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

Xinyan Chen^{*1}, Renrui Zhang^{*†1}, Dongzhi Jiang¹, Aojun Zhou¹
Shilin Yan, Weifeng Lin¹, Hongsheng Li^{‡1,2,3}

¹CUHK MMLab ²Shanghai AI Laboratory ³CPII under InnoHK

{chenxyxy06, 1700012927zrr}@gmail.com
hsli@ee.cuhk.edu.hk

^{*}Equal Contribution [†]Project Leader [‡]Corresponding Author

Abstract

Chain-of-Thought (CoT) has widely enhanced mathematical reasoning in Large Language Models (LLMs), but it still remains challenging for extending it to multimodal domains. Existing works either adopt a similar textual reasoning for image input, or seek to interleave visual signals into mathematical CoT. However, they face three key limitations for math problem-solving: *reliance on coarse-grained box-shaped image regions, limited perception of vision encoders on math content, and dependence on external capabilities for visual modification*. In this paper, we propose **MINT-CoT**, introducing **M**athematical **I**nterleaved **T**okens for **C**hain-**o**f-**T**hought visual reasoning. MINT-CoT adaptively interleaves relevant visual tokens into textual reasoning steps via an Interleave Token, which dynamically selects visual regions of any shapes within math figures. To empower this capability, we construct the MINT-CoT dataset, containing 54K mathematical problems aligning each reasoning step with visual regions at the token level, accompanied by a rigorous data generation pipeline. We further present a three-stage MINT-CoT training strategy, progressively combining text-only CoT SFT, interleaved CoT SFT, and interleaved CoT RL, which derives our MINT-CoT-7B model. Extensive experiments demonstrate the effectiveness of our method for effective visual interleaved reasoning in mathematical domains, where MINT-CoT-7B outperforms the baseline model by +34.08% on MathVista, +28.78% on GeoQA, and +23.2% on MMStar, respectively. Our code and data are available at <https://github.com/xinyan-cxy/MINT-CoT>.

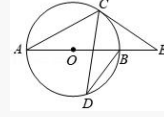
1 Introduction

Chain-of-Thought (CoT) [66, 32] has emerged as an effective strategy for enhancing the reasoning capabilities of Large Language Models (LLMs) [49, 51, 62, 72, 81, 39] by generating sequential rationales in their responses. In Multimodal Large Language Models (MLLMs) [50, 33, 88, 18, 20], CoT also plays a significant role [84] across various tasks involving image [41, 86, 40, 22, 17, 25], video [38, 4, 71, 14], and 3D [70, 24, 58, 21]. It enables MLLMs to reason over both textual and visual inputs, serving as a bridge that connects visual perception with abstract reasoning tasks.

However, despite these advances, applying CoT in mathematical reasoning with visual contexts remains challenging. Existing MLLMs mainly generate text-only reasoning steps for multimodal math problems [84, 85, 60, 79], simply adopting similar textual reasoning for image input. Nevertheless,

Question:

In the given diagram, circle O has line segment AB as its diameter and CD as a chord. A tangent passing through point C intersects the extension of AB at point E , and angle E measures 42° . What is the measure of angle CDB ? Choices: A: 22° B: 24° C: 28° D: 48°

**Text-only CoT Reasoning:**

Step 1: Since angle $E = 42^\circ$, therefore angle $A = 42^\circ$. ✗
 Step 2: Since AB is the diameter of circle O , angle $ACB = 90^\circ$. Therefore, angle $B = 180^\circ - 42^\circ - 90^\circ = 48^\circ$.
 Step 3: Since AB is the diameter of circle O , angle $CDB = \text{angle } B = 48^\circ$. ✗
 Answer: D ✗

Lack of visual information leads to perception error.

Box-shaped Visual CoT Reasoning:

Step 1: Since angle $E = 42^\circ$, angle $CEB = 42^\circ$. ✓
 Step 2: Since AB is the diameter of circle O , angle $ACB = 90^\circ$. ✓
 Step 3: Therefore, angle $CDB = \text{angle } CEB = 42^\circ$. ✗
 Answer: D ✗

Box-level cues are too coarse for math concepts.

Visual Interleaved CoT Reasoning (Ours):

Step 1: Connect OC , $OC \perp CE$. ✓
 Step 2: Angle $COE = 180^\circ - 90^\circ - 42^\circ = 48^\circ$. ✓
 Step 3: Since $OC = OD$, angle $CDB = \text{angle } ODC = 1/2 \text{ angle } BOC = 24^\circ$. ✓
 Answer: B ✓

Token-level interleaved CoT is fine-grained and effective.

Figure 1: Comparison of three CoT reasoning methods: text-only CoT reasoning, box-shaped visual CoT reasoning and our visual interleaved CoT reasoning methods. (1) Text-only CoT lacks visual information, causing perception errors in mathematical reasoning. (2) Box-level cues are too coarse to capture complex visual structures in mathematical images. (3) Token-level interleaved CoT accurately identifies fine-grained visual regions to support reasoning.

due to the limited capability in perceiving math images, this strategy often fails to accurately interpret visual information within the CoT process, leading to reasoning errors.

Recent approaches have attempted to interleave visual content within reasoning steps through mechanisms such as bounding box selection and image cropping [55, 26, 76]. While effective in general visual scenarios, these methods still face three key limitations when extended to multimodal mathematical reasoning:

1. *Reliance on coarse-grained box-shaped image regions:* Recent advances introduce visual information into the CoT process by selecting image regions through bounding box-based methods. Visual-CoT [55], Visual SKETCHPAD [26], and VPT [76] all operate on box-shaped image regions, employing strategies such as bounding box generation, iterative masking, cropping, or re-encoding. However, as shown in Figure 1, these approaches all rely on bounding box-based cropping. While such box-level cues are effective in domains like object detection, where objects are typically isolated, they are too coarse-grained to capture the complex structures in mathematical images, where visual information is not discrete but highly interconnected. As a result, box-shaped selection tends to interleave too many irrelevant or misleading visual tokens, impairing the accuracy of mathematical reasoning.
2. *Limited perception of vision encoders on math content:* Some methods, like ICoT [16], adopt attention-based token selection to identify relevant visual tokens during reasoning without requiring additional training. These approaches rely heavily on visual features extracted by the vanilla vision encoders without specific tuning. However, as noted in MAVIS [83], mainstream vision encoders, which are primarily based on CLIP [54] or SigLIP [78], are pre-trained on natural images with general scenes, making mathematical images out-of-distribution. As a result, such methods often struggle to accurately locate relevant visual regions in complex mathematical tasks.

3. *Dependence on external capabilities for visual modification:* Other approaches attempt to enhance visual reasoning by dynamically generating new visual content or modifying existing images. MVoT [36] is built upon a unified autoregressive MLLM [59] to generate images as part of the CoT process, but it is only applicable to spatial planning tasks. Meanwhile, Visual SKETCHPAD requires external tools to draw on the original image in geometry-related tasks. These approaches depend on external capabilities, either requiring large-scale data to train the understanding model for generation, or relying on external tools with additional inference over the modified images, which leads to numerous extra costs.

