Enhancing Scientific Visual Question Answering through Multimodal Reasoning and Ensemble Modeling

Prahitha Movva
University of Massachusetts Amherst
Amherst, MA, USA
prahitha.movva03@gmail.com

Naga Harshita Marupaka
University of Southern California
Los Angeles, CA, USA
nagaharshitamarupaka@gmail.com

Abstract

Technical reports and articles often contain valuable information in the form of semi-structured data like charts, and figures. Interpreting these and using the information from them is essential for downstream tasks such as question answering (QA). Current approaches to visual question answering often struggle with the precision required for scientific data interpretation, particularly in handling numerical values, multi-step reasoning over visual elements, and maintaining consistency between visual observation and textual reasoning. We present our approach to the SciVQA 2025 shared task, focusing on answering visual and non-visual questions grounded in scientific figures from scholarly articles.

We conducted a series of experiments using models with 5B to 8B parameters. Our strongest individual model, InternVL3, achieved ROUGE-1 and ROUGE-L F1 scores of **0.740** and a BERTScore of **0.983** on the SciVQA test split. We also developed an ensemble model with multiple vision language models (VLMs). Through error analysis on the validation split, our ensemble approach improved performance compared to most individual models, though InternVL3 remained the strongest standalone performer. Our findings underscore the effectiveness of prompt optimization, chain-of-thought reasoning and ensemble modeling in improving the model's ability in visual question answering.

1 Introduction

Scientific literature communicates complex ideas not only through text but also through carefully designed visual elements including charts, graphs, diagrams, and technical illustrations. These visualizations serve as dense information carriers, encoding quantitative relationships, experimental results, architectural designs, and conceptual frameworks that are essential for scientific understanding. The ability to automatically interpret and reason about these visual elements represents a critical challenge in advancing scientific AI systems.

The task of Visual Question Answering (VQA) over scientific figures presents unique challenges that distinguish it from general-domain VQA. Scientific visualizations demand mathematical precision, often requiring exact numerical extraction and calculation. They involve complex compositional reasoning across multiple visual elements, and frequently contain domain-specific conventions, symbols, and representations that require specialized understanding (Ishmam et al., 2024). Furthermore, scientific figures often embed multiple layers of information, including raw data points, derived trends, statistical relationships, and comparative analyses.

Current VQA models, while showing impressive performance on general datasets, often struggle with the precision and reasoning depth required for scientific applications (Kabir et al., 2024). Common failure modes include visual grounding errors, where models misinterpret chart elements or scales; compositional reasoning failures, where multi-step logical processes break down; and consistency issues between visual observations and textual explanations (Tanjim et al., 2025; Thawakar et al., 2025).

This paper presents our approach to the SciVQA Shared Task[1] [2] (Borisova et al., 2025), focusing on QA over scientific visualizations. The task involves answering closed-ended visual (i.e., addressing visual attributes such as colour, shape, size, height, etc.) and non-visual (not addressing figure visual attributes) questions. We leverage the reasoning and visual understanding capabilities of VLMs, and employ task-specific Chain-of-Thought (CoT) (Wei et al., 2023) prompting techniques to retrieve and summarize relevant information from the visual-

[1] https://sdproc.org/2025/scivqa.html
[2] https://huggingface.co/datasets/katebor/SciVQA

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 252–262

izations. Our approach involved testing multiple prompt variants and selecting optimal configurations based on validation performance. To support reproducibility and future research, we make the code publicly available on GitHub[3].

Our main contributions are:

- A systematic ensemble strategy with figure type specific model selection based on comprehensive validation analysis.
- Optimized prompt engineering templates tailored to different question answer pair and figure type combinations.

2 Related Work

Recent advancements in chart-based QA have focused on various approaches to understanding and generating responses about visualizations. ChartLlama (Han et al., 2023) and UniChart (Masry et al., 2023) demonstrate the benefits of chart-specialized language models, showing improved performance in both chart captioning and QA tasks. These works often rely on explicit chart structure parsing as a preprocessing step, achieving strong results on synthetic chart datasets.

Chart-based Reasoning (Carbune et al., 2024) and LlamaV-o1 (Thawakar et al., 2025) propose decomposed reasoning traces and transfer of LLM capabilities to visual settings. Our work builds on these insights by emphasizing structured reasoning and adopt step-level supervision to encourage coherent and faithful intermediate reasoning in visual contexts.

