DomainCQA: Crafting Knowledge-Intensive QA from Domain-Specific Charts

Yujing Lu[1*], Ling Zhong[1*], Jing Yang[1*], Weiming Li[1*], Peng Wei[2], Yongheng Wang[1], Manni Duan[1†], Qing Zhang[1†]

[1]Zhejiang Lab, Hangzhou, China
[2]National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China
{luyujing, zhongling, yangjing0128, liwm, wangyh, duanmanni, qing.zhang}@zhejianglab.org,
weipeng01@nao.cas.cn

Abstract

Chart Question Answering (CQA) evaluates Multimodal Large Language Models (MLLMs) on visual understanding and reasoning over chart data. However, existing benchmarks mostly test surface-level parsing, such as reading labels and legends, while overlooking deeper scientific reasoning. We propose DomainCQA, a framework for constructing domain-specific CQA benchmarks that emphasize both visual comprehension and knowledge-intensive reasoning. It integrates complexity-aware chart selection, multitier QA generation, and expert validation. Applied to astronomy, DomainCQA yields AstroChart, a benchmark of 1,690 QA pairs over 482 charts, exposing persistent weaknesses in fine-grained perception, numerical reasoning, and domain knowledge integration across 21 MLLMs. Fine-tuning on AstroChart improves performance across fundamental and advanced tasks. Pilot QA sets in biochemistry, economics, medicine, and social science further demonstrate DomainCQA's generality. Together, our results establish DomainCQA as a unified pipeline for constructing and augmenting domain-specific chart reasoning benchmarks.

Code — https://github.com/LingZhong01/DomainCQA
Datasets —
https://huggingface.co/datasets/yangjing0128/AstroChart
Extended version — https://arxiv.org/abs/2503.19498

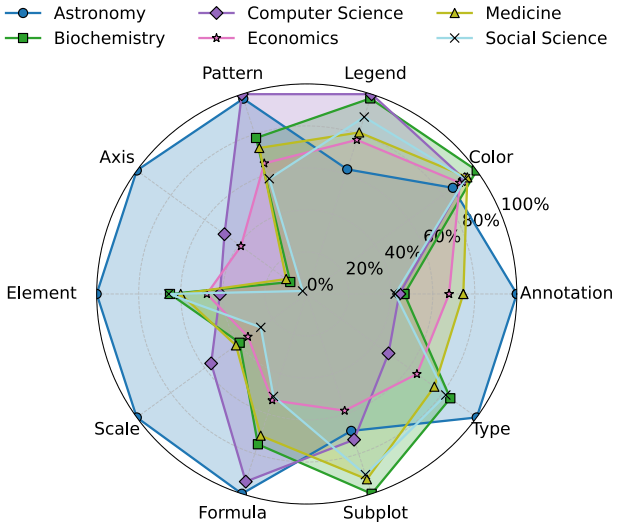

Figure 1: Radar plot of chart complexity across domains by comparing various visual design features, computed from 500 sampled charts per domain. Each axis represents a normalized design element contributing to overall chart complexity (formally defined later as the Chart Complexity Vector, or CCV). The domain-specific differences motivate our complexity-aware chart selection strategy.

Introduction

The success of Multimodal Large Language Models (MLLMs) has sparked growing interest in their ability to process and analyze scientific charts, which play a crucial role in conveying complex research data (Team et al. 2024; OpenAI 2024). Among various chart-related tasks, Chart Question Answering (CQA) has emerged as a fundamental challenge, requiring MLLMs to extract, interpret, and reason about chart-based information in response to natural language queries.

Although recent benchmarks in CQA, such as ChartQA (Masry et al. 2022), PlotQA (Methani et al. 2020), CharXiv (Wang et al. 2024), and SciCap (Hsu, Giles, and Huang 2021), have greatly advanced the field, all of them are deliberately *knowledge-agnostic*. Their question-answer (QA) pairs probe a model's ability to parse axes, legends and visual layouts, yet never require *domain–specific reasoning*. Consequently, we still do not know whether modern MLLMs can truly integrate visual cues *and* scientific knowledge.

Simply extending existing benchmark–building pipelines is inadequate for two reasons: (1) **Chart selection:** current pipelines choose charts either randomly or by ad–hoc manual curation, overlooking the fact that the mix of visual elements differs sharply from one scientific field to another, as Figure 1 shows that astronomy charts emphasize annotation and formula usage, biochemistry charts lean on color and subplot, etc. In short, the charts selected in these benchmarks are *not domain-representative*; (2) **Question design:** exist-

*Equal contribution.
†Corresponding author.

ing CQA datasets still focus on superficial visual cues; they rarely ask questions that demand domain knowledge, for instance, in astronomy charts correlating oscillation-frequency histograms with stellar classifications or interpreting how a redshift-magnitude scatter plot reflects cosmic expansion. In short, the questions designed in these benchmarks are *not knowledge-intensive*.

To address the two gaps identified above, we present **DomainCQA**, a framework for building domain-specific CQA benchmarks that integrates chart selection, QA generation, and expert QA validation into one seamless process. We encode each candidate chart with a 10-dimensional *Chart Complexity Vector (CCV)* and apply non-parametric Gibbs sampling to select the subset of charts used for questions that test basic understanding. For domain knowledge probing, we propose a chart abstract selector using chain-of-thought (CoT) reasoning to identify the most representative chart, along with a voting validator that enhances robustness through cross-model majority voting. Across both chart pools, we construct two tiers of QA: *Fundamental QA (FQA)* and *Advanced QA (AQA)*, and pass every QA pair through a multi-stage human review. The resulting benchmark is both domain-representative in its visuals and genuinely knowledge-intensive in its questions.

