

Table 18, the minimal cosine similarity drift strongly supports the CLIP-like semantic stability assumption, confirming that AGDI's gains stem from exploiting intrinsic model properties.

References

[1] Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. GalleryGPT: Analyzing paintings with large multimodal models. In *ACM MM*, pages 7734–7743, 2024. 1

[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 *ICCV*, pages 4291–4301, 2019. 1

[3] Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. In *Findings of ACL*, pages 19683–19704, 2025. 3

[4] 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. 3

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2

[6] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, 2004. 2

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 1

[8] Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58 (12):8555–8566, 2020. 3

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[11] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 2

[12] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with graph visual question answering. In *ECCV*, 2025. 3

[13] 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. 1

[14] Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[15] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[16] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5

[17] Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. FinVis-GPT: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*, 2023. 3

[18] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[19] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016. 1