CURV: Enhancing Chart Understanding Through Curriculum Visual Grounded Reasoning

Xuehang Guo[1], **Pingyue Zhang**[2], **Ruiyi Zhang**[3], **Zhenhailong Wang**[4], **Hanrui Lyu**[2], **Heng Ji**[4], **Tong Sun**[3], **Qingyun Wang**[1], **Manling Li**[2]

[1]William & Mary [2]Northwestern University [3]Adobe [4]UIUC

Abstract

Chart question answering (CQA) requires multimodal large language models (MLLMs) to integrate visual comprehension with logical reasoning, yet current models struggle with accurate visual grounding and coherent reasoning chains. While extrinsic chain-of-thought prompting and visual cues significantly improve performance, current MLLMs lack intrinsic visual grounded reasoning capabilities, leading to inaccurate perception and reasoning disconnected from visual evidence. To address these limitations, we propose CURV, a curriculum learning framework that develops intrinsic visual reasoning capabilities by reformulating CQA as multi-step visual grounded reasoning, where each step coordinates logical reasoning with dynamic visual grounding through spatial attention concentration. To assist model learning, we further introduce CCQA, a three-level curriculum dataset with scalable synthetic generation across diverse chart types and reasoning patterns. Our curriculum systematically progresses from basic single-operation reasoning to complex multi-chart compositional tasks. Experiments demonstrate that CURV achieves up to ↑ 20.50% improvements over baselines and is generalizable to real-world benchmarks (up to ↑ 12.30%) and out-of-domain multimodal reasoning tasks (up to ↑ 10.20%), validating the effectiveness of internalizing visual reasoning with dynamic grounding for enhanced chart understanding capabilities.

1 Introduction

How do humans tackle multimodal problems? Inspired by cognitive theories (Baddeley et al., 1974; Johnson-Laird, 1983; Barsalou, 2008; Grant & Spivey, 2003), humans **decompose** complex tasks into stepwise reasoning chains, **interleave** each step with dynamic visual grounding, and **compose** these grounded steps into a coherent solution (Fig. 1). Chain-of-thought (CoT) reasoning has demonstrated its effectiveness in decomposing problems into stepwise inferences (Xu et al., 2024; Zhang et al., 2025a). This ability becomes more critical in multimodal reasoning, where multimodal large language models (MLLMs) are expected to integrate visual and textual information (Fan et al., 2025; Zhang et al., 2025b) while visual perception errors contribute to the majority of multimodal reasoning failures (Wang et al., 2025c). Without external support such as explicit CoT prompting or visual cues (Fig. 1), MLLMs struggle with accurate visual grounded reasoning (Wang et al., 2025d;b;a).

This challenge is particularly evident in chart question answering (CQA), where models need to faithfully interpret complex geometric structures, spatial relationships, and quantitative patterns to derive correct answers. As a result, CQA requires models to accurately perceive visual details, perform step-by-step reasoning over interconnected components,

Correspondence to: Manling Li (manling.li@northwestern.edu), Qingyun Wang (qwang16@wm.edu), Xuehang Guo (xguo15@wm.edu)



Figure 1: **From Extrinsic Assistance to Intrinsic Visual Grounded Reasoning.** Inspired by human ways of thinking, we present CURV to internalize extrinsic **CoT prompting** and **visual guidance** to intrinsic capabilities, enabling models to perform **visual grounded reasoning** through dynamically shifting focuses across targeted image regions.

and dynamically shift focus across different chart regions (Chen et al., 2025). However, existing MLLMs exhibit several fundamental limitations in this setting (§2 & A): **(1) Decomposition:** *They struggle to decompose complex problems into coherent chains of reasoning, often producing inconsistent or logically flawed intermediate steps (Fig. 12-13)*; **(2) Interleaved Visual Reasoning:** *They show limitations in accurately grounding individual reasoning steps in the visual input, such as misreading chart values or attending to correct regions (Fig. 12)*; and **(3) Composition:** *They exhibit difficulties in integrating logical reasoning with visual grounding across multiple steps into a coherent, interleaved chain, leading to a disconnect between what is perceived, reasoned, and concluded (Fig. 13)*. Collectively, these limitations lead to inaccurate perception and reasoning that is disconnected from the visual evidence, ultimately causing errors even when the necessary information is present.

To address these limitations, we propose **CURV**, a curriculum learning (CL) framework that develops intrinsic visual grounded reasoning capabilities in MLLMs. Our approach reformulates CQA as multi-step reasoning processes where each step couples logical reasoning with dynamic visual grounding. Instead of relying on extrinsic assistance, CURV enables models to internalize the ability through dynamically focusing on relevant chart regions while maintaining coherent reasoning chains across steps. Curriculum learning progresses from single-operation reasoning to complex multi-operation compositions, allowing models to gradually develop both visual perception accuracy and reasoning sophistication. Our main contributions include:

- We propose **CURV**, a curriculum learning framework that develops MLLMs' intrinsic visual reasoning capabilities by progressively shifting visual attention along reasoning, transitioning from basic single-operation tasks to complex nested reasoning (§3).
- We introduce a scalable synthetic CQA data generation method (§4) that systematically increases task complexity via nested reasoning chains, enabling efficient data creation.
- We present *Curriculum Chart Question Answering* (**CCQA**) (§4) that supports curriculum learning across four different generation modes, three visual grounding strategies, and multi-level metrics for comprehensive performance evaluation (§2 & 5).
- Our experiments demonstrate that CoT reasoning with visual grounding provides models with step-by-step alignment between visual perception and logical reasoning, leading to notable improvements across different chart types, complexity, and domains (§5).

2 What Prohibits MLLMs From Chart Understanding Success?

2.1 Preliminary Analysis of CQA Failures

What are the bottlenecks that hinder MLLMs from correctly understanding chart images?

Mode	Error Analysis (*count*)			Acc (%)	Δ_{acc} (%)
	Vision	Reasoning	Answer		
A	–	–	**34**	43.33	–
VA	0	**18**	5	61.67	↑ 18.33
RA	**17**	6	5	53.33	↑ 10.00
RVA	0	5	3	**86.67**	↑ **43.33**

Table 1: **CQA Preliminary.** Preliminary results on GPT-4o using accuracy (%) and relative improvements Δ_{acc} (%).

We evaluate GPT-4o (OpenAI, 2024) on 60 CQA samples randomly selected from CharXiv with four modes: **(1) Answer only (A):** MLLM directly generates the answer; **(2) Vision + Answer (VA):** MLLM is provided with human-annotated visual information to give the answer; **(3) Reason + Answer (RA):** MLLM is prompted to first generate CoT reasoning, followed by the final answer; **(4) Reason + Vision + Answer (RVA):** MLLM is provided with the same visual information as in VA, prompted to first generate CoT reasoning and then the final answer. Surprisingly (Tab. 1), GPT-4o achieves significantly higher scores (↑ 10.00%) when prompted to generate reasoning without visual cues (RA). Moreover, combining perception with reasoning (RVA) yields the best performance (↑ 43.33%). These highlight *MLLMs' lack of logical decomposition and visual reasoning capabilities*.

2.2 Enhancing Chart Reasoning Through Dynamic Visual Grounding

How to improve MLLMs' intrinsic visual reasoning capabilities?



Figure 2: **Performance Across Reasoning Depths and Modes.** We evaluate GPT-4o (*upper*) and Qwen2.5-VL-7B (*lower*) on 1,800 randomly selected samples from CCQA (evenly distributed across $D_i, 1 \leq i \leq 3$).

Motivated by the effectiveness of extrinsic CoT prompting and visual guidance (Tab. 1), we aim to internalize these capabilities within MLLMs (Fig. 1). To concretize this approach, we expand our preliminary exploration to visual reasoning with grounded focuses (Tab. 2) using CCQA (§4) and two models: Qwen2.5-VL-7B (Qwen, 2025) and GPT-4o (OpenAI, 2024). Similarly, we examine model performance under four modes using ground-truth reasoning and visual grounding. Despite increased curriculum difficulty ($1 \leq D \leq 3$), both models achieve higher performance when equipped with either reasoning or visual assistance, with performance further improved when both are combined. Consistently, extrinsic assistance enhances chart understanding through structured reasoning and visual grounding, motivating our core hypothesis: *MLLMs can internalize these capabilities through visual grounded reasoning* (§3 & 4).

3 CURV: Chart Reasoning with Dynamic Visual Grounding

3.1 Problem Formulation

Problem Definition. Given a chart image $\mathcal{I} \in \mathbb{R}^{H \times W \times C}$ and a question Q, the goal of CQA is to generate the answer A. However, current MLLMs directly learn the mapping:

$$f_\theta : (\mathcal{I}, Q) \rightarrow A \tag{1}$$

Not only does this direct mapping approach lack an intermediate reasoning structure that enables accurate visual perception and robust visual understanding, but it also fails to effectively and dynamically ground reasoning chains in visual space.

Our Approach. We propose to decompose this problem into a multi-step reasoning process with dynamic visual grounding. Specifically, we reformulate the CQA task as:

$$f_\theta : (\mathcal{I}, Q) \rightarrow \{(R_1, V_1), (V_2, V_2), \ldots, (R_T, V_T)\} \rightarrow A \tag{2}$$

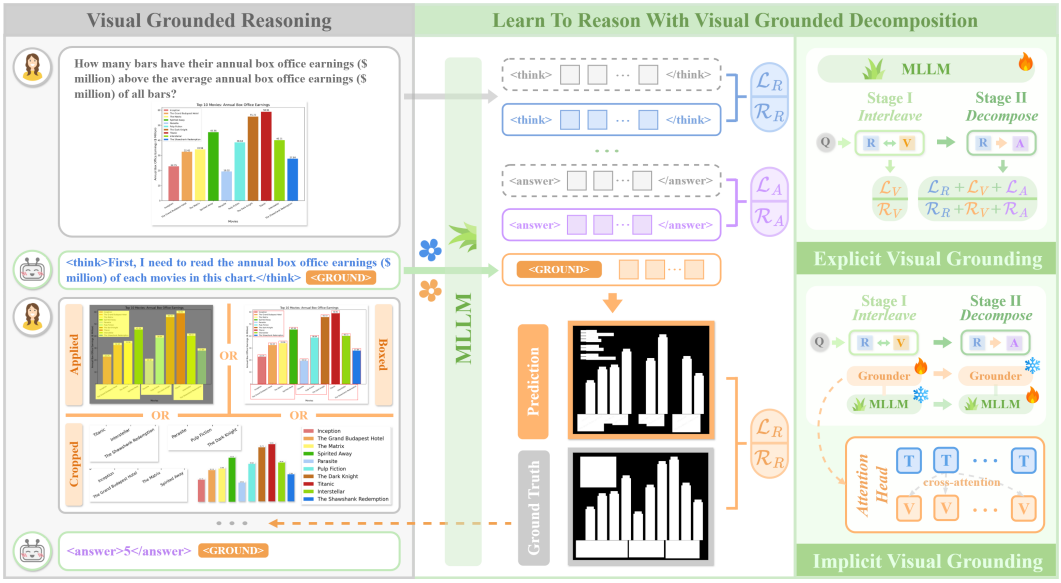


Figure 3: **CURV Overview.** Different visual grounding (§D) and training (§G) approaches to develop intrinsic **visual reasoning** capabilities via **multi-step visual grounded reasoning**.

where R_t represents the t-th reasoning step in natural language, V_t denotes visual regions grounding R_t in the visual space, T is the total number of reasoning steps to reach A, and the sequence $\{(R_t, V_t)\}_{t=1}^{T}$ forms a structured progressive visual reasoning chain.

3.2 Multi-Step Reasoning With Dynamic Visual Grounding

We design a two-stage multi-step curriculum learning framework (CURV) that enables MLLMs to develop intrinsic progressive reasoning capabilities with dynamic visual grounding, moving beyond extrinsic assistance toward self-contained visual reasoning (Fig. 1). To further strengthen this process, we implement *explicit* and *implicit* visual grounding (§G.5).

Stage I: Visual Grounding (RV). In the first stage, the model learns to establish multimodal correspondences between reasoning flows and associated visual focuses. Specifically, given the chart image $\mathcal{I}$ and question Q, at time t, the model learns to predict the visual focus V_t that grounds R_t onto $\mathcal{I}$:

$$V_t = f_\theta^{(S1)}\left(\mathcal{I},\, Q,\, \{R_{t'}, V_{t'}\}_{t'=1}^{t-1},\, R_t\right) \tag{3}$$

where prior grounding pairs $\{(R_{t'}, V_{t'})\}_{t'=1}^{t-1}$ provide context for consistent grounding across steps. The training objective supervises the predicted visual focuses:

$$\mathcal{L}^{(S1)} = \sum_{t=1}^{T} \mathcal{L}_V(V_t,\, V_t^*) \tag{4}$$

where $\mathcal{L}_V(V_t, V_t^*)$ denotes the grounding loss between the predicted visual focus V_t and the ground-truth V_t^* at step t.

Stage II: Interleaved Visual Reasoning (RVA). Building upon the visual grounding capability acquired in *Stage I*, the model learns to *proactively leverage* visual grounding as feedback during reasoning. Concretely, at each step t, the visual focus V_t produced alongside reasoning R_t is applied to the input chart image $\mathcal{I}$ to construct the grounded visual state $\mathcal{I}_t$ via one of the visual grounding methods (Fig. 3). This grounded image $\mathcal{I}_t$ is then provided as additional visual input when generating the *next* reasoning step, enabling the model to dynamically shift its visual attention in alignment with the evolving logical reasoning:

$$(R_t, V_t) = f_\theta^{(S2)}\left(\mathcal{I},\, Q,\, \{R_{t'},\, V_{t'} \to \mathcal{I}_{t'}\}_{t'=1}^{t-1}\right) \tag{5}$$

where $V_{t'} \to \mathcal{I}_{t'}$ denotes the new state mapped into $\mathcal{I}_{t'}$. *Stage II*'s training objective is:

$$\mathcal{L}^{(S2)} = \sum_{t=1}^{T} \lambda_R \mathcal{L}_R(R_t,\, R_t^*) + \lambda_V \sum_{t=1}^{T} \mathcal{L}_V(V_t,\, V_t^*) + \lambda_A \mathcal{L}_A(A,\, A^*) \tag{6}$$

where $\mathcal{L}_R(R_t, R_t^*)$ is the reasoning supervision loss at step t, $\mathcal{L}_V(V_t, V_t^*)$ is the grounding loss at step t, $\mathcal{L}_A(A, A^*)$ is the final answer prediction loss, and λ_R, λ_V, and λ_A balance the visual grounded reasoning objective against answer supervision. As such, *Stage II* develops the model's intrinsic progressive visual reasoning capabilities, enabling it to: (1) *dynamically focus* on relevant chart regions by updating visual grounding at each step, (2) *ground logical reasoning* in concrete visual evidence, and (3) *maintain cross-step coherence* through joint conditioning on both the textual reasoning history and the evolving visual states.

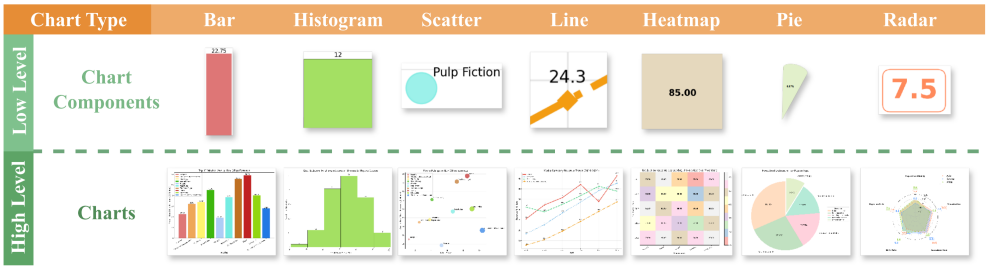


Figure 4: **From Low-Level Chart Components To High-Level Charts.** We decompose all types of charts into low-level components to endow MLLMs with both foundational chart understanding abilities and adaptive generalizabilities to high-level complexities.

3.3 Curriculum Visual Grounded Reasoning

Visual Grounding. In addition to *explicit* and *implicit* visual reasoning (§G.5), we propose three visual grounding strategies (§D), including **applied**, **boxed**, and **cropped**, to dynamically shift visual focuses during reasoning (Fig. 3).

Reasoning Depth. To formalize reasoning complexity in CQA, we introduce two distinct but complementary concepts that characterize the reasoning process:

- **Number of Reasoning Steps (T):** The total number of CoT reasoning steps $\{R_t\}_{t=1}^{T}$ a model goes through to reach the final answer A. Multiple reasoning steps may operate at the same logical complexity level (*i.e.*, *reasoning depth tier* below) while contributing different pieces of information toward the solution.
- **Tier of Reasoning Depth (D):** The maximum number of nested logical functions required to solve the task, corresponding to the deepest level of functional composition in the reasoning chain. Formally, for a question requiring nested functions $f_1(f_2(...(f_D(x))))$, the reasoning depth is D. This metric captures the inherent logical complexity of a problem, independent of how many intermediate steps a model uses to reach A.

Curriculum Learning. Charts are a unique form of data organized through structured relationships among fundamental meta-elements (Fig. 4). Through two-stage ***training curriculum***, we incorporate ***chart curriculum*** to guide the model from understanding low-level visual components to high-level chart structures (Fig. 4) across diverse chart types. With gradually increased task difficulty along both reasoning and visual dimensions, we define data curriculum consisting of three *levels* (§C.2) across five fine-grained *tiers* (§C.3).

CURV Variants. CURV accommodates multiple design choices along two orthogonal dimensions: First, it supports three visual grounding strategies (§D) that control the representation and granularity of visual grounding. We further implement *explicit* visual grounding to develop MLLMs' intrinsic visual grounded reasoning capabilities (§3.2), while performing ablations of *implicit* grounding (§G.5) to validate the critical role of visual grounding in CURV as intermediate visual-reasoning alignment guidance rather than an ultimate learning objective (§5.4). Second, CURV is compatible with different training paradigms, including *SFT*, *reinforcement learning (RL)*, and *their combination*, enabling flexible instantiations under varying optimization objectives and resource constraints (§5.3 & G.4).

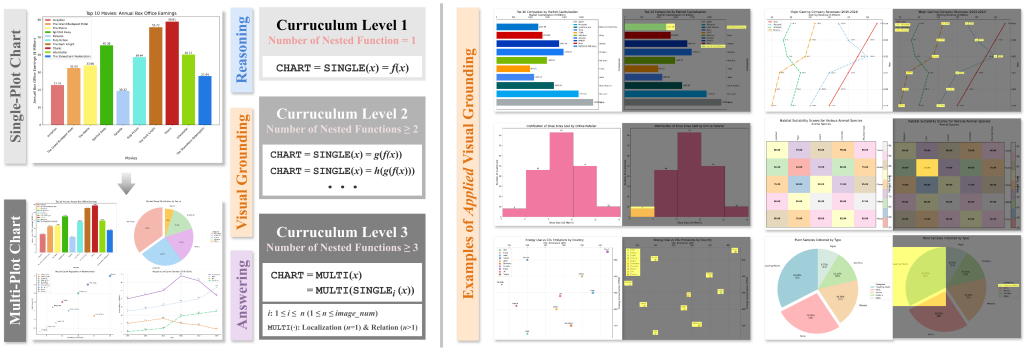


Figure 5: **Multi-Level Curriculum Construction.** We construct CCQA through reasoning decomposition, interleaving visual reasoning, and reasoning chain composition. Examples of **applied** visual grounding is shown on the right. More examples can be found in § D.

4 CCQA: Curriculum Chart Question Answering

4.1 Dataset Construction Principles

Supporting CURV (§3), we introduce *Curriculum Chart Question Answering* (**CCQA**), a curriculum learning dataset to progressively develop visual reasoning capabilities. Our dataset construction presents three core principles with rigorous quality control (§C.5):

Progressive Complexity. We implement three-level curriculum through systematic variation in reasoning depth (D), chart complexity, and operation sophistication (Fig. 5).

Interleaved Visual Reasoning. Each reasoning step R_t is paired with corresponding ground-truth V_t^* and binary masks M_t^*, enabling direct alignment of visual grounding.

Template-Based Accuracy. We employ synthetic templates (§C.1) to ensure data accuracy and systematic coverage of reasoning patterns, effectively supporting a stable progression of curriculum. As shown below, all chart-specific features are replaced by plotting data:

QUESTION: What is the < `y_axis_title` > of the < `object_singular` >?

4.2 Curriculum Chart Question Answering

Chart Types. We include 7 chart types to endow CCQA with high visual diversity (Fig. 6 & Tab. 4): *bar*, *histogram*, *scatter*, *line*, *heatmap*, *pie*, and *radar*.

Data Category. We define 30 distinct domain categories (Tab. 4), employing GPT-4o (OpenAI, 2024) to generate plotting data for chart drawing (Fig. 20).

QA Types. Our curriculum templates (§C.2) cover various operations (Tab. 5) in atomic or nested forms across different chart components and subplots.

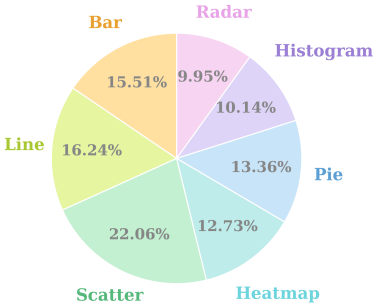


Figure 6: **Chart Type Distribution**

Curriculum with Meta Learning. Our dataset construction implements a *meta-learning* paradigm that maximizes visual reasoning generalization while minimizing visual overfitting (§C.4). With only 30 unique charts for each chart type (*i.e.*, 7 types × 30 categories), diverse ***query-reason-ground-answer*** quadruplets are derived from each image via systematic template instantiation (§C.2). Aiming to foster MLLMs' intrinsic visual grounded reasoning abilities rather than memorizing specific visual appearances, we generate multiple CQA instances $\{(Q_k, \{D_d, B_d^*\}_{d=1}^{D}, A_k)\}_{k=1}^{K}$ ($K \gg 1$) for each base image $\mathcal{I}_j$ to encourage the model to learn robust visual reasoning skills transferable across diverse chart appearances, data distributions, contexts and domains, as well as task complexities.

