
Supplementary Material for MINT-CoT: Enabling Interleaved Visual Tokens in Mathematical Chain-of-Thought Reasoning

Anonymous Author(s)

Affiliation

Address

email

A Appendix

A.1 Overview

We organize our supplementary material as follows.

- Related Work
 - MLLMs for Mathematics
 - Visual Chain of Thought
- Dataset Details
 - Dataset Example
 - Dataset Statistic
- Theoretical Details of Interleaved CoT RL
- Additional Implementation Details
- Additional Quantitative Results
 - Results on Other Benchmark
 - Projector Ablation
- Additional Qualitative Results

A.2 Related work

MLLMs for Mathematics Recent advancements in MLLMs [25, 18, 1] have shown impressive capabilities in various vision-language tasks. However, even powerful models like GPT-4V [25] and Qwen2-VL [31] fail to demonstrate satisfying performance on existing visual mathematical benchmarks [3, 21, 20, 38]. Various specialized approaches [11, 39, 15, 7, 22, 30, 26] have emerged to enhance visual mathematical reasoning. Current approaches mostly focus on enriching the multimodal math data. G-LLaVA [11] extends the LLaVA architecture with geometric reasoning capabilities by augmenting the current dataset. Math-LLaVA [30] enlarges the data scope with the introduced MathV360K dataset. MAVIS [39] first identifies the critical issue of the vision encoder and empowers it with the mathematical capability. Then it further develops an automated system for generating mathematical visual datasets at scale. Reverse Chain-of-Thought (R-CoT) [7] introduces the Geometry Generation Chain for creating geometric images with more accurate descriptions.

Visual Chain of Thought With advancements of various visual reasoning tasks [20, 36, 16], visual chain of thought has been emerging as an effective method to enhance accuracy. One line of the

method relies on textual CoT to conduct multimodal analysis [9, 23, 4, 37, 8, 34]. For example, R1-V[4] extends the paradigm of DeepSeek R1 [13] to generate a comprehensive text CoT to analyze the visual information before providing the final answer. Another line of method explicitly incorporates multimodal elements in the rational [28, 24, 33, 14, 17]. Visual CoT [28] and Chain-of-Spot [19] propose to crop the region of high interest on the image and integrate it into the CoT process. Chain-of-Image [24] and Visual SKETCHPAD [14] introduce auxiliary tools to generate helpful diagrams for mathematical or geometric problem-solving. Although these methods demonstrate competitive performance, they are limited to rigid image cropping or dependence on external tools. Recently, ICoT [12] leverages the attention map of the MLLM to select the relevant visual tokens to compose the multimodal rational. However, this approach relies solely on attention scores on the image feature maps, which have been shown to be insufficiently informative for mathematical scenarios [39].

A.3 Dataset Details

Dataset Example We present examples from our MINT-CoT Dataset in Figures 1 to 3, where the yellow highlights indicate the interleaved grid indices, and the blue highlights denote the key words in each reasoning step.

Dataset Statistic We provide the key statistics of MINT-CoT Dataset in Table 1. This dataset comprises 54,031 data points derived from the mathematical portion of the Mulberry-260k dataset.

Table 1: **Key statistics of the MINT-CoT dataset.**

Statistic	Value
Total data points	54,031
Data points containing Interleave Tokens (interleaved data points)	52,142
Average number of Interleave Tokens per interleaved data point	2.80
Maximum number of Interleave Tokens in a single interleaved data point	12
Average number of selected indices per interleaved data point	19.91
Average number of selected indices per Interleave Token	7.10
Minimum number of selected indices in a single Interleave Token	1
Maximum number of selected indices in a single Interleave Token	140

46

A.4 Theoretical Details of Interleaved CoT RL

Following the standard GRPO framework [29], we integrate GRPO into our approach. Specifically, similar to \mathcal{L}_{CE} in Stage 2, we apply a policy loss \mathcal{L}_{GRPO_text} to textual tokens:

$$\mathcal{L}_{GRPO_text} = -\mathbb{E}_{\{Y_j\}_{j=1}^G \sim P_{\theta_{old}}(\cdot|I,T)} \left[\frac{1}{G} \sum_{j=1}^G \frac{1}{|T_j|} \sum_{t \in T_j} \left\{ \frac{P_{\theta}(y_{j,t}|y_{j,<t},I,T)}{P_{\theta_{old}}(y_{j,t}|y_{j,<t},I,T)} \cdot \hat{A}_{j,t} - \beta D_{KL}[P_{\theta} \parallel P_{ref}] \right\} \right], \quad (1)$$

where $\hat{A}_{j,t}$ is the advantage detailed in Section 2.3, P_{ref} is a reference policy that serves as a regularization target, and $D_{KL}[P_{\theta} \parallel P_{ref}]$ penalizes deviation from this reference distribution to encourage stable updates. The min and clip operations are omitted for brevity.