Therefore, to address these challenges, we aim to propose a fine-grained, efficient visual interleaved CoT method to enhance the mathematical reasoning capabilities of MLLMs. In this paper, we introduce **MINT-CoT**, an approach of **M**athematical **I**Nterleaved **T**oken selection for **C**hain-of-**T**hought reasoning, which facilitates multimodal reasoning by interleaving relevant visual regions within reasoning steps. At the core of the MINT-CoT is the Interleave Token, a special token generated through the next-token prediction process. During reasoning, MINT-CoT automatically identifies and incorporates the most relevant visual tokens from the original image at each reasoning step. This is achieved by computing similarity scores between the output hidden states of the Interleave Token and all visual tokens, in order to identify the tokens most relevant to the mathematical concept at the current step. These selected visual tokens are then dynamically integrated into the textual reasoning steps, enabling the flexible selection of visual regions throughout the CoT process. In this way, the interleaved regions of mathematical images are not restricted to box-shaped areas but can flexibly include geometric shapes, line segments, coordinates, and other elements.

To enable effective training of MINT-CoT, we construct the MINT-CoT dataset, a 54K visual interleaved reasoning dataset. Each data point contains reasoning steps paired with the indices of selected tokens corresponding to the mathematical concepts involved in each step. We source mathematical problems from the Mulberry-260K dataset [74] to construct text-only CoT reasoning format, then annotate the reasoning steps with corresponding image regions through a four-step pipeline: (1) dividing images into grid-indexed regions, (2) mapping recognized text elements to grid indices via OCR-based text localization, (3) extracting key words, and (4) assigning visual regions to these key words using an advanced MLLM. This process creates a visual interleaved CoT reasoning dataset providing token-level supervision for training models to interleave visual content into reasoning steps.

Building on the MINT-CoT framework and MINT-CoT dataset, we design a progressive training strategy, the MINT-CoT training strategy, that incrementally improves MLLMs’ ability with three training stages: (1) Text-only CoT Training, (2) Interleaved CoT SFT, and (3) Interleaved CoT RL. Through this training strategy, we train a MINT-CoT-7B model with the capability of mathematical visual interleaved CoT reasoning. Extensive experiments demonstrate the superiority of our proposed approach. Specifically, our method achieves absolute improvement of +32.59% on MathVista [43], +26.92% on GeoQA [5], and +23.2% on MMStar [7] benchmark compared to the baseline model.

Our main contributions are as follows:

- We propose MINT-CoT, which uses the Interleave Token to interleave fine-grained visual tokens within reasoning steps, enhancing multimodal mathematical reasoning.
- We construct the MINT-CoT dataset, a 54K dataset for multimodal mathematical reasoning, offering fine-grained alignment between textual rationales and visual inputs. We develop an automated pipeline to generate visual interleaved CoT data annotated with token indices.
- We develop a progressive three-stage MINT-CoT training strategy, to improve interleaved mathematical reasoning. Extensive experiments validate the efficiency of our method.

2 Related work

MLLMs for Mathematics. Recent advancements in MLLMs [50, 41, 2, 31] have shown impressive capabilities in various vision-language tasks. However, even powerful models like GPT-4V [50] and Qwen2-VL [63] fail to demonstrate satisfying performance on existing visual mathematical benchmarks [5, 44, 43], as highlighted by MathVerse [82]. Various specialized approaches [15, 83, 28, 9, 45, 57, 53] have emerged to enhance visual mathematical reasoning. Current approaches mostly focus on enriching the multimodal math data. G-LLaVA [15] extends the LLaVA architecture

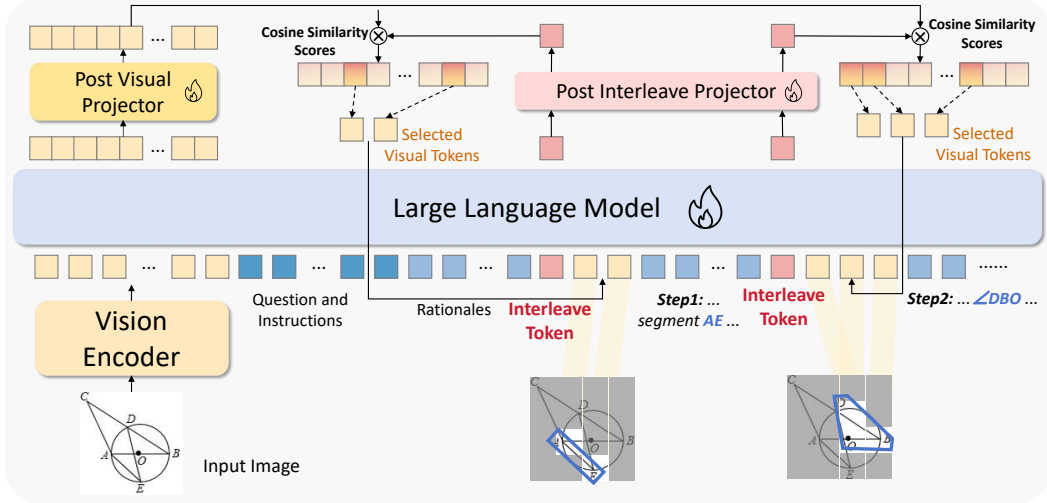


Figure 2: **Overview of the MINT-CoT framework.** During CoT reasoning, MINT-CoT generates an Interleave Token before each reasoning step and computes the similarity scores between embeddings projected by the decoder-side visual projector and the interleave projector. Based on these similarity scores, relevant visual tokens are selected, and the model infers with these selected visual tokens.

with geometric reasoning capabilities by augmenting the current dataset. Math-LLaVA [57] enlarges the data scope with the introduced MathV360K dataset. MAVIS [83] 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) [9] introduces the Geometry Generation Chain for creating geometric images with more accurate descriptions.

Visual Chain of Thought. With advancements of various visual reasoning tasks [43, 77, 30, 75, 69], visual chain of thought has been emerging as an effective method for both image generation [23, 29, 61, 87] and understanding [52, 74, 60] tasks. Our work focuses on leveraging it for reasoning on images, where two distinct methods have emerged. One line of the method relies on textual CoT to conduct multimodal analysis [11, 46, 6, 79, 10, 73]. For example, R1-V [6] extends the paradigm of DeepSeek R1 [19] 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 rationale [55, 47, 67, 26, 35]. Visual CoT [55] and Chain-of-Spot [42] propose to crop the region of high interest on the image and integrate it into the CoT process. Chain-of-Image [47] and Visual SKETCHPAD [26] 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 [16] leverages the attention map of the MLLM to select the relevant visual tokens to compose the multimodal rationale. However, this approach relies solely on attention scores on the image feature maps, which have been shown to be insufficiently informative for mathematical scenarios [83].

3 Method

To address the challenges of multimodal CoT in mathematical reasoning, we propose MINT-CoT. In this section, we first introduce the framework of MINT-CoT in Section 3.1. Then we introduce the MINT-CoT dataset and provide a detailed discussion of the dataset generation method in Section 3.2. Finally, we present the progressive MINT-CoT training strategy in Section 3.3.