Model	Size	Level 1					Level 2					Level 3				
		@M	@0.0	@0.05	@0.1	@0.2	*@M*	@0.0	@0.05	@0.1	@0.2	*@M*	@0.0	@0.05	@0.1	@0.2
Close-Source MLLMs																
GPT-4o	-	57.64	54.07	62.00	65.36	70.50	34.04	33.75	44.93	50.54	57.96	22.14	22.29	30.25	34.04	39.14
GPT-4.1-mini	-	70.86	67.43	76.29	78.79	79.93	37.61	36.54	47.18	52.93	60.86	26.14	25.46	33.11	37.79	42.32
Open-Source Baselines																
Gemma-3	4B	38.21	28.64	32.79	37.00	41.29	18.07	12.86	18.82	23.64	28.29	11.43	9.64	13.14	15.54	18.96
Llama-3.2-V	11B	44.86	38.29	40.86	42.86	47.07	23.43	18.25	22.46	25.29	29.39	16.57	14.25	16.54	18.29	20.29
InternVL3	1B	20.54	16.38	20.53	24.40	29.68	8.51	7.44	12.61	15.37	20.65	6.76	5.87	8.31	10.09	10.99
	2B	33.53	32.52	37.52	42.52	47.29	13.11	13.22	19.89	27.14	35.55	10.69	11.18	14.71	18.72	24.90
	8B	46.79	44.29	52.14	57.71	63.00	25.75	25.57	35.36	41.64	50.79	17.84	17.48	24.57	28.93	34.68
Qwen2.5-VL	3B	45.25	43.52	51.54	56.25	61.54	22.75	22.86	31.14	37.21	45.36	16.18	16.00	21.89	25.82	31.71
	7B	54.21	50.79	60.43	64.64	69.29	28.68	28.82	39.93	45.54	52.61	19.01	19.38	26.71	31.50	36.93
Ours (Stage II)																
Applied (InternVL3)	1B	25.79	21.36	25.29	31.43	37.86	10.57	9.25	12.89	17.18	22.21	7.11	7.36	9.25	11.00	12.93
	2B	42.64	41.79	48.71	55.36	62.14	18.68	19.29	25.79	32.93	41.50	11.39	12.57	16.61	20.75	26.32
	8B	58.86	54.64	66.29	69.14	71.79	34.47	33.90	49.67	56.47	63.91	18.87	19.84	27.99	31.97	37.41
Applied (Qwen2.5-VL)	3B	54.21	51.21	59.50	65.00	68.50	25.86	27.14	39.18	47.25	55.18	16.86	17.32	23.93	28.61	33.11
	7B	65.79	59.14	71.93	75.29	78.29	36.82	34.79	50.82	56.18	62.75	21.04	21.11	30.21	34.25	39.39
Boxed (Qwen2.5-VL)	3B	58.07	51.79	61.00	65.71	69.86	25.96	25.32	37.75	45.89	53.86	16.92	16.82	22.89	27.21	33.29
	7B	59.79	52.79	71.93	74.57	76.64	33.79	30.32	49.75	55.89	62.21	20.14	18.04	28.89	32.32	36.29
Cropped (Qwen2.5-VL)	3B	49.71	46.79	54.86	60.07	65.21	20.68	21.11	32.61	40.36	50.82	15.00	18.89	20.96	24.71	30.25
	7B	58.93	56.57	67.71	72.93	76.71	30.04	29.71	40.43	46.71	53.79	18.68	18.71	26.79	30.36	35.11
Ours (Stage I + II)																
Applied (InternVL3)	1B	28.64	26.79	31.57	36.36	43.79	14.18	12.86	16.25	22.86	27.43	10.39	10.64	13.46	15.79	17.75
	2B	47.86	46.93	54.71	60.57	67.79	23.36	23.61	28.79	36.57	43.36	14.89	15.82	20.29	24.79	31.39
	8B	67.71	62.57	70.57	75.14	77.29	39.68	37.79	52.54	58.71	65.04	24.57	22.96	30.54	34.79	40.32
Applied (Qwen2.5-VL)	3B	60.64	57.21	63.71	69.57	73.79	30.11	29.61	41.82	48.86	57.64	17.96	18.21	24.25	29.07	34.50
	7B	**69.86**	**66.79**	**75.29**	**78.50**	**80.43**	**40.21**	**38.64**	**54.00**	**60.04**	**66.89**	**26.11**	**24.25**	**32.21**	**36.96**	**42.54**
Boxed (Qwen2.5-VL)	3B	58.50	54.71	64.21	72.57	74.79	27.64	26.86	38.75	46.86	55.36	17.43	17.54	23.50	27.96	33.46
	7B	63.29	59.57	72.14	75.86	77.00	35.32	33.57	51.54	57.54	64.36	23.57	22.14	31.96	35.04	38.25
Cropped (Qwen2.5-VL)	3B	53.07	50.79	57.71	63.57	69.50	23.29	23.54	34.29	41.43	51.54	17.14	18.04	22.46	26.11	33.04
	7B	62.14	60.71	69.57	74.71	77.86	32.25	31.82	46.32	50.57	57.36	20.11	20.36	28.46	33.86	37.11

Table 2: **Performance On CCQA.** We compare performance across different curriculum levels using five accuracy metrics ***acc***@X where X is MLLM (@M) or ranges (§ 5.2).

Generalize To Real-World Charts & Domains. Bridging visual perception and logical reasoning, we construct CCQA to underline the significance of enhancing MLLMs' understanding of fundamental visual components and spatial features for accurate visual reasoning (§G.8). CURV finetuned on CCQA are also applicable to real-world chart understanding and out-of-domain benchmarks (§5.2), validating not only the effectiveness but also the adaptability and generalizability of our approach (§A.1 & C.4).

5 Experiments

5.1 Setup

Baseline. We use two *close-source* MLLMs, GPT-4o (OpenAI, 2024) and GPT-4.1-mini (OpenAI, 2025), and seven *open-source* MLLMs, Llama-3.2-Vision (AI, 2024), Gemma-3 (Google, 2024), InternVL3 (OpenGVLab, 2025), and Qwen2.5-VL (Qwen, 2025) with different sizes, as comparison baselines.

Model. We employ five MLLMs as base models, finetuned through CURV for enhanced visual-grounded chart reasoning: Qwen2.5-VL (3B and 7B) and InternVL-3 (1B, 2B, and 8B).

Data. We split CCQA into non-overlapped *training* and *test* sets. All models are evaluated on the *test* sets unseen for finetuned models. Implementation details can be found in §G.1.

5.2 Evaluation

Evaluation Data. We evaluate on CCQA (§4), CQA, and out-of-domain benchmarks:

- **CCQA:** Highlighting the significance of learning chart basics through increased task difficulty, we evaluate models on three *test* sets of CCQA covering three increased curriculum levels, respectively.

- **Chart Benchmarks:** To assess the adaptability of MLLMs to chart understanding benchmarks, we extend our evaluation to *ChartMuseum* (*visual subset*) (Tang et al., 2025), *CharXiv* (*reasoning subset*) (Wang et al., 2024), *ChartQA* (Masry et al., 2022), and *ChartQAPro* (Masry et al., 2025).
- **Out-of-Domain Benchmarks:** Generalizing to different domains of multimodal reasoning, we extend our evaluation to multi-discipline multimodal reasoning tasks, including *MathVista* (Lu et al., 2023) and *MMMU-Pro* (Yue et al., 2025).

Evaluation Metrics. We evaluate different aspects of responses through complementary metrics that take both textual outputs and visual grounding into consideration (§F):

(1) *Reasoning*: For reasoning evaluation, we employ two complementary approaches:

Micro Evaluation (*acc@mic*): We evaluate reasoning via a combination of five *micro* metrics: ROUGE-L (Eq. 17), BLEU (Eq. 18), METEOR (Eq. 20), BERTSCORE (Eq. 19), and COSINE SIMILARITY (Eq. 21).

Model	Size	Chart Benchmarks				Out-of-Domain	
		ChartQA	ChartQA-Pro	CharXiv	ChartMuseum	MathVista	MMMU-Pro
Baselines							
InternVL3	1B	41.68	11.48	15.70	10.01	35.80	9.94
	2B	66.96	20.84	24.30	15.02	55.90	16.36
	8B	74.56	30.56	36.20	24.42	68.80	26.99
Qwen2.5-VL	3B	62.32	17.02	19.50	12.21	56.10	21.45
	7B	72.48	29.77	32.50	21.62	64.60	28.21
Ours (Stage I + II)							
Applied (InternVL3)	1B	51.56	15.86	18.90	11.21	45.93	12.95
	2B	69.60	23.97	28.10	16.12	62.78	18.27
	8B	**78.28**	32.39	**40.20**	**26.23**	**75.37**	30.81
Applied (Qwen2.5-VL)	3B	72.36	24.74	31.80	19.32	66.30	22.77
	7B	75.04	**32.80**	36.70	25.12	72.41	**32.08**

Table 3: **Performance on Chart and Out-of-Domain Benchmarks.** MLLM-as-judge ***acc@MLLM*** (§ 5.2) on baselines and CURV.

Macro Evaluation (*acc@mac*): We employ GPT-4.1-mini as the judge for *macro*-level evaluation by assigning quality scores based on three evaluation criteria (§F.2).

(2) *Visual Grounding*: We leverage Intersection-over-Union (IoU) variants CIOU (Eq. 22) and GIOU (Eq. 23), where GIOU (Rezatofighi et al., 2019) is generalized IoUs and CIOU evaluates the cumulative intersection over the cumulative unions (Zheng et al., 2019).

(3) *Answer*: We define CQA accuracy as the mean *answer* accuracy across all testing samples.

MLLM as Judge (*acc@MLLM*): We employ GPT-4.1-mini (OpenAI, 2025) as the judge to evaluate answer accuracy through *True-or-False* assessment (*pass*@1). More details in §F.1.

Rule as Judge (*acc@range*): To mitigate potential biases introduced by *MLLM-as-judge* (Dorner et al., 2024; Li et al., 2024a), we introduce rule-based evaluation metrics (Algorithm 1) to assess answer accuracy through systematic parsing and rubric judgment. In particular, it incorporates four *range* criteria that capture different levels of strictness (§F.1), including the *absolute* accuracy (*acc*@0.0) and three progressively *relaxed* thresholds (*acc*@0.05, *acc*@0.1, *acc*@0.2).

5.3 Main Results On Chart & Multimodal Understanding

Performance on CCQA. Compared with the baselines (Tab. 2), our finetuned models achieve consistently higher accuracy across all six metrics, with absolute gains of up to 14.85% and 20.92% for single-stage and two-stage training, respectively. The highest performance is achieved by CURV@Applied (Qwen2.5-VL-7B), with up to 15.65% higher than its base model. Compared with GPT models, it achieves up to 12.22% higher accuracy, demonstrating the effectiveness of CURV in enhancing MLLMs' visual reasoning abilities.

Performance on Complex Chart Understanding. Although trained solely on single-plot charts ($1 \leq D < 3$), finetuned models generalize to multi-plot charts ($D \geq 3$). As shown in Tab. 2, CURV@Applied (Qwen2.5-VL-7B) achieves up to ↑7.10% improvements, and CURV@Applied (InternVL-8B) also shows ↑6.73% across all metrics.

Performance on Chart Benchmarks. Aiming for CURV to be not only adaptable across different task complexities but also generalizable to real-world chart comprehension, we

extend our evaluation to CQA benchmarks (§5.2). Results in Tab. 3 highlight the strong generalizability of CURV, with improvements ≥↑1.20% across four chart benchmarks.

Generalizability to Out-of-Domain Multimodal Reasoning. The advantage of CURV remains consistent in out-of-domain multimodal reasoning across diverse multimodal reasoning categories (§5.2), attaining up to ↑10.20% accuracy improvements (Tab. 3).

Generalizability to Different Training Paradigms. CURV is readily applicable to other training paradigms (§G.4) to enhance MLLMs' intrinsic visual reasoning capabilities. As shown in Tab. 9, applying CURV to *reinforcement learning* yields up to 12.58% improvement, while combining SFT and RL further increases gains to 17.04%. Despite even higher performance with RL, its substantially higher computational overhead (Tab. 8) contrasts with the more favorable performance-efficiency trade-off of SFT.

5.4 Ablations on Visual Grounded Reasoning

Benefits of Explicit over Implicit Visual Grounded Reasoning. Comparing CURV with *explicit* and *implicit* visual grounded reasoning (§G.5), Fig. 7 demonstrates that *explicit* visual grounded reasoning consistently outperforms its *implicit* counterpart (↑ 8.78%). This advantage highlights the value of visual grounding as an integral intermediate vision-reasoning bridge, rather than the ultimate learning objective, to effectively develop and enhance MLLMs' intrinsic visual reasoning capabilities.

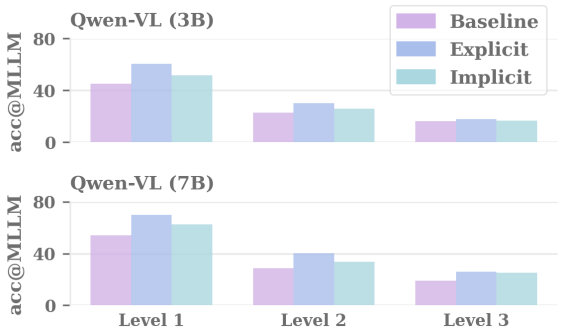


Figure 7: **Explicit vs. Implicit**

CURV@*Applied* Presents More Effective Visual Grounding. Implementing three visual grounding methods (§3 & §D), results in Tab. 2 unveils that **boxed** grounding stays less beneficial than directly highlighting regions of focus through **applied** masking, despite its simplicity and straightforwardness. On the other hand, although restricted by the trade-off between zoom-in resolution and computation overhead (§G.2), **cropped** grounding showcases its strengths (Tab. 2) with up to ↑ 7.93% improvements on *acc@MLLM* and ↑ 9.92% gains on *acc*@0.0, despite a sixteen-fold reduction in resolution.

Effectiveness of Curriculum Learning. Comparing across task difficulty levels (§G.6), results (Fig. 30) show that CL consistently improves performance over standard training, which remains inferior to CURV and exhibits diminished gains as task complexity increases.

Consistent Gains across Chart Types. As shown Fig. 8 (*top*), we compare the accuracy improvements (Δ_{acc}) of CURV@*Applied* (Qwen2.5-VL-7B) over its base model across chart types. *Bar* charts benefit the most, followed by *line* plots and *heatmaps*. The remaining chart types also demonstrate positive gains, with *radar* charts contributing the least, reflecting their lower prevalence than other chart types. Furthermore, the comparison among CURV@*Applied*, @*Boxed*, and @*Cropped* (Fig. 8) shows different distributions across chart types, suggesting different effects across visual grounding strategies (§D).



Figure 8: **Ablation on Chart Types.** We compare Δ_{acc} among chart types.

Foundational Learning Drives More Balanced Gains. While all curriculum levels improve overall accuracy, training on *levels* 1+2 yields the most consistent and substantial gains across different levels (Fig. 9). In contrast, relying solely on *level* 3 is less effective, highlighting the significance of foundational learning in establishing adaptive and generalizable visual reasoning capabilities to support more robust chart understanding across varying difficulty levels (§G.8).

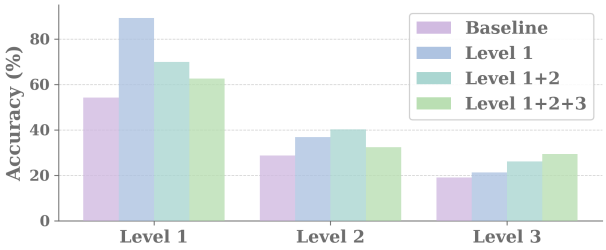


Figure 9: **Foundational Learning In Chart Understanding.** We compare Qwen2.5-VL-7B with its *applied* model trained on curriculum levels 1, $1+2$, and $1+2+3$, respectively.

CURV Enhances Multifaceted Visual Reasoning Abilities. CURV notably improves MLLMs' performance, even on the most challenging *level* 3 *multi-chart QA* (Tab. 2). Despite the inherent complexity gap between single-chart (*levels* 1-2: *accuracy* up to 80.43%) and multi-chart scenarios (*level* 3: *accuracy* $\leq$ 42.54%), CURV narrows this gap by enhancing both *localization* accuracy and *relational* understanding across different chart subplots (§C.2). The operational breakdown in Fig. 10 shows that CURV leads to clear improvements in **localization** ($\uparrow$ 4.68%), enabling models to more accurately identify relevant chart subplots within complex layouts. In addition, CURV also effectively improves **relation** understanding ($\uparrow$ 2.42%), facilitating more reliable cross-chart reasoning over dispersed visual evidence. Collectively, these gains suggest that CURV strengthens MLLMs' multifaceted visual reasoning abilities to tightly couple visual evidence with logical reasoning. This extends effectively beyond simple CQA settings to more complex multi-chart scenarios and out-of-domain reasoning, indicating that the benefits of CURV arise from improved reasoning grounded in visual structure rather than task-specific adaptation.

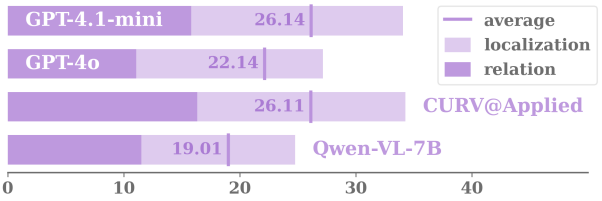


Figure 10: **Performance on Multi-Chart Understanding.** Top-4 accuracy scores on curriculum *level* 3.

6 Conclusion

In this work, we present CURV (§3), a curriculum learning framework that develops intrinsic visual reasoning capabilities via progressive multi-step visual grounded reasoning. To support model learning, we construct CCQA (§4) with three progressive curriculum levels. Through systematic experiments (§5 & G), results demonstrate that tightly interleaving reasoning with visual grounding consistently improves performance across curriculum levels and generalizes effectively to chart understanding and out-of-domain multimodal reasoning. Our work establishes a foundation for developing self-contained visual reasoning capabilities in MLLMs, moving beyond extrinsic assistance toward intrinsic grounded visual reasoning.

Acknowledgments

The authors acknowledge William & Mary Research Computing for providing computational resources and/or technical support that have contributed to the results reported within this paper. This work used DeltaAI at NCSA through allocation CIS260012 and CIS230280 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

Ethics Statement

In this work, we introduce a curriculum learning framework with the dataset CCQA constructed through meta-learning supported CQA creation. Other evaluation benchmarks, including chart understanding and multimodal reasoning, are publicly available and do not contain personally identifiable information or sensitive content. Our methods are designed for research and educational purposes, and we do not foresee direct misuse. With every step being effectively controlled, we positively believe that our work does not violate any ethical standards.

References

Meta AI. Llama 3.2 11b vision instruct, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct.

A. Baddeley, G. Hitch, and D. Baddeley. Working memory. In *In G. H. Bower (Ed.), The psychology of learning andmotivation: Advances in research and theory*, volume 8, pp. 47–89, 1974. URL https://api.semanticscholar.org/CorpusId:201393857.

L. Barsalou. Grounded cognition. *Annual review of psychology*, 59:617–45, 2008. URL https://www.ncbi.nlm.nih.gov/pubmed/17705682.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Xinyan Chen, Renrui Zhang, Dongzhi Jiang, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2506.05331.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 192–210, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.11. URL https://aclanthology.org/2024.naacl-long.11/.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/pdf/2401.10935.pdf.

Google DeepMind. Gemini 2.5 flash, june 2025. URL https://deepmind.google/models/gemini/flash/.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. *ArXiv*, abs/2410.13341, 2024. URL https://api.semanticscholar.org/CorpusId:273404208.

Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R. Fung. Unveiling the lack of LVLM robustness to fundamental visual variations: Why and path forward. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20222–20242, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1037. URL https://aclanthology.org/2025.findings-acl.1037/.

Google. Gemma 3 4b (italian), 2024. URL https://huggingface.co/google/gemma-3-4b-it.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for gui agents. *ArXiv*, abs/2410.05243, 2024. URL https://api.semanticscholar.org/CorpusID:273186286.

Elizabeth R. Grant and Michael J. Spivey. Eye movements and problem solving: Guiding attention guides thought. *Psychological Science*, 14(5):462–466, 2003. doi: 10.1111/1467-9280.02454. URL https://pubmed.ncbi.nlm.nih.gov/12930477/.

Philip Nicholas Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge, Mass. : Harvard University Press, 1983.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350, 2025. URL https://api.semanticscholar.org/CorpusId:276903810.

Sang-Woo Kim, Dongsu Zhang, Minseok Choi, and Sang goo Lee. Simplot: Enhancing chart question answering by distilling essentials, 2024. URL https://arxiv.org/abs/2405.00021.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. *ArXiv*, abs/2410.15393, 2024a. URL https://api.semanticscholar.org/CorpusId:273501717.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. Synthesize step-by-step: Tools, templates, and llms as data generators for reasoning-based chart vqa, 2024b. URL https://arxiv.org/abs/2403.16385.

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, 2023. URL https://arxiv.org/abs/2310.02255.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. Chartgemma: Visual instruction-tuning for chart reasoning in the wild, 2024. URL https://arxiv.org/abs/2407.04172.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151,

Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.978. URL https://aclanthology.org/2025.findings-acl.978/.

OpenAI. Hello gpt-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. Introducing gpt-4.1, april 2025. URL https://openai.com/index/gpt-4-1/.

OpenGVLab. Internvl3: Advancing open-source multimodal models with native multimodal pretraining, 2025. URL https://internvl.github.io/blog/2025-04-11-InternVL-3.0/.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogcom: A visual language model with chain-of-manipulations reasoning. *arXiv preprint arXiv:2402.04236*, 2024. URL https://arxiv.org/abs/2402.04236.

Qwen. Qwen2.5 technical report, 2025. URL https://arxiv.org/pdf/2412.15115.

S. H. Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 658–666, 2019. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8953982.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024. URL https://arxiv.org/abs/2403.16999.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, Ramya Namuduri, Bodun Hu, Juan Diego Rodriguez, Puyuan Peng, and Greg Durrett. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444*, 2025. URL https://arxiv.org/abs/2505.13444.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.153. URL https://aclanthology.org/2023.acl-long.153/.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a. doi: 10.48550/arXiv.2507.07999.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b. doi: 10.48550/arXiv.2506.11991.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning, 2025c. URL https://arxiv.org/abs/2507.06448.

Zhenhailong Wang, Joy Hsu, Xingyao Wang, Kuan-Hao Huang, Manling Li, Jiajun Wu, and Heng Ji. Visually descriptive language model for vector graphics reasoning. *Transactions on Machine Learning Research*, May 2025d. URL https://arxiv.org/abs/2404.06479. Preprint available at arXiv:2404.06479.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024. URL https://arxiv.org/abs/2406.18521.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022. URL https://arxiv.org/abs/2201.11903. arXiv:2201.11903.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, Si Qin, Lars Lidén, Qingwei Lin, Huan Zhang, Tongxing Zhang, Jianbing Zhang, Dongmei Zhang, and Jianfeng Gao. Gui-actor: Coordinate-free visual grounding for gui agents. *ArXiv*, abs/2506.03143, 2025. URL https://api.semanticscholar.org/CorpusID:279118510.

Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13326–13365, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL https://aclanthology.org/2024.acl-long.720/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 736–748. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Bohan Zhang, Xiaokang Zhang, Jing Zhang, Jifan Yu, Sijia Luo, and Jie Tang. CoT-based synthesizer: Enhancing LLM performance through answer synthesis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6286–6303, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.315. URL https://aclanthology.org/2025.acl-long.315/.

Lingling Zhang, Muye Huang, QianYing Wang, Yaxian Wang, Wenjun Wu, and Jun Liu. Got-cqa: Graph-of-thought guided compositional reasoning for chart question answering, 2024a. URL https://arxiv.org/abs/2409.02611.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning, 2024b. URL https://arxiv.org/abs/2410.16198.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.82. URL https://aclanthology.org/2025.acl-long.82/.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie

Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images, 2025c. URL https://arxiv.org/abs/2508.11630.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, and Tsung-Yi Lin. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. *arXiv preprint arXiv:2503.22020*, 2025. URL https://arxiv.org/abs/2503.22020.

Hanwen Zheng, Sijia Wang, Chris Thomas, and Lifu Huang. Chartformer + qdcat: Advancing chart question answering with robust chart component recognition, 2024. URL https://arxiv.org/abs/2407.21038.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. *ArXiv*, abs/1911.08287, 2019. URL https://api.semanticscholar.org/CorpusId:208158250.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning, 2025.