To enable more flexible and effective selection of visual tokens, we further apply a \mathcal{L}_{GRPO_vis} to the scaled similarity scores $\alpha_{j,\tau}^{(i)}$, which are derived from the interactions between Interleave tokens and input visual tokens in the j -th chain of reasoning steps. Let M_j denote the number of input visual tokens in j -th chain, and N_j denote the the number of Interleave tokens in j -th chain. Formally, the loss is defined as:

$$\mathcal{L}_{GRPO_vis} = -\mathbb{E}_{\{Y_j\}_{j=1}^G \sim P_{\theta_{old}}(\cdot|I,T)} \left[\frac{1}{G} \sum_{j=1}^G \frac{1}{N_j M_j} \sum_{i,\tau=1}^{N_j, M_j} \left\{ \frac{P_{\theta}(\alpha_{j,\tau}^{(i)}|y_{j,<\tau},I,T)}{P_{\theta_{old}}(\alpha_{j,\tau}^{(i)}|y_{j,<\tau},I,T)} \cdot \hat{A}_j - \beta D_{KL}[P_{\theta} \parallel P_{ref}] \right\} \right]. \quad (2)$$

The final policy loss is defined as the sum of both losses, with the \mathcal{L}_{GRPO_vis} rescaled by a weighting factor λ :

$$\mathcal{L}_{GRPO} = \mathcal{L}_{GRPO_text} + \lambda \cdot \mathcal{L}_{GRPO_vis}. \quad (3)$$

By computing this combined loss, we enhance both token selection and inference capabilities using Interleave tokens.

Table 2: **Combined results on the Mathematics section of MMStar benchmark.** We evaluate MINT-CoT-7B, the baseline model and the state-of-the-arts.

Model	MMStar-Math
Qwen2.5-VL-7B-Instruct [2]	66.8
InternVL2-8B [6]	66.8
R1-VL-7B [37]	68.4
Mulberry [35]	66.8
Open-R1-Multimodal [10]	59.2
Qwen2-VL-7B-Instruct [32] (Baseline)	46.4
MINT-CoT-7B	69.6
Δ over the Baseline Model	+ 23.2

Table 3: **Ablation study on different training stages.** We evaluate the baseline model, Text-only CoT SFT, Interleaved CoT SFT and Interleaved CoT RL on the Mathematics section of MMStar benchmark.

Model	MMStar-Math
Baseline [32]	46.4
+ Text-only CoT SFT	67.6
+ Interleaved CoT SFT	68.00
+ Interleaved CoT RL	69.6

62 A.5 Additional Implementation Details

63 We use Qwen2-VL-7B [32] as the base MLLM model in our experiments. Each of the two projectors,
64 $P_{\text{interleave}}$ and P_{vis} , is implemented as a single linear layer. We uniformly set the threshold $\theta = 0.7$ to
65 filter the similarity scores. The hyper-parameter γ to scale the similarity is set to $1/0.07$ following
66 CLIP [27]. The training procedure consists of three stages: (1) *Text-only CoT Training*, where we train
67 for 2 epochs on the MINT-CoT dataset without applying the interleaving strategy, using a learning
68 rate of $5.0\text{e-}6$ and a batch size of 64, following the configuration of Mulberry [35]; (2) *Interleaved*
69 *CoT SFT*, where we train for 3 epochs on the MINT-CoT dataset with a learning rate of $1\text{e-}6$ and a
70 batch size of 64; and (3) *Interleaved CoT RL*, where we train for 700 steps on the MINT-CoT dataset,
71 using a group size $G = 4$, a weighting factor $\lambda = 0.02$, a learning rate of $1\text{e-}6$ and a batch size of
72 16. During training, all model parameters, including the Interleave Token and projector layers, are
73 unfrozen, except for the vision encoder, which remains fixed. Finally, the resulting model is named
74 MINT-CoT-7B.