As a concrete application of DomainCQA, we construct **AstroChart**, the first CQA benchmark for astronomy. Leveraging our pipeline, we select 482 representative charts and generate $1,690$ QA pairs. Of these, $1,509$ are FQA pairs that test the understanding of the chart itself, while 181 are AQA pairs that require extra astronomical knowledge beyond the chart. Evaluating 21 state-of-the-art MLLMs on AstroChart exposes three persistent weaknesses: (i) chart reasoning – inferring trends and relationships from visual encodings; (ii) numerical computation – extracting values and performing arithmetic reliably; and (iii) domain-fact integration – combining chart evidence with astronomy-specific knowledge. Fine-tuning these models on data generated by DomainCQA yields notable gains, confirming the framework's value for both evaluation and data creation.

Beyond astronomy, we create pilot sets in biochemistry, economics, medicine and social science, each with domain specific charts and QA pairs showing that DomainCQA generalizes well across disciplines. These results confirm that the framework effectively addresses the two key gaps: selecting representative charts and generating knowledge-intensive questions.

Our key contributions are as follows: (1) DomainCQA, a three-phase framework for building domain-specific CQA benchmarks; (2) CCV, a 10-dimensional descriptor that captures domain-dependent visual traits and guides chart selection; (3) Chart abstracts, defined as charts summarizing articles' main findings, are ideal anchors for knowledge-intensive question generation; (4) AstroChart, the first CQA benchmark for astronomy; we evaluate 21 state-of-the-art (SOTA) MLLMs in zero-shot and fine-tuned settings to probe their domain-specific chart understanding.

Related Work

MLLMs for Chart Understanding Recent progress in MLLMs has substantially advanced chart understanding. Proprietary models such as GPT-4o (OpenAI 2024), Claude 3.5 (Anthropic 2024), Qwen-VL (Qwen Team 2023), and Gemini-2.5 (Google 2025) have demonstrated strong multimodal reasoning capabilities. Meanwhile, open-source MLLMs are rapidly evolving, offering accessible and customizable alternatives. Many models primarily focus on enhancing general vision-language ability through improved alignment, stronger representations, and more efficient inference, which improves performance on chart-related tasks. Notable examples include LLaVA (Liu et al. 2023b, 2024c,d), mPLUG-Owl (Ye et al. 2023a,b, 2024), SPHINX (Liu et al. 2024a), InternVL (Dong et al. 2024), CogVLM (Zhipu AI 2024), MiniCPM (OpenBMB 2024), and Pixtral (Mistral AI 2024a). In contrast, other models are specifically fine-tuned on chart-related tasks to better support structured data understanding, such as UniChart (Masry et al. 2023), Matcha (Liu et al. 2023a), ChartAssistant (Meng et al. 2024), and TinyChart (Zhang et al. 2024).

Benchmarks for CQA Evaluation A CQA benchmark consists of two key components: charts and corresponding QA pairs, both essential for evaluating a model's chart comprehension capabilities (Huang et al. 2025). Early datasets like DVQA (Kafle et al. 2018) and FigureQA (Kahou et al. 2018) utilized synthetic charts alongside templated QA pairs, whereas later efforts such as PlotQA (Methani et al. 2020), LEAF-QA (Chaudhry et al. 2020), and LEAF-QA++(Singh and Shekhar 2020) incorporated real numerical data with synthetic visualizations. More recent benchmarks, such as ChartQA(Masry et al. 2022), OpenCQA (Kantharaj et al. 2022), and MMC-Benchmark (Liu et al. 2024b), introduced charts sourced from real-world datasets. Among these, OpenCQA pioneered open-ended CQA tasks. The growing capabilities of LLMs have enabled recent studies such as SciGraphQA(Li and Tajbakhsh 2023), ChartX(Xia et al. 2025), and CharXiv(Wang et al. 2024) to generate more diverse QA pairs. Nevertheless, existing benchmarks mainly focus on general or broad scientific domains and lack the domain-specific focus required for detailed chart interpretation.

DomainCQA Framework

DomainCQA (see Figure 2) offers a systematic framework to build domain-specific CQA benchmarks that test both general visual understanding and specialized reasoning. It defines two types of QA tasks:

- **Fundamental QA (FQA)**: testing chart comprehension via basic visual reasoning like label recognition, color differentiation, and simple comparisons.
- **Advanced QA (AQA)**: requiring domain knowledge beyond the chart, including interpreting specialized symbols, terms, or concepts.

Together, questions from both tasks enable the benchmark to evaluate chart understanding across a spectrum from surface-level comprehension to discipline-specific insight.