Contents

1 Introduction **1**

2 What Prohibits MLLMs From Chart Understanding Success? **2**
- 2.1 Preliminary Analysis of CQA Failures . . . 2
- 2.2 Enhancing Chart Reasoning Through Dynamic Visual Grounding . . . 3

3 CURV: Chart Reasoning with Dynamic Visual Grounding **3**
- 3.1 Problem Formulation . . . 3
- 3.2 Multi-Step Reasoning With Dynamic Visual Grounding . . . 4
- 3.3 Curriculum Visual Grounded Reasoning . . . 5

4 CCQA: Curriculum Chart Question Answering **6**
- 4.1 Dataset Construction Principles . . . 6
- 4.2 Curriculum Chart Question Answering . . . 6

5 Experiments **7**
- 5.1 Setup . . . 7
- 5.2 Evaluation . . . 7
- 5.3 Main Results On Chart & Multimodal Understanding . . . 8
- 5.4 Ablations on Visual Grounded Reasoning . . . 9

6 Conclusion **10**

A Preliminary Exploration & Validation **18**
- A.1 Preliminary Exploration On Motivations . . . 18
- A.2 Preliminary Exploration on CQA Challenges . . . 18
- A.3 Preliminary Validation on CCQA . . . 19

B Related Work **27**

C Dataset Construction **27**
- C.1 Data Structure . . . 29
- C.2 Multi-Level Curriculum . . . 30
- C.3 Fine-Grained Curriculum Tiers . . . 31
- C.4 Meta-Learning Supported Curriculum Learning . . . 32
- C.5 Quality Control . . . 32

D Visual Grounding Strategies **33**
- D.1 **Applied**: Grounding Through Dynamic Visual Focus Highlighting . . . 33
- D.2 **Boxed**: Grounding Through Dynamic Visual Box Guides . . . 34

D.3 **Cropped**: Grounding Through Dynamic Visual Focus Zooming 34

E Generation Mode **37**
E.1 Mode **A** 37
E.2 Mode **VA** 37
E.3 Mode **RA** 38
E.4 Mode **RVA** 38

F Evaluation Metrics **40**
F.1 Evaluation of Answers 40
F.2 Evaluation of Reasoning 41
F.3 Evaluation of Visual Grounding 43
F.4 Evaluation Mode 44

G In-Depth Analysis **44**
G.1 Implementation Details. 44
G.2 Grounding Method & Computation Cost 44
G.3 The Role of Visual Grounding: From Extrinsic Assistance To Intrinsic Abilities 44
G.4 Visual Grounded Reasoning via Reinforcement Learning 45
G.5 Explicit vs. Implicit Visual Grounded Reasoning 46
G.5.1 Explicit Visual Grounding 46
G.5.2 Implicit Visual Grounding 47
G.6 Curriculum Learning: Effectiveness of CL in Chart Understanding 48
G.7 Two-Stage Learning: Strength of Training Curriculum 48
G.8 Foundational Learning: Trade-off between Robustness & Adaptability . . . 49
G.9 Broader Baselines: Comparison Against Chart Specialist & Tool-Use Agents 49

H Multi-Step Reasoning With Dynamic Visual Grounding **50**
H.1 Challenges In Multi-Step Visual Grounding 50
H.2 Inference Failure 51
H.3 Inference Success 51

I Limitations & Future Work **52**

A Preliminary Exploration & Validation

A.1 Preliminary Exploration On Motivations

Building on the cognitive perspectives that humans solve multimodal problems through **decomposition**, **interleaved visual reasoning**, and **composition** (Fig. 1), we conduct preliminary studies to examine where current MLLMs fall short (§2).

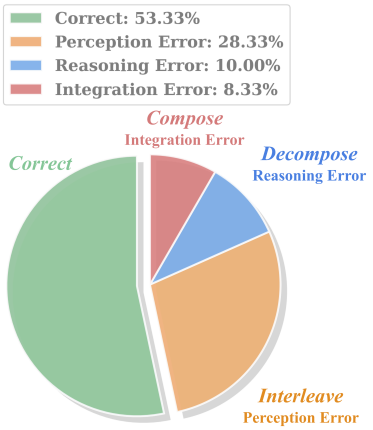


Figure 11: **Error Analysis** (§ 2)

Concretely, we evaluate GPT-4o on 60 CharXiv ((Wang et al., 2024)) samples, categorizing the root causes of CQA failures into three classes: *reasoning errors* (**decomposition**), *perception errors* (**interleaved visual reasoning**), and *integration errors* (**composition**). As shown in Fig. 11, perception errors (28.33%) emerge as the dominant reason of failures, reflecting the difficulty of accurately grounding reasoning steps in fine-grained chart details. Beyond perception, models also exhibit weaknesses in decomposition (10.00%), struggling to break down complex problems into coherent chains of reasoning. Additionally, they also show deficiencies in composition (8.33%), failing to integrate grounded visual evidence into a coherent, interleaved chain of thought. These findings reveal systematic shortcomings in human-inspired reasoning stages, motivating our design of CURV that enhances MLLMs' intrinsic visual grounded reasoning capabilities by simulating human cognitive process of **decomposing**, **interleaving**, and **composing** toward a solution.

Inspired by recent findings that models can effectively learn from low-level features (Wang et al., 2025d), we concretize the notion of **decomposition** in two complementary forms: (1) *visual decomposition*, where each chart is decomposed into low-level components (Fig. 4) to to guide MLLMs' attention toward fine-grained and informative details; and (2) *reasoning decomposition*, where each chart understanding problem is decomposed into a structured chain of reasoning steps to help MLLMs enhance their logical reasoning capacities. Building on these decompositions, we incorporate **interleaving** insights into our design, enabling reasoning to be interleaved with dynamically shifting visual focuses. Accordingly, the **composition** process integrates all intermediate learning in a coherent chain: from *reasoning composition* that consolidates step-wise reasoning into a coherent logical chain, to *visual composition* that progressively aggregates low-level visual interpretations into holistic chart comprehension.

Together, these elements form the foundation of our **curriculum learning** design (§3), which standardizes two dimensions of progression: (1) *curriculum CQA reasoning difficulty*, controlled by increasing the number of nested functions; and (2) *curriculum chart visual complexity*, controlled by increasing the number of low-level components, chart types, and chart subplots. We further support our design with **meta-learning** (§C.4) to endow MLLMs with adaptability and generalizability in the face of varying chart types, context domains, and task complexity.

A.2 Preliminary Exploration on CQA Challenges

To investigate the underlying causes of failures in chart understanding, we employ five MLLMs, including three *close-source* models (GPT-4.1-mini (OpenAI, 2025), GPT-4o (OpenAI, 2024), Gemini-2.5-Flask (DeepMind, 2025)) and two *open-source* models (Qwen2.5-VL-3B and Qwen2.5-VL-7B (Qwen, 2025)) to identify the root causes of their failures. Specifically, we analyze their CQA outputs case-by-case on different CQA benchmarks (§5.2), noticing several key patterns in MLLMs' CQA failures:

- **Reasoning Accuracy & Consistency:** While prompting MLLMs to do CoT reasoning can guide them toward correct answers in some cases, we still observe notable visual reasoning failures. For example, in Fig. 12, Qwen2.5-VL-3B misaligns line colors with their corresponding labels at the beginning, which propagates this misperception through subsequent reasoning and results in an incorrect answer. On the other hand, in Fig. 13, GPT-4o fails to excluded "*Loki*" despite having correctly identified it in earlier steps, unveiling the inconsistency in its evolution of reasoning. Another form of reasoning inconsistency emerges in recursive self-correction, where it may occur repeatedly throughout the model's reasoning process, ultimately producing inconsistent or divergent answers (*e.g.*, Gemini-2.5-Flash in Fig. 13).
- **Visual Grounding Accuracy:** MLLMs exhibit significant challenges in precisely capture visual details from the chart images. For example, in Fig. 12, GPT-4o inaccurately estimates the W_H value of the red "fi" point as approximately 0.105, while the true value is significantly less than 0.1, residing just above 0.0.
- **Visual Reasoning Effectiveness:** In complex reasoning tasks requiring the integration of multiple visual regions and reasoning steps, MLLMs often struggle to effectively link visual attention with logical reasoning. For instance (Fig. 13), although GPT-4o and GPT-4o-mini both perceive accurately in their initial perception, they exhibit distinct failures in subsequent reasoning: GPT-4o incorrectly includes "*Roar*" while GPT-4o-mini fails to incorporate "*Loki*".

A.3 Preliminary Validation on CCQA

In validating our proposed curriculum learning benchmark, CCQA (§4), we employ the same five MLLMs as our preliminary exploration (§A.2), examining case studies on five fine-grained difficulty tiers (§C.3) of the three curriculum levels (§C.2):

- **Tier 1: Single-Plot Reasoning ($D = 1$).** The example in Fig. 14 is a *Statistics-Mean* query (Tab. 5) that MLLMs often fail to correctly answer. Among all three faiure cases, Qwen2.5-VL-3B and GPT-4o-mini fail at accurately perceive the numbers from the chart image, while Qwen2.5-VL-7B encounters calculation errors despite correct visual understanding.
- **Tier 2: Single-Plot Reasoning ($D = 2$).** The example in Fig. 15 defines a specific *Subset* (Tab. 5) that poses significant obstacles for most MLLMs in arriving at the correct answer. Except for Gemini-2.5-Flask, all other four models struggle to accurately identify the required chart components.
- **Tier 3: Single-Plot Reasoning ($D \geq 3$).** The example in Fig. 16 involves multiple operators (Tab. 5) that unfolds notable challenges for most MLLMs in both visual reasoning and mathematical calculation. Except for GPT-4o, all other four models struggle to accurately calculate the mean, leading to wrong final answers.
- **Tier 4: Multi-Plot Reasoning (Localization).** The example in Fig. 17 involves both *Localization* and *Statistics-Sum* (Tab. 5). All five models fail to correctly perceive the visual information, resulting in wrong intermediate reasoning and incorrect final answer.
- **Tier 5: Multi-Plot Reasoning (Relation).** The example in Fig. 18 additionally pay attention to the *Relation* (Tab. 5) of different subplots. Likewise, the incorrect visual reasoning eventually lead to wrong final answers for all five models.

The diverging trends of failure cases (*increase*) and success cases (*decrease*) (Fig. 14-18) demonstrate the effectiveness of our curriculum design for progressively increased complexity tiers, meanwhile highlighting the challenges that persist in chart understanding across all reasoning depths and task complexity levels.

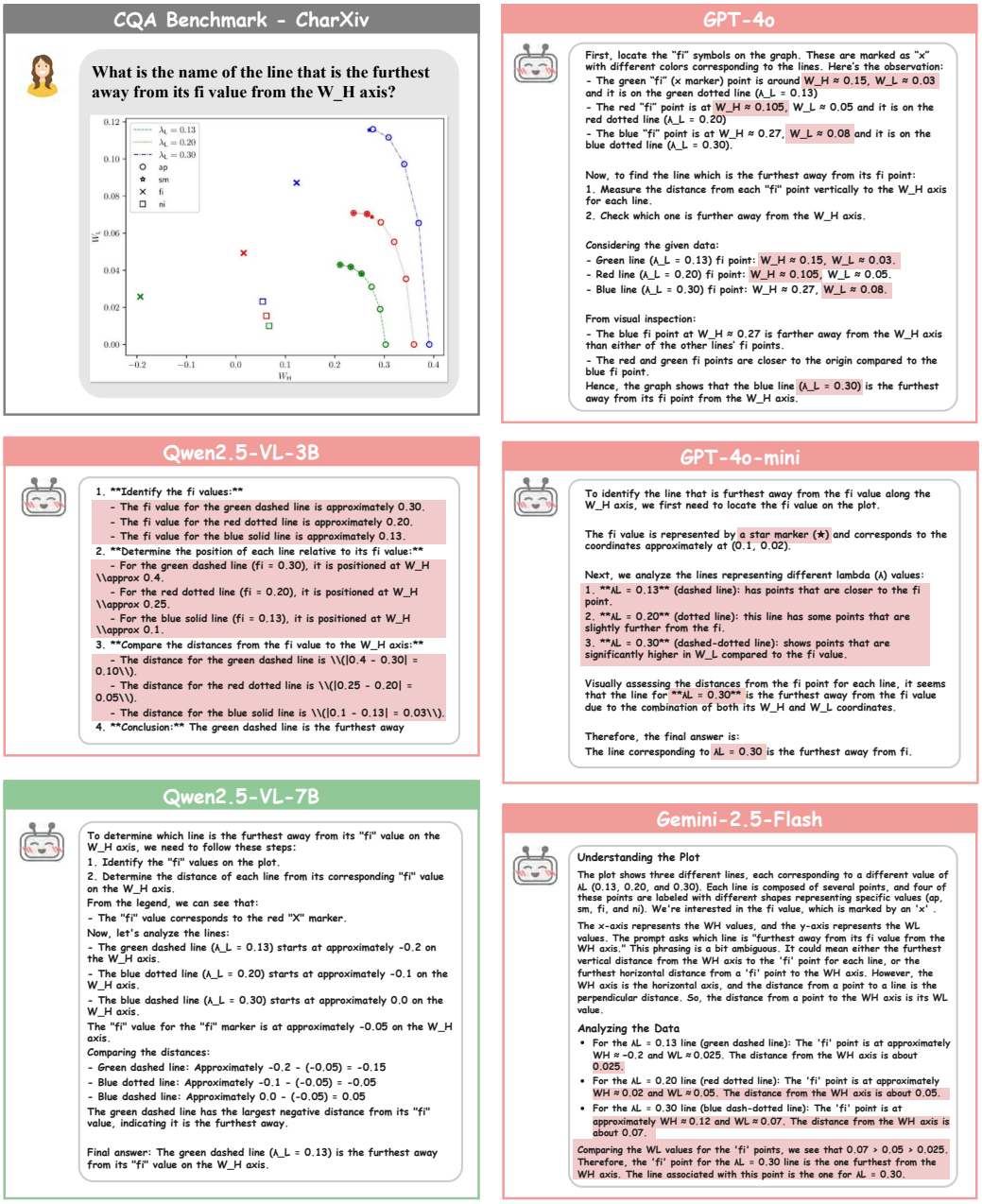


Figure 12: **Examples of Preliminary Exploration on CharXiv.** We conduct case study on CharXiv samples by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

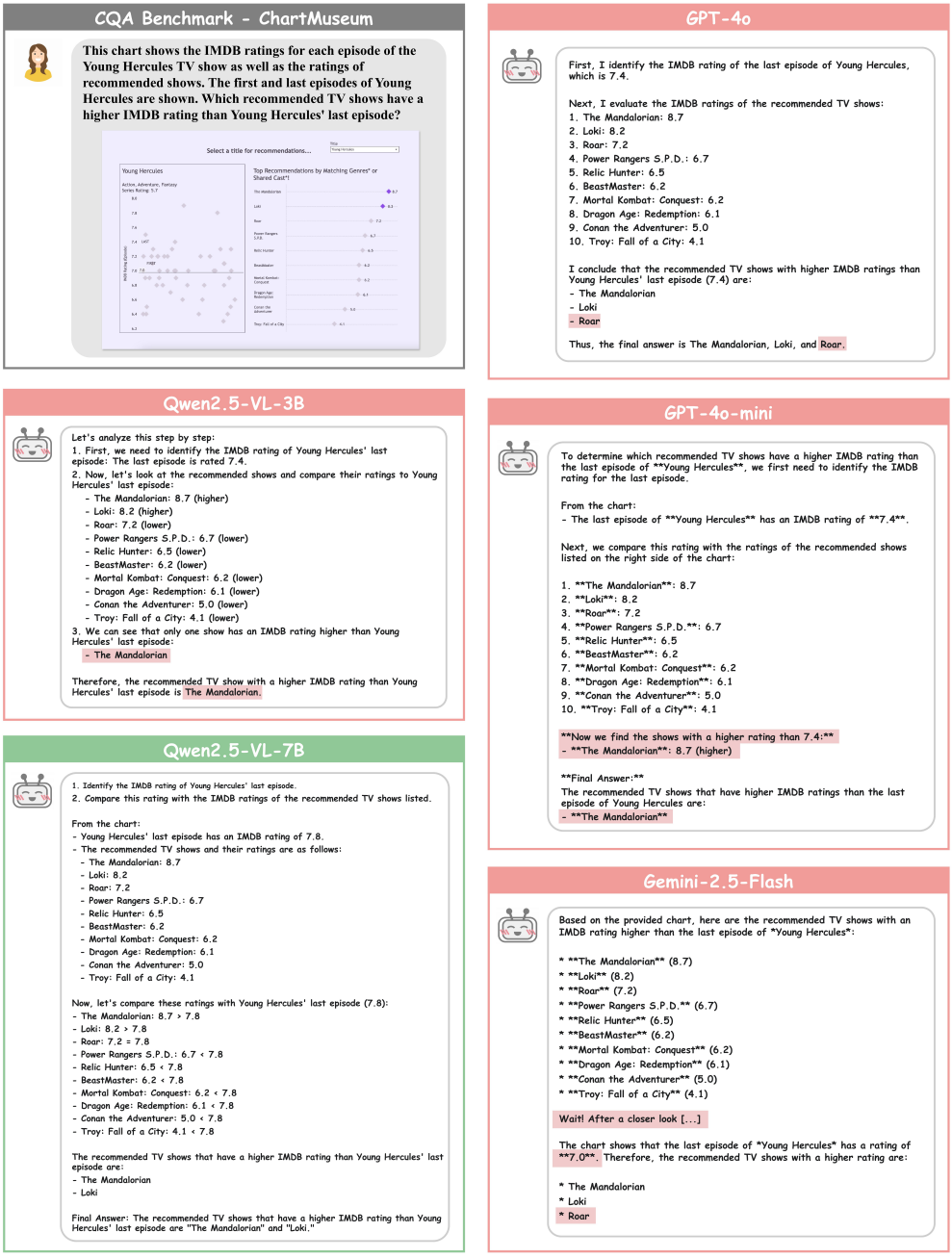


Figure 13: **Examples of Preliminary Exploration on ChartMuseum.** We conduct case study on ChartMuseum samples by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

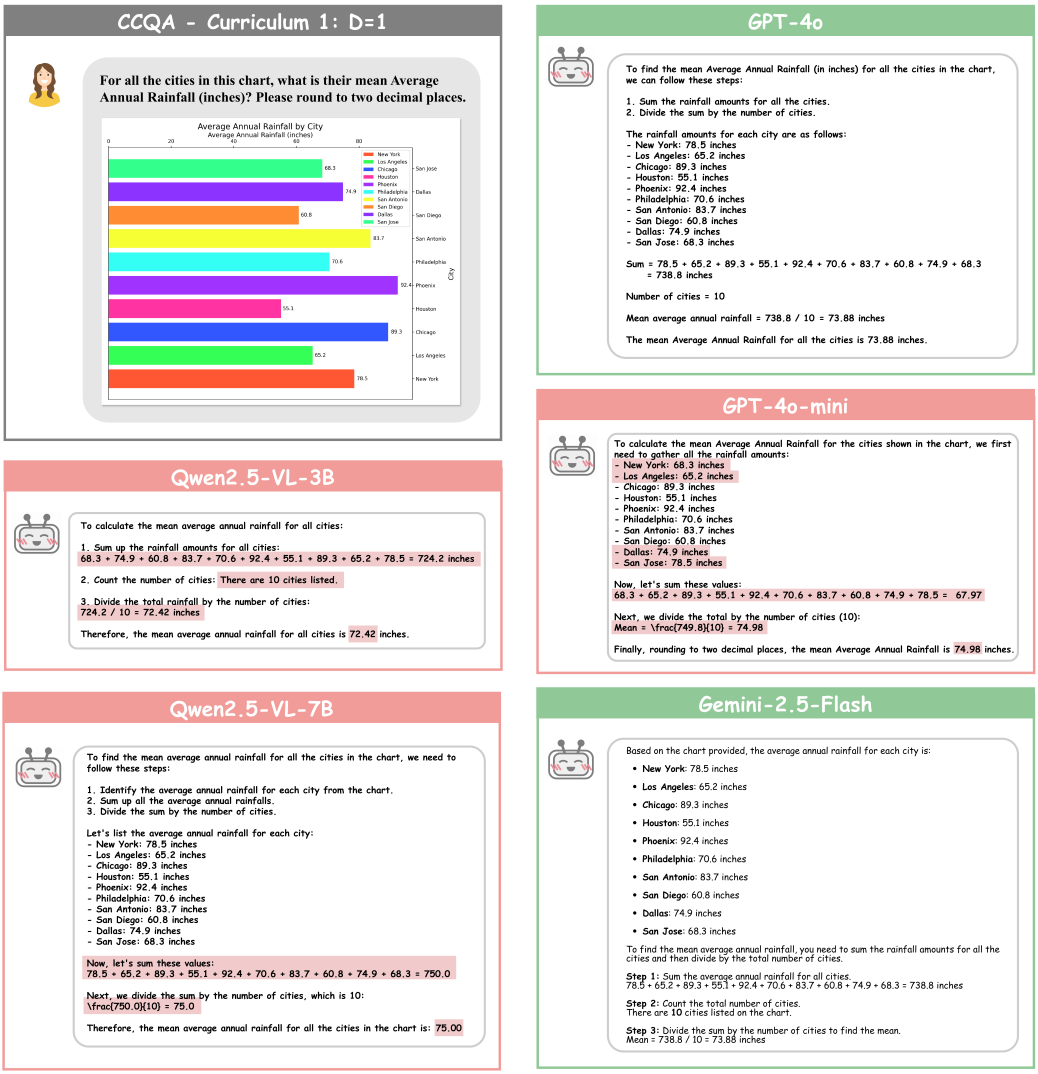


Figure 14: **Examples of Validation On Curriculum Level 1 ($D = 1$).** We validate our CCQA on *Tier 1* (§ C.3), i.e., *Curriculum Level 1* ($D = 1$), by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

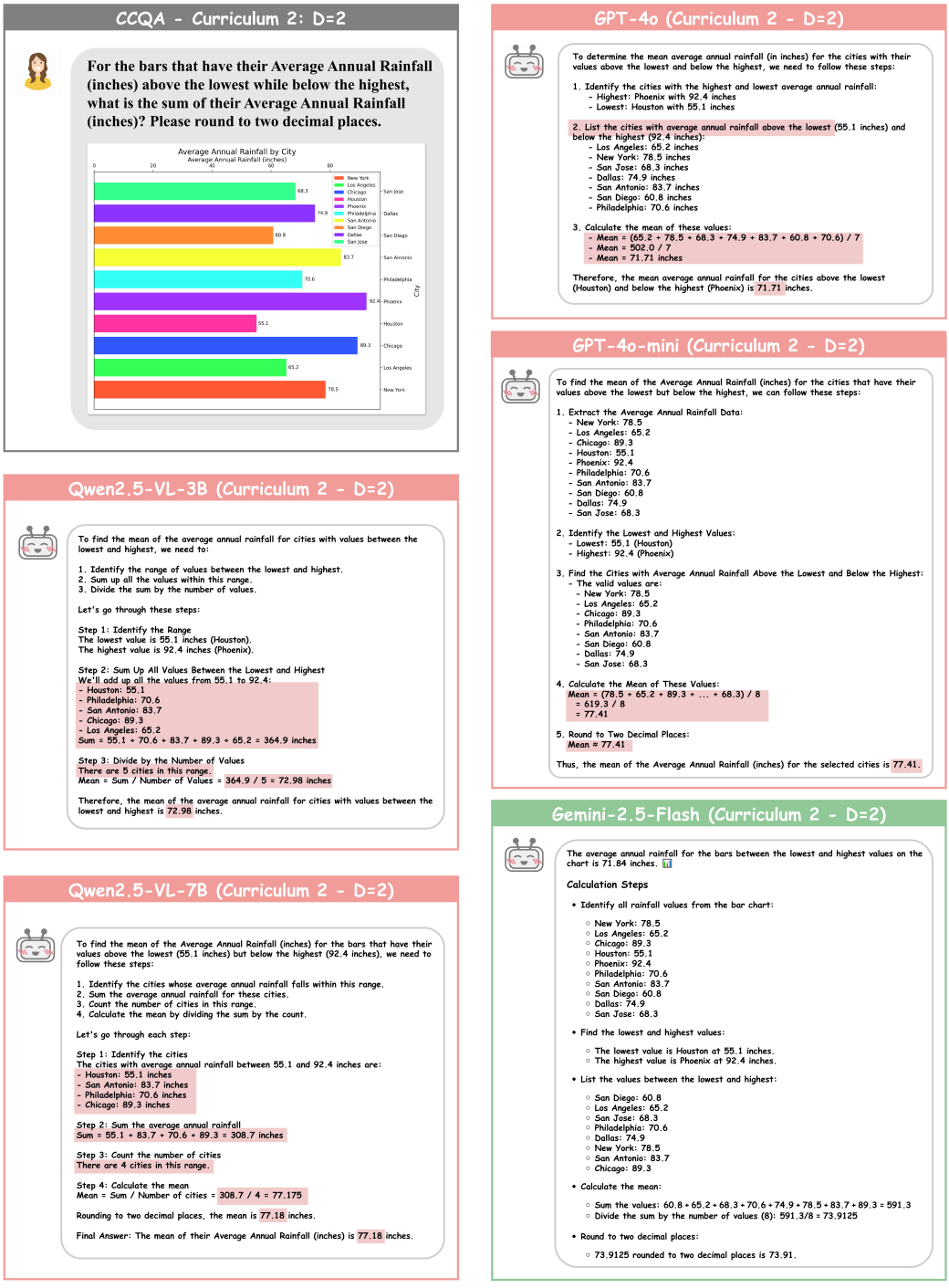
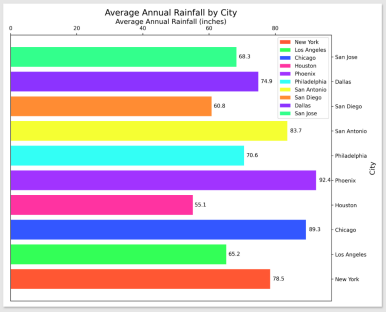


Figure 15: **Examples of Validation On Curriculum Level 2 ($D \geq 2$).** We validate our CCQA on *Tier 2* (§ C.3), i.e., *Curriculum Level 2* ($D \geq 2$), by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

CCQA - Curriculum 2: D=3

What is the lowest Average Annual Rainfall (inches) among bars that have their Average Annual Rainfall (inches) above the average Average Annual Rainfall (inches) of all bars? Please round to two decimal places.