75 For Bounding Box CoT SFT, we use the MINT-COT dataset and extract the minimal enclosing
76 rectangle that covers the index positions of all labels as the ground truth bounding box to train the
77 model. We train 2 epochs with a learning rate of $1\text{e-}6$ and a batch size of 64. And during inference, it
78 interleaves the minimal enclosing rectangle that covers all the selected tokens. For Original Image CoT
79 SFT, however, we enforce the concatenation of the entire image at the beginning of each step during
80 both training and inference. We train only 1 epoch with a learning rate of $1\text{e-}6$ and a batch size of 64,

81 A.6 Additional Quantitative Results

82 **Results on Other Benchmark** In addition to the quantitative results in Section 3.2, we provide
83 results of our method on the Mathematics section of MMStar [5] benchmark. As shown in Table 2,
84 MINT-CoT-7B outperforms the baseline model by +23.2%, and performs better than the state-of-the-
85 arts. Furthermore, we present the results of an ablation study on different training stages in Table 3.
86 We find that the performance improves at each training stage, validating the efficiency of our three
87 training stages.

88 **Projector Ablation** We conduct an ablation study on the post interleave projector $P_{\text{post_intlv}}$ and
89 the post visual projector $P_{\text{post_vis}}$ on the Interleaved CoT SFT stage. Both projectors were initially
90 implemented as single-layer linear layers. We first remove both projectors entirely, and then replace

Table 4: **Ablation study on the post interleave projector and the post visual projector.** We compare three configurations: without projectors, with single-layer linear projections, and with two-layer MLPs.

Configuration	Layer Number	All	GEO	ALG	GPS	TQA
w.o. projectors	–	64.44	63.68	64.42	63.94	66.13
w. projectors	1	67.78	66.51	67.79	67.31	69.35
	2	65.18	63.21	65.54	63.94	69.35

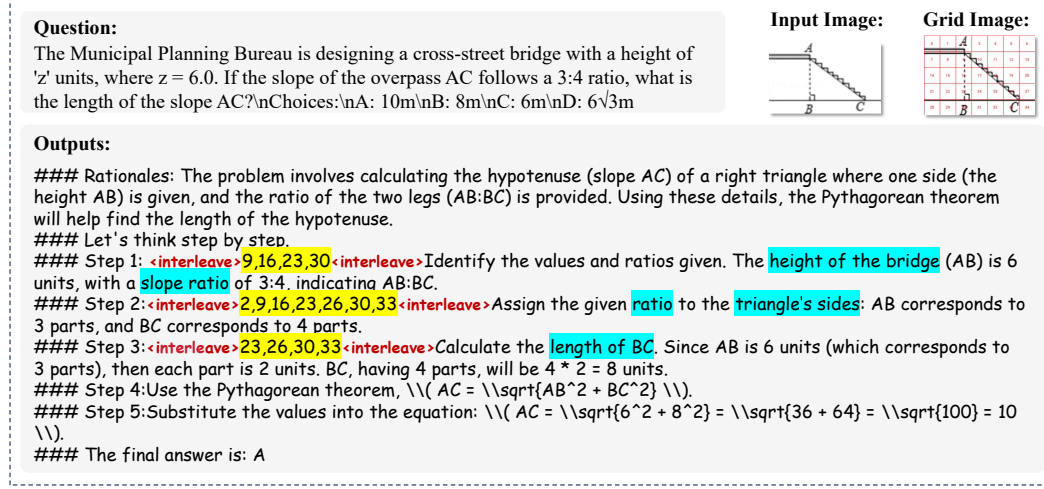


Figure 1: An example from MINT-CoT dataset.

91 them with two-layer MLPs using GELU activation. Both configurations are trained for three epochs.
 92 The results on the mathematical subset of MathVista are shown in Table 4, in which we find that the
 93 initial configuration as single-layer linear layers performs the best over all primary tasks.