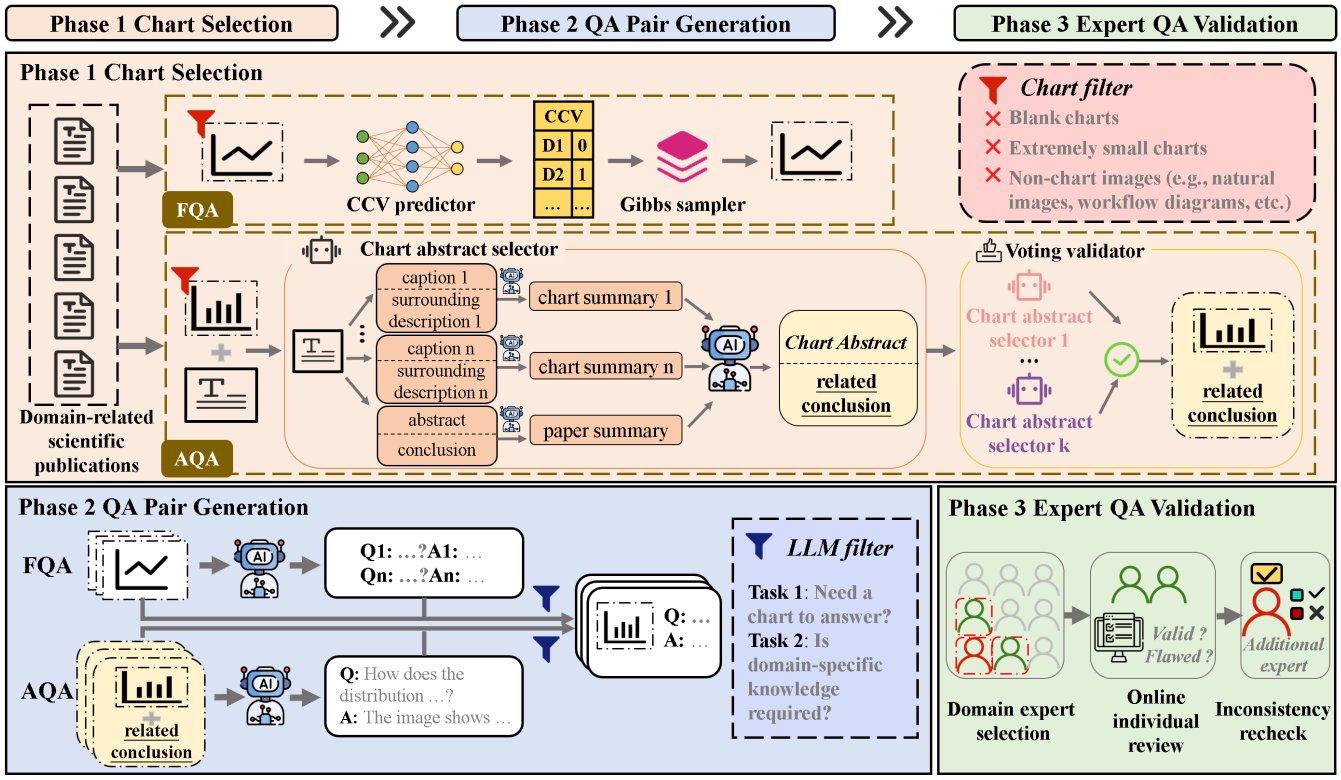

Figure 2: Overview of the DomainCQA framework for constructing domain-specific CQA benchmarks. The pipeline consists of three stages: Chart Selection, QA Pair Generation and Expert QA Validation. The resulting benchmarks support evaluation of both visual comprehension and knowledge-intensive reasoning.

Chart Selection

DomainCQA selects the charts separately for the FQA pairs and the AQA pairs, ensuring that each set is aligned with the specific evaluation requirements of its type of question.

Charts for FQA Our goal is to build an FQA-chart pool whose visual variety matches the unknown, true distribution of charts in a scientific domain. We operate on a pre-compiled corpus of domain charts and focus on two ingredients: a Chart Complexity Vector (CCV) that embeds each chart in a ten-dimensional feature space, and a non-parametric Gibbs sampler that draws a subset whose joint CCV statistics closely match those of the corpus.

Each CCV dimension measures a distinct aspect of visual difficulty, such as plot elements, color diversity, annotation density, and visual clutter. We train a ResNet-50 classifier on an annotated subset to predict the ten CCV attributes for the remaining charts, yielding 10-dimensional representations that capture domain-specific patterns (see Appendix A for more details on CCV).

Random sampling disregards the structured distribution of visual complexity within each domain, producing samples that do not faithfully reflect domain-specific patterns. Instead, we treat the CCV collection as an empirical distribution and perform non-parametric Gibbs sampling(Casella and George 1992) to preserve marginal distributions and inter-dimensional dependencies (see Appendix B for the Gibbs sampling pseudocode).

Let $\mathcal{C} = \{\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(N)}\} \subset \mathbb{R}^{10}$ be the set of CCV vectors for all candidate charts. Each $\mathbf{c}^{(n)} = (c_1^{(n)}, \ldots, c_{10}^{(n)})$ encodes the visual, structural, and interpretive attributes of a chart. At each iteration t, we:

1. Randomly choose a dimension $k_t \in \{1, \ldots, 10\}$;
2. Sample a target value $\zeta \sim \hat{p}_{k_t}$, the empirical marginal distribution of dimension k_t;
3. Search for a new chart $\mathbf{c}^{(t)} \in \mathcal{C}$ that best matches the current state $\mathbf{c}^{(t-1)}$ on the remaining 9 dimensions and is closest to ζ in dimension k_t.

The resulting chart subset approximates the latent domain distribution in CCV space and serves as our FQA chart pool.

Charts for AQA Selecting charts for AQA requires more than visual diversity, instead it demands charts that meaningfully reflect domain knowledge. A naive approach would be to reuse charts from the FQA set and pose domain-specific questions on them. However, this often results in noisy inputs that dilute question quality. Many visually complex charts are tangentially related to the paper's core findings, making them poor candidates for knowledge-intensive tasks.

We address this by targeting a chart that directly reflects a paper's main scientific conclusions, commonly referred to as *chart abstract*. To identify them, we design a lightweight two-stage LLM-based method: a *chart abstract* selector that

leverages CoT to identify the chart most relevant to a paper's abstract and conclusion, and a voting validator, which aggregates reasoning outputs from multiple LLMs via cross-model majority voting to enhance selection reliability (see Appendix C for the pseudocode of AQA chart selection).

This approach yields semantically meaningful charts that support deeper reasoning, as demonstrated by our later experiments across five scientific domains (see Sec. 5.5). QA pairs constructed from chart abstracts consistently outperform those from our FQA method in both domain relevance and QA validity.