3.1 MINT-CoT

Previous CoT approaches in MLLMs mainly generate text-based reasoning steps, which are not explicitly grounded in visual features and therefore struggle with mathematical reasoning that involves

visual details. We formulate this CoT reasoning process as:

$$\{s^{(1)}, s^{(2)}, \dots, s^{(k)}\}, \text{answer} = \text{LLM}(V, \text{TextEncoder}(T)), \quad (1)$$

where $V = \text{VisionEncoder}(I) = \{v_\tau\}_{\tau=1}^N$ denotes the visual feature extracted from the input image I , and each v_τ represents the τ -th visual token generated by the vision encoder. T denotes the input mathematical question and instructions, $\{s^{(i)}\}$ is the sequence of textual reasoning steps generated by the model, and answer is the final answer. Recent advancements attempt to incorporate multimodal reasoning steps in the CoT process. However, current coarse-grained methods only focus on selecting box-shaped visual regions; how to adaptively select the visual content in alignment with each textual reasoning step remains an open question. We thus propose the **MINT-CoT** framework and introduce an Interleave Token to help MLLMs select visual tokens from the visual feature V . The overview of the MINT-CoT framework is illustrated in Figure 2.

Interleave Token. An Interleave Token is a special token generated prior to each reasoning step. It is used to select visual tokens that are relevant to the mathematical concepts involved in that step (e.g., “line segment AB”, “angle DOC”), thereby facilitating the reasoning process. When an Interleave Token is output in step i , its output hidden state $h_{\text{post_intlv}}^{(i)}$ is projected via a post interleave projector $P_{\text{post_intlv}}$, while all the output hidden states of the visual tokens $h_{\text{post_vis}}$ are projected via a post visual projector $P_{\text{post_vis}}$. The cosine similarity between the two projected embeddings is first computed and then scaled by a learnable parameter γ :

$$\alpha^{(i)} = \gamma \cdot \cos \left(P_{\text{post_intlv}}(h_{\text{post_intlv}}^{(i)}), P_{\text{post_vis}}(h_{\text{post_vis}}) \right). \quad (2)$$

Each tokens’ similarity score $\alpha_\tau^{(i)}$ is then compared against a predefined threshold θ , and visual tokens with scores above this threshold are selected:

$$\{v^{(i)}\} = \{v_\tau^{(i)} \mid \alpha_\tau^{(i)} > \theta\}. \quad (3)$$

The selected tokens $\{v^{(i)}\}$ are interleaved into the reasoning process at step i . In this way, the important visual regions are interleaved into the model, prior to each textual step, enhancing visual perception and improving reasoning accuracy.

Inference with Interleaved Visual Tokens. With the selected visual tokens $\{v^{(i)}\}$ obtained at each reasoning step, MINT-CoT interleaves both visual content and text-based reasoning steps throughout the inference process, ultimately producing the final answer. Formally, this process extends the standard CoT formulation in Eq. 1 as:

$$\{v^{(1)}, s^{(1)}, v^{(2)}, s^{(2)}, \dots, v^{(k)}, s^{(k)}\}, \text{answer} = \text{LLM}(V, \text{TextEncoder}(T)). \quad (4)$$

This interleaved token selection mechanism enables the model to explicitly ground visual evidence throughout the reasoning chain, thereby facilitating visual interleaved CoT reasoning for solving multimodal mathematical problems.

3.2 Dataset Curation

To empower MINT-CoT capabilities for MLLMs, we develop a data generation pipeline that automatically generates mathematical visual interleaved data annotated with selected token indices, and obtain 54K samples for model training. To construct the text-only cot format of our dataset, we begin by selecting mathematical problems from the Mulberry-260K dataset [74], which was created using Collective Monte Carlo Tree Search and demonstrates strong performance on reasoning tasks. Specifically, we extract the “### Rationale” and “### Steps” sections from the dataset as the reference reasoning steps for our task. Using these sections alongside the corresponding images, we follow a four-step data construction process, as shown in Figure 3:

1. **Grid Images.** To obtain the indices of visual tokens for subsequent token index annotation in textual reasoning steps, we divide the original images into grid cells. Following the patch-splitting strategy used in vision encoders such as Vision Transformer [12], each image is partitioned into a grid, and a unique index is assigned to each cell. These grid cells and their indices are subsequently overlaid onto the original images to produce *grid-indexed images*.

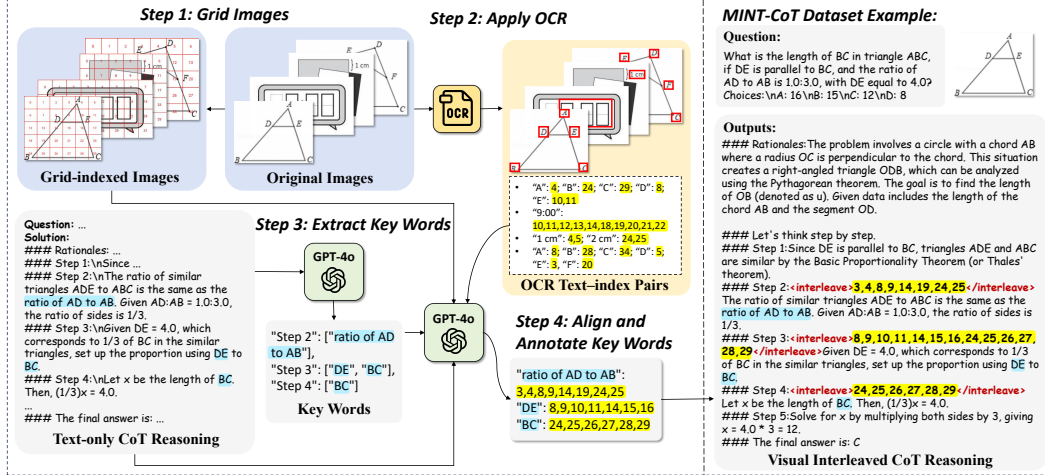


Figure 3: **Data generation pipeline.** *Step 1: Grid Images.* We divide each image into grid cells and assign index values to each cell. *Step 2: Apply OCR.* We use PaddleOCR to recognize textual elements and associate them with corresponding grid indices. *Step 3: Extract Key Words.* We employ GPT-4o to extract key words from each reasoning step. *Step 4: Align and Annotate Key Words.* We use GPT-4o to annotate each key word with the grid indices, and get the final visual interleaved CoT reasoning steps.

2. **Apply OCR.** Then, to more accurately map token indices onto textual reasoning steps, we apply PaddleOCR [37] to recognize textual elements in the original images. And we align the bounding boxes of the detected text with their corresponding grid indices, thereby constructing “OCR text-index” pairs.
3. **Extract Key Words.** Certain mathematical concepts often play a significant role in each reasoning step. Selecting visual tokens closely related to these concepts can improve reasoning accuracy. Therefore, we employ GPT-4o [12] to extract *key words* from each reasoning step. Since the extracted key words are used in the subsequent annotation with visual indices, they are extracted only when a reasoning step contains links to visual tokens.
4. **Align and Annotate Key Words.** Finally, given the grid-indexed images, the “#### Rationale” and “#### Steps” sections, the “OCR text-index” pairs, and the extracted key words, we prompt GPT-4o to annotate each key word with the corresponding grid indices. These *annotated indices* are subsequently inserted into the reasoning steps associated with their corresponding key words, resulting in a visual-interleaved CoT reasoning dataset.

Through this process, we construct a dataset of 54K samples, where the reasoning steps are annotated with corresponding grid indices. As shown in the right column of Figure 3, each data point consists of a mathematical problem and an image as input, with the corresponding visual interleaved CoT response as output. This dataset serves as the foundation for training the MINT-CoT models. Further details are provided in Appendix A.2.

3.3 Training strategy

Building on the previously introduced MINT-CoT framework and dataset, we now describe the corresponding MINT-CoT training strategy, which consists of three stages: (1) Text-only CoT Training, (2) Interleaved CoT SFT, and (3) Interleaved CoT RL.