94 A.7 Additional Qualitative Results

95 In addition to Section 3.4, we provide more qualitative results of the baseline model Qwen2-VL-7B-
 96 Instruct and our proposed model MINT-CoT-7B in Figures 4 to 6

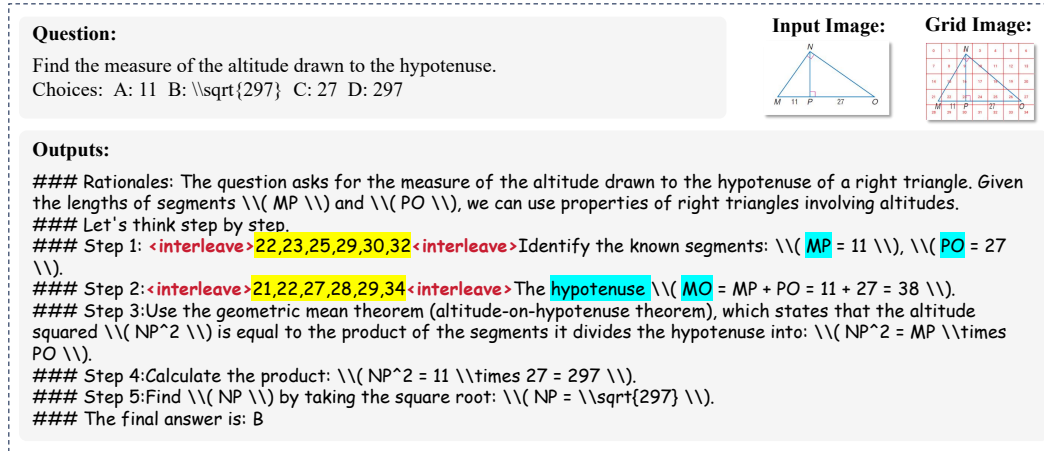


Figure 2: An example from MINT-CoT dataset.

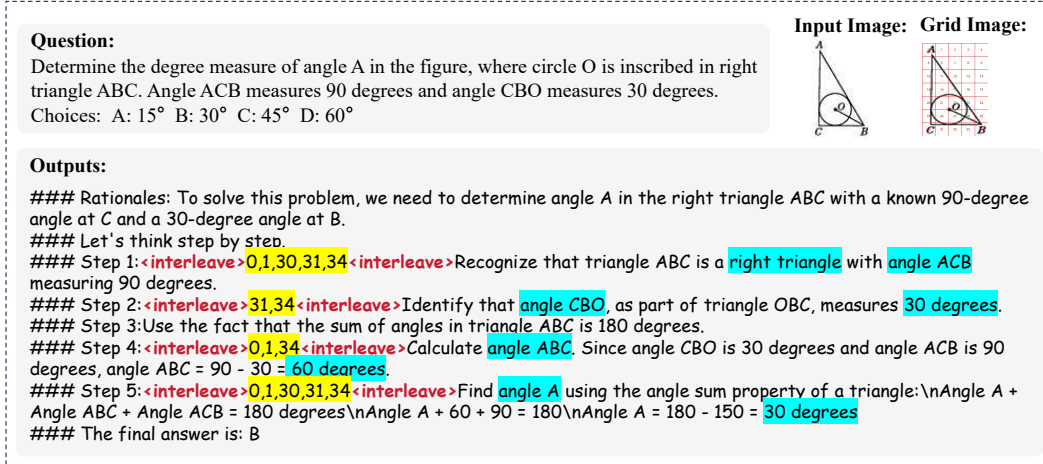
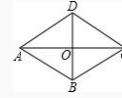


Figure 3: An example from MINT-CoT dataset.

Question:

As shown in the figure, in diamond ABCD, the diagonals AC and BD intersect at O. It is known that BD = 6 and AC = 8. The perimeter of diamond ABCD is ().
Choices: A: 40 B: 20 C: 10 D: 5