QA Pair Generation

From selected charts, we design two types of questions: FQA and AQA. FQA covers four categories of tasks: *Visual* (recognizing graphical elements), *Data* (retrieving and computing values), *Inference* (inferring patterns and relations), and *Chart Description* (summarizing the visual content). AQA is formulated as a knowledge-based inference (*KB-Inference*) task, requiring integration of external scientific knowledge with visual content.

To ensure quality, we apply a secondary LLM-based validation filter to all generated QA pairs. This verifier checks two key criteria: (1) whether the QA pair is grounded in the visual content of the chart, and (2) for AQA, whether it requires domain-specific knowledge to answer (see Appendices D and E for prompt templates and validation criteria.)

Expert QA Validation

To ensure benchmark quality, each QA pair undergoes expert review to validate both its clarity and factual correctness. Reviewers label each item as either: *Valid* (the question is well-posed and the answer is accurate); *Flawed* (the question is ambiguous, misleading, or the answer is incorrect). All QA pairs are independently assessed by domain experts. Disagreements are resolved through additional review rounds until consensus is reached.

AstroChart: A Benchmark for Astronomy

We present a complete benchmark instantiation, AstroChart, in the astronomy domain, comprising $1,690$ QA pairs grounded in 482 charts. We also conducted partial experiments in other domains to validate key steps, chart selection and QA pair generation for AQA, as detailed in Evaluation.

Chart Selection To construct the FQA chart portion of AstroChart, we collected figures from arXiv astronomy papers published between 2007 and 2023. A ResNet-18 classifier, trained to detect non-scientific or low-quality visuals, was used to filter out irrelevant figures. For each remaining chart, we computed its CCV and applied non-parametric Gibbs sampling to select 305 charts whose CCV distribution approximates the overall domain distribution, ensuring a diverse and representative subset.

To assess the representativeness of our selected charts, we further compared the visual complexity of AstroChart with existing CQA benchmarks. Specifically, we computed the CCV for each chart in several public datasets (CharXiv, ChartQA, OpenCQA, PlotQA), and summed the ten CCV dimensions to obtain an overall complexity score. As illustrated in Figure 3, the charts in these benchmarks are mainly clustered in the 1 to 4 range, indicating simple visual structures. In contrast, AstroChart centers around 4 to 6 and exhibits a broader spread across the complexity spectrum. This suggests that AstroChart offers richer and more varied visual content, better aligned with real-world scientific charts.

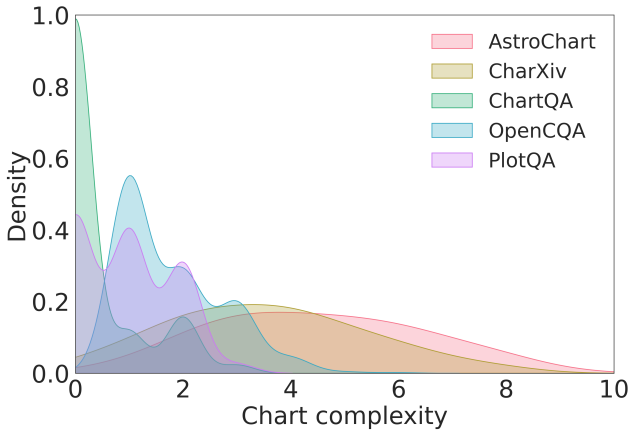

Figure 3: Chart complexity calculated from CCVs across benchmarks, where AstroChart shows a broader and higher complexity distribution than other benchmarks, with more domain-specific charts in the 6–10 range (see Appendix A.2 for CCV score details).

For AQA, we targeted the high-impact literature by selecting the top 1% most-cited articles each year in the six main subfields of astronomy (See Appendix F). After applying the same filtering process, we identify *chart abstracts* using a consensus-based approach from GPT-4o and Claude 3.5. This results in 178 high-quality charts suitable for domain-specific reasoning.

In total, AstroChart includes 482 distinct charts with one in both (see Appendix G for visualizations).

QA Pair Generation We employ Claude 3.5 to generate QA pairs using category-specific prompts, ensuring that each question is well aligned with its associated chart. To refine quality, GPT-4o is used to automatically filter out QA pairs that either lack a clear connection to the chart or do not require external domain knowledge for answering.

We further assessed the reliability of GPT-4o's filtering by comparing its judgments against human annotations on 200 randomly sampled QA pairs. Beyond achieving 96.5% overall accuracy, GPT-4o demonstrated substantial agreement with human reviewers, with a Cohen's Kappa (Cohen 1960) of 0.77, indicating strong consistency in identifying deletable items. Most discrepancies were conservative false positives, underscoring GPT-4o's cautious filtering style and practical reliability at scale (see Appendix H for details).

The final dataset comprises $1,690$ QA pairs, including $1,509$ FQA pairs and 181 AQA pairs, as summarized in Table 1 (see Appendix I for examples).

Type	Category	Aspect	Count
FQA	Visual	Color	211
		Style	133
		Text	213
		Layout	45
	Data	Point	130
		Interval	102
		Calculation	84
	Inference		289
	Chart Description		302
AQA	KB-Inference		181
Total			**1690**

Table 1: Distribution of question types in AstroChart

Expert QA Validation We conducted a comprehensive verification of the entire AstroChart benchmark to ensure its accuracy and reliability. A team of eight astronomy experts independently reviewed all $1,690$ QA pairs using our custom online assessment platform, with a total annotation time exceeding 160 hours. Each pair of QAs was evaluated by two randomly assigned reviewers and any disagreements were resolved through additional review rounds until consensus was reached (details in Appendix J). This rigorous expert validation process reinforces the credibility of AstroChart as a high-quality benchmark for evaluating MLLMs in astronomical chart understanding.