Other relevant works include SPIQA (Pramanick et al., 2025) and MathVista (Lu et al., 2024), which evaluate visual reasoning in scientific domains. MathVista particularly focuses on precise numerical and symbolic interpretation in mathematical visualizations, similar to our emphasis on scientific accuracy.

In the realm of prompt engineering and model alignment, (Zhan et al., 2025) proposed SPRI (Situated-PRInciples), a framework that automatically generates context-specific guiding principles for each input query to improve model alignment. Their approach demonstrates that instance-specific principles can outperform generic ones, which informs our ensemble methodology that combines prompt engineering with multiple VLMs.

Motivated by recent advances in prompt rewriting (Tanjim et al., 2025), we explore instruction tuning and prompt optimization to enhance model adherence to scientific QA formats. Unlike previous work that focuses on architectural innovations requiring additional training, our approach focuses on ensemble strategies and prompt optimization for maximum performance on scientific VQA tasks.

3 Dataset

The SciVQA dataset comprises scientific figures from ACL Anthology and arXiv papers. Each figure is annotated with seven question-answer pairs and associated metadata including captions, figure IDs, figure types (e.g., compound, line graph, bar chart, scatter plot), and QA pair types, with dataset splits and distributions detailed in Section A and Tables 3, 4, and 5.

4 Methodology

Our system integrates three key components:

- systematic prompt optimization for different figure types
- strategic ensemble modeling, and
- post-processing for answer standardization.

We utilized the vLLM (Kwon et al., 2023) engine for maximum compute utilization during inference. A40 instances were sufficient for 7B models, while 8B models required A100 GPUs. CoT inference required approximately twice the computation time due to the two-level reasoning process, but provided significant quality improvements.

4.1 Model Selection

To inform model selection and ensure alignment with the target domain, we referred to the performance of recent models on established multimodal QA benchmarks analogous to SciVQA (Borisova et al., 2025), including ChartQA (Masry et al., 2022), MathVista (Lu et al., 2024), ChartXiv (Wang et al., 2024). VLMs in the 5–8B parameter range demonstrated competitive performance on these leaderboards, achieving results comparable to significantly larger models with 32–72B parameters.

According to the InternVL3 technical report (Zhu et al., 2025), the models InternVL3-8B and Qwen2.5-VL-7B performed well on tasks such as OCR, chart, and document understanding, specifically on datasets like ChartXiv and ChartQA. Additionally, a fine-tuned version of the Qwen2.5-VL-7B Instruct model (Bai et al.,

[3] https://github.com/NagaHarshita/Infyn-SciVQA

2025), as reported in the Bespoke technical report [4], demonstrates competitive performance on ChartXiv, ChartQA, and EvoChart, achieving results comparable to InternVL3-8B. Based on these observations, we chose the following four VLMs, namely, InternVL3-8B, Qwen2.5-VL-7B Instruct, Bespoke MiniChart 7B, and Phi-4 Multimodal Instruct for our task.

InternVL3-8B (Zhu et al., 2025) features an advanced vision encoder architecture tailored for complex visual understanding, unlike Qwen2.5-VL which uses a standard ViT encoder (Zhu et al., 2025). The model supports high-resolution image processing capabilities essential for interpreting detailed charts, and incorporates multi-scale feature extraction to enable both global comprehension and fine-grained numerical reading. It is particularly robust when handling overlapping text and visually dense layouts commonly found in scientific figures. **Qwen2.5-VL-7B Instruct** (Bai et al., 2025) exhibits strong mathematical reasoning, although its performance is limited by the underlying vision encoder. **Bespoke MiniChart 7B**, trained with DPO (Rafailov et al., 2024), benefits from improved chain-of-thought reasoning for chart understanding tasks, but lacks architectural features suited for complex scientific visualizations. Finally, **Phi-4 Multimodal Instruct** (5.6B) (Microsoft et al., 2025) offers general multimodal capabilities but is not specifically optimized for scientific content.