Stage 1: Text-only CoT SFT. To enable the MLLM to adopt a general reasoning format, we first train the base model using the text-only CoT reasoning data in MINT-CoT dataset, without visual interleaving. This stage serves as a foundation for subsequent interleaved training.

Stage 2: Interleaved CoT SFT. In the second stage, we aim to train the model to select visual tokens using the Interleave Token and adapt to reasoning with interleaved visual content. The model

is fine-tuned with a loss that jointly optimizes both textual reasoning and visual alignment. As introduced in Eq. 4, the output sequence of MINT-CoT alternates between sets of selected visual tokens $v^{(i)}$ and textual reasoning steps $s^{(i)}$, followed by the final answer:

$$\{v^{(1)}, s^{(1)}, v^{(2)}, s^{(2)}, \dots, v^{(k)}, s^{(k)}\}, \text{answer} \sim P_\theta(\cdot \mid I, T), \quad (5)$$

We first apply a cross-entropy loss to textual tokens at positions $\mathbf{T} \subset \{1, 2, \dots, T\}$ covering all segments $\{s^{(i)}\}$ and the answer, while conditioning on the full preceding sequence. Let $Y = \{y_1, y_2, \dots, y_T\}$ denotes the full sequence of output tokens. Specifically, the loss for predicting the next textual token is defined as:

$$\mathcal{L}_{\text{CE}} = - \sum_{t \in \mathbf{T}} \log P_\theta(y_t \mid y_{<t}, I, T) \quad (6)$$

We do not supervise the cross-entropy loss for predicting the Interleave token. Instead, we manually concatenate it at each step, and during inference, we concatenate the Interleave Token whenever the “### Step” marker is generated. To supervise the interleaved visual tokens, we apply a binary cross-entropy loss on the scaled cosine similarity scores α introduced in Eq. 2 with ground-truth labels $X \in \{0, 1\}$:

$$\mathcal{L}_{\text{BCE}} = - \sum_{i=1}^N \sum_{j=1}^L \left(X_{ij} \log \sigma(\alpha_{ij}) + (1 - X_{ij}) \log(1 - \sigma(\alpha_{ij})) \right), \quad (7)$$

where N is the number of Interleaved Tokens in a batch, L is the length of input visual tokens, and $\sigma(\cdot)$ denotes the sigmoid function. The final training objective is defined as the sum of both losses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{BCE}}. \quad (8)$$

This combined loss guides the model to jointly align visual tokens and perform interleaved reasoning.

Stage 3: Interleaved CoT RL. To move beyond supervised annotations, we aim to enable the model to autonomously explore more flexible and effective selection of visual tokens guided by reasoning objectives, and enhance its ability to perform interleaving CoT reasoning. Reinforcement learning provides a natural framework for this goal. To this end, we extend the Group Relative Policy Optimization (GRPO) [56] framework to our MINT-CoT training strategy. For a group of reasoning chains with group size G , we compute answer correctness as the reward $r \in \{0, 1\}$ and define the advantage via group-wise comparison as $\hat{A}_j = \frac{r_j - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$, where r_j indicates if the j -th chain of steps in a group yields the correct answer. The policy loss for the generated tokens is then formulated as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{\{Y_j\}_{j=1}^G} \left[\frac{1}{G} \sum_{j=1}^G \left(\frac{P_\theta(Y_j)}{P_{\theta_{\text{old}}}(Y_j)} \hat{A}_j - \beta D_{\text{KL}}[P_\theta \parallel P_{\text{ref}}] \right) \right], \quad (9)$$

where P_{ref} is a reference policy that serves as a regularization target. This stage further strengthens the model’s reasoning ability with visual interleaved content, ultimately resulting in MINT-CoT-7B. Additional theoretical details of this training stage are provided in Appendix A.3.

4 Experiments

In this section, we first introduce the experimental settings in Section 4.1. Then, we discuss the quantitative results and ablation study in Section 4.2 and Section 4.3 respectively. Finally, we present the qualitative results in Section 4.4.

4.1 Experimental Settings

Implementation Details. We build on Qwen2-VL-7B [64] and train our model in three stages with a combination of SFT and RL on the MINT-CoT dataset. All model parameters except the vision encoder are updated. Full implementation details are provided in Appendix A.4.

Test Benchmark. We evaluate MINT-CoT on three mathematical benchmarks: GeoQA [5], MathVista [43] and MMStar [7]. GeoQA is a benchmark of geometric problems with annotated solution programs. To evaluate on GeoQA, we follow R1-V [6] and Hint-GRPO [27] using the Geo170K test set [15], the English version of the GeoQA benchmark. MathVista is a benchmark designed to

Table 1: **Combined quantitative results on MathVista.** We evaluate MINT-CoT-7B, the baseline model, and state-of-the-art general and reasoning MLLMs on the mathematical subset of MathVista. MINT-CoT significantly outperforms the baseline model and achieves superior performance compared to open-source reasoning models. **Bold** and underlined results indicate the best and second-best among open-source models, respectively.

Model	#Params	MathVista-Math				
		All	GEO	ALG	GPS	TQA
<i>Closed-Source Model</i>						
GPT-4o [48]	–	66.67	63.68	67.04	63.46	77.42
Claude-3.5 Sonnet [1]	–	67.41	65.09	67.79	65.38	74.19
<i>Open-Source General Model</i>						
LLaVA-OneVision-Qwen2-7b-ov [34]	7B	67.04	69.34	67.04	69.71	58.06
InternVL2-8B [8]	8B	62.59	62.26	62.92	62.50	62.90
InternVL2-8B-MPO [65]	8B	68.52	68.87	68.91	69.71	64.52
DeepSeek-VL2 [68]	4.5B	65.56	63.68	65.54	63.94	70.97
Qwen2.5-VL-7B-Instruct [3]	7B	66.66	65.56	66.29	65.87	69.35
<i>Open-Source Reasoning Model</i>						
Open-R1-Multimodal [13]	7B	54.81	52.36	54.68	53.37	59.68
R1-VL-7B [80]	7B	69.63	68.87	69.66	69.71	69.35
Mulberry [74]	7B	68.52	67.92	68.54	68.75	67.74
MM-Eureka [46]	7B	<u>72.59</u>	<u>71.22</u>	<u>72.66</u>	<u>72.60</u>	72.58
Qwen2-VL-7B-Instruct [64] (Baseline)	7B	41.11	35.85	41.57	36.54	56.45
MINT-CoT-7B	7B	73.70	74.53	73.78	75.00	69.35
Δ over the Baseline Model		+32.59	+38.63	+32.21	+38.46	+12.9

Table 2: **Combined quantitative results of on GeoQA.** We evaluate MINT-CoT-7B, the baseline model and the state-of-the-arts.

Model	GeoQA
Qwen2.5-VL-7B-Instruct [3]	43.50
R1-V [6]	<u>59.00</u>
Open-R1-Multimodal [13]	48.67
Hint-GRPO [27]	55.31
Qwen2-VL-7B-Instruct [64] (Baseline)	37.80
MINT-CoT-7B	64.72
Δ over the Baseline Model	+26.92

Table 3: **Combined results on the mathematical subset of MMStar.** We evaluate MINT-CoT-7B, the baseline model and the state-of-the-arts.