Evaluation

To assess the utility and difficulty of AstroChart, we design three experiments. First, we benchmark 21 SOTA MLLMs under a zero-shot setting to assess their capabilities across question categories. Second, we construct a training set using the same pipeline as AstroChart (excluding expert validation), fine-tune a representative model, and test its performance on both AstroChart and other benchmarks to assess generalization. Third, we compare AstroChart with CharXiv to evaluate relative difficulty. Finally, we also verify that the DomainCQA framework can produce high-quality AQA pairs in other scientific domains.

Setup and Metrics

Zero-Shot Setup We evaluated 21 MLLMs, including both proprietary and open-source variants. Proprietary models were accessed via API, and open-source models were run locally on a single Nvidia A100-80GB GPU. Under the zero-shot protocol, each model received only the chart and its corresponding question, without any in-context examples or prior training. Four astronomy researchers were also invited to establish a human baseline by answering 10% of questions from each category using the same prompts as the models to ensure fairness.

Fine-Tuning Setup To evaluate training effectiveness, we constructed a fine-tuning dataset using the same pipeline as AstroChart, omitting the final expert QA validation step. This yielded 9,857 training and 8,729 validation scientific charts, from which we generated 86,681 and 21,738 QA pairs, respectively. We fine-tuned an open-source model, MiniCPM-V2.6-8B, using the parameter-efficient LoRA (Hu et al. 2021) method. Training was conducsted on 8 Nvidia A100-80GB GPUs with BF16 mixed precision and DeepSpeed ZeRO-2 (Rajbhandari et al. 2021) optimization for scalability and efficiency.

Evaluation Metrics We assess the accuracy of model outputs for both numerical and open-ended questions (details in Appendix K). For numerical responses, we computed relative error normalized by the axis range for retrieval tasks, and required an exact match for derivation tasks such as counting or arithmetic. For open-ended responses, an LLM judge (DeepSeek-V3) assigned scores from 0 to 1 based on relevance, correctness, and completeness, following Liu et al. (2023c). To verify scoring reliability, we compared DeepSeek-V3's scores with human annotations on 176 samples, yielding a Pearson correlation of 0.816, Spearman correlation of 0.817, and MAE of 0.096. ROUGE-L (Lin 2004), BLEU-4 (Papineni et al. 2002), and L3Score (Pramanick, Chellappa, and Venugopalan 2024) show similar trends to LLM scoring (see Appendix L).

Benchmarking 21 MLLMs on AstroChart

We report the performance of 21 MLLMs on AstroChart across FQA and AQA categories, as shown in table 2.

In FQA, models performed strongly on visual understanding tasks—top performers such as Gemini-2.5-Pro and GPT-4o achieved over 85% accuracy across categories like color, style, and layout, indicating mature capabilities in recognizing and interpreting visual elements. In contrast, data-centric tasks, especially those involving interval comparison and numerical calculation, remained more challenging. Although leading models exceeded 60% on interval questions, calculation accuracy typically stayed below 50%, exposing a gap in quantitative reasoning.

For AQA, which focuses on knowledge-based inference, performance declined further. Even top models scored below 75%, showing the challenge of integrating chart evidence with astronomy knowledge. In the human baseline, researchers achieved only 39%, far lower than leading VLMs, suggesting that even experts face limits beyond their subfields. These results confirm AstroChart as a valuable benchmark for assessing MLLMs' scientific reasoning ability.

Fine-Tuning a Representative MLLM

To further assess AstroChart's value as a training resource, we fine-tuned MiniCPM-V2.6-8B, the strongest performer among mid-sized open-source models, using a training set generated by the same pipeline as AstroChart (excluding expert validation). As shown in Table 2, the fine-tuned model achieves consistent improvements across all FQA and AQA categories, with an overall gain of 5.02%, confirming the effectiveness of our training data in enhancing both visual understanding and scientific reasoning.

To evaluate generalization, we tested the fine-tuned MiniCPM on three existing CQA benchmarks: CharXiv, ChartQA, and MMC-Benchmark. As shown in Table 3, performance changes are minimal, i.e., some metrics slightly