4.2 Prompt Optimization

We crafted task-specific prompts that incorporate captions, figure types, and QA pair types. Prompt variants included explicit CoT cues, multiple correct answer hints, and image-caption-context fusion, with performance differences noted across QA pair types. Additionally, we set two baseline models (both using InternVL3): one using a general prompt without specifying the expected output format, and another with explicit formatting instructions stating that the output should be either a number or a single sentence. Baseline 1 refers to a general prompt without formatting constraints, while Baseline 2 uses explicit answer formatting instructions (exact prompts provided in the Appendix in Table 6). The structured format ensures consistency and enables automated evaluation of both reasoning quality and final answers. All the components described below (e.g., Base Prompt, Compound Images Prompt, Figure Type Prompt, etc.), and in the Tables 7 and 8 are combined into a single, composite prompt to ensure that all possible aspects of the task are considered.

4.2.1 Single Prompt

We developed an initial prompt that includes the figure caption, question-answer pair type classification, and task-specific instructions that elicit reasoning, with exact prompts detailed in Table 7.

Prompt Used: *Base Prompt + Compound Images Prompt + Figure Type Prompt + Question + Binary Prompt + Choice Prompt*

4.2.2 CoT and Rethink

Incorporating Chain-of-Thought (CoT) (Wei et al., 2023) and Rethink mechanisms (Wang et al., 2025) where models regenerate answers with self-correction significantly enhances performance, particularly for math-intensive and ambiguous examples. Prompts are designed to elicit reflective thinking, with final answers distinctly highlighted using structured XML tags ($< reasoning >$ and $< answer >$), as detailed in Table 8.

Step 1 Prompt Used: *Step 1 Base Prompt + Compound Images Prompt*

Step 2 Prompt Used: *Step 2 Base Prompt + Figure Type Prompt + Binary Prompt + Choice Prompt*

4.3 Ensemble

Based on a comprehensive validation analysis (see Section B and Table 9 in Appendix for detailed model performance across figure types), we implemented a figure-type-aware ensemble approach in which each model was assigned to chart types aligned with its demonstrated strengths. Specifically, Qwen2.5-VL was selected for scatter plots, confusion matrices, trees, and graphs, given its relative effectiveness on relational and structural visualizations. Bespoke MiniChart was applied to pie charts, bar charts, architecture diagrams, neural networks, and box plots, leveraging its finetuning for specialized chart comprehension. Meanwhile, Phi-4 was assigned to line charts, tables, histograms, vector plots, and illustrative diagrams, where it showed comparatively better performance. Although this targeted ensemble method yielded competitive results, it was ultimately outperformed by InternVL3, which demonstrated robust and consistent accuracy across all figure types.

[4] https://www.bespokelabs.ai/blog/bespoke-minichart-7b

Model	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R	BS-F1	BS-P	BS-R
InternVL3	**0.730**	**0.744**	**0.732**	**0.729**	**0.743**	**0.731**	**0.981**	**0.983**	**0.980**
Qwen2.5-VL	0.619	0.621	0.641	0.618	0.620	0.641	0.970	0.967	0.973
Bespoke	0.636	0.641	0.647	0.634	0.640	0.645	0.975	0.975	0.976
Phi-4	0.532	0.531	0.596	0.531	0.529	0.595	0.950	0.944	0.956
Ensemble	0.646	0.651	0.660	0.645	0.650	0.658	0.974	0.974	0.976

Table 1: Comparison across ROUGE (R1, RL) and BERTScore (BS) metrics (F1, Precision, Recall) on **validation (without CoT)** after applying post-processing.

Model	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R	BS-F1	BS-P	BS-R
InternVL3	**0.740**	**0.754**	**0.739**	**0.740**	**0.754**	**0.738**	**0.983**	**0.985**	**0.982**
Qwen2.5-VL	0.695	0.699	0.714	0.694	0.698	0.713	0.975	0.973	0.977
Bespoke	0.709	0.716	0.716	0.708	0.715	0.715	0.979	0.979	0.979
Phi-4	0.562	0.566	0.578	0.561	0.565	0.578	0.969	0.966	0.970
Ensemble	0.735	0.744	0.744	0.734	0.743	0.743	0.979	0.978	0.980
InternVL3	0.727	0.739	0.728	0.727	0.738	0.727	0.982	0.983	0.981
Qwen2.5-VL	0.633	0.633	0.658	0.633	0.632	0.658	0.972	0.969	0.975
Bespoke	0.652	0.657	0.664	0.651	0.656	0.663	0.976	0.976	0.977
Phi-4	0.544	0.540	0.600	0.543	0.540	0.600	0.954	0.948	0.960
Ensemble	0.705	0.714	0.710	0.704	0.713	0.709	0.979	0.979	0.979
Baseline 1	0.180	0.164	0.498	0.180	0.163	0.496	0.834	0.812	0.857
Baseline 2	0.700	0.707	0.710	0.699	0.707	0.710	0.977	0.977	0.978

Table 2: Comparison across ROUGE (R1, RL) and BERTScore (BS) metrics (F1, Precision, Recall) on **test with CoT (top) and without CoT (bottom)** after applying post-processing.