Model	MMStar-Math
Qwen2.5-VL-7B-Instruct [3]	66.8
InternVL2-8B [8]	66.8
R1-VL-7B [79]	<u>68.4</u>
Mulberry [74]	66.8
Open-R1-Multimodal [13]	59.2
Qwen2-VL-7B-Instruct [64] (Baseline)	46.4
MINT-CoT-7B	69.6
Δ over the Baseline Model	+23.2

integrate challenges from diverse mathematical and visual tasks. As our paper targets specifically mathematical problems, we extract the mathematical subsets (FunctionQA, Geometry3K, GeoQA+, GEOS, and UniGeo), i.e., ‘MathVista-Math’ in Table 1, and report accuracy scores across four primary tasks: geometry reasoning (GEO), algebraic reasoning (ALG), geometry problem solving (GPS), and textbook question answering (TQA). MMStar is a multi-modal benchmark covering different core capabilities and detailed axes. For evaluation, we also extract the mathematical capability dimension, referred to as “MMStar-Math”.

4.2 Quantitative Results

Comparison with the Baseline. As shown in Table 1 for the results of mathematical subsets of MathVista, our MINT-CoT-7B achieves an improvement of up to +32.59% over the baseline, and improves a lot on all four primary tasks. This strongly demonstrates the effectiveness of our MINT-CoT framework and training strategy. Table 2 presents the results on the GeoQA benchmark, where our MINT-CoT-7B outperforms the baseline model by +26.92%. Similarly, in Table 3, MINT-CoT-7B outperforms the baseline model by +23.2% on MMStar-Math, validating the efficiency of MINT-CoT on geometry problems.

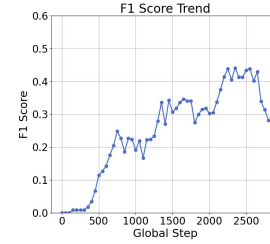
Table 4: **Ablation study on different training stages.** We evaluate the three progressive training stages on different benchmarks.

Model	MMStar-Math	GeoQA	MathVista-Math				
			All	GEO	ALG	GPS	TQA
Baseline [64]	46.4	37.80	41.11	35.85	41.57	36.54	56.45
+ Text-only CoT SFT	67.6	59.02	64.07	64.15	64.04	64.42	62.90
+ Interleaved CoT SFT	<u>68.0</u>	<u>62.07</u>	<u>67.78</u>	<u>66.51</u>	<u>67.79</u>	<u>67.31</u>	69.35
+ Interleaved CoT RL	69.6	64.72	73.70	74.53	73.78	75.00	69.35

Table 5: **Ablation study of different interleaving methods on GeoQA and MathVista-Math.** Our Interleaved CoT SFT achieves the highest improvement on both benchmarks, demonstrating the effectiveness of our interleaved token selection method.

Model	GeoQA	MathVista-Math				
		All	GEO	ALG	GPS	TQA
Original	37.80	41.11	35.85	41.57	36.54	56.45
Text-only CoT SFT	59.02	64.07	<u>64.15</u>	64.04	<u>64.42</u>	62.90
Original Image CoT SFT	61.41	40.37	38.68	40.82	39.42	43.54
Bounding Box CoT SFT	<u>61.80</u>	<u>65.56</u>	63.21	<u>65.54</u>	63.94	70.97
Interleaved CoT SFT (Ours)	62.07	67.78	66.51	67.79	67.31	<u>69.35</u>

Figure 4: F1 score plot of visual token selection during Interleaved CoT SFT.



Comparison with State-of-the-arts. We also compare our model with state-of-the-art MLLMs, including closed-source model, open-source models, and open-source reasoning models. Specifically, for open-source reasoning models, we choose recent works like R1-VL-7B [79], MM-Eureka [46] and Open-R1-Multimodal [13]. As shown in Table 1, our model achieves the highest overall accuracy on the MathVista mathematical subsets, outperforming both open-source reasoning models and general models, and surpassing the best-performing open-source MLLM by +1.11% as well as closed-source models, demonstrating strong capabilities in mathematical reasoning. On geometry reasoning, geometry problem solving and algebraic reasoning, MINT-CoT-7B outperforms state-of-the-art models by +3.31%, +1.12%, and +2.4%, respectively. However, for textbook question answering, our performance is slightly below MM-Eureka. On the GeoQA benchmark, as shown in Table 2, our model outperforms the state-of-the-art models by +5.72%. In Table 3, MINT-CoT-7B also outperforms the state-of-the-art by +1.2% on MMStar-Math, further demonstrating its capability in geometry reasoning.

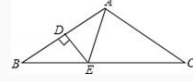
4.3 Ablation Study

Training Stage Ablation. We conduct an ablation study on the different training stages of MINT-CoT, as described in Section 3.3. The results on different benchmarks are presented in Table 4. The Text-only CoT SFT stage improves performance by +21.2% on MMStar-Math, +21.22% on GeoQA, and +22.96% on MathVista-Math, as it helps the model learn the general reasoning format illustrated in the left column of Figure 3. The Interleaved CoT SFT stage further boosts performance by +0.4% on MMStar-Math, +3.05% on GeoQA, and +3.71% on MathVista-Math across all primary tasks by enabling the model to interleave visual tokens into textual reasoning steps. Finally, the Interleaved CoT RL stage enhances performance by an additional +1.6% on MMStar-Math, +2.65% on GeoQA, and +5.92% on MathVista-Math through reinforcement learning, which enables the model to reason more effectively with interleaved tokens.

Interleaving Method Ablation. We conduct an ablation study on the interleaving method used in the Interleaved CoT SFT stage, with the results presented in Table 5. Starting with the model trained in the Text-only CoT SFT stage, we simply interleave the original image into each reasoning step without the use of projectors or the Interleave token structure, which we refer to as “Original Image CoT SFT”. We find that, on MathVista-Math, the performance of Original Image CoT SFT significantly decreases compared to Text-only CoT SFT. On the GeoQA benchmark, it also underperforms our Interleaved CoT SFT. This decline is likely due to the interleaving of excessive unrelated visual tokens during reasoning. Furthermore, we train a model that uses the Interleave token to select a

Question:

Given triangle ABC where $AB = AC$ and angle BAC measures 100.0 degrees, and the perpendicular bisector of AB intersects AB and BC at points D and E, what is the measure of angle BAE? Represent the measure of angle BAE as θ . Choices: A: 80° B: 60° C: 50° D: 40°



Qwen2vl-7B-Instruct

To solve this problem, we first need to visualize the triangle and the given information:\n\n1. Triangle ABC is isosceles with $AB = AC$ and angle $BAC = 100.0$ degrees.\n2. The perpendicular bisector of AB intersects AB at point D and intersects BC at point E.\n\nSince $AB = AC$, triangle ABC is isosceles as well. The perpendicular bisector of AB implies that **AE is the median to the hypotenuse of triangle ABC** (because D is the midpoint of AB). \n\nAlso, in an isosceles triangle, if a median is drawn to the hypotenuse, it creates two right-angled triangles. Let's denote the measure of angle BAE as θ . Due to the properties of the isosceles triangle and the perpendicular bisector, **angle BAE (θ) should be half of the apex angle BAC**. \n\nSince $\angle BAC = 100.0$ degrees, $\theta = 100.0 / 2 = 50.0$ degrees. \n\nTherefore, the correct answer is C. ❌

MINT-CoT-7B

Rationales: The problem involves a geometry question about angles in an isosceles triangle with certain line intersections. To find angle BAE, consider properties of isosceles triangles and angle bisectors.
Let's think step by step.