Model	FQA											AQA	All/1690
	Visual/602					Data/316				Infer./289	Chart Desc./302	KB-Infer./181	
	All	Color	Style	Text	Layout	All	Point	Interval	Calculation				
Human Baseline(10% Sample)	98.60	98.54	98.40	98.63	99.53	96.40	98.62	93.86	96.50	91.82	70.00	39.00	85.56
Proprietary Multimodal Large Language Models													
Gemini-2.5-Pro(Google 2025)	**88.22**	87.37	**87.70**	**90.23**	**84.67**	**72.66**	**81.22**	**75.10**	56.43	81.31	81.09	**73.65**	**81.30**
Gemini-2.5-flash(Google 2025)	87.21	87.04	87.04	88.73	81.78	64.15	68.34	63.40	**58.57**	**82.01**	**82.65**	72.49	79.62
GPT-4o(OpenAI 2024)	86.23	**88.92**	84.15	85.31	**84.67**	53.19	53.78	60.35	43.57	75.40	80.96	73.04	75.84
Qwen-VL-Max(Qwen Team 2023)	83.13	87.79	76.78	83.43	77.78	50.96	52.27	56.55	42.14	75.16	76.62	68.23	72.99
Open-source Multimodal Large Language Models													
TinyChart-3B(Zhang et al. 2024)	29.41	47.75	25.15	13.71	27.11	12.22	18.80	9.39	5.48	23.94	1.56	20.83	19.36
Llava1.5-7B(Liu et al. 2024c)	31.04	49.39	27.70	14.34	33.33	8.47	8.88	7.98	8.45	42.53	13.94	45.36	27.26
Llava1.6-Mistral-7B(Liu et al. 2024d)	46.45	61.36	41.96	33.00	50.89	13.77	17.74	14.18	7.14	49.24	23.84	48.23	36.97
Qwen-VL-Chat-7B(Qwen Team 2023)	44.47	55.73	41.04	32.68	54.44	10.47	16.11	6.72	6.31	38.89	22.05	45.19	33.23
Janus-Pro-7B(DeepSeek 2024)	66.69	74.74	67.26	56.62	74.67	32.27	35.10	38.09	20.83	56.37	51.23	54.75	54.45
MiniCPM-V2.6-8B(OpenBMB 2024)	70.31	75.92	61.89	71.92	61.78	33.30	34.87	43.74	18.21	55.16	55.60	54.20	56.44
InternVL3-8B(OpenGVLab 2025)	66.64	72.72	62.04	62.23	71.78	35.41	38.89	42.87	20.95	54.33	49.57	54.09	54.30
mPLUG-Owl2-8.2B(Ye et al. 2023b)	28.54	39.95	27.70	16.48	32.00	9.48	11.15	12.20	3.57	38.41	9.37	42.21	24.70
Pixtral-12B(Mistral AI 2024a)	79.27	83.00	75.70	78.26	76.89	51.54	53.63	60.64	37.26	71.90	78.74	69.28	71.66
Llava1.6-Vicuna-13B(Liu et al. 2024d)	49.45	66.43	44.74	34.84	50.89	13.23	17.77	10.40	9.64	44.36	23.44	50.77	37.30
SPHINX-v2-13B(Liu et al. 2024a)	31.68	48.40	29.41	18.26	21.11	7.23	13.36	1.47	4.76	37.27	6.13	44.25	24.84
Llama4-Maverick-17B(Meta 2025)	84.27	86.01	78.59	86.20	**83.56**	56.30	55.14	58.27	**55.71**	77.02	76.42	**74.64**	75.37
CogVLM2-19B(Zhipu AI 2024)	66.29	74.81	54.52	64.04	71.78	29.27	29.82	37.48	18.45	51.90	54.90	50.66	53.20
Gemma-3-27B(Gemma 2025)	69.93	69.30	68.44	69.06	80.89	37.21	38.22	47.63	22.98	58.72	66.23	62.54	60.44
Llava1.6-Yi-34B(Liu et al. 2024d)	50.63	66.34	44.37	37.93	53.56	18.19	17.60	25.30	10.48	47.09	36.19	55.36	41.89
Qwen2.5-VL-72B(Qwen Team 2025)	83.21	85.31	77.04	86.34	76.22	53.46	54.57	56.36	48.21	72.46	77.52	68.34	73.20
Pixtral-large-124B(Mistral AI 2024b)	**86.11**	**86.76**	**82.59**	**88.22**	82.44	**59.38**	**63.67**	**63.51**	47.74	**78.65**	**80.93**	70.83	**77.23**
Fine-tuned													
MiniCPM-V2.6-8B-fine-tuned	78.15↑	81.08↑	76.26↑	76.76↑	76.00↑	37.47↑	37.66↑	47.78↑	24.64↑	56.30↑	60.89↑	57.02↑	61.46↑

Table 2: Accuracy (%) on the AstroChart benchmark. "Infer." denotes Inference, and "Chart Desc." denotes Chart Description, and "KB-Infer." denotes KB-Inference. Bold numbers indicate the best-performing model among proprietary and open-source MLLMs, respectively (see Appendix M for model architecture details).

increase while others slightly drop. This suggests that the model has not overfitted AstroChart and that its learned reasoning skills remain largely transferable across domains.

Difficulty Comparison with CharXiv

Figure 4 compares model performance on AstroChart and CharXiv. We choose CharXiv for this comparison because, among existing benchmarks, it contains charts with the second-highest overall visual complexity after AstroChart (see Figure 3). To ensure fairness, we randomly sample 1,600 QA pairs from CharXiv to match AstroChart in size. Despite this, we observe a consistent performance drop across multiple MLLMs on AstroChart, highlighting its greater difficulty. All evaluations follow a unified metric framework for consistency.

This gap stems not only from complex visual structures but also from domain-specific questions that require deeper scientific reasoning rather than shallow visual interpretation.

CharXiv			
	Descriptive	Reasoning	Overall
MiniCPM-V2.6-8B	54.06	29.15	49.08
MiniCPM-V2.6-8B–fine-tuned	54.58	29.30	**49.52**↑
ChartQA			
	Human	Augmented	Overall
MiniCPM-V2.6-8B	57.12	83.84	70.48
MiniCPM-V2.6-8B-fine-tuned	58.24	80.80	69.52↓
MMC-Benchmark			
	MCQ	T/F	Overall
MiniCPM-V2.6-8B	66.46	77.56	74.40
MiniCPM-V2.6-8B-fine-tuned	62.38	77.42	73.28↓

Table 3: Performance of MiniCPM-V2.6-8B before and after fine-tuning on various CQA benchmarks.

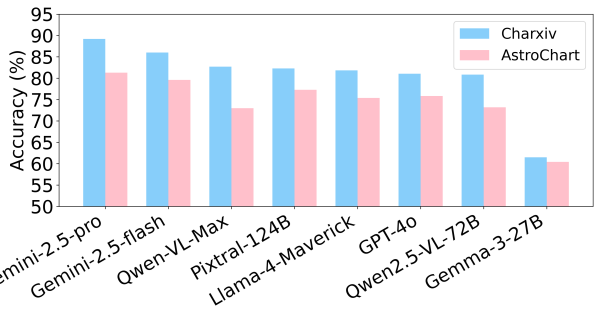

Figure 4: Performance comparison of MLLMs on Charxiv and AstroChart.