4.4 Postprocessing

Our post-processing pipeline involved two key modifications to improve answer quality and evaluation metrics. First, all $|end|$ tags were removed from generated responses to ensure clean output format. Then, for questions where the reasoning process determined insufficient information to give a valid response, outputs were standardized to "It is not possible to answer this question based only on the provided data." regardless of the initial model output. Model outputs, after applying post-processing, were evaluated on both the test set (Table 2) and validation set (Table 1) using BERTScore and ROUGE metrics.

4.5 Results

Our final system, based on an ensemble approach, was submitted to the challenge leaderboard and ranked 5th. Table 10 in Appendix shows the top-7 rankings as on the leaderboard.

Chain-of-Thought Performance: CoT prompting achieved consistent improvements across all VLMs on the test set, with gain in scores for complex multi-step reasoning questions. CoT with rethinking mechanisms demonstrated the most stable performance across different question types.

Model Scale Impact: Larger parameter models consistently outperformed smaller variants on the test set, confirming the correlation between model capacity and reasoning quality in scientific visual question answering.

Comparative Performance: InternVL3 achieved the highest individual model performance, outperforming other individual models by at least +0.30 ROUGE-1 F1 score on the test split.

5 Conclusion and Future Work

This work demonstrates that advanced vision encoding architecture, combined with systematic prompt engineering, provides a highly effective approach to scientific visual question answering. High-quality visual understanding is critical for strong performance. Our approach establishes a strong baseline for the SciVQA dataset with a ROUGE-1 and ROUGE-L F1 score of 0.740.

5.1 Future Directions

Advanced Reasoning Techniques: Exploring advanced prompting techniques such as Tree-of-Thought (Yao et al., 2023), leveraging Mixture-of-Experts (MoE) (Shazeer et al., 2017) models tailored to different question-answer pairs, and incorporating expert-critic ensembles for answer reranking holds significant potential for enhancing overall performance.

Scalability and Model Improvements: Running inference using larger models or fine-tuning the current models on this dataset, enabled by increased computational resources, is expected to yield substantial performance gains.

Quality of the Dataset: Addressing the identified dataset quality issues from Appendix C by standardizing data formatting (e.g., multi-correct answers, numerical representations, and answer formatting) and conducting a comprehensive review of the gold standard annotations, to ensure more accurate evaluations. These improvements will enhance dataset accuracy, provide fairer evaluations, and help ensure more reliable model performance comparisons in subsequent iterations of this task.

Limitations

Due to computational resource constraints, experiments were primarily limited to 7B model variants. Larger model variants and fine-tuning using the entire train split would most definitely yield superior results.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Ekaterina Borisova, Nikolas Rauscher, and Georg Rehm. 2025. SciVQA 2025: Overview of the first scientific visual question answering shared task. In *Proceedings of the 5th Workshop on Scholarly Document Processing (SDP)*, Vienna, Austria.

Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. 2024. Chart-based reasoning: Transferring capabilities from llms to vlms.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Md. Farhan Ishmam, Md. Sakib Hossain Shovon, M.F. Mridha, and Nilanjan Dey. 2024. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, 106:102270.

Raihan Kabir, Naznin Haque, Md Saiful Islam, and Marium-E-Jannat. 2024. A comprehensive survey on visual question answering datasets and algorithms.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2025. Spiqa: A dataset for multimodal question answering on scientific papers.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer.

Md Mehrab Tanjim, Ryan A. Rossi, Mike Rimer, Xiang Chen, Sungchul Kim, Vaishnavi Muppala, Tong Yu, Zhengmian Hu, Ritwik Sinha, Wei Zhang, Iftikhar Ahamath Burhanuddin, and Franck Dernoncourt. 2025. Exploring rewriting approaches for different conversational tasks.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Hongli Zhan, Muneeza Azmat, Raya Horesh, Junyi Jessy Li, and Mikhail Yurochkin. 2025. Spri: Aligning large language models with context-situated principles.