GPT-4o (Curriculum 2 - D=3)

Step 1: Calculate the average rainfall of all cities:
The rainfall data for the cities are:
- San Jose: 68.3 inches
- Dallas: 74.9 inches
- San Diego: 60.8 inches
- San Antonio: 83.7 inches
- Philadelphia: 70.6 inches
- Phoenix: 92.4 inches
- Houston: 55.1 inches
- Chicago: 89.3 inches
- Los Angeles: 65.2 inches
- New York: 78.5 inches

Sum of all rainfall:
68.3 + 74.9 + 60.8 + 83.7 + 70.6 + 92.4 + 55.1 + 89.3 + 65.2 + 78.5 = 738.8 inches

Number of cities = 10

Average rainfall:
738.8 inches / 10 cities = 73.88 inches

Step 2: Identify cities with rainfall above the average of 73.88 inches:
Cities with rainfall above 73.88 inches:
- Dallas: 74.9 inches
- San Antonio: 83.7 inches
- Phoenix: 92.4 inches
- Chicago: 89.3 inches
- New York: 78.5 inches

Step 3: Find the lowest rainfall among the identified cities:
Rainfall values for above-average cities:
- Dallas: 74.9 inches
- San Antonio: 83.7 inches
- Phoenix: 92.4 inches
- Chicago: 89.3 inches
- New York: 78.5 inches

By comparing these values, the lowest rainfall among them is:
Dallas: 74.9 inches

Qwen2.5-VL-3B (Curriculum 2 - D=3)

To determine the lowest average annual rainfall among bars with an average above the overall average, we need to follow these steps:

1. Calculate the overall average annual rainfall:
   - Sum up all the rainfall values: 68.3 + 74.9 + 60.8 + 83.7 + 70.6 + 92.4 + 55.1 + 89.3 + 65.2 + 78.5 = 749.2 inches
   - Divide by the number of cities (10): 749.2 / 10 = 74.92 inches

2. Identify the cities with an average above this value:
   - New York: 78.5 inches
   - Chicago: 89.3 inches
   - Los Angeles: 65.2 inches

3. Among these cities, find the one with the lowest rainfall:
   - New York: 78.5 inches
   - Chicago: 89.3 inches
   - Los Angeles: 65.2 inches

The city with the lowest average annual rainfall among those above the overall average is Los Angeles, with 65.2 inches.

GPT-4o-mini (Curriculum 2 - D=3)

1. Extract the Average Annual Rainfall data:
   - New York: 68.3
   - Los Angeles: 65.2
   - Chicago: 89.3
   - Houston: 55.1
   - Phoenix: 92.4
   - Philadelphia: 70.6
   - San Antonio: 83.7
   - San Diego: 60.8
   - Dallas: 80.0
   - San Jose: 78.5

2. Calculate the average Annual Rainfall of all bars:
   - Sum of all rainfall: \(68.3 + 65.2 + 89.3 + 55.1 + 92.4 + 70.6 + 83.7 + 60.8 + 80.0 + 78.5 = 883.9 \)
   - Number of cities: 10
   - Average rainfall = \(\frac{883.9}{10} = 88.39 \)

3. Identify the values above the average (88.39):
   - Phoenix: 92.4
   - Chicago: 89.3
   - San Antonio: 83.7 (not above the average)
   - New York: 68.3 (not above the average)
   - Dallas: 80.0 (not above the average)
   - San Diego: 60.8 (not above the average)
   - Houston: 55.1 (not above the average)
   - Philadelphia: 70.6 (not above the average)
   - Los Angeles: 65.2 (not above the average)
   - San Jose: 78.5 (not above the average)

The cities with rainfall above the average are:
- Phoenix: 92.4
- Chicago: 89.3

4. Determine the lowest Average Annual Rainfall among those:
   - Lowest among Phoenix (92.4) and Chicago (89.3) is Chicago with 89.3.

5. Final Answer:
   - The lowest Average Annual Rainfall among bars that have their Average Annual Rainfall above the average is 89.30 inches.

Qwen2.5-VL-7B (Curriculum 2 - D=3)

To find the lowest Average Annual Rainfall (inches) among bars, we need to go through the following steps:

Step 1: Calculate the average Average Annual Rainfall (inches)
First, we sum up the Average Annual Rainfall (inches) for all cities:
68.3 + 74.9 + 60.8 + 83.7 + 70.6 + 92.4 + 55.1 + 89.3 + 65.2 + 78.5 = 732.5

Next, we divide this sum by the number of cities (10):
\\[\\text{Average} = \\frac{732.5}{10} = 73.25 \\]

Step 2: Identify cities with Average Annual Rainfall (inches) above the average
We compare each city's Average Annual Rainfall (inches) to 73.25 inches:
- New York: 78.5 (above)
- Los Angeles: 65.2 (below)
- Chicago: 89.3 (above)
- Houston: 55.1 (below)
- Phoenix: 92.4 (above)
- Philadelphia: 70.6 (above)
- San Antonio: 83.7 (above)
- San Diego: 60.8 (below)
- Dallas: 74.9 (above)
- San Jose: 68.3 (below)

The cities with Average Annual Rainfall (inches) above the average are:
- New York
- Chicago
- Phoenix
- Philadelphia
- San Antonio
- Dallas

Step 3: Find the lowest Average Annual Rainfall (inches) among those cities
From the list above, we identify the lowest value:
- New York: 78.5
- Chicago: 89.3
- Phoenix: 92.4
- Philadelphia: 70.6
- San Antonio: 83.7
- Dallas: 74.9

The lowest value among these is 70.6 inches.
Thus, the lowest Average Annual Rainfall (inches) among bars that have their Average Annual Rainfall (inches) above the average is 70.6 inches.

Gemini-2.5-Flash (Curriculum 2 - D=3)

Calculating the Average of All Bars

To find the average of all the bars, we sum all the values and divide by the number of values (which is 10).

(78.5 + 65.2 + 89.3 + 55.1 + 92.4 + 70.6 + 83.7 + 60.8 + 74.9 + 68.3)/10
638.8/10 = 63.88

The average annual rainfall for all cities is 63.88 inches.

Identifying Bars Above the Average

Next, we identify all the cities that have an average annual rainfall greater than 63.88 inches:

- New York: 78.5 inches
- Los Angeles: 65.2 inches
- Chicago: 89.3 inches
- Phoenix: 92.4 inches
- Philadelphia: 70.6 inches
- San Antonio: 83.7 inches
- Dallas: 74.9 inches
- San Jose: 68.3 inches

Finding the Lowest Value Among Those Bars

Finally, from this new list of values, we find the lowest one.

The values are: 78.5, 65.2, 89.3, 92.4, 70.6, 83.7, 74.9, and 68.3.

The lowest value in this group is 70.6, which is the average annual rainfall for Philadelphia.

Figure 16: **Examples of Validation On Curriculum Level 2 ($D \geq 3$).** We validate our CCQA on *Tier 3* (§ C.3), i.e., *Curriculum Level 2* ($D \geq 3$), by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

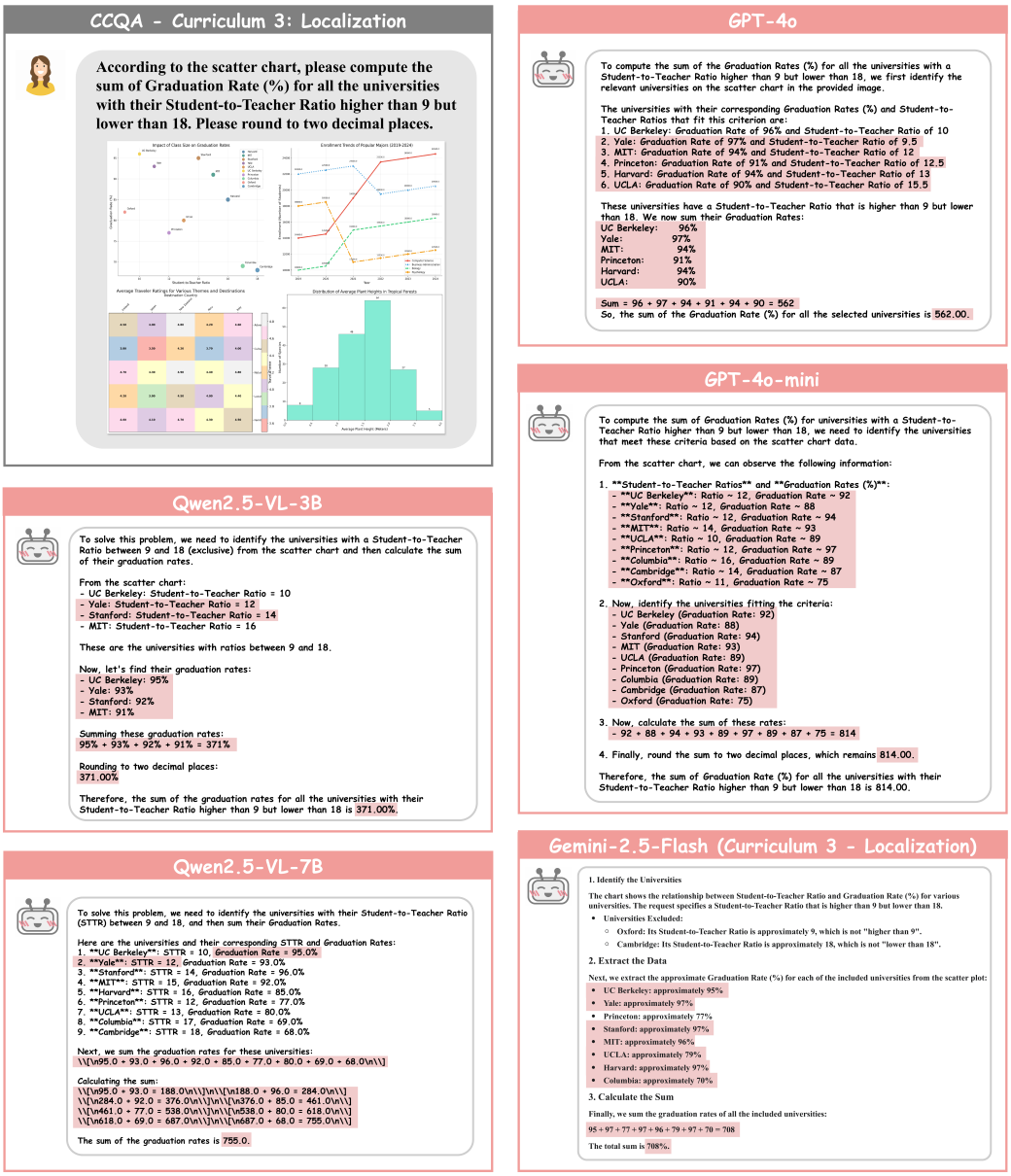


Figure 17: **Examples of Validation On Curriculum Level 3 ($D \geq 3$).** We validate our CCQA on *Tier 4* (§ C.3), i.e., *Curriculum Level 3* ($D \geq 3$), by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

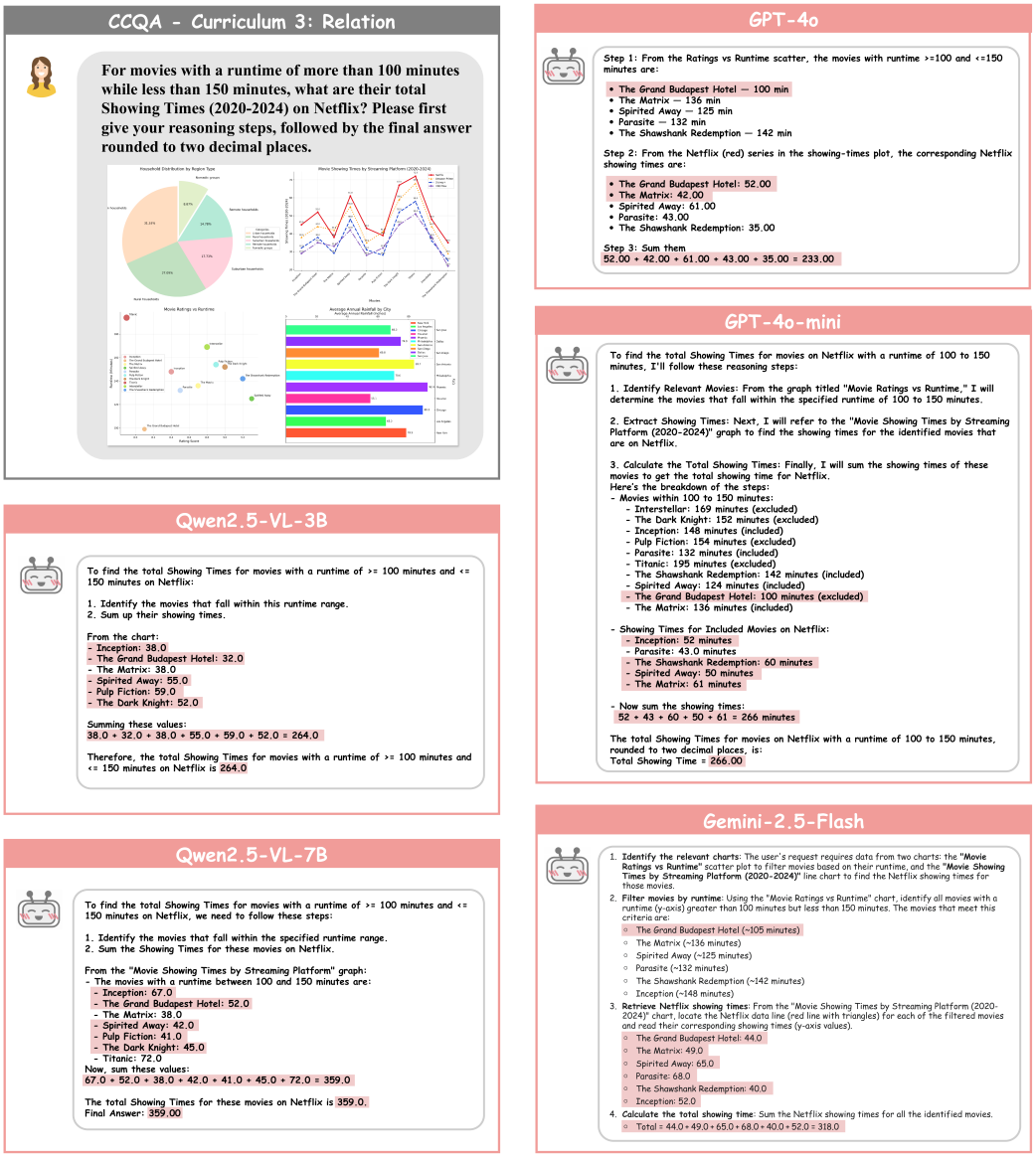


Figure 18: **Examples of Validation On Curriculum Level 3 ($D \geq 3$).** We validate our CCQA on *Tier 5* (§ C.3), i.e., *Curriculum Level 3* ($D \geq 3$), by prompting the model to generate CoT reasoning and final answer. CoT reasoning prompt is omitted in the input for clarify.

B Related Work

Chain-of-Thought Reasoning. CoT reasoning has emerged as foundations for enhancing the interpretability and performance of large language models (LLMs). Even prompting LLMs to perform CoT reasoning before answering can improve performance (Wei et al., 2022; Wang et al., 2023). CoT reasoning is particularly beneficial for MLLMs in complex visual reasoning tasks where attentions are interrelated to both visual and textual features (Zhang et al., 2024b; Shao et al., 2024; Zhao et al., 2025; Qi et al., 2024). Recent work enhances CoT via rationale-augmented training (Zhang et al., 2024b) and feedback-based refinement (Chen et al., 2024), as well as incorporating simple visual operations into reasoning (Qi et al., 2024). However, most approaches remain largely text-centric or loosely grounded, failing to tightly integrate visual evidence into each reasoning step.

Visual Grounding. Visual grounding aligns language with visual regions, enabling spatially-aware localization of entities, attributes, and relationships. Through *explicit* visual localizations (Gou et al., 2024) or *implicit* coordinate-free grounding (Kang et al., 2025), recent advances in segmentation (Carion et al., 2025) and GUI grounding (Cheng et al., 2024; Wu et al., 2025) significantly improve text-visual alignment. Nevertheless, grounding is typically regarded as a standalone objective or terminal prediction, rather than an integral component of the reasoning process. In contrast, our work leverages visual grounding as an intermediate and iterative medium within the reasoning chain, where visual changes dynamically inform subsequent reasoning steps and vice versa, shifting visual grounding from a passive alignment objective to an active component in interleaved multimodal reasoning.

Multimodal Chart Understanding. CQA represents a specialized task that requires accurate understanding of structured visual representations and complex reasoning over visual and textual elements. Recent benchmarks have focused on real-world chart complexity and diversity, such as ChartQA (Masry et al., 2022), ChartQA-Pro (Masry et al., 2025), ChartMuseum (Tang et al., 2025), CharXiv (Wang et al., 2024), etc. Prior work addresses CQA via structured pipelines (Kim et al., 2024), compositional reasoning (Zhang et al., 2024a), improved chart representation (Zheng et al., 2024), and synthetic data generation (Li et al., 2024b). However, these methods often decouple perception from reasoning or rely on static representations, limiting adaptive visual dynamics throughout reasoning.

C Dataset Construction

Chart Metadata. Extending our introduction to CCQA (§4), we elaborate on the seven types of charts, the domain categories of the source plotting data, and fundamental operators that support multi-layer nested functions, as summarized in Tables 4-5.

Data Augmentation. To effectively support curriculum learning with meta-learning insights (§C.4), we design a comprehensive set of chart-specific data augmentation strategies implemented through chart rendering functions. These augmentations introduce controlled variability in both the structural layout and visual presentation of charts, thereby enhancing model generalization across diverse chart types. Specifically, we consider the following transformations: (1) chart rotation at different angles (e.g., 0°, 30°, 45°, 60°, 75°, 90°); (2) orientation adjustments between vertical and horizontal layouts; (3) axis placement variations, such as shifting the x- and/or y-axis among left, right, top, and bottom; (4) color setting across chart elements; (5) legend positioning (top, bottom, center, left, or right) and visibility; (6) label positions and visibili(ty such as axis labels, tick labels, and numeric annotations); and other augmentation strategies tailored for specific types of charts (e.g., 'explode' settings for pie charts, marker styles for scatter plots, etc.) Collectively, these augmentation strategies form a systematic approach for generating richly diverse chart appearances using a small set of metadata, ensuring robustness and adaptability of models trained under meta-learning supported curriculum learning.

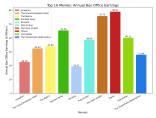
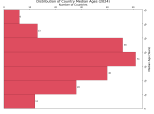
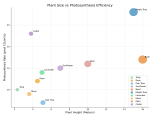
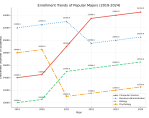
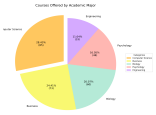
Category	Description
Chart Types	Bar  Histogram  Scatter  Line  Heatmap  Pie  Radar 
Domain Categories	1. Media & Entertainment 2. Geography & Demography 3. Education & Academia 4. Business & Industry 5. Major & Course 6. Animal & Zoology 7. Plant & Botany 8. Biology & Chemistry 9. Food & Nutrition 10. Space & Astronomy 11. Sale & Merchandise 12. Market & Economy 13. Sports & Athletics 14. Computing & Technology 15. Health & Medicine 16. Energy & Environment 17. Travel & Expedition 18. Arts & Culture 19. Communication & Collaboration 20. Language & Linguistics 21. History & Archaeology 22. Weather & Climate 23. Transportation & Infrastructure 24. Psychology & Personality 25. Materials & Engineering 26. Philanthropy & Charity 27. Fashion & Apparel 28. Parenting & Child Development 29. Architecture & Urban Planning 30. Gaming & Recreation

Table 4: **Metadata of Chart Plotting.** We employ seven commonly used chart types across 30 different domain categories to construct the meta images for CCQA.

Operator	Description
Read	Read or estimate the value of chart component(s) that meet given requirement(s)
Statistics - Sum/Mean/Median	Calculate the sum/mean/median of a group of chart components that meet given requirement(s)
Statistics - Count	Count the number of chart components that meet given requirement(s)
Extrema - Value - Min/Max	Calculate the minimum/maximum value (which may be combined with nested functions, *e.g.*, the minimum mean value of two groups of chart components) of chart components that meet given requirement(s)
Extrema - Position	Localize chart component(s) that meet given requirement(s), *e.g.*, the leftmost bar in the bar chart
Sort - Ascending/Descending	Sort a group of chart components that meet given requirement(s)
Compare - Value/Diff/Position	Compare the value/difference/position of two groups of chart components based on the given requirement(s)
Filter	Filter chart component(s) based on the given requirement(s)
Threshold	Identify chart component(s) based on the given threshold condition(s)
Subset	Identify the subset of chart component(s) that satisfy the specified requirement(s)
Localization	Localize specific chart components and/or subplots
Relation	Understand relations between or among different chart components and/or subplots

Table 5: **Foundational Operators.** Our CCQA incorporates 12 basic operators to query different aspects of chart components, facilitating comprehensive understanding of each chart elements.