- ### Step 1:  Since triangle ABC is isosceles with $AB = AC$, angles ABC and ACB are equal.
- ### Step 2:  Calculate angle ABC and ACB. The sum of angles in triangle ABC is 180 degrees. Therefore, angle ABC = angle ACB = $(180 - \text{angle BAC}) / 2 = (180 - 100) / 2 = 40$ degrees.
- ### Step 3:  The perpendicular bisector of AB implies that point E is equidistant from A and B, making $BE = AE$. The bisector makes angle BAE an angle in triangle BAE.
- ### Step 4:  Apply angular considerations: Consider that both AE and BE are equal, and with angle BAC = 100 degrees and angle ABC = 40 degrees, angle BAE will share properties of an isosceles triangle and angle relationships.
- ### The final answer is: D ✓



Cosine Similarity Scores

Figure 5: Qualitative results of Qwen2-VL-7B-Instruct and MINT-CoT-7B. MINT-CoT-7B demonstrates improved CoT reasoning capability by interleaving fine-grained visual tokens. There is also a visualization of the similarity scores for the Interleaved Token generated during Step 4.

rectangular region of visual tokens at each reasoning step, referred to as “Bounding Box CoT SFT”. As shown in the table, this approach underperforms our Interleaved CoT SFT on both benchmarks, except for the TQA task, and even underperforms the Text-only CoT SFT on GEO and GPS tasks in MathVista-Math. These results demonstrate the effectiveness of our token selection method for mathematical reasoning tasks.

4.4 Qualitative Results

We present the qualitative results of the baseline model Qwen2-VL-7B-Instruct and our proposed model MINT-CoT-7B, as shown in Figure 5. Compared to the baseline, MINT-CoT-7B demonstrates a more coherent reasoning format and is capable of selecting and interleaving relevant visual tokens during inference. More qualitative results of our model are shown in Appendix A.6. Moreover, we provide a plot of the average F1 score between the selected visual tokens and ground truth visual tokens in each reasoning step during the Interleaved CoT SFT stage, as shown in Figure 4. For the Interleaved CoT RL stage, we do not report an F1 score plot due to the absence of ground truth visual token indices for online inference. As shown in the plot, the F1 score exhibits a fluctuating upward trend during training, demonstrating that the accuracy of visual token selection is increasing during the Interleaved CoT SFT training strategy.

5 Conclusion

In this paper, we first propose MINT-CoT, a method for enhancing multimodal mathematical reasoning by interleaving fine-grained visual tokens into CoT. We use the novel Interleave Token to automatically select visual tokens for each reasoning step. Then, we introduce the MINT-CoT dataset and a four-step dataset generation pipeline. Finally, we present the MINT-CoT training strategy, which includes Text-only CoT Training, Interleaved CoT SFT and Interleaved CoT RL, enhancing the MLLMs’ ability to reason over interleaved visual tokens. Our experiments with the obtained MINT-CoT-7B model demonstrate significant improvements across various benchmarks.

Acknowledgements

This study was supported in part by National Key R&D Program of China Project 2022ZD0161100, in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, in part by NSFC-RGC Project N_CUHK498/24, and in part by Guangdong Basic and Applied Basic Research Foundation (No. 2023B1515130008, XW).

References

- [1] Sonnet Anthropic. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet.
- [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*, abs/2308.12966, 2023.
- [3] 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.
- [4] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang 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.
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

- [13] 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.
- [14] Chaoyou Fu, Yuhang Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [15] 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.
- [16] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought, 2025.
- [17] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [18] Google Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [19] 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.
- [20] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [21] Zilu Guo, Hongbin Lin, Zhihao Yuan, Chaoda Zheng, Pengshuo Qiu, Dongzhi Jiang, Renrui Zhang, Chun-Mei Feng, and Zhen Li. Pisa: A self-augmented data engine and training strategy for 3d understanding with large models. *arXiv preprint arXiv:2503.10529*, 2025.
- [22] Ziyu Guo, Ray Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of llms on multi-modal scientific problems. *arXiv preprint arXiv:2503.10627*, 2025.
- [23] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Peng Gao, Hongsheng Li, and Pheng-Ann Heng. Can we generate images with cot? let’s verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025.
- [24] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.
- [25] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.
- [26] 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.
- [27] Qihan Huang, Long Chan, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo, 2025.
- [28] Zihan Huang, Tao Wu, Wang Lin, Shengyu Zhang, Jingyuan Chen, and Fei Wu. Autogeo: Automating geometric image dataset creation for enhanced geometry understanding. *arXiv preprint arXiv:2409.09039*, 2024.
- [29] Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. *arXiv preprint arXiv:2505.00703*, 2025.

- [30] 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. Mmcot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency, 2025.
- [31] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.
- [32] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] 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, 2025.
- [37] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system, 2022.
- [38] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- [39] Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. Adaptive classifier-free guidance via dynamic low-confidence masking. *arXiv preprint arXiv:2505.20199*, 2025.
- [40] 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. *ECCV 2024*, 2023.
- [41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujie Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv preprint arXiv:2501.04686*, 2025.

- [46] 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.
- [47] Fanxu Meng, Haotong Yang, Yiding Wang, and Muhan Zhang. Chain of images for intuitively reasoning. *arXiv preprint arXiv:2311.09241*, 2023.
- [48] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codisputi, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O’Connell, Ian O’Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming

Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunningham, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiye Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024.

- [49] OpenAI. Chatgpt. <https://chat.openai.com>, 2023.
- [50] OpenAI. GPT-4V(ision) system card, 2023.
- [51] OpenAI. Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/>, 2024.
- [52] OpenAI. Introducing openai o1, 2024., 2024.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] 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.
- [57] 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.
- [58] Yiwen Tang, Zoey Guo, Zhuohao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Xuelong Li, et al. Exploring the potential of encoder-free architectures in 3d llms. *arXiv preprint arXiv:2502.09620*, 2025.
- [59] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [60] Qwen Team. Qvq-72b-preview. <https://huggingface.co/Qwen/QVQ-72B-Preview>, 2025. Accessed: 2025-05-13.
- [61] Chengzhuo Tong, Ziyu Guo, Renrui Zhang, Wenyu Shan, Xinyu Wei, Zhenghao Xing, Hongsheng Li, and Pheng-Ann Heng. Delving into rl for image generation with cot: A study on dpo vs. gpo. *arXiv preprint arXiv:2505.17017*, 2025.
- [62] 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.

- [63] 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.
- [64] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024.
- [65] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [67] Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind’s eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [68] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.
- [69] Han Xiao, Guozhi Wang, Yuxiang Chai, Zimu Lu, Weifeng Lin, Hao He, Lue Fan, Liuyang Bian, Rui Hu, Liang Liu, et al. Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. *arXiv preprint arXiv:2505.21496*, 2025.
- [70] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023.
- [71] Shilin Yan, Jiaming Han, Joey Tsai, Hongwei Xue, Rongyao Fang, Lingyi Hong, Ziyu Guo, and Ray Zhang. Crosslmm: Decoupling long video sequences from llms via dual cross-attention mechanisms. *arXiv preprint arXiv:2505.17020*, 2025.
- [72] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.
- [73] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.
- [74] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*, 2024.
- [75] Junyan Ye, Dongzhi Jiang, Jun He, Baichuan Zhou, Zilong Huang, Zhiyuan Yan, Hongsheng Li, Conghui He, and Weijia Li. Blink-twice: You see, but do you observe? a reasoning benchmark on visual perception. *arXiv preprint arXiv:2510.09361*, 2025.