Qwen2vl-7B-Instruct

The quadrilateral ABCD is a rhombus since its diagonals are perpendicular. \n\n[asy]\npair A,B,C,D,O;\nA=(0,4);\nB=(6,0);\nC=(-3*sqrt(3), 3);\nD=(0, 4);\nO=(0, 0);\ndraw(A--B--C--D--cycle, black+linewidth(1));\ndraw(A--C, black+linewidth(1));\ndraw(B--D, black+linewidth(1));\ndraw(O--A, black+linewidth(1));\ndraw(O--B, black+linewidth(1));\ndraw(O--C, black+linewidth(1));\ndraw(O--D, black+linewidth(1));\nlabel("\$A\$", A, N);\nlabel("\$B\$", B, E);\nlabel("\$C\$", C, dir(315));\nlabel("\$D\$", D, N);\nlabel("\$O\$", O, S);\n[/asy]\n\nLet's label the intersection point of the diagonals O. \n\nSince OA = OC = 4 and OB = OD = 3, we have OB = 3 and BD = 6. \n\nSo, the perimeter of the rhombus ABCD is 4OA + 4OB = 4 * 4 + 4 * 3 = 16 + 12 = 28. \n\nTherefore, the answer is 28. \n\nSince the question asks for the perimeter of the diamond ABCD, the answer is 20. \n\nTherefore, the answer is B. ❌

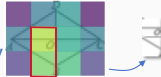
MINT-CoT-7B

Rationales:

To find the perimeter of the rhombus, the length of all sides needs to be determined. Since the diagonals bisect each other at right angles in a rhombus, the sides can be found using the Pythagorean theorem. \n\n#### Let's think step by step.

Step 1:

Cosine Similarity Scores



Identify the properties of the rhombus. In a rhombus, the diagonals bisect each other at right angles. Therefore, AO = OC = AC/2 = 8/2 = 4 and BO = OD = BD/2 = 6/2 = 3.

Step 2:

Cosine Similarity Scores



Use the Pythagorean theorem to calculate the side length (AD) of the rhombus. The hypotenuse is AD, with AO = 4 and BO = 3. \n\n[math display="block">AD = \sqrt{AO^2 + BO^2} = \sqrt{4^2 + 3^2} = \sqrt{16 + 9} = \sqrt{25} = 5

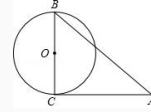
Step 3: Calculate the perimeter of the rhombus. Since all sides are equal in a rhombus, the perimeter is 4 times the side length. \n\nPerimeter = 4 * 5 = 20

The final answer is: B ✅

Figure 4: Comparison between Qwen2-VL-7B-Instruct and MINT-CoT-7B.

Question:

As shown in the figure, the radius of circle O is 4.0, BC is the diameter, AC is the tangent of circle O, and AC = 6.0, then AB = ()
 Choices: A: 4 B: 6 C: 10 D: 12

**Qwen2vl-7B-Instruct**

Because AC is tangent to circle O, it is perpendicular to OC. By defining the length of AC as x, we can solve for segment AB using the Pythagorean theorem in a right triangle. Start by understanding the Pythagorean theorem. Apply the Pythagorean theorem to a right triangle with segments AC and AB as the legs and segment BC as the hypotenuse. Given that AO is the radius (4.0), we use the Pythagorean theorem in triangle OAB: $AB = \sqrt{(OA^2 + OB^2)} = \sqrt{(4.0^2 + 4.0^2)}$ $AB = \sqrt{16 + 16}$ $AB = \sqrt{32}$ $AB \approx 6.4$ So the answer is approximate 6.4. It seems there is no exact option with this value, but the closest one would be C) 6. However, given the calculation, the length AB should be a bit more than the radius of circle O. Therefore, the answer is C:6. ✗

MINT-CoT-7B

Rationales: To determine AB, understand the properties of the circle and the right-angled triangle formed by the tangent AC and the diameter BC. Use the Pythagorean theorem in triangle ABC.

Let's think step by step.

Step 1:

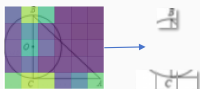
Cosine Similarity Scores



Identify the known lengths: radius of the circle (OC) is 4, and AC, the tangent, is 6.

Step 2:

Cosine Similarity Scores



Recognize that BC is a diameter of the circle. Since the radius is 4, BC (being the diameter) is twice the radius: BC = 8.

Step 3:

Cosine Similarity Scores



Since AC is a tangent and BC is a diameter, consider triangle ABC. Triangle ABC is a right triangle with AC perpendicular to BC.

Step 4: Apply the Pythagorean theorem in triangle ABC: $AB^2 = AC^2 + BC^2$.

Step 5: Substitute the known values into the Pythagorean theorem: $AB^2 = 6^2 + 8^2$.

Step 6: Calculate the squares: $6^2 = 36$ and $8^2 = 64$.

Step 7: Sum the squares: $AB^2 = 36 + 64 = 100$.