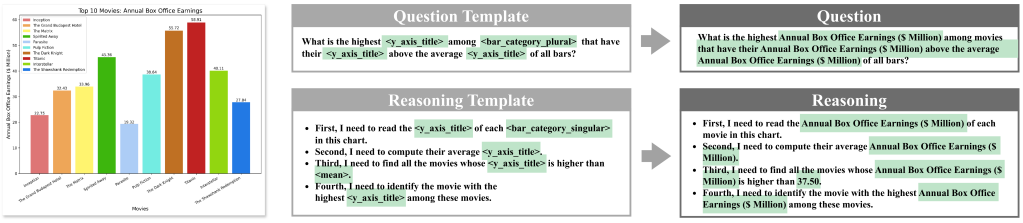


Figure 19: **From Template To CQA Data.** A template-based data generation example that illustrates how question and reasoning templates are converted to CQA data based on the chart data.

C.1 Data Structure

Our CCQA (§4) encompass seven basic chart types, including *bar chart, histogram, scatter plot, line chart, heatmap, pie chart, and radar chart* (Tab. 4). Only the chart plotting data (*i.e.*, the value and label of each chart component, along with the axis and image titles) are generated by GPT-4o (see an example in Fig. 20). We construct the CQA data of each chart type and curriculum level across 30 domain categories (Tab. 4). The number of samples for each chart type is influenced by chart features (e.g., scatter plots depend on both X- and Y-axis features, whereas heatmaps depend on cell values and labels), CQA types (Tab. 4, 5), and properties

of the source plotting data (e.g., the number of bars, scatter points, or cells; variations in label angles; etc.). To ensure high data quality, we implement template-based CQA generation that guarantees not only the diversity of CQA tasks but also the accuracy and reliability of intermediate reasoning, visual grounding, and final answers.

Specifically, all question-answer pairs in CCQA, together with their corresponding reasoning steps and dynamic visual grounding coordinates, are generated using human-defined templates and functions (§4.1). An example is shown in Fig. 19 to illustrate the template-based CQA data generation process.

GPT-4o Prompt For Plotting Data Generation

Can you help with heatmap chart plotting data generation for my heatmap plotting?

Data Example
This plotting data example below is an input data dictionary for the first domain category "1 - Media & Entertainment".

```python
HEATMAP_DICT = {
   "heatmap_data": [[80.5, 20.5, 55.5, 90.0, 60.0], [30.5, 75.0, 45.5, 60.0, 80.0], [90.0, 10.5, 80.0, 20.0, 70.5], [80.0, 60.5, 30.0, 90.0, 59.5], [77.5, 80.5, 66.5, 25.5, 95.5]],
   "heatmap_category": {"singular": "rating score", "plural": "rating scores"},
   "x_labels": ["Titanic", "Star Wars", "The Lord of the Rings", "Snow White and the Seven Dwarfs", "The Lion King"],
   "y_labels": ["2005", "2010", "2015", "2020", "2025"],
   "x_label": "Movies",
   "y_label": "Rating Year",
   "image_title": "Correlation Between Movie Ratings and Rating Year",
}
```

Your Task
For the second domain category "2 - Geography & Demography", can you help generate a new plotting data using the same dictionary structure and data format?

Important Requirements
* Please generate the new plotting data specifically for the new domain category.
* Each data entry should be unique and should simulate real-world circumstances of the new domain category.
* Please be careful about `heatmap_category`, which will be used as how to generally call the heatmap cells. For example, `sentence=f"these heatmap cells represent different {HEATMAP_DICT['heatmap_category']['plural']}"` will become "these heatmap cells represent different *movie rating scores*".
* Please response in the following format:

```python
HEATMAP_DICT = <your new data dictionary>
```

```python
HEATMAP_DICT = {
   "heatmap_data": [
      [70.5, 82.0, 65.0, 90.0, 75.0],
      [60.0, 55.0, 85.0, 80.0, 70.0],
      [55.0, 78.0, 88.0, 60.0, 65.0],
      [80.0, 68.0, 72.0, 85.0, 90.0],
      [95.0, 75.0, 60.0, 70.0, 80.0]
   ],
   "heatmap_category": {"singular": "urbanization index", "plural": "urbanization indices"},
   "x_labels": ["Tokyo", "New York", "Paris", "São Paulo", "Lagos"],
   "y_labels": ["1990", "2000", "2010", "2020", "2030"],
   "x_label": "Cities",
   "y_label": "Year",
   "image_title": "Urbanization Indices Over Time in Major Global Cities"
}
```

Figure 20: **Example of Plotting Data Generation Prompt.** We prompt GPT-4o to generate the source plotting data, which will be used as the input for chart drawing functions. This example is for the heatmap plotting data generation of the second domain category (Tab. 4).

C.2 Multi-Level Curriculum

Implementing curriculum learning to progressively increase reasoning difficulty across three distinct levels (Fig. 5), each level targets at specific aspects of visual reasoning development (Tab. 5): *read*, *statistics*, *extrema*, *sorting*, *comparison*, *filtering*, *thresholding*, *subset constraints*, *localization*, and *relation*. The knowledge transfer between levels contributes to the increase of task complexity (§C).

LEVEL 1: Foundational Single-Operation Reasoning. LEVEL 1 establishes fundamental chart understanding capabilities (*reasoning depth*: $D_1 = 1$) with single-operation reasoning processes on single-plot charts:

$$\texttt{SINGLE}(x) = f(x) \tag{7}$$

where $f(\cdot)$ represents basic operations such as direct value reading, simple arithmetic, and elementary comparisons within a single-plot chart. Hereby, the ***reasoning depth*** of a CQA data sample, denoted as D_l, is defined as the number of nested operations for curriculum level l.

Accordingly, the template structure of LEVEL-1 CQA data follow the format:

$$\text{TEMPLATE}_1 = \{Q, \{R_t, B_t^*\}_{t=1}^{T_1}, A\} \tag{8}$$

where Q is the question, $T_1 \geq 1$ is the number of reasoning steps, R_t is the t-th reasoning step, B_t^* is the corresponding ground-truth visual grounding with reasoning depth $D_1 = 1$, and A is the final answer.

LEVEL 2: Multi-Operation Reasoning. LEVEL 2 introduces compositional reasoning (*reasoning depth*: $D_2 > 1$) through nested operations on single charts:

$$\texttt{SINGLE}(x) = F(f(x)) \tag{9}$$

where $F(\cdot)$ represents composite operations applied to $f(x)$, *e.g.*, $F(f(x)) = h(g(f(x)))$. This level requires models to perform sequential reasoning and visual grounding where each step builds upon previous computations. Consequently, LEVEL 2 templates extend to multi-step reasoning:

$$\text{TEMPLATE}_2 = \{Q, \{R_t, B_t^*\}_{t=1}^{T_2}, A\} \tag{10}$$

where $T_2 \geq 2$ is the number of reasoning steps, each reasoning step R_t progresses through nested operations with reasoning depth $D_2 \geq 2$, and B_t^* is the corresponding ground-truth visual grounding for the t-th step.

LEVEL 3: Complex Multi-Chart Reasoning LEVEL 3 represents the most challenging scenarios that involve complex reasoning ($D_3 \geq 2$) across multiple subplots and chart types:

$$\texttt{MULTI}(x) = \texttt{MULTI}(\texttt{SINGLE}_i(x)) \tag{11}$$

where i conforms to $1 \leq i \leq n$ and $1 \leq n \leq \textit{subplot\_num}$. LEVEL 3 templates thereby incorporate multi-step and cross-chart dependencies:

$$\text{TEMPLATE}_3 = \{Q, \{R_t, B_t^*, C_t\}_{t=1}^{T_3}, A\} \tag{12}$$

where $T_3 \geq 3$ is the number of reasoning steps, each reasoning step R_t progresses through complex nested operations with reasoning depth $D_3 \geq 3$, B_t^* is the corresponding ground-truth visual grounding, and C_t indicates the chart index for the t-th reasoning step. Specifically, our multi-plot reasoning incorporates both localization $(n-1)$ and relation $(n > 1)$ operations across multiple charts.

C.3 Fine-Grained Curriculum Tiers

We construct our three-level curriculum dataset (Fig. 5) based on reasoning depth and chart complexity (§C.2). According to their fine-grained problem-solving difficulty, we categorize them into five curriculum tiers:

- **Tier 1: Curriculum Level 1 ($D = 1$).** In *Tier 1*, all CQA data correspond to queries about single-plot chart image input. Reasoning is limited to one depth level, *i.e.*, single-function reasoning (Eq. 7, $D_1 \geq 1$), to derive the final answer.
- **Tier 2: Curriculum Level 2 ($D = 2$).** In *Tier 2*, all CQA data correspond to queries about single-plot chart image input. Reasoning requires two depth levels, *i.e.*, constructed through two nested functions (Eq. 9, $D_2 \geq 2$), to derive the final answer.

- **Tier 3: Curriculum Level 2 ($D \geq 3$).** In *Tier 3*, all CQA data correspond to queries about single-plot chart image input, with reasoning that involves three or more depth levels, *i.e.*, constructed through three or more nested functions (Eq. 9, $D_2 \geq 3$), to derive the final answer.
- **Tier 4: Curriculum Level 3 - Localization ($D \geq 3$).** In *Tier 4*, all CQA data correspond to queries about multi-plot chart image input, with reasoning that involves three or more depth levels, *i.e.*, constructed through three or more nested functions (Eq. 11, $D_3 \geq 3$), to derive the final answer. While different from single-plot charts, multi-plot CQA tasks in *Tier 4* involves the precise localization of target subplot(s) that directly yield the answer.
- **Tier 5: Curriculum Level 3 - Relation ($D \geq 3$).** CQA tasks in *Tier 5* are similar to the constitution of *Tier 4*, corresponding to queries about multi-plot charts with reasoning that involves three or more depth levels, *i.e.*, constructed through three or more nested functions (Eq. 11, $D_3 \geq 3$). The key distinction is that, while *Tier 4* emphasizes precise localization of target subplot(s), *Tier 5* additionally demands the modeling of relations across the identified subplots.

C.4 Meta-Learning Supported Curriculum Learning

Our curriculum learning design (§3) is reinforced through meta-learning, which provides a principled way to structure both data and task complexity. Specifically, we leverage meta-learning through the following aspects:

1. **Domain diversity as meta-tasks.** We construct CCQA using 30 domain categories (Tab. 4), where each category contributes one source plotting data, and thus one chart image. This structured diversity provides a wide range of meta-tasks that expose MLLMs to domain-generalizable visual reasoning.
2. **Chart-type variability as meta-structures.** We employ 7 fundamental chart types (Tab. 4) to visualize the 30 domain datasets. Multiplying 30 plotting datasets by 7 chart types yields 210 unique chart-structure metadata, based on which the entire dataset is systematically constructed. This ensures that each domain is represented across diverse chart structures, promoting cross-task adaptation.
3. **Operator set as meta-functions.** To support multi-layer nested reasoning, we define 12 fundamental operators (Tab. 5). These operators serve as compositional primitives for constructing multi-level CQA tasks. By progressively increasing the depth of nesting, we control CQA difficulty level, thereby enabling MLLMs to gradually acquire higher-order reasoning capabilities.
4. **Decomposition as learning scaffolds.** Following decomposition insights (§A.1), we disentangle each task into *visual decomposition* (low-level chart components) and *reasoning decomposition* (singular operations across nested functions). This scaffolding allows MLLMs to incrementally learn fine-grained visual perception and step-wise logical reasoning, supporting the high-level composition of accurate chart understanding.
5. **Meta-learning for transferability.** Beyond dataset construction, our design leverages meta-learning to encourage transferability across chart types, domains, and reasoning depths. By repeatedly exposing MLLMs to varied meta-tasks with systematically controlled complexity, we enable them to acquire generalizable strategies rather than overfitting to particular chart types or reasoning templates. Our meta-learning implementation strengthens the robustness of curriculum learning by aligning it with principles of adaptation and generalization.

Together, these design principles ensure that our curriculum learning is not only systematic but also meta-learnable, allowing MLLMs to progressively integrate visual and reasoning competencies across tasks of increasing complexity.

C.5 Quality Control

To ensure high data fidelity for both training and evaluation, we enforce rigorous quality control throughout the construction of CCQA:

First, we implement the meta-learning-driven construction pipeline in which each chart type is defined by human-authored meta-functions. These meta-functions explicitly encode the data generation logic and reasoning structure, with configurable chart attributes such as the number of chart elements, spatial arrangements, structural relationships, shapes and color settings, etc. This design enables precise and controllable synthesis of visual reasoning samples while preserving data diversity across multiple chart visual dimensions, such as colors, shapes, sizes, positions, etc. Prior to inclusion in the dataset, all generated instances are pre-executed to verify functional correctness and internal consistency.

Second, we perform manual validation of the visualization metadata, including ground-truth visual grounding masks and the corresponding step-by-step reasoning annotations associated with each metadata instance. In addition to manually verifying the 30 *meta images* for each type of charts, we randomly sample 10 visual reasoning instances of each category (*i.e.*, $10 \times 30 \times 7$ in total) to validate accurate vision-reasoning alignment (Tab. 6). This process ensures that every visual grounding step accurately corresponds in multimodal spaces, and that the resulting reasoning chains are coherent, interpretable, and faithful to the underlying data.

Critically, we restrict the role of LLM (GPT-4o) to generating randomized numerical parameters within predefined domain categories for chart construction, which will only be used as inputs in plotting functions as visual grounded reasoning chain templates. That is, these parameters are deterministically incorporated into plotting functions and template-based reasoning chains, pre-computed in predefined equations prior to data instantiation. As a result, the use of LLMs is ensured to exhibit no effect on the correctness of ground-truth answers, thereby preserving the overall reliability and integrity of the dataset.

Human Validation	**Bar**	**Histogram**	**Scatter**	**Line**	**Heatmap**	**Pie**	**Radar**
Meta Image	30	30	30	30	30	30	30
CQA Instance	10	10	10	10	10	10	10

Table 6: **Human Validation on Data Quality**

D Visual Grounding Strategies

We propose three visual grounding strategies — *applied*, *boxed*, and *cropped* (Fig. 3, Tab. 7) — to enable *dynamic* visual focus navigation throughout the evolution of multi-step reasoning. All three strategies follow the same **RVA** process where the model generates reasoning steps accompanied by grounded bounding box coordinates, while the "*dynamic*" nature refers to how the visual focus adaptively changes as the train of thoughts progresses. Each strategy implements a distinct grounding mechanism for directing the model's visual attention to corresponding image regions of focus while maintaining coherent reasoning flow. On the other hand, these strategies also represent different trade-offs between visual clarity, computational efficiency, and reasoning precision (Tab. 7).

D.1 Applied: Grounding Through Dynamic Visual Focus Highlighting

Method. The *applied* grounding strategy directly underlines the predicted regions of focus through semi-transparent yellow highlighting overlays. As the reasoning progresses, the yellow highlighting adaptively shifts to emphasize different visually focused regions mirroring each reasoning step, while preserving full visual context.

Specifically, our *applied* grounding strategy implements visual grounding through semi-transparent highlighting overlays that mask visual focuses:

$$I'_{vis,t} = I_{orig} \odot (1 - \kappa \cdot M_{focus,t}) + \kappa \cdot H_{yellow} \odot M_{focus,t} \tag{13}$$

Method	Low Computation	High Precision	Full Context	No Occlusion	Multi-Region	Easy Integration	Easy Comprehension
applied	✓	✗	✓	✗	✓	✓	✓
boxed	✓	✗	✓	✓	✓	✓	✓
cropped	✗	✓	✓	✓	✓	✗	✗

Table 7: **Comparative Analysis of Visual Grounding Strategies.** We evaluate each grounding strategy across seven key dimensions that are significant for effective visual reasoning. ✓ indicates the strategy possesses the advantage, while ✗ indicates limitation. *Boxed* grounding achieves the best overall balance, *applied* grounding provides clear interpretability and highlighting with moderate trade-offs, and *cropped* grounding maximizes precision at the cost of computational efficiency and understanding straightforwardness.

where I_{orig} is the original input image, $M_{focus,t}$ is the binary mask derived from the list of bounding boxes $\{B_{t,i}\}_{i=1}^{N_t}$ where each $B_{t,i} = [x_{min}, y_{min}, x_{max}, y_{max}]$ and N_t is the number of bounding boxes at reasoning step t, H_{yellow} is the highlight color (*i.e.*, yellow), κ controls transparency, and $\odot$ denotes element-wise multiplication.

Examples. Fig. 21 shows an example for multi-step CoT reasoning with *applied* visual grounding.

D.2 Boxed: Grounding Through Dynamic Visual Box Guides

Method. The *boxed* grounding strategy straightforwardly guides visual attention by adding red rectangular borders to the focus regions. These red boxes dynamically relocate and resize along with the evolution of the reasoning chain, emphasizing current regions of focus through explicit visual boundaries.

Particularly, the border guides are generated by drawing rectangular outlines at the specified coordinates:

$$I'_{vis,t} = \textsc{Box}(I_{orig}, \{B_{t,i}\}_{i=1}^{N_t}, C_{red}, \tau) \tag{14}$$

where I_{orig} is the original input image, $\{B_{t,i}\}_{i=1}^{N_t}$ is the list of bounding boxes at reasoning step t where each $B_{t,i} = [x_{min}, y_{min}, x_{max}, y_{max}]$, $\textsc{Box}(\cdot)$ draws colored rectangular border lines for each specified regions of focus, C_{red} is the border line color (*i.e.*, red), and τ is the thickness of border lines.

Examples. Fig. 22 shows an example for multi-step CoT reasoning with *boxed* visual grounding.

D.3 Cropped: Grounding Through Dynamic Visual Focus Zooming

Method. The *cropped* grounding strategy localizes corresponding regions of focus by zooming in, extracting and presenting the focused sub-regions as separate zoomed images alongside the full chart. As reasoning evolves, different cropped regions are dynamically generated and presented, enabling detailed examination of the specific components relevant to each reasoning step.

Therefore, the cropping operation extracts sub-regions using array indexing based on the bounding box coordinates:

$$I'_{vis,t} = \{I_{orig}, \{\textsc{Crop}(I_{orig}, B_{t,i})\}_{i=1}^{N_t}\} \tag{15}$$

$$\text{where } \textsc{Crop}(I_{orig}, B_{t,i}) = I_{orig}[y_{min} : y_{max}, x_{min} : x_{max}] \tag{16}$$

Here, I_{orig} is the original chart image, $\{B_{t,i}\}_{i=1}^{N_t}$ is the list of bounding boxes at reasoning step t, and the model processes both the full context I_{orig} and multiple zoomed crops $\{\text{CROP}(I_{orig}, B_{t,i})\}_{i=1}^{N_t}$ simultaneously.

Examples. Fig. 23 shows an example for multi-step CoT reasoning with *cropped* visual grounding.

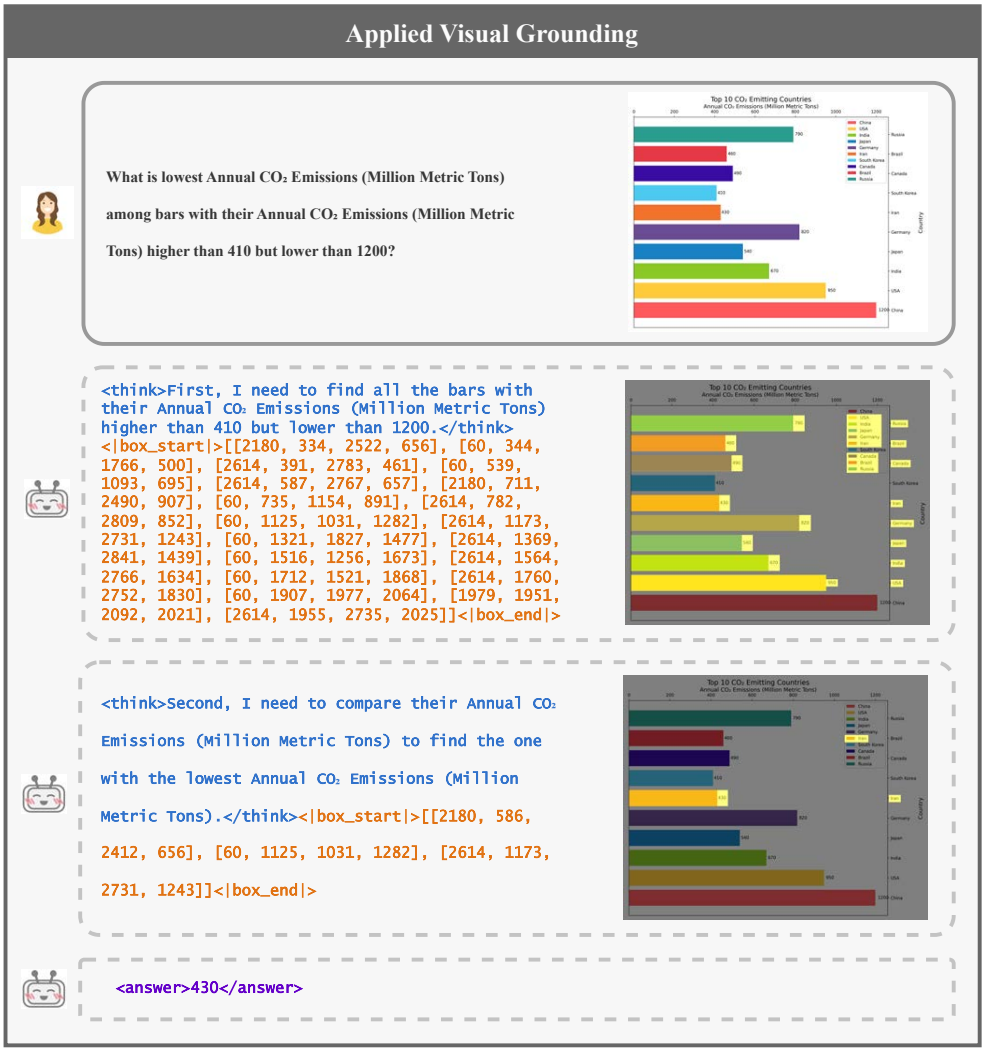


Figure 21: **Example of Applied Visual Grounding.** The *applied* visual grounding method directly accentuates the regions of focus through semi-transparent yellow highlighting overlays.