- [76] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Introducing visual perception token into multi-modal large language model. *arXiv preprint arXiv:2502.17425*, 2025.
- [77] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [78] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.
- [79] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.
- [80] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization, 2025.
- [81] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *ICLR 2024*, 2024.
- [82] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.
- [83] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. Mavis: Mathematical visual instruction tuning with an automatic data engine, 2024.
- [84] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [85] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023.
- [86] 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.
- [87] Le Zhuo, Liangbing Zhao, Sayak Paul, Yue Liao, Renrui Zhang, Yi Xin, Peng Gao, Mohamed Elhoseiny, and Hongsheng Li. From reflection to perfection: Scaling inference-time optimization for text-to-image diffusion models via reflection tuning. *arXiv preprint arXiv:2504.16080*, 2025.
- [88] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The three main conclusions and the scope of our paper are detailed in the final section of the introduction and summarized in the abstract.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations of the work in Appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All the assumptions and theorems are clearly stated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The architecture details are provided in the Methodology and Experiment Section for reproducing the main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data are currently available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The training and test details are provided in the Experiment Section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We conduct experiments only once and report the accuracy of the best model, and it would be too computationally expensive to conduct the pre-training multiple times.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The compute resources are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Our research conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: There is no societal impact of the work performed.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers or websites that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We details the data and model in the experiment section, appendix, and GitHub.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: This paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Appendix

A.1 Overview

We organize our supplementary material as follows.

- Dataset Details
 - Dataset Example
 - Dataset Statistic
- Theoretical Details of Interleaved CoT RL
- Additional Implementation Details
- Additional Ablation Study
 - Text-only Method Ablation
 - Results on Other Benchmark
 - Projector Ablation
- Additional Qualitative Results
- Limitations

A.2 Dataset Details

Dataset Example We present examples from our MINT-CoT Dataset in Figures 6 to 8, 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 6. This dataset comprises 54,031 data points derived from the mathematical portion of the Mulberry-260k dataset.

Table 6: **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

A.3 Theoretical Details of Interleaved CoT RL

Following the standard GRPO framework [56], we integrate GRPO into our approach. Specifically, similar to $\mathcal{L}_{\text{CE}}$ in Stage 2, we apply a policy loss $\mathcal{L}_{\text{GRPO_text}}$ to textual tokens:

$$\mathcal{L}_{\text{GRPO_text}} = -\mathbb{E}_{\{Y_j\}_{j=1}^G \sim P_{\theta_{\text{old}}}(\cdot|I, T)} \left[\frac{1}{G} \sum_{j=1}^G \frac{1}{|\mathbf{T}_j|} \sum_{t \in \mathbf{T}_j} \left\{ \frac{P_{\theta}(y_{j,t}|y_{j,<t}, I, T)}{P_{\theta_{\text{old}}}(y_{j,t}|y_{j,<t}, I, T)} \cdot \hat{A}_{j,t} - \beta D_{\text{KL}}[P_{\theta} \parallel P_{\text{ref}}] \right\} \right], \quad (10)$$

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 $D_{\text{KL}}[P_{\theta} \parallel P_{\text{ref}}]$ 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 $\mathcal{L}_{\text{GRPO_vis}}$ 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 N_j denote the the number of

Table 7: **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

reasoning steps in j -th chain, and $M_j^{(i)}$ denote the number of visual tokens interleaved in the i -th reasoning step in the j -th chain. Formally, the loss is defined as:

$$\mathcal{L}_{\text{GRPO_vis}} = -\mathbb{E}_{\{Y_j\}_{j=1}^G \sim P_{\theta_{\text{old}}}(\cdot|I,T)} \left[\frac{1}{G} \sum_{j=1}^G \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{1}{M_j^{(i)}} \sum_{\tau=1}^{M_j^{(i)}} \left\{ \frac{P_{\theta}(\alpha_{j,\tau}^{(i)}|y_{j,<\tau},I,T)}{P_{\theta_{\text{old}}}(\alpha_{j,\tau}^{(i)}|y_{j,<\tau},I,T)} \cdot \hat{A}_j - \beta D_{\text{KL}}[P_{\theta} \parallel P_{\text{ref}}] \right\} \right]. \quad (11)$$

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

$$\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\text{GRPO_text}} + \lambda \cdot \mathcal{L}_{\text{GRPO_vis}}. \quad (12)$$

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

A.4 Additional Implementation Details

We use Qwen2-VL-7B [64] as the base MLLM model in our experiments. Each of the two projectors, $P_{\text{interleave}}$ and P_{vis} , is implemented as a single linear layer. We uniformly set the threshold $\theta = 0.7$ to filter the similarity scores. The hyper-parameter γ to scale the similarity is set to $1/0.07$ following CLIP [54]. The training procedure consists of three stages: (1) *Text-only CoT Training*, where we train for 2 epochs on the MINT-CoT dataset without applying the interleaving strategy, using a learning rate of $5.0\text{e-}6$ and a batch size of 64, following the configuration of Mulberry [74]; (2) *Interleaved CoT SFT*, where we train for 3 epochs on the MINT-CoT dataset with a learning rate of $1\text{e-}6$ and a batch size of 64; and (3) *Interleaved CoT RL*, where we train for 700 steps on the MINT-CoT dataset, 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 16. During training, all model parameters, including the Interleave Token and projector layers, are unfrozen, except for the vision encoder, which remains fixed. Finally, the resulting model is named MINT-CoT-7B.

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

A.5 Additional Ablation Study

Text-only Method Ablation We conduct an ablation study using a text-only training pipeline. This involves first applying a Text-only CoT SFT stage to the baseline model, followed by a Text-only CoT RL stage using the GRPO framework. As shown in Table 8, MINT-CoT still outperforms this setting by 2.96% on the MathVista mathematical subsets, showcasing the superiority of our Interleaved CoT approach.

Results on Other Benchmark In addition to the quantitative results in Section 3.2, we provide results of the different training stages of our method on the vision-only input setting of the Mathematics section of MMMU-Pro [77] benchmark. As shown in Table 9, MINT-CoT-7B outperforms the baseline model by +5%. and the performance improves with the Interleaved CoT approach, validating the efficiency of our method.

Table 8: Ablation study of the text-only training method.

Model	MathVista-Math
Baseline	41.11
+ Text-only CoT SFT	64.07
+ Text-only CoT RL	70.74
+ Interleaved CoT SFT	67.78
+ Interleaved CoT RL	73.70

Table 9: Ablation study of different training stages on the Mathematics section of the MMMU-Pro benchmark.