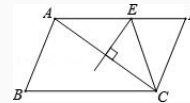
Step 8: Take the square root to find AB: $AB = \sqrt{100} = 10$.

The final answer is: C ✓

Figure 5: Comparison between Qwen2-VL-7B-Instruct and MINT-CoT-7B.

Question:

In ABCD, given that AB < BC, the perpendicular bisector of diagonal AC intersects AD at point E. Connect CE. If the perimeter of ABCD is 200cm, what is the perimeter of triangle CDE?
 Choices: A: 200.0 B: 400.0 C: 150.0 D: 100.0

**Qwen2vl-7B-Instruct**

The correct answer is D.

By symmetry, BC = 50 cm, AD = 50 cm, so AE = 25 cm. From the right triangle AEC, $AC^2 - AE^2 = CE^2$, so $AC^2 - AE^2 = CE^2 = AD^2 - AE^2$. Therefore, CE = 25 cm. Therefore, the perimeter of triangle CDE = 75 cm. ✗

MINT-CoT-7B

Rationales: The problem involves understanding the geometric placement of point E and the relations between various sides and diagonals in the quadrilateral. It requires determining the perimeter of the triangle CDE given the perimeter of ABCD.

Let's think step by step.

Step 1: Given the perimeter of ABCD is 200 cm, we need to explore the relationships among the sides of the quadrilateral and the triangle formed by extending certain segments.

Step 2:

Cosine Similarity Scores



The perpendicular bisector of AC indicates a symmetrical property: EA = EC. The intersection at E means that E is equidistant from A and C.

Step 3: Since AB < BC, consider potential equalities or relationships with the line segment created by the intersection of AD and the perpendicular bisector at E.

Step 4: Express the perimeter of triangle CDE: CD + DE + EC. Given EC = EA due to E being on the perpendicular bisector of AC, explore possible equalities or simplifications.

Step 5: Using the perimeter constraint and the properties outlined, determine the lengths of the sides of triangle CDE and thus its total perimeter.

The final answer is: D ✓

Figure 6: Comparison between Qwen2-VL-7B-Instruct and MINT-CoT-7B.

References

- [1] 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*, abs/2308.12966, 2023.
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.
- [3] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *ArXiv*, abs/2105.14517, 2021.
- [4] Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>, 2025. Accessed: 2025-02-02.
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.
- [6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.
- [7] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*, 2024.
- [8] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Open-vlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement, 2025.
- [9] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkan Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.
- [10] EvolvingLMMs-Lab. open-r1-multimodal: A fork to add multimodal model training to open-r1. <https://github.com/EvolvingLMMs-Lab/open-r1-multimodal>, 2025. Accessed: 2025-05-13.
- [11] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjuan 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.
- [12] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought, 2025.
- [13] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [14] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024.
- [15] Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometry image dataset creation for enhanced geometry understanding. *arXiv preprint arXiv:2409.09039*, 2024.

- [16] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency, 2025.
- [17] Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. *arXiv preprint arXiv:2501.07542*, 2025.
- [18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [19] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. *arXiv preprint arXiv:2403.12966*, 2024.
- [20] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.
- [21] 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. In *Annual Meeting of the Association for Computational Linguistics*, pages 6774–6786, 2021.
- [22] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.
- [23] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.
- [24] Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*, 2023.
- [25] OpenAI. GPT-4V(ision) system card, 2023.
- [26] Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multi-math: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*, 2024.
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [28] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.
- [29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.
- [30] 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.
- [31] 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.

- 193 [32] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing
194 Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men,
195 Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-
196 language model’s perception of the world at any resolution, 2024.
- 197 [33] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei.
198 Mind’s eye of llms: Visualization-of-thought elicits spatial reasoning in large language models.
199 In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- 200 [34] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu,
201 Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized
202 multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*,
203 2025.
- 204 [35] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie
205 Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like
206 reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*,
207 2024.
- 208 [36] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
209 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal
210 understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF*
211 *Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- 212 [37] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng
213 Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group
214 relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- 215 [38] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun
216 Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see
217 the diagrams in visual math problems? In *European Conference on Computer Vision*, pages
218 169–186. Springer, 2024.
- 219 [39] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhao
220 Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and
221 Hongsheng Li. Mavis: Mathematical visual instruction tuning with an automatic data engine,
222 2024.