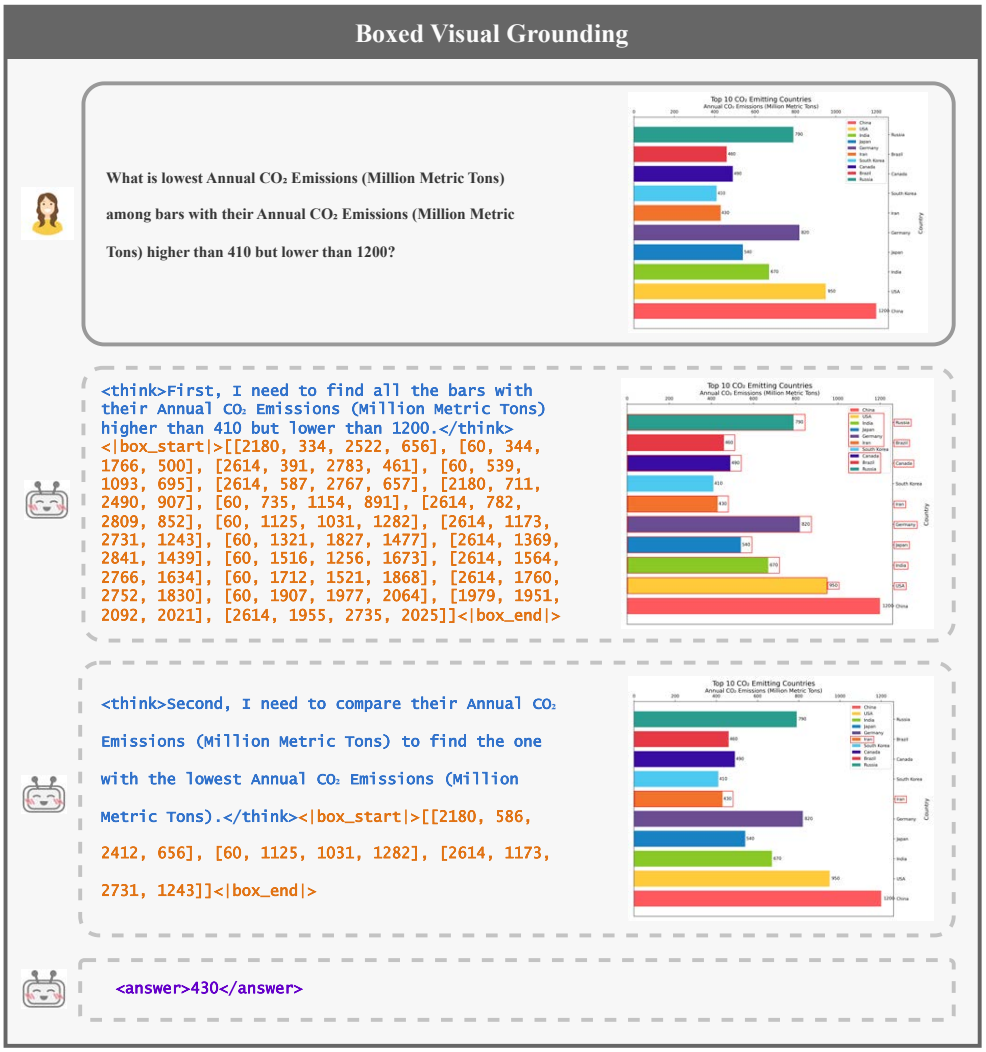


Figure 22: **Example of Boxed Visual Grounding.** The *boxed* visual grounding method directly accentuates the regions of focus through semi-transparent yellow highlighting overlays.

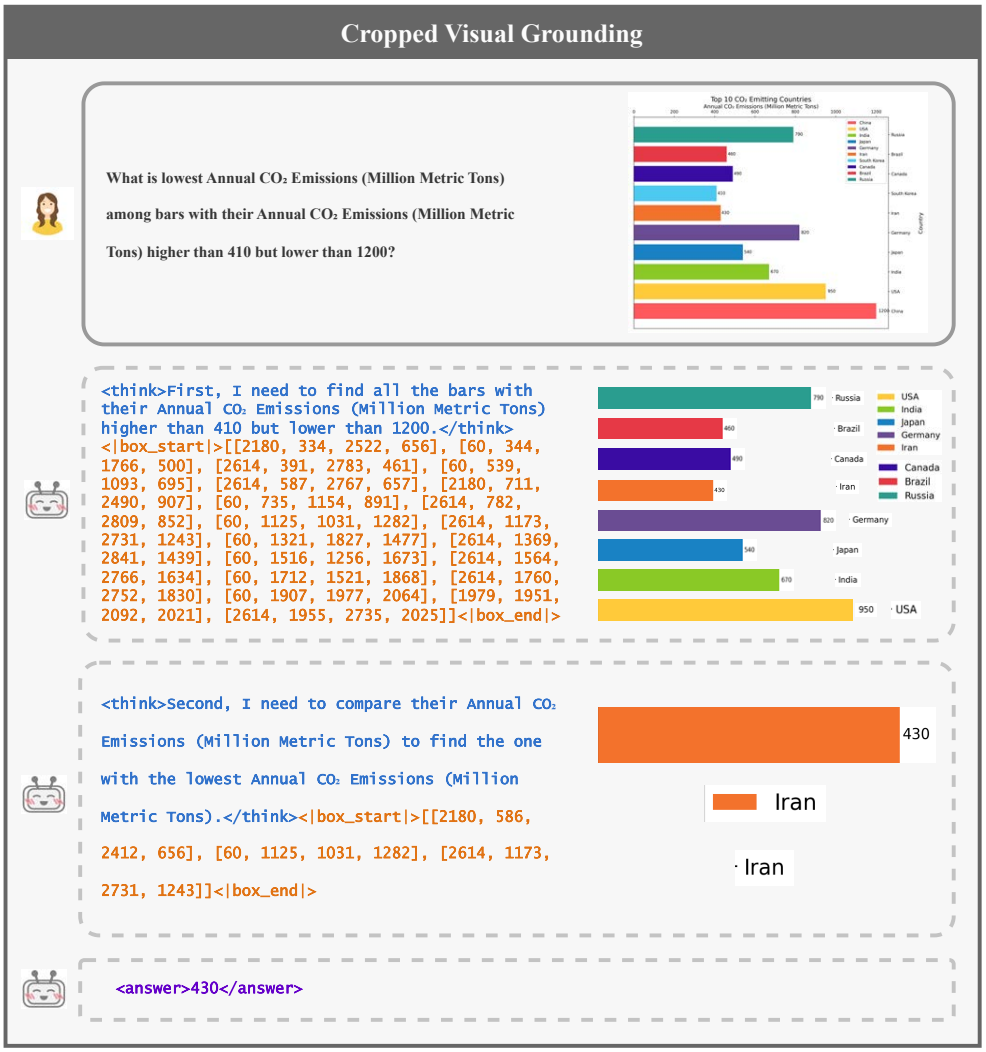


Figure 23: **Example of Cropped Visual Grounding.** The *cropped* visual grounding method directly accentuates the regions of focus through semi-transparent yellow highlighting overlays.

E Generation Mode

E.1 Mode A

Fig. 25 shows an example of generation mode **A**, where the model directly outputs the *answer* without intermediate reasoning and visual grounding.

E.2 Mode VA

Fig. 24 shows an example of generation mode **VA**, through which the model first generates its intermediate *visual grounding* via *applied* grounding method, followed by its final *answer*. For clarity, the input instructions for **VA** generation are omitted in the figure. To save space, the user's intermediate responses are shown as smaller images on the right of each model response, corresponding to the model's response on the left in each turn of the multi-step interaction.

E.3 Mode RA

Fig. 26 shows an example of generation mode **RA**, through which the model first generates its intermediate *CoT reasoning*, followed by its final *answer*. The model is prompted to produce its CoT reasoning and final answer in a single-turn manner. For clarity, the input instructions of CoT reasoning and answering for **RA** mode are omitted in the figure.

E.4 Mode RVA

Fig. 27 shows an example of generation mode **RVA**, where the model first produces its intermediate *reasoning* with *visual grounding*, followed by its final *answer*. Similar to Fig. 24, for clarity, the input instructions for **RVA** mode generation are omitted in the figure. To save space, the user's intermediate responses are shown as smaller images on the right of each model response, corresponding to the model's response on the left in each turn of the multi-step **RVA** reasoning.

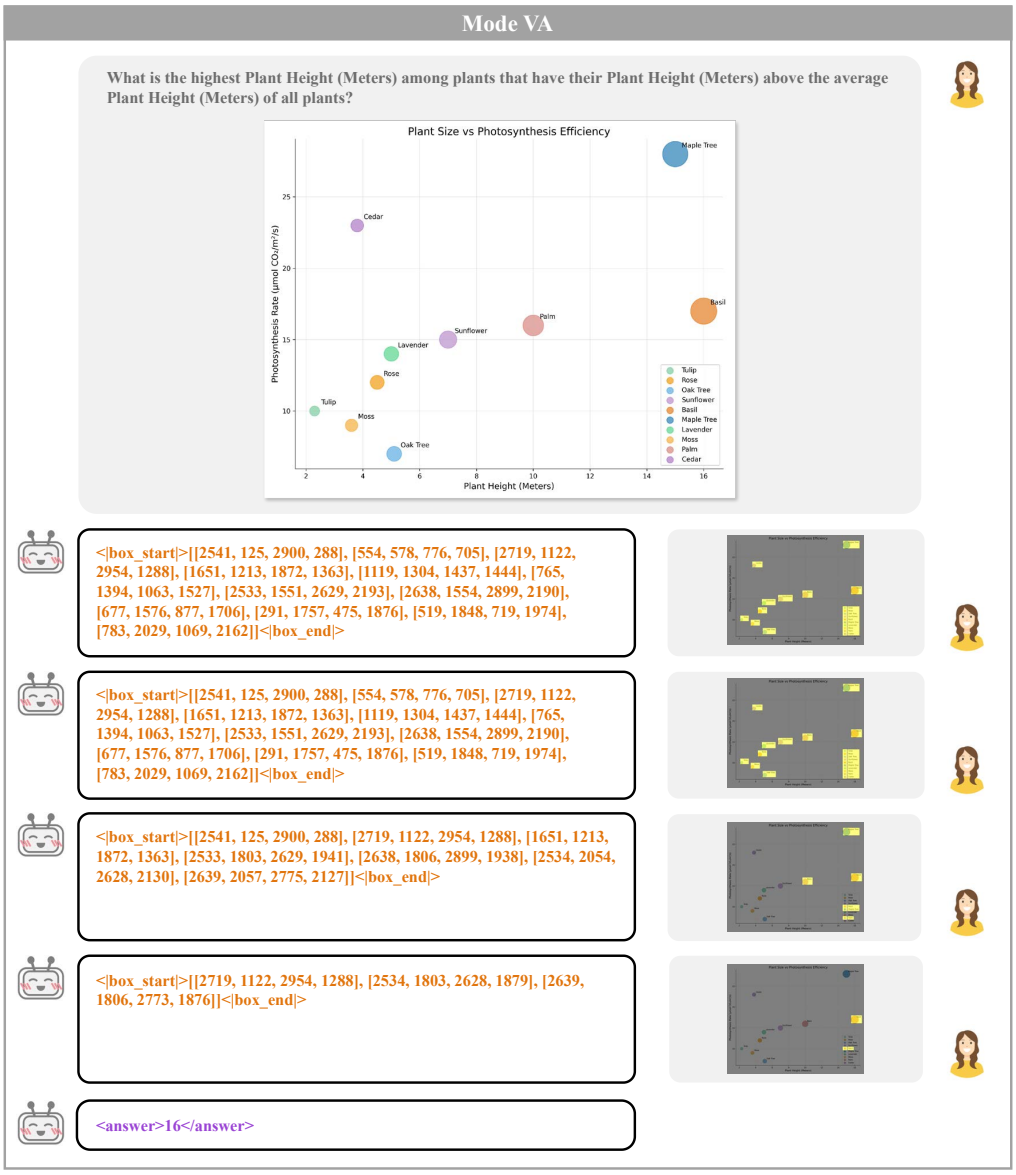
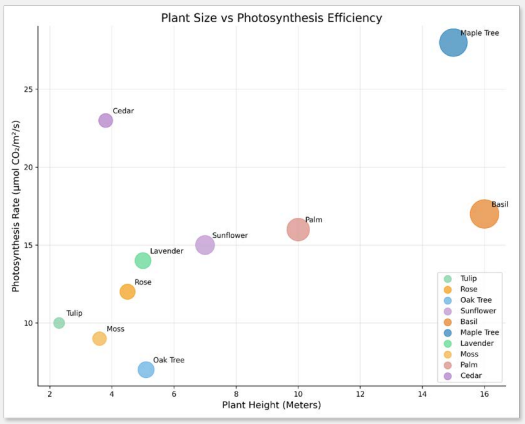


Figure 24: **Example of Generation Mode VA.** A CQA example resolved through generation mode **VA**.

Mode A

What is the highest Plant Height (Meters) among plants that have their Plant Height (Meters) above the average Plant Height (Meters) of all plants?

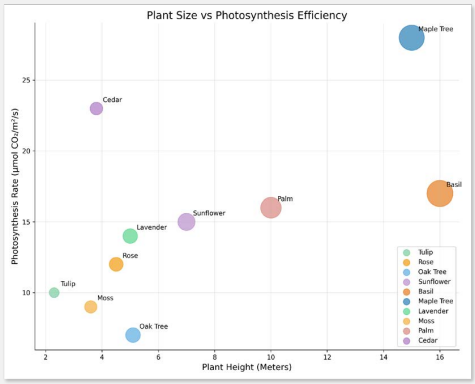


<answer>16</answer>

Figure 25: **Example of Generation Mode A.** A CQA example resolved through generation mode **A**.

Mode RA

What is the highest Plant Height (Meters) among plants that have their Plant Height (Meters) above the average Plant Height (Meters) of all plants?



<think>
First, I need to read the Plant Height (Meters) of each plant in this chart.
Second, I need to compute the average Plant Height (Meters) of all plants.
Third, I need to find all the plants whose Plant Height (Meters) is higher than 7.23.
Fourth, I need to identify the plant with the highest Plant Height (Meters) among these plants.
</think>

<answer>16</answer>

Figure 26: **Example of Generation Mode RA.** A CQA example resolved through generation mode **RA**.

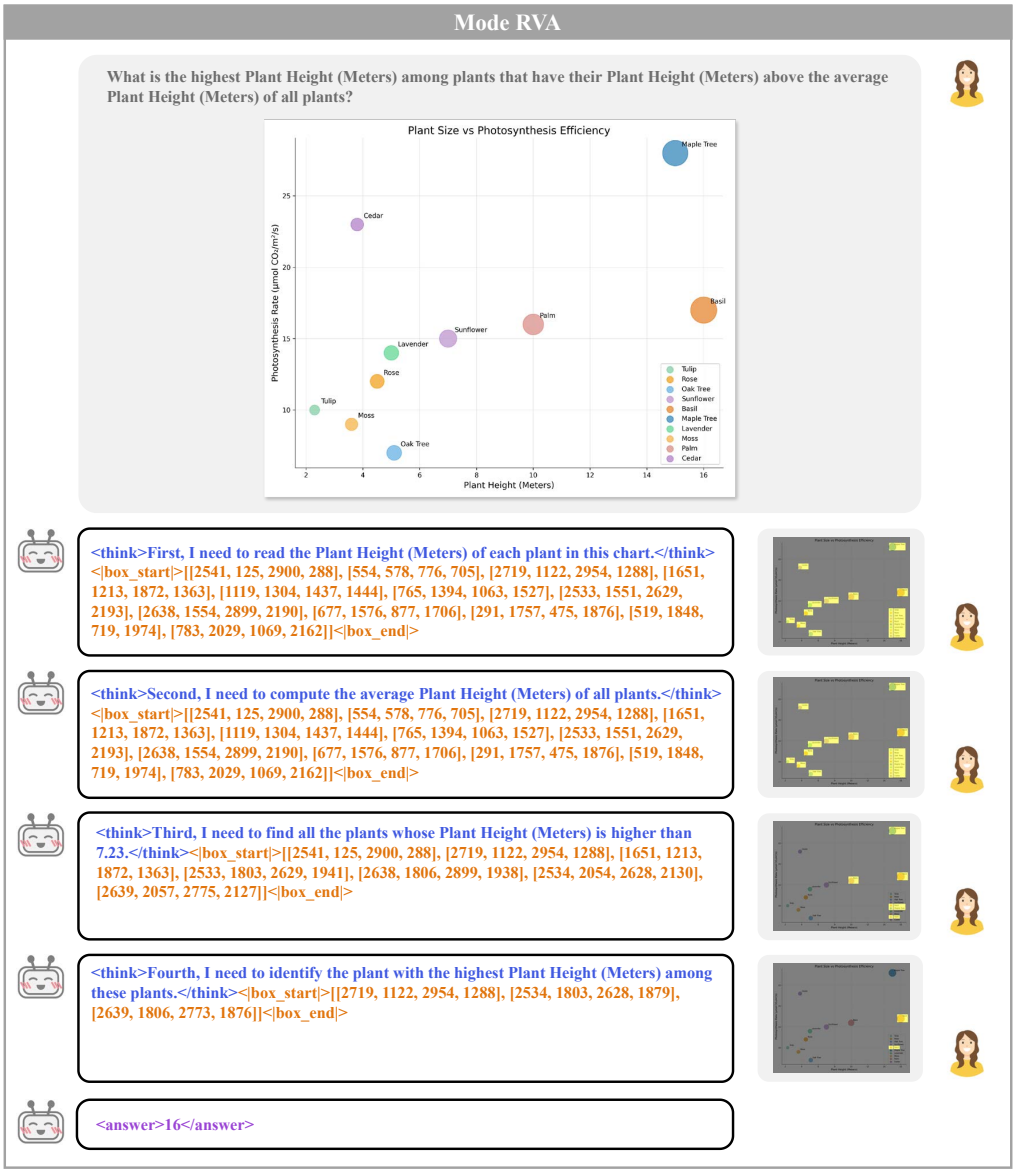


Figure 27: **Example of Generation Mode RVA.** A CQA example resolved through generation mode RVA.

F Evaluation Metrics

To elaborate more details in §5.2, our evaluation incorporates multiple complementary metrics to assess different aspects of model performance.

F.1 Evaluation of Answers

In pursuit of accurate evaluation on multimodal datasets that contain both multi-choice and free-form responses, we compute answer accuracy by comparing model outputs with their corresponding ground-truth answers. Aiming for more comprehensive assessment, we employ two complementary evaluation approaches: MLLM-as-judge for semantic understanding and rule-based evaluation for systematic accuracy measurement. The overall accuracy score for each dataset is calculated as the mean accuracy across all test samples.

MLLM-Based Answer Evaluation. For MLLM-based answer evaluation, we employ GPT-4.1-mini as the *judge*, guided by the prompt shown in Fig. 28. Each model response undergoes MLLM-as-judge evaluation to extract the model answer content, ensuring consistent comparison with ground truth. The *judge* performs a *True*-or-*False* assessment by evaluating whether the model response semantically matches the ground truth, accounting for variations in phrasing and presentation while maintaining semantic equivalence.

Rule-Based Answer Evaluation. To mitigate potential biases introduced by using MLLMs as judges (§5.2), we complement the MLLM-as-judge approach with a systematic rule-based evaluation (Algorithm 1). This rule-based method assesses answer accuracy through predefined parsing and judgment rules, incorporating both strict and relaxed error tolerance through four range criteria: *absolute accuracy* (*acc*@0.0) and three progressively relaxed thresholds (*acc*@0.05, *acc*@0.1, *acc*@0.2).

Algorithm 1 Rule-Based Answer Evaluation with Tolerance Ranges

Require: Ground truth answer gt, predicted answer $pred$, choices C (optional), tolerance ranges $R = \{0.0, 0.05, 0.1, 0.2\}$
Ensure: Accuracy scores $acc@r$ for each $r \in R$

$gt \leftarrow \text{CLEAN}(gt)$, $pred \leftarrow \text{CLEAN}(pred)$
$answer\_type \leftarrow \text{DETECTTYPE}(gt)$
if $gt = \emptyset$ **and** $pred = \emptyset$ **then**
  return $acc@r = 1.0$ for all $r \in R$
end if
if $answer\_type =$ "multi-choice" **then**
  $gt\_list \leftarrow \text{PARSECHOICES}(gt, C)$
  $pred\_list \leftarrow \text{PARSECHOICES}(pred, C)$
  if $gt\_list = pred\_list$ **then**
    return $acc@r = 1.0$ for all $r \in R$
  else
    $match\_rate \leftarrow \frac{|gt\_list \cap pred\_list|}{|gt\_list|}$
    for $r \in R$ **do**
      $acc@r \leftarrow \mathbf{1}[match\_rate \geq r]$
    end for
  end if
else if $answer\_type \in \{\text{int}, \text{float}\}$ **then**
  $gt\_num \leftarrow \text{EXTRACTNUMBER}(gt)$
  $pred\_num \leftarrow \text{EXTRACT}(pred)$
  $acc@0.0 \leftarrow \text{EXTRACTNUMBER}(pred\_num, gt\_num)$
  for $r \in \{0.05, 0.1, 0.2\}$ **do**
    $lower \leftarrow gt\_num \times (1 - r)$
    $upper \leftarrow gt\_num \times (1 + r)$
    $acc@r \leftarrow \mathbf{1}[lower \leq pred\_num \leq upper]$
  end for
else
  $exact\_match \leftarrow \text{GRADEANSWER}(pred.lower(), gt.lower())$
  $substring\_match \leftarrow \mathbf{1}[|gt| > 5 \text{ and } gt.lower() \in pred.lower()]$
  $acc@r \leftarrow max(exact\_match, substring\_match)$ for all $r \in R$
end if
return $acc@r$ for all $r \in R$

F.2 Evaluation of Reasoning

To comprehensively evaluate model reasoning, we implement both micro- and macro-level assessments (§5.2). Our micro-level evaluation relies on five metrics (Eq. 17 - 21), providing the semantic similarity assessment of model reasoning. The final score (*acc@mic*) is the average across all five metrics. At the macro level (*acc@mac*), we leverage GPT-4.1-mini as the *judge*, which rates the quality of model reasoning on a $0 - 10$ scale based on three criteria: (1) *visual understanding and grounding*, (2) *logical coherence and multimodal integration*, and (3) *alignment with ground-truth reasoning*. While micro-level evaluation focuses on fine-grained

similarity between ground-truth and model reasoning, macro evaluation provides a holistic judgment of reasoning quality through MLLM-as-judge (Fig. 29).

LLM As Judge For Answer Evaluation

Given a Question, a Model Response, and its Ground-truth Answer, determine whether the Model's reponse is correct.
The Question can be either multi-choice or free-form, please treat them separately:

* 1. **Multi-Choice Question:**
The Model Response is correct only if it **exactly matches** the correct choice letter (e.g., "A", "B", "C", or "D") after necessary normalization.
Follow these instructions carefully:
(1) If the Model Response is a number (e.g., "2", "3", etc.), map it to the corresponding option letter based on its order in the Question (e.g., 1 → A, 2 → B, etc.).
(2) Ignore irrelevant text, explanations, or format differences. Extract the core predicted answer.
(3) Compare the final normalized response with the Ground-truth Answer letter.

* 2. **Free-Form Question:**
The Model Response is correct if it fully aligns with the meaning and key information of the Ground-truth Answer.
Respond with True if the Model Response is correct and False otherwise.

Question: {question}
Model Response: {model_answer}
Ground-truth Answer: {gt_answer}

Response Format:
* Analysis: First extract the model's answer, then explain the comparison. Please enclose your analysis in: <analysis>your analysis here</analysis>
* Extracted Model Answer: Please ensure the format of the extracted answer is consistent with the format of the ground-truth answer, but the content of the extracted answer should be exactly what the model responsed. Please enclose the extracted model answer in: <model>extracted model answer</model>
* Judge: Return "True" only for exact matches, otherwise "False". Please enclose your True-or-False judge in <judge>True or False</judge>

<analysis>...</analysis>
<model>...</model>
<judge>...</judge>

Figure 28: **LLM-As-Judge For Answer Evaluation.** We employ GPT-4.1-mini as the judge to assess model answer accuracy using the prompt shown in this figure.

LLM As Judge For Reasoning Evaluation

You are an expert evaluator tasked with assessing the quality of reasoning chains in Chart Question Answering (CQA) tasks.
Given an Image, a Question, a Model's Reasoning, and the Ground-truth Reasoning, evaluate ONLY the Model's reasoning process, NOT the final answer.