Domain	Relevance (R/A)	Validity (R/A)
Astronomy	3.27 / 3.84	0.73 / 0.76
Biochemistry	3.48 / 3.80	0.86 / 0.98
Economics	3.48 / 3.78	0.95 / 0.98
Medicine	3.21 / 3.77	0.89 / 0.96
Social Science	3.23 / 3.46	0.95 / 0.93

Table 4: Expert validation scores for QA pairs generated from randomly sampled charts from the FQA pool (R) vs. AQA-selected charts (A).

Evaluation of AQA Generation on Domains

To evaluate the generalizability of DomainCQA across disciplines, we conduct a pilot study in four additional scientific domains: biochemistry, economics, medicine, and social science. While the full benchmark includes both FQA and AQA components, we focus on AQA, which selects chart abstracts and generates knowledge-intensive questions. In contrast, FQA involves domain-aware sampling and requires minimal downstream evaluation. This study examines whether AQA can reliably identify knowledge-centric charts and generate high-quality, domain-relevant QA pairs across diverse fields.

Domain experts independently assess each QA pair along two dimensions. Domain relevance is scored on a 1–5 scale, with higher scores indicating deeper and more precise use of domain-specific knowledge beyond what is directly shown in the chart. QA validity is scored as 1 (correct), 0 (cannot determine), or -1 (incorrect), based on clarity of the question and factual correctness of the answer.

As shown in Table 4, AQA-generated QA pairs generally receive higher scores in both relevance and validity compared to those from randomly sampled FQA charts across all domains (see Appendix O for rating criteria). This expert validation confirms the robustness and adaptability of the DomainCQA methodology, supporting its application to a broad range of scientific fields.

Discussion

Limitations Revealed by AstroChart AstroChart highlights key weaknesses in current MLLMs when handling scientific charts. Most models do well on visual tasks like identifying layouts or chart types, but struggle with detailed perception, especially in distinguishing similar colors or reading small labels. Their numerical reasoning is also weak, that is, models often misread axis values or return full axis ranges instead of specific intervals. On calculation tasks, such issues are made worse by OCR errors and limited math skills.

AQA evaluation reveals deeper challenges in domain understanding. Many models give vague, generic responses, confuse scientific ideas, or misuse technical terms. This shows a clear gap in vision-language alignment and the lack of embedded scientific knowledge. A major reason is that most vision-language pretraining relies on generic image–caption pairs, which fail to expose models to the structured layouts and domain-specific terminology found in scientific charts (see Appendix N for failure cases).

Effectiveness of DomainCQA Our results demonstrate the effectiveness of DomainCQA as both a benchmark construction framework and a practical training pipeline. By reusing the same generation methodology to build a fine-tuning set without targeting specific weaknesses, we cover challenging tasks like data interpretation, visual discrimination, and domain-informed inference. Fine-tuning on this dataset consistently improves performance on both FQA and AQA tasks, showing the QA pairs' informativeness and training value. The fine-tuned model also performs well on external benchmarks such as CharXiv, ChartQA, and MMC-Benchmark, indicating it has not overfit to AstroChart and that its reasoning skills transfer across domains. Moreover, DomainCQA can be easily applied to other scientific fields, highlighting its generalizability as a domain-independent CQA construction pipeline.

Conclusion & Future Work

Conclusion We present **DomainCQA**, a structured methodology for building domain-specific chart QA benchmarks, and demonstrate its effectiveness through *AstroChart*, the first CQA benchmark for astronomy. AstroChart captures both basic chart understanding and domain-informed reasoning. Through extensive evaluation of 21 MLLMs, we reveal consistent weaknesses in chart understanding, especially when models integrate visual features with domain-specific knowledge. In addition to AstroChart, we apply DomainCQA to four scientific fields, such as biochemistry, economics, medicine, and social science, conducting pilot AQA studies with expert validation. These results confirm the generality and effectiveness of our methodology in producing high-quality, relevant, and challenging QA pairs. Furthermore, using data generated by DomainCQA for fine-tuning significantly improves MLLM performance across diverse chart reasoning tasks without overfitting, highlighting the training utility of our pipeline.

Future Work Building on our preliminary exploration across multiple scientific domains, we plan to extend DomainCQA into a broader suite of benchmarks in multiple scientific domains. Our long-term goal is to establish DomainCQA as a standard framework for chart-based scientific reasoning in real-world MLLM applications.

Ethical Statement

This work does not involve human or animal subjects. All data used in this work are chart-based and originate from publicly available scientific publications. These materials were accessed solely for research purposes, and no proprietary, confidential, or human-related information is involved. No ethical concerns were identified in the construction of the benchmarks and experiments.

Acknowledgments

We sincerely thank the anonymous reviewers and contributing researchers for their valuable feedback. This research was supported by the National Natural Science Foundation of China (U22A2032), the Leading Innovation and Entrepreneurship Team of Zhejiang Province of China (Grant No. 2023R01008), Zhejiang Provincial Science and Technology Plan Project (2023C01120), Key R&D Program of Zhejiang (2024SSYS0012), and the China Manned Space Project (CMS-CSST-2025-A21).

References

Anthropic. 2024. Claude 3 Model Family: Opus, Sonnet, Haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Casella, G.; and George, E. I. 1992. Explaining the Gibbs Sampler. *The American Statistician*, 46: 167–174.

Chaudhry, R.; Shekhar, S.; Gupta, U.; Maneriker, P.; Bansal, P.; and Joshi, A. 2020. LEAF-QA: Locate, Encode & Attend for Figure Question Answering. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 3501–3510.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

DeepSeek. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525.

Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Wei, X.; Zhang, S.; Duan, H.; Cao, M.; Zhang, W.; Li, Y.; Yan, H.; Gao, Y.; Zhang, X.; Li, W.; Li, J.; Chen, K.; He, C.; Zhang, X.; Qiao, Y.; Lin, D.; and Wang, J. 2024. InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model. arXiv:2401.16420.

Gemma, G. T. 2025. Gemma 3 technical report. arXiv:2503.19786.

Google, G. T. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Hsu, T.-Y.; Giles, C. L.; and Huang, T.-H. 2021. SciCap: Generating Captions for Scientific Figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3258–3264. Association for Computational Linguistics.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Huang, K.-H.; Chan, H. P.; Fung, M.; Qiu, H.; Zhou, M.; Joty, S.; Chang, S.-F.; and Ji, H. 2025. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5): 2550–2568.

Kafle, K.; Price, B.; Cohen, S.; and Kanan, C. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5648–5656.

Kahou, S. E.; Michalski, V.; Atkinson, A.; Kadar, A.; Trischler, A.; and Bengio, Y. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. arXiv:1710.07300.

Kantharaj, S.; Do, X. L.; Leong, R. T.; Tan, J. Q.; Hoque, E.; and Joty, S. 2022. OpenCQA: Open-ended Question Answering with Charts. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 11817–11837. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Li, S.; and Tajbakhsh, N. 2023. SciGraphQA: A Large-Scale Synthetic Multi-Turn Question-Answering Dataset for Scientific Graphs. arXiv:2308.03349.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, D.; Zhang, R.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; Zhang, K.; Shao, W.; Xu, C.; He, C.; He, J.; Shao, H.; Lu, P.; Qiao, Y.; Li, H.; and Gao, P. 2024a. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 32400–32420. PMLR.

Liu, F.; Piccinno, F.; Krichene, S.; Pang, C.; Lee, K.; Joshi, M.; Altun, Y.; Collier, N.; and Eisenschlos, J. 2023a. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12756–12770. Toronto, Canada: Association for Computational Linguistics.

Liu, F.; Wang, X.; Yao, W.; Chen, J.; Song, K.; Cho, S.; Yacoob, Y.; and Yu, D. 2024b. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1287–1310. Mexico City, Mexico: Association for Computational Linguistics.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024c. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 26286–26296.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024d. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023c. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522.

Masry, A.; Do, X. L.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279. Dublin, Ireland: Association for Computational Linguistics.

Masry, A.; Kavehzadeh, P.; Do, X. L.; Hoque, E.; and Joty, S. 2023. UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14662–14684. Singapore: Association for Computational Linguistics.

Meng, F.; Shao, W.; Lu, Q.; Gao, P.; Zhang, K.; Qiao, Y.; and Luo, P. 2024. ChartAssistant: A Universal Chart Multimodal Language Model via Chart-to-Table Pre-training and Multitask Instruction Tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 7775–7803. Bangkok, Thailand: Association for Computational Linguistics.

Meta. 2025. Llama 4: A New Era of Natively Multimodal AI Innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-06-26.

Methani, N.; Ganguly, P.; Khapra, M. M.; and Kumar, P. 2020. PlotQA: Reasoning over Scientific Plots. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Mistral AI. 2024a. Pixtral 12B. arXiv:2410.07073.

Mistral AI. 2024b. Pixtral Large: A 124B Open-Weights Multimodal Model. Mistral AI blog and model page.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

OpenBMB. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800.

OpenGVLab. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. arXiv:2504.10479.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P.; Charniak, E.; and Lin, D., eds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pramanick, S.; Chellappa, R.; and Venugopalan, S. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37: 118807–118833.

Qwen Team, A. G. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Qwen Team, A. G. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Rajbhandari, S.; Ruwase, O.; Rasley, J.; Smith, S.; and He, Y. 2021. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, 1–14.

Singh, H.; and Shekhar, S. 2020. STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3275–3284.

Team, R.; Ormazabal, A.; Zheng, C.; de Masson d'Autume, C.; Yogatama, D.; Fu, D.; Ong, D.; Chen, E.; Lamprecht, E.; Pham, H.; Ong, I.; Aleksiev, K.; Li, L.; Henderson, M.; Bain, M.; Artetxe, M.; Relan, N.; Padlewski, P.; Liu, Q.; Chen, R.; Phua, S.; Yang, Y.; Tay, Y.; Wang, Y.; Zhu, Z.; and Xie, Z. 2024. Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models. arXiv:2404.12387.

Wang, Z.; Xia, M.; He, L.; Chen, H.; Liu, Y.; Zhu, R.; Liang, K.; Wu, X.; Liu, H.; Malladi, S.; Chevalier, A.; Arora, S.; and Chen, D. 2024. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. arXiv:2406.18521.

Xia, R.; Zhang, B.; Ye, H.; Yan, X.; Liu, Q.; Zhou, H.; Chen, Z.; Ye, P.; Dou, M.; Shi, B.; Yan, J.; and Qiao, Y. 2025. ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning. arXiv:2402.12185.

Ye, J.; Xu, H.; Liu, H.; Hu, A.; Yan, M.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2024. mPLUG-Owl3: Towards Long Image-Sequence Understanding in Multi-Modal Large Language Models. arXiv:2408.04840.

Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; Jiang, C.; Li, C.; Xu, Y.; Chen, H.; Tian, J.; Qian, Q.; Zhang, J.; and Huang, F. 2023a. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. arXiv:2304.14178.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; and Zhou, J. 2023b. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration. arXiv:2311.04257.

Zhang, L.; Hu, A.; Xu, H.; Yan, M.; Xu, Y.; Jin, Q.; Zhang, J.; and Huang, F. 2024. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning. arXiv:2404.16635.

Zhipu AI. 2024. CogVLM2: Visual Language Models for Image and Video Understanding. arXiv:2408.16500.