Model	MMMU-Pro-V
Baseline	25
+ Text-only CoT SFT	25
+ Interleaved CoT SFT	30
+ Interleaved CoT RL	30

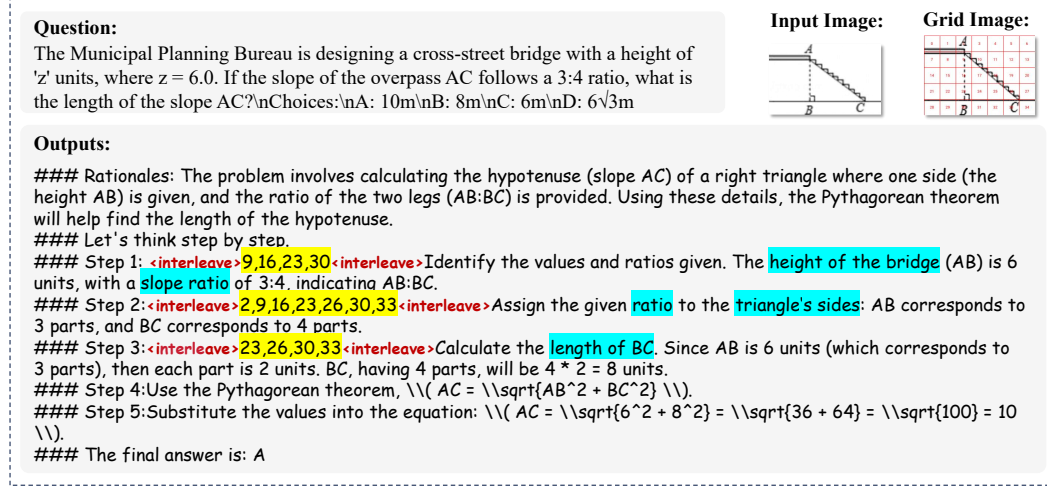


Figure 6: An example from MINT-CoT dataset.

Projector Ablation We conduct an ablation study on the post interleave projector $P_{\text{post_intlv}}$ and the post visual projector $P_{\text{post_vis}}$ on the Interleaved CoT SFT stage. Both projectors were initially implemented as single-layer linear layers. We first remove both projectors entirely, and then replace them with two-layer MLPs using GELU activation. Both configurations are trained for three epochs. The results on the mathematical subset of MathVista are shown in Table 7, in which we find that the initial configuration as single-layer linear layers performs the best over all primary tasks.

A.6 Additional Qualitative Results

In addition to Section 3.4, we provide more qualitative results of the baseline model Qwen2-VL-7B-Instruct and our proposed model MINT-CoT-7B in Figures 9 to 11.

A.7 Limitations

In our dataset curation pipeline, the use of GPT-4o for dataset annotation still has some computational cost. Moreover, alternative reinforcement learning strategies beyond our method now remain underexplored.

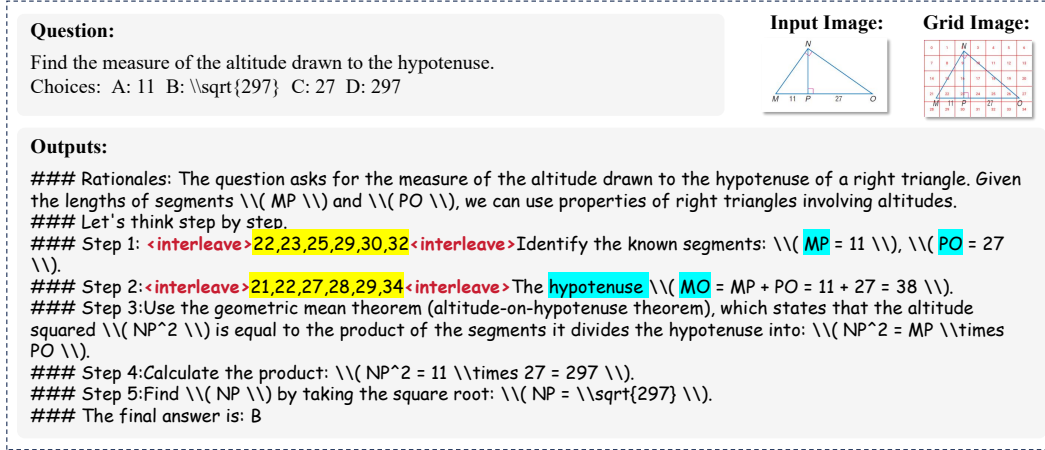


Figure 7: An example from MINT-CoT dataset.

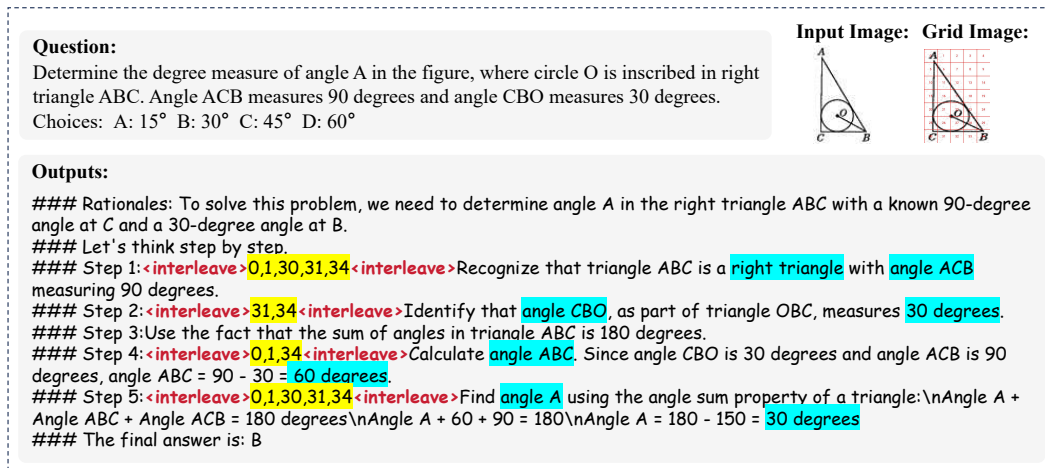
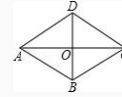


Figure 8: An example from MINT-CoT dataset.

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 $AD = \sqrt{AO^2 + BO^2} = \sqrt{4^2 + 3^2} = \sqrt{16 + 9} = \sqrt{25} = 5$ \n\n

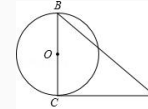
Step 3: Calculate the perimeter of the rhombus. Since all sides are equal in a rhombus, the perimeter is 4 times the side length. \n\nPerimeter $= 4 * 5 = 20$ \n\n

The final answer is: B ✔️

Figure 9: 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

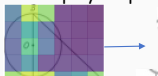
**Qwen2vl-7B-Instruct**

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: \n\n1. Start by understanding the Pythagorean theorem. \n\n2. Apply the Pythagorean theorem to a right triangle with segments AC and AB as the legs and segment BC as the hypotenuse. \n\nGiven that AO is the radius (4.0), we use the Pythagorean theorem in triangle OAB: \n\n $AB = \sqrt{OA^2 + OB^2} = \sqrt{4.0^2 + 4.0^2} = \sqrt{16 + 16} = \sqrt{32}$ \n\n $AB \approx 6.4$ \n\nSo 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. \n\nTherefore, 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: $AB^2 = AC^2 + BC^2$ \n\n

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

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

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

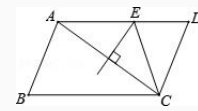
Step 8: Take the square root to find AB: $AB = \sqrt{100} = 10$ \n\n

The final answer is: C ✔️

Figure 10: 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

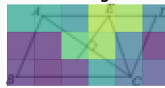
Rationales: The problem involves understanding the geometric placement of point E and the relations between various sides and diagonals in the quadrilateral. It requires determining the perimeter of the triangle CDE given the perimeter of ABCD.

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