Evaluation Criteria:
1. **Visual Understanding and Grounding (0-10)**: Does the reasoning demonstrate accurate perception and interpretation of visual elements in the image? Are visual observations correctly identified, described, and used as evidence? Check for:
- Accurate identification of chart elements, components, labels, values, legends, and spatial relationships
- Correct interpretation of visual context and relevant details
- Appropriate grounding of reasoning in observable visual evidence
- Absence of hallucinated or misperceived visual elements

2. **Logical Coherence and Multimodal Integration (0-10)**: Does the reasoning logically connect visual observations with the question requirements? Are the steps coherent and does each conclusion follow from visual evidence and prior reasoning? Evaluate:
- Logical flow from visual observations to conclusions
- Appropriate integration of visual and textual information
- Sound inferential steps that bridge perception and reasoning
- Absence of logical fallacies or non-sequiturs

3. **Alignment with Ground-Truth Reasoning (0-10)**: How well does the model's reasoning align with the provided ground-truth reasoning in terms of approach, key insights, and logical structure? Evaluate:
- Similarity in problem-solving approach and methodology
- Coverage of the same key reasoning steps and insights as the ground-truth
- Alignment in the logical progression and chain of thought
- Whether the model identifies and addresses the same critical visual elements as the ground-truth
- Overall coherence with the expected reasoning pattern, even if expressed differently

Input:

* Image: [The image will be provided to you for reference]

* Question: [question]

* Model Reasoning:
[model_reasoning]

* Ground-truth Reasoning:
[gt_reasoning]

Instructions:
- Carefully examine the provided image and understand its contents
- Compare the Model Reasoning against both the image content and Ground-truth Reasoning
- Focus exclusively on the reasoning process, NOT the final answer
- Be objective and consistent in your scoring across different types of visual content
- Use the full 0-10 scale for each criterion, and each score should be an integer, NOT float
- Implement your evaluation by strictly following the three criteria above

Response Format:

<analysis>
Provide a detailed analysis of the Model Reasoning across all three criteria:
1. Visual Understanding and Grounding: [Assess accuracy of visual perception, object identification, and use of visual evidence]
2. Logical Coherence and Multimodal Integration: [Evaluate logical flow and integration of visual and textual information]
3. Alignment with Ground-Truth Reasoning: [Compare the model's approach and insights with ground-truth reasoning, assessing similarity in methodology and coverage of key points]

Compare with ground-truth reasoning and justify your scores based on visual accuracy, reasoning quality, and alignment with expected reasoning patterns.
</analysis>

<judge>score1, score2, score3</judge>

Where score1 = Visual Understanding and Grounding, score2 = Logical Coherence and Multimodal Integration, score3 = Alignment with Ground-Truth Reasoning

Response Format:
* Analysis: Explain xxx. Please enclose your analysis in: <analysis>your analysis here</analysis>
* Judge: Return three scores respectively for three criteria, separated with comma. Please enclose your judge scores in <judge>score1, score2, score3</judge>

<analysis>...</analysis>
<judge>...</judge>

Figure 29: **LLM-as-Judge For Reasoning Evaluation.** We employ GPT-4.1-mini as the judge to evaluate model reasoning using the prompt shown in this figure. Prompt is restricted to smaller sizes to save space.

Micro-Evaluation: Reasoning Similarity. We employ five metrics to measure the semantic similarity between ground-truth and model reasoning, including ROUGE-L (Eq. 17), BLEU (Eq. 18), METEOR (Eq. 20), BERTSCORE (Eq. 19), and COSINE SIMILARITY (Eq. 21).

$$\text{ROUGE} = \text{ROUGE-L} = \frac{2 \cdot P_{lcs} \cdot R_{lcs}}{P_{lcs} + R_{lcs}} \tag{17}$$

$$\text{BLEU} = \text{BLEU-4} = \text{BP} \cdot \exp\left(\sum_{n=1}^{4} w_n \log p_n\right) \tag{18}$$

where P_{lcs} and R_{lcs} are precision and recall of longest common subsequences.

$$\text{BERTSCORE} = \frac{1}{|\mathcal{S}_p|} \sum_{s_i \in \mathcal{S}_p} \max_{s_j \in \mathcal{S}_g} \frac{\mathbf{v}_i^T \mathbf{v}_j}{|\mathbf{v}_i||\mathbf{v}_j|} \tag{19}$$

where $\mathcal{S}_p$ and $\mathcal{S}_g$ are the sets of predicted and ground-truth reasoning tokens respectively, and $\mathbf{v}_i, \mathbf{v}_j$ are their corresponding BERT contextual embeddings.

$$\text{METEOR} = \frac{(1+\eta_1) \cdot P_r \cdot R_r}{\eta_1 \cdot P_r + R_r} \cdot \left(1 - \eta_2 \cdot \left(\frac{c}{u_m}\right)^{\eta_3}\right) \tag{20}$$

$$\text{COSINE} = \frac{\mathbf{e}_p \cdot \mathbf{e}_g}{|\mathbf{e}_p||\mathbf{e}_g|} \tag{21}$$

where P_r and R_r are precision and recall of reasoning tokens, u_m is the number of matched unigrams, c is the number of chunks, and we define $\eta_1 = 0.9$, $\eta_2 = 0.5$, and $\eta_3 = 3$ as hyperparameters controlling the weight of recall, penalty magnitude, and penalty sharpness, respectively; and $\mathbf{e}_p$ and $\mathbf{e}_g$ are the embedding vectors of predicted and ground-truth reasoning steps respectively.

Macro-Evaluation: Reasoning Quality. The quality of model reasoning is evaluated through MLLM-as-judge assessment. Specifically, we employ GPT-4.1-mini as the *judge*, guided by the prompt shown in Fig. 29, to assign a quality score on a $0-10$ scale based on three criteria, including *visual understanding and grounding*, *logical coherence and multimodal integration*, and *alignment with ground-truth reasoning*. The final quality score for each dataset is calculated as the mean score across all test samples.

- **Criterion 1: Visual Understanding and Grounding.** Reasoning accuracy in identifying, interpreting, and grounding reasoning in visual elements, meanwhile without introducing hallucinated details
- **Criterion 2: Logical Coherence and Multimodal Integration.** Logical progression throughout the entire reasoning chain, with appropriate integration of multimodal information.
- **Criterion 3: Alignment with Ground-Truth Reasoning.** Consistency with ground-truth reasoning, especially in terms of problem-solving approach, key insights, and logical structure, even if expressed differently.

F.3 Evaluation of Visual Grounding

As introduced in (§5.2), we employ two IoU variants, CIOU (Eq. 22) and GIOU (Eq. 23), as the primary evaluation metrics for visual grounding assessment.

$$\text{CIoU} = \text{IoU} - \frac{\rho^2(c_p, c_g)}{d^2} \tag{22}$$

$$\text{GIoU} = \text{IoU} - \frac{|A_c - A_u|}{A_c} \tag{23}$$

where $\rho^2(c_p, c_g)$ is the squared distance between predicted and ground-truth centroids, d is the diagonal of the enclosing box, A_c is the enclosing area, and A_u is the union area.

F.4 Evaluation Mode

Furnishing models with the capabilities to reason through dynamic visual grounding, we employ different generation modes (§5.1) to support comparable evaluations:

- **Mode A**: *Answer-Only* mode where MLLMs are prompted to directly generate the final answer.
- **Mode RA:** *Reason-Answer* mode where MLLMs first go through the intermediate reasoning process, followed by the final answer.
- **Mode VA:** *Vision-Answer* mode where MLLMs first generate their visual grounding coordinates, followed by the final answer.
- **Mode RVA:** *Reason-Vision-Answer* mode where MLLMs first go through the reasoning process with dynamic visual grounding, and then generate the final answer.

G In-Depth Analysis

G.1 Implementation Details.

We train each model for 3 epochs with an initial learning rate $lr = 1e-4$ using `cosine` scheduler. The ratio of training and validation is set to `train:val=9:1`. Employing two NVIDIA 80G H100 GPUs, our model training is powered by LoRA for memory efficiency. For hyperparameter settings, *Stage I* supervises visual grounding only ($\lambda_V = 1.0$, $\lambda_R = 0.0$, $\lambda_A = 0.0$), and *Stage II* supervises reasoning, grounding, and answer jointly (Eq. 6). Implementing different generation modes, *reasoning* (**R**) is enclosed within <`think`></`think`>, *visual grounding* (**V**) is enclosed within <|`box_start`|><|`box_end`|> for explicit visual grounding while represented as <`GROUND`> in implicit visual ground (§G.5), and the *final answer* (**A**) is enclosed within <`answer`></`answer`>.

G.2 Grounding Method & Computation Cost

Employing zoom-in visual grounding, the *cropped* grounding method requires substantially larger memory at the same resolution. To mitigate this cost, we reduce the training resolution of *cropped* grounding to 128×128, thereby maintaining a comparable computational overhead. Despite the resolution degradation, reasoning with **cropped** visual grounding achieves notably higher accuracy than the baseline (up to 4.72% improvement on CCQA) and performs competitively with the other two grounding methods (Tab. 2). These results highlight the effectiveness of zoom-in visual enhancement, albeit at the expense of increased computational cost when aiming for higher performance.

Method	Resolution	D_{max}	T_{max}	SFT	RL	Inference
Applied	448×448	4	5	$2 \times 80GB$	$\geq 8 \times 80GB$	$1 \times 80GB$
Boxed	448×448	4	5	$2 \times 80GB$	$\geq 8 \times 80GB$	$1 \times 80GB$
Cropped	128×128	4	5	$2 \times 80GB$	$\geq 8 \times 80GB$	$1 \times 80GB$

Table 8: **The Computation & Configuration Of Different Grounding Method.** This table summarizes the visual computation requirements and parameter configuration.

G.3 The Role of Visual Grounding: From Extrinsic Assistance To Intrinsic Abilities

A critical finding from our experiments reveals the fundamental distinction between the utility of multi-step visual reasoning during training versus inference (Tab. 11). While incorporating visual grounding in the training process significantly enhances models' intrinsic visual reasoning capabilities, directly applying the same multi-step approach during inference can paradoxically degrade performance due to error accumulation (Fig. H.2 & §H.1).

Training Benefits of Visual Grounding. Our curriculum learning approach with visual grounding supervision effectively teaches models to develop stronger intrinsic representations for chart understanding. By learning to align reasoning steps with visual focuses during training, models internalize the ability to focus on relevant image components, leading to improved performance even when generating direct answers without explicit visual grounding steps.

Inference Challenges with Multi-Turn Visual Grounding. On the other hand, when models are required to explicitly generate visual grounding coordinates during inference (i.e., *Mode* **RVA**), performance degrades in comparison with direct answer generation (*Mode* **A**) and reasoning without grounding (*Mode* **RA**). This degradation stems from two primary factors: (1) *Cumulative grounding errors*: Inaccurate bounding box predictions in early reasoning steps propagate and compound errors in subsequent steps; (2) *Reasoning-grounding misalignment*: Discrepancies between intended visual focus and actual predicted coordinates lead to reasoning based on incorrect visual regions.

Power of Intrinsic Visual Reasoning Capabilities. Results in Tab. 11 demonstrate that visual grounding serves as an effective *training signal* rather than an *inference mechanism*. Our curriculum learning with visual supervision enables models to learn better intrinsic visual-textual alignments, which manifest as improved performance in direct answer generation Tab. 2. However, explicitly requiring visual grounding during inference introduces additional complexity and error sources that outweigh the potential benefits. Nevertheless, compared with baselines, our finetuned models manage to achieve remarkably higher performance in not only *Mode* **A**, but also *Modes* **RA**, **VA**, and **RVA**.

G.4 Visual Grounded Reasoning via Reinforcement Learning

To further validate the generality of CURV, we extend our two-stage training framework to reinforcement learning (RL), reformulating training objectives as specialized reward signals that directly incentivize accurate visual grounding, faithful reasoning, and correct answer generation.

Stage I: Visual Grounding. In *Stage I*, we define visual grounding reward $\mathcal{R}_V$ that evaluates the geometric quality of the predicted visual focus V_t against the ground-truth focus region V_t^* using CIoU (Eq. 22) and GIoU (Eq. 23):

$$\mathcal{R}_V(V_t, V_t^*) = \lambda_{\text{CIoU}}\, \mathcal{R}_{\text{CIoU}}(V_t, V_t^*) + \lambda_{\text{GIoU}}\, \mathcal{R}_{\text{GIoU}}(V_t, V_t^*), \tag{24}$$

where $\lambda_{\text{CIoU}} = \lambda_{\text{GIoU}} = 0.5$. As such, *Stage I* training objective becomes:

$$\mathcal{R}^{(S1)} = \sum_{t=1}^{T} \mathcal{R}_V(V_t, V_t^*). \tag{25}$$

Stage II: Interleaved Visual Reasoning Reward. In *Stage II*, the training objective extends to the full interleaved visual grounded reasoning chain with a composite reward:

$$\mathcal{R}^{(S2)} = \lambda_R \sum_{t=1}^{T} \mathcal{R}_R(R_t, R_t^*) + \lambda_V \sum_{t=1}^{T} \mathcal{R}_V(V_t, V_t^*) + \lambda_A\, \mathcal{R}_A(A, A^*) + \lambda_F\, \mathcal{R}_F, \tag{26}$$

where λ_R, λ_V, λ_A, and λ_F are weighting coefficients balancing the four reward components. Each component is defined as follows:

Reasoning reward $\mathcal{R}_R(R_t, R_t^*)$ measures the semantic similarity between the predicted reasoning step R_t and its ground-truth counterpart R_t^* as a rubric-based aggregate of ROUGE-L (Eq. 17), BERTSCORE (Eq. 19), and COSINE SIMILARITY (Eq. 21):

$$\text{SIMILARITY} = \lambda_{rouge} \cdot \text{ROUGE} + \lambda_{bert} \cdot \text{BERT} + \lambda_{cosine} \cdot \text{COSINE} \tag{27}$$

Model	Level 1					Level 2					Level 3				
	@M	@0.0	@0.05	@0.1	@0.2	*@M*	@0.0	@0.05	@0.1	@0.2	*@M*	@0.0	@0.05	@0.1	@0.2
Baseline															
Baseline	45.25	43.52	51.54	56.25	61.54	22.75	22.86	31.14	37.21	45.36	16.18	16.00	21.89	25.82	31.71
Ours (RL)															
Stage II	53.64	51.71	57.43	61.36	65.00	24.61	25.39	35.21	41.04	51.57	16.21	18.86	25.25	29.32	33.96
Stage I+II	55.14	53.23	61.37	67.38	70.54	26.25	28.46	40.35	48.05	57.94	19.07	22.53	27.43	31.43	35.58
Ours (SFT)															
Stage II	54.21	51.21	59.50	65.00	68.50	25.86	27.14	39.18	47.25	55.18	16.86	17.32	23.93	28.61	33.11
Stage I+II	60.71	57.21	63.71	69.57	73.79	30.11	29.61	41.82	48.87	57.46	17.98	18.21	24.25	29.07	34.50
Ours (SFT + RL)															
Stage II	55.86	53.64	60.79	66.43	69.29	27.96	28.79	40.75	48.64	57.25	20.64	21.39	26.79	30.25	34.93
Stage I+II	62.29	59.79	65.86	72.21	73.79	29.36	29.74	42.57	50.39	59.39	22.43	23.39	29.25	32.43	36.89

Table 9: **CURV with Reinforcement Learning on CCQA.** Employing Qwen2.5-VL-3B as the base model, we evaluate the performance of its baseline and finetuned versions on CCQA across three curriculum levels via five accuracy evaluation metrics ***acc*@*X*** where *X* is MLLM (abbreviated as @*M*) or ranges (§ 5.2). All settings use CURV@Applied throughout training and inference.

where λ_{rouge}, λ_{bert}, and λ_{cosine} are weighting coefficients respectively set to 0.3, 0.4, 0.3 during *Stage II* training.

Grounding reward $\mathcal{R}_V(V_t, V_t^*)$ evaluates geometric alignment of the predicted visual focus at each reasoning step, as defined in Eq. 24.

Answer reward $\mathcal{R}_A(A, A^*)$ is a binary exact-match signal that returns 1 if the predicted final answer A matches the ground-truth A^* and 0 otherwise.

Format reward $\mathcal{R}_F$ is a binary compliance signal that returns 1 if the model output adheres to the required structured format, ensuring output parsability throughout training.

G.5 Explicit vs. Implicit Visual Grounded Reasoning

We introduce two designs for integrating visual grounding into the multi-step reasoning framework: *explicit* and *implicit* visual grounded reasoning. Both approaches are based on our two-stage training framework CURV (Fig. 3), but differ fundamentally in *how* the MLLM is encouraged to attend to chart regions across reasoning steps.

G.5.1 Explicit Visual Grounding

In *explicit* visual grounded reasoning, the MLLM is trained end-to-end to produce multi-step reasoning with dynamically changed visual focuses. Concretely, in *Stage I*, given the chart image $\mathcal{I}$ and question Q, the model learns to dynamically ground logical reasoning in corresponding visually focused regions by predicting bounding boxes V_t for each reasoning step R_t across t steps of reasoning (Eqs. 3-4). Each predicted bounding box is represented as a normalized coordinate tuple $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ over the image canvas, and the grounding loss $\mathcal{L}_V$ directly supervises the model's spatial predictions V_t against ground-truth focus regions V_t^*. In *Stage II*, the model transitions to interleaved visual grounded reasoning (RVA). At each step t, the bounding box V_t is applied to the original chart image $\mathcal{I}$ through one of our visual grounding strategies (§D) to construct the grounded visual state $\mathcal{I}_t$, which is then provided as an additional visual input at the next reasoning step to implement the $V_{t'} \rightarrow \mathcal{I}_{t'}$ mapping (Eq. 5). Note that during *training*, we use ground-truth V_t^* as intermediate visual augmentation; while in *inference*, the model continues its reasoning based on its last-step V_t.

Under this design, the MLLM jointly learns to directly articulate its visual focus as explicit coordinate predictions and dynamically augment its reasoning by linking each R_t with corresponding visually focused regions. The training objective in Eq. 6 jointly supervises reasoning, grounding, and final answer quality, with λ_R, λ_V, and λ_A balancing the three components. The key characteristic of explicit visual grounded reasoning is that the MLLM

internalizes both roles on its own, inherently encouraged to dynamically shift attention alongside reasoning.

G.5.2 Implicit Visual Grounding

In implicit visual grounded reasoning design, visual grounding is performed by a dedicated *grounder module* that operates alongside the MLLM, implicitly affecting MLLM reasoning.

Grounder Module. The grounder is a lightweight grounding head G_ϕ that operates on the model's internal representations to produce patch-level attention maps over the chart image. Specifically, it takes as input (1) the visual patch embeddings $\mathbf{E}^{\text{v}} \in \mathbb{R}^{N_v \times d}$, extracted from the frozen vision encoder with N_v patches and hidden dimension d, and (2) the decoder hidden state $\mathbf{h}_t \in \mathbb{R}^d$ at a designated <GROUND> token position appended after each reasoning step R_t. As such, the grounder computes a patch-level attention map:

$$\mathbf{a}_t = g_\phi(\mathbf{E}^{\text{v}}, \mathbf{h}_t) \in \Delta^{N_v - 1} \tag{28}$$

where Δ^{N_v-1} denotes the (N_v-1)-simplex (i.e., the output is a distribution over visual patches). Internally, g_ϕ applies self-attention over $\mathbf{E}^{\text{v}}$ to capture contextual relationships among patches, followed by separate projection networks for the visual and decoder representations, and a scaled dot-product operation to compute attention logits. The training objective for g_ϕ combines KL divergence (Eq. 30), binary cross-entropy (Eq. 31), Dice loss (Eq. 32), and cross-entropy (Eq. 33), all computed against ground-truth binary focus masks $M_t^* \in \{0,1\}^{N_v}$:

$$\mathcal{L}_V(V_t, V_t^*) = \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{BCE}}\mathcal{L}_{\text{BCE}} + \lambda_{\text{Dice}}\mathcal{L}_{\text{Dice}} + \lambda_{\text{CE}}\mathcal{L}_{\text{CE}} \tag{29}$$

where each term supervises a distinct aspect of the predicted attention map $\mathbf{a}_t$ against the ground-truth binary focus mask M_t^*:

$$\mathcal{L}_{\text{KL}} = \sum_{i=1}^{N_v} \tilde{M}_{t,i}^* \log \frac{\tilde{M}_{t,i}^*}{\mathbf{a}_{t,i} + \epsilon} \tag{30}$$

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N_v} \sum_{i=1}^{N_v} \left[M_{t,i}^* \log(\mathbf{a}_{t,i} + \epsilon) + (1 - M_{t,i}^*) \log(1 - \mathbf{a}_{t,i} + \epsilon) \right] \tag{31}$$

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \sum_{i=1}^{N_v} \mathbf{a}_{t,i} M_{t,i}^*}{\sum_{i=1}^{N_v} \mathbf{a}_{t,i} + \sum_{i=1}^{N_v} M_{t,i}^* + \epsilon} \tag{32}$$

$$\mathcal{L}_{\text{CE}} = -\log \mathbf{a}_{t, \arg\max_i M_{t,i}^*} \tag{33}$$

where $\tilde{M}_t^* = M_t^* / (\sum_i M_{t,i}^* + \epsilon)$ is the row-normalized ground-truth mask treated as a target probability distribution for the KL term, ϵ is a small constant for numerical stability, and $\mathbf{a}_{t,i}$ denotes the predicted attention weight for the i-th visual patch. $\mathcal{L}_{\text{KL}}$ penalizes distributional divergence between the predicted attention map and the normalized ground-truth mask; $\mathcal{L}_{\text{BCE}}$ enforces patch-level binary classification independently across all patches; $\mathcal{L}_{\text{Dice}}$ optimizes region overlap between the predicted and ground-truth focus regions; and $\mathcal{L}_{\text{CE}}$ treats the patch with the highest ground-truth activation as the target class, encouraging the model to concentrate probability mass on the most salient region. The loss weights λ_{KL}, λ_{BCE}, λ_{Dice}, and λ_{CE} are set to 0.25 uniformly in our experiments.

Stage I: Grounder Training. In *Stage I*, the MLLM parameters θ are kept frozen while the grounder G_ϕ is trained to predict accurate visual focuses for each reasoning step. The decoder hidden state $\mathbf{h}_t$ at <GROUND> position is extracted and passed to G_ϕ together with $\mathbf{E}^{\text{v}}$. The training objective follows Eq. 4, with $\mathcal{L}_V$ computed as in Eq. 29.

Stage II: MLLM Training with Visual Grounded Feedback. In *Stage II*, the grounder G_ϕ is frozen and the MLLM parameters θ are finetuned on the full interleaved reasoning objective. The model is trained to produce reasoning chains of the form:

$$\underbrace{R_1}_{\text{step 1}} < \texttt{GROUND} > \underbrace{R_2}_{\text{step 2}} < \texttt{GROUND} > \cdots \underbrace{R_t}_{\text{step } t} < \texttt{GROUND} > \underbrace{A}_{\text{answer}} < \texttt{GROUND} > \tag{34}$$

where <GROUND> tokens serve as anchors at which the grounder produces attention maps $\{\mathbf{a}_t\}_{t=1}^T$, which are used to construct grounded visual states $\{\mathcal{I}_t\}_{t=1}^T$ through our visual grounding strategies (§D). However, as the MLLM generates reasoning steps R_t without explicit coordinate predictions, the visual grounding signal remains implicit. As such, the model is encouraged to internally shift its internal attention focus at each reasoning step, revealing less effective visual augmentation than *explicit* visual grounded reasoning (§5.4). This performance gap also highlights differences between visual grounded reasoning tasks that leverage learn-to-focus dynamics to enhance reasoning and visual grounding tasks aimed at accurately localizing visual focus.

G.6 Curriculum Learning: Effectiveness of CL in Chart Understanding

To validate the effectiveness of curriculum learning, we further compare CURV with the baseline model (*Qwen2.5-VL-7B*) in both untrained and non-curriculum training settings. As shown in Fig. 30, although training without curriculum learning improves performance on CCQA, it consistently underperforms as compared to CURV trained with either *Stage I* or *Stage I+II*. This gap becomes more pronounced as task complexity increases (i.e., levels $1 \rightarrow 3$), with diminishing performance gains observed in the non-curriculum setting. These results highlight the effectiveness of curriculum learning in guiding the model to learn from fundamental chart elements and gradually develop intrinsic visual reasoning capabilities. This structured learning process not only improves the model's adaptability to more complex chart understanding tasks but also enhances its generalization to out-of-domain multimodal reasoning.

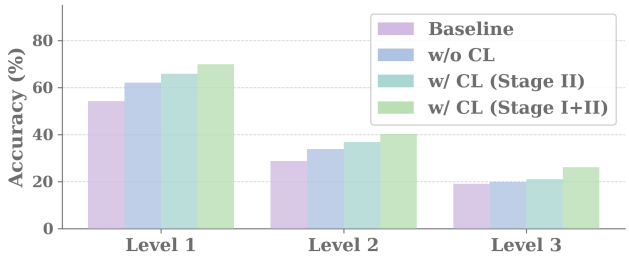


Figure 30: **Effects of Curriculum Learning.** Using Qwen2.5-VL-7B as the base model, we compare its performance across different training settings.

G.7 Two-Stage Learning: Strength of Training Curriculum

In addition to data curriculum (§G.6), CURV also implements the two-stage curriculum training paradigm from internalizing to enhancing MLLM visual reasoning capabilities (§3). In *Stage I*, MLLM learns vision-reasoning alignment through decoupled visual grounding, establishing explicit associations between visual evidence and reasoning logics. In *Stage II*, MLLM is further trained to perform interleaved visual reasoning, leveraging the grounding capabilities acquired in *Stage I* to support

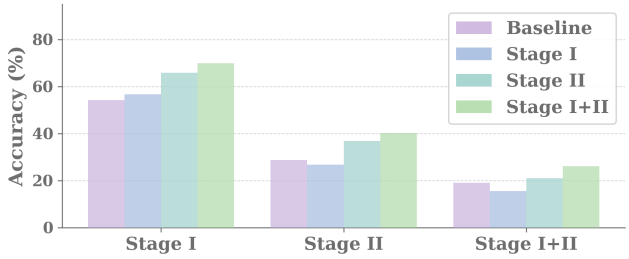


Figure 31: **Strength of Curriculum Training.** Using Qwen2.5-VL-7B as the base model, we compare its performance across different training settings.

Method	Level 1		Level 2		Level 3	
	acc@M	*acc@0.0*	*acc@M*	*acc@0.0*	*acc@M*	*acc@0.0*
Base Models						
Qwen2.5-VL-7B	54.21	50.79	28.68	28.82	19.01	19.38
Qwen3-VL-4B	48.07	47.28	26.86	26.61	20.68	20.04
Chart Specialist						
ChartGemma (Masry et al., 2024)	28.71	21.36	12.46	8.25	7.71	8.86
Tool-Use Agents						
Thyme (Zhang et al., 2025c)	56.57	55.57	24.39	24.96	18.07	18.54
DeepEyes (Zheng et al., 2025)	58.43	57.21	30.04	29.29	20.75	20.29
Ours						
CURV (Qwen3-VL-4B)	**64.21**	**63.79**	**34.86**	**34.57**	**23.82**	**23.25**
CURV (Qwen2.5-VL-7B)	**69.86**	**66.79**	**40.21**	**38.64**	**26.11**	**24.25**

Table 10: **Comparison Against Chart Specialist & Tool-Use Agents on CCQA.** We evaluate CURV against a newer base model (Qwen3-VL-4B), a chart-specialist baseline (ChartGemma), and tool-augmented agents (Thyme, DeepEyes) across three curriculum levels.

step-by-step reasoning over visual inputs. To validate the effectiveness of this design, we compare the performance among Qwen2.5-VL-7B, CURV@*Stage I*, CURV@*Stage II*, and CURV@*Stage I+II* (Fig. 31). Results show that CURV@*Stage I+II* consistently outperforms both single-stage variants, indicating the advantage of combining decoupled alignment with interleaved reasoning via two-stage curriculum training.

In contrast, CURV@*Stage I* alone improves performance on *level* 1 simple CQA while degrades on harder single-chart and multi-chart understanding tasks. This suggests that decoupled grounding, while effective for learning basic visual-reasoning associations, is insufficient for supporting compositional reasoning that requires integrating multiple pieces of visual evidence. Conversely, CURV@*Stage II* alone fails to fully realize these benefits, implying that interleaved reasoning without well-established grounding leads to weaker visual-reasoning alignment.

Collectively, these findings highlight that the strength of CURV lies in its staged curriculum training design: *Stage I* builds foundational visual grounding ability, while *Stage II* enables the model to operationalize this grounding ability into coherent, step-by-step visual reasoning. Their combination results in more robust and generalizable visual reasoning capacity across varying task complexities.

G.8 Foundational Learning: Trade-off between Robustness & Adaptability

Fig. 9 reveals a critical finding that strongly validates our curriculum learning design (§4). While training on *level* 1 alone provides solid foundational performance (↑35.08% on *level* 1), progressive training on *levels* 1+2 demonstrates the optimal learning accumulation, achieving the best overall performance across all difficulty *levels* (↑ 15.65% on *level* 1, ↑ 11.53% on *level* 2, ↑ 7.10% on *level* 3). However, extending training to include *level* 3 unfolds a concerning trade-off between robustness and adaptability: While notably improves complex reasoning performance (↑ 12.82% on *level* 3), it significantly degrades foundational reasoning abilities, only higher than baseline by ↑ 5.93% and ↑ 0.03% on *levels* 1 and 2, respectively.

G.9 Broader Baselines: Comparison Against Chart Specialist & Tool-Use Agents

We broaden our comparison along three axes: a newer base model `Qwen3-VL-4B` (Yang et al., 2025), a chart-specialist baseline `ChartGemma` (Masry et al., 2024), and tool-augmented agents that invoke external visual tools `Thyme` (Zhang et al., 2025c) and `DeepEyes` (Zheng et al., 2025). All methods are evaluated on our CCQA test set across the three curriculum levels

using both *acc@M* and the absolute accuracy *acc@0.0* (§5). As shown in Tab. 10, CURV is generalizable to newer backbones, improving over its base model (Qwen3-VL-4B) by up to ↑16.14%. This validates that the improvements from CURV are not tied to a specific base model, but instead reflect a transferable strengthening of intrinsic visual-grounded reasoning capabilities. Moreover, CURV also outperforms chart specialist and tool-augmented agents built on the same Qwen2.5-VL-7B backbone, providing further evidence to the effectiveness and robustness of CURV in the face of varying chart complexity levels.

H Multi-Step Reasoning With Dynamic Visual Grounding

Model	Size	CCQA											
		Level 1				Level 2				Level 3			
		Reasoning		Grounding	Answer	Reasoning		Grounding	Answer	Reasoning		Grounding	Answer
		acc@mac	*acc@mic*	mIoU	*acc@mac*	*acc@mac*	*acc@mic*	mIoU	*acc@mac*	*acc@mac*	*acc@mic*	mIoU	*acc@mac*
Baselines													
GPT-4o	-	53.17	51.02	22.03	50.00	53.32	53.60	13.01	24.00	50.20	46.86	11.14	21.50
Qwen2.5-VL	3B	44.90	39.07	37.12	36.00	38.72	40.08	29.73	14.00	32.17	35.47	27.43	12.00
	7B	48.53	41.70	48.17	45.00	40.49	40.68	32.25	21.50	35.82	38.10	32.26	17.50
Ours													
Applied	3B	50.50	53.63	47.17	50.00	45.79	45.66	40.59	19.00	41.95	38.76	34.60	15.00
(Qwen2.5-VL)	7B	**56.13**	**54.01**	**57.55**	**52.00**	**49.87**	**47.17**	45.72	23.00	**46.05**	**38.83**	**39.90**	19.50
Boxed	3B	50.23	49.90	45.55	46.00	40.70	42.74	38.11	14.50	36.98	36.97	33.37	13.50
(Qwen2.5-VL)	7B	51.27	48.20	50.88	49.00	42.97	43.44	43.09	22.50	41.33	37.45	34.52	18.00
Cropped	3B	42.17	46.68	51.22	39.00	40.52	42.69	45.82	14.50	39.63	35.86	34.27	13.00
(Qwen2.5-VL)	7B	50.33	50.79	51.33	47.00	44.97	43.75	**50.10**	**24.00**	42.43	38.51	39.19	19.00

Table 11: **Performance Evaluation for RVA Mode Inference.** Employing the set of evaluation metrics (§ 5.2), we assess model reasonnig, visual grounding, and final answer, respectively.

H.1 Challenges In Multi-Step Visual Grounding

We leverage CCQA, randomly selecting 500 samples to evaluate model inference through **RVA** mode (§5.1). Training MLLMs with explicit reasoning and visual grounding as intermediate outputs effectively enhances model's intrinsic visual reasoning capabilities (§5.3). This step-by-step visual reasoning guides the model to decompose complex tasks into structured reasoning chains through dynamic attention grounding. With intermediate grounding naturally supporting more coherent reasoning trajectories, this in turn enhances model's ability to establish interleaved thinking-perception correspondences. Aligning with human visual reasoning, *decomposed* reasoning chains effectively help models to develop and strengthen their intrinsic visual reasoning capabilities.

Different from learning, during inference, human visual reasoning is rather a *composed* process that interleaves logical reasoning with visual comprehension, while *compositing* all intermediate steps into a coherent chain of thought. In contrast, inference in **RVA** exposes the fragility of step-wise generation: once an intermediate step is flawed, whether by incorrect calculation or inaccurate visual comprehension, the error propagates through the chain, breaking the balance between perception and reasoning that eventually leads to incorrect final answers (§H.1). Therefore, *it can be an effective way of learning, while may not be as useful in inference*.

Tab. 11 summarizes the evaluation results of **RVA** inference. Fintuned models achieve noticeable improvements across reasoning (up to 10.23% absolute gain), grounding (up to 9.38% absolute gain), and answering (up to 14% absolute gain). Beyond these numerical results, qualitative inspection (§H.3) reveals distinct behavioral patterns where training with step-by-step visual grounding encourages systematic reasoning chains with sharper object localization, showcasing stronger alignment with human-like reasoning trajectories.

Despite these improvements, however, the answering performance remains lower than that of CURV when using the same base model and grounding method (Tab. 2). This indicates that, while **RVA** training can effectively enhance intrinsic visual reasoning capabilities, **RVA** inference magnifies the vulnerability to intermediate error accumulation.

Build upon our discussions above, we present the challenge for MLLMs in CQA through mode **RVA**. Although **RVA** training notably enhances model's intrinsic visual reasoning capabilities, **RVA** inference reveals critical limitations. During **RVA** inference, cumulative

intermediate errors can cascade through the reasoning chain, ultimately compromising the model's ability to produce correct answers. For example, Fig. 32 shows a CQA example that is responded to by GPT-4o and Qwen2.5-VL-7B, respectively. Both GPT-4o and Qwen2.5-VL-7B fail to generate the correct answer due to their disrupted reasoning chains that evolve through their incorrect visual grounding.



Figure 32: **Challenge of Inference in Mode RVA.** Tested on GPT-4o and Qwen2.5-VL, this example illustrates the challenge MLLMs face in performing inference in **RVA**.

H.2 Inference Failure

Fig. 35 illustrates examples of model inference failures in **RVA**. In both cases, the model fails to properly ground its reasoning in the chart, leading to inaccurate extraction and misinterpretation of visual information. Arising in early reasoning steps, these visual comprehension inaccuracies can propagate through the reasoning chain, ultimately resulting in incorrect question answering.

H.3 Inference Success

Example 1 - Mode A: Figures 33 & 34 exhibit examples on chart question answering in mode **A**, where the baseline Qwen2.5-VL-7B fails to generate the correct answer, while CURV (Qwen2.5-VL-7B) finetuned through *applied* grounding succeeds. Fig. 33 (a) is a simple value reading problem ($D = 1$), where the baseline model fails to localize the exact queried chart component. Fig. 33 (b) consists of two nested functions ($D = 2$), where the baseline model fails to localize the queried bar in the given subset of countries. Fig. 33 (c) further enhance the CQA complexity ($D = 3$), involving three nested functions across reasoning, visual grounding, and interleaved calculation that the baseline model fails to correctly response. Different from Fig. 33 that query about a single chart, each CQA sample in Fig. 34 involves multiple charts that significantly complicates question answering. The baseline model fails in Fig. 33 (a) ($D = 3$) as it requires the localization of the exact chart subplot, the required subset, as well as the Y-axis value reading. Fig. 33 (b) increases the

CQA difficulty ($D = 4$) by including not only accurate localization of chart components, but also extremia comparison of both bar values and spatial positions. Fig. 33 (c) ($D = 5$) presents further enhanced complexity by involving relations across different charts. This relational chart understanding making the problem solving more challenging, unveiling the significance of accurate visual reasoning in tackling complex CQA tasks.

Example 2 - Mode RVA: Fig. 36 presents two examples of successful inference in **RVA** mode. The bar chart example on the left shows reasoning with accurate visual grounding. The heatmap example on the right shows a case where the grounding is not exact but falls close to the regions of focus, also leading to the correct answer.

I Limitations & Future Work

In this work, we propose CURV (§3), a curriculum learning framework that develops intrinsic visual grounded reasoning capabilities in MLLMs by reformulating chart question answering as multi-step visual grounded reasoning with dynamic spatial attention. To support model learning, we introduce CCQA (§4), a three-level curriculum dataset with scalable synthetic generation across diverse chart types and reasoning patterns. Results demonstrate that CURV demonstrates notable and consistent improvements across chart understanding and out-of-domain multimodal reasoning benchmarks (§5.3). Nevertheless, we acknowledge a few limitations that we aim to investigate further in our future work. First, CCQA encompasses seven common chart types, while less conventional visualization forms, such as Sankey diagrams, treemaps, and geographic maps, are not currently covered, potentially limiting the applicability of CCQA to future work with more specialized training scenarios. Second, our framework and dataset focus exclusively on English-language charts and questions, which may restrict generalizability to multilingual contexts where chart-based communication is equally prevalent. In future work, we aim to address these limitations by extending CCQA to a broader range of chart types and multilingual settings. Beyond these, we also aim to explore agentic chart understanding, where models can better leverage external knowledge, tools, and multi-agent collaboration to complement the intrinsic visual reasoning capabilities established in this work.

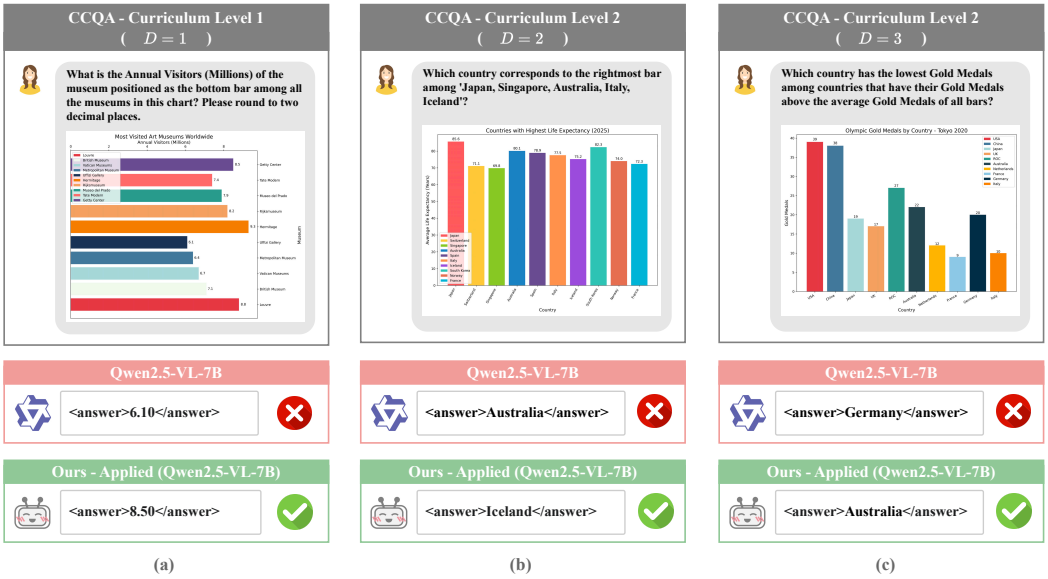


Figure 33: **Success Examples On Curriculum Levels 1-2.** This figure shows three examples on curriculum levels 1-2 of CCQA, where the baseline Qwen2.5-VL-7B fails while our CURV using *applied* grounding succeeds.

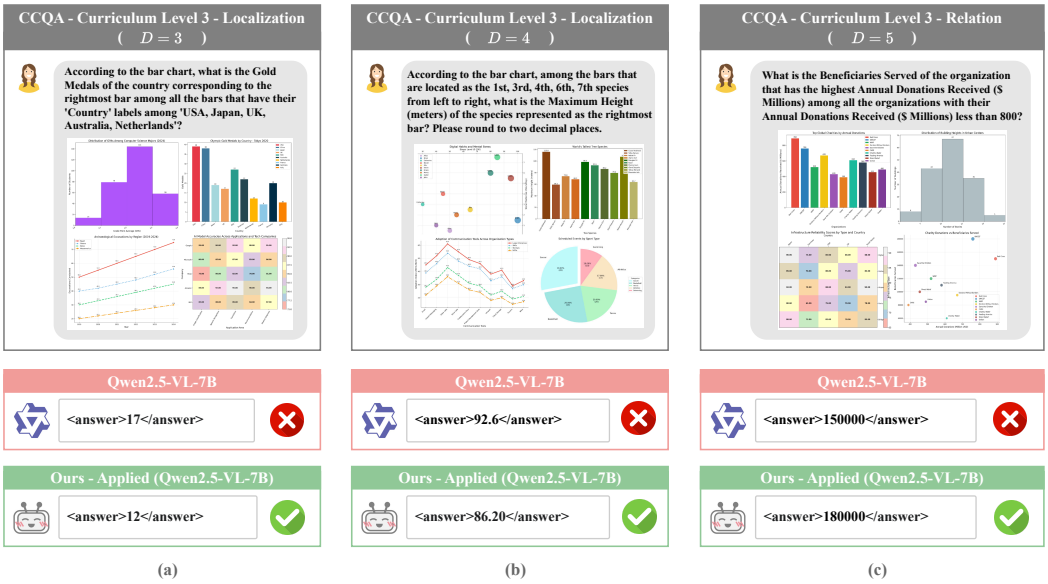


Figure 34: **Success Examples On Curriculum Level 3.** This figure shows three examples on the curriculum level 3 of CCQA, where the baseline Qwen2.5-VL-7B fails while our CURV using *applied* grounding succeeds.

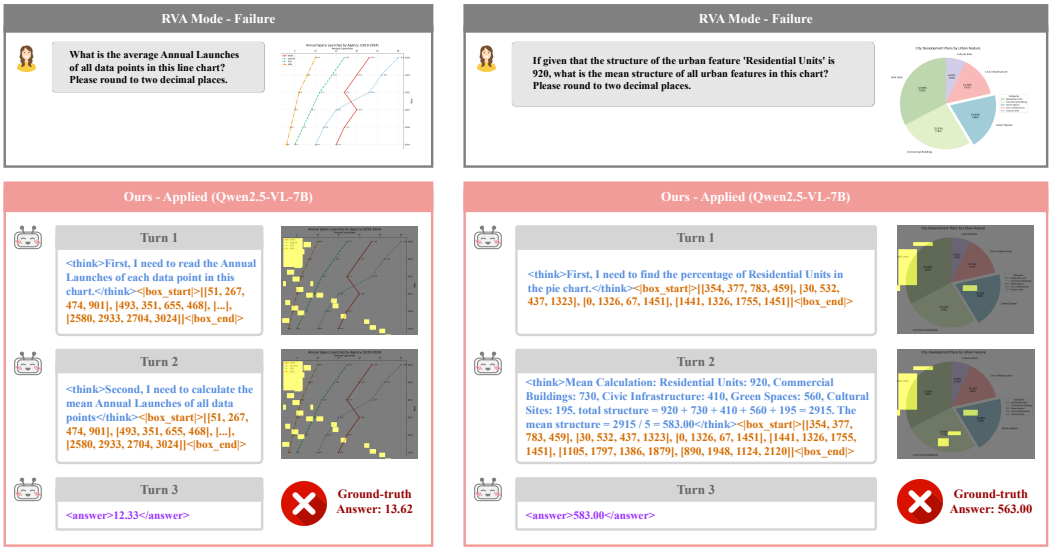


Figure 35: **Failure Examples Through Mode RVA.** This figure shows two examples of RVA inference on CCQA, where the model fails to give correct answers.

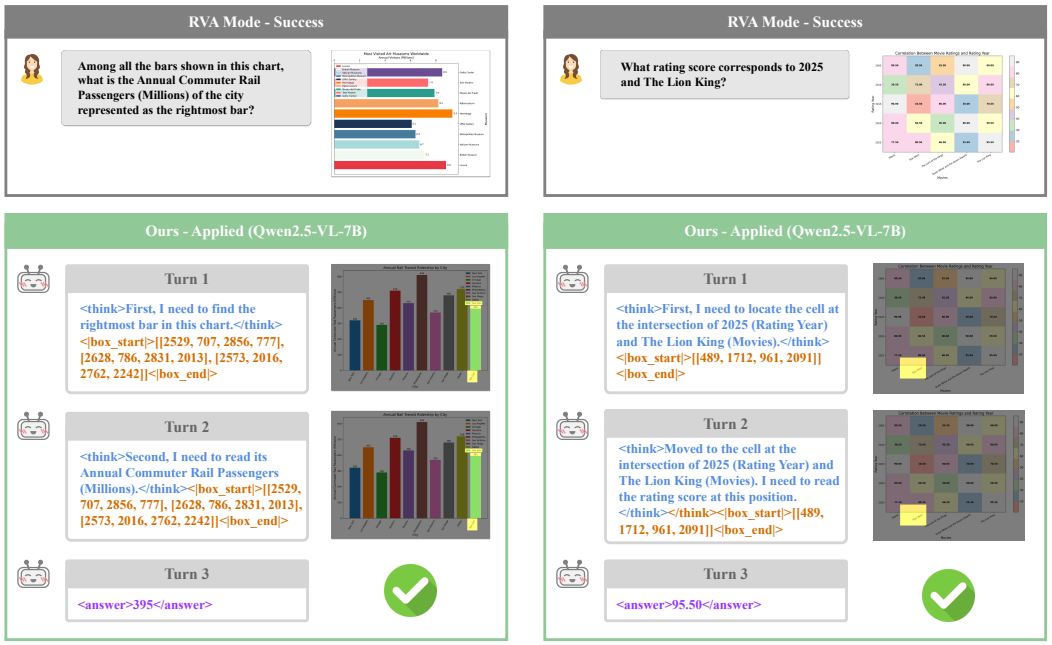


Figure 36: **Success Examples Through Mode RVA.** This figure shows three examples of RVA inference on CCQA, where the baseline Qwen2.5-VL-7B fails while our CURV using *applied* grounding succeeds.