Qihao Zhu Runxin Xu Junxiao Song Mingchuan Zhang Y.K. Li Y. Wu Daya Guo Zhihong Shao, Peiyi Wang. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.

Appendix

A Dataset Distribution

The dataset is biased toward line charts (66%), requiring strong numerical reading capabilities. Additionally, a high percentage of non-visual questions (around 60%) highlights the need for reasoning that goes beyond visual features.

Dataset Split	Samples
Train	15,120
Validation	1,680
Test	4,200

Table 3: Dataset split with number of samples.

QA Type	Answer Set		Samples
Closed -ended	Infinite	Visual	1,079
		Non-visual	2,172
	Finite	Binary & Visual	1,124
		Binary & Non-visual	3,219
		Non-binary & Visual	1,751
		Non-binary & Non-visual	3,615
Unanswerable			2,160

Table 4: QA Pair Type Categorization in the Train split.

Figure Type	Samples
Line Chart	10,007
Tree	924
Scatter Plot	735
Graph	553
Bar Chart	525
Architecture Diagram	504
Pie Chart	497
Neural Networks	462
Confusion Matrix	427
Box Plot	133
Histogram	77
Other	77

Table 5: Figure Type Distribution in the Train split.

B Error Analysis

We conducted error analysis manually on the validation dataset because gold answers are available for it. We only selected incorrect predictions and categorized them into three primary types:

Visual Misinterpretations: Issues with feature extraction from images, including:

- Comparing sub-figures on different scales
- Misunderstanding axis starting points
- Difficulties with overlapping text
- Challenges with low-resolution images

Numerical Misalignments: Precision issues in numerical extraction and calculation, often stemming from visual ambiguity in chart elements.

Flawed Reasoning: Instances where logical progression was incorrect despite proper visual observation, or cases where correct reasoning led to incorrect answers due to misalignment with reference answer formats.

Notably, some failures occurred in compound charts and arose from misinterpreted visual cues or insufficient numerical precision. The use of more powerful vision encoders could address many of these issues and further improve performance.

C Dataset Quality Issues

During our validation analysis, we identified several systematic inconsistencies in the gold standard annotations that may impact evaluation reliability:

Format Inconsistencies:

- Multi-correct answers appear in inconsistent formats: ["A", "B"] instead of the expected A,B format
- Numerical representations vary between word form ("three") and digit form ("3") within similar contexts
- Answer formatting lacks standardization across question types

Annotation Errors: We identified potential annotation errors through manual inspection. For example: instance_id 09dab5a715034cebb2a62f0f1c2a75c9 gold answer is "52,3%" but visual inspection suggests that the correct answer should be "3%" or "3", which our models corrects predict.

These inconsistencies suggest that reported performance metrics may underestimate true model capabilities, as models may be penalized for providing correct answers that don't match inconsistent

gold standards. A comprehensive gold standard review and standardization would benefit future iterations of this shared task.

D Alternative Approaches Evaluated

Majority Voting: Simple majority voting across models showed minimal improvement over InternVL3 alone, confirming the quality-over-quantity principle.

Fine-tuning: Limited computational resources prevented extensive fine-tuning, but initial experiments suggested that InternVL3's pre-trained capabilities were already well-suited for the task. To enhance performance, we performed supervised fine-tuning (SFT) using LoRA (Hu et al., 2021), targeting all linear layers and the vision encoder, on a subset of 5,000 training samples for the Bespoke MiniChart model, which yielded a slight performance improvement. We are currently extending this approach by applying Group Relative Policy Optimization (GRPO)-based (Zhihong Shao, 2024) fine-tuning in a similar fashion.

E Tables

Baseline	Prompt Content
1	You are a helpful assistant. Give the concise answer for the context given below. The caption of the figure is mentioned as, [caption]. The question for the figure is, [question]
2	Answer the question with only the raw numerical value or single word/phrase, omitting all units, context words, and explanatory text. The caption of the figure is mentioned as, [caption]. The question for the figure is, [question]

Table 6: Baseline Prompts

Prompt Type	Prompt Content
Base	Answer the question with only the raw numerical value or single word/phrase, omitting all units, context words, and explanatory text. The caption of the figure is mentioned as, [caption].
Compound Images	This is a compound figure containing multiple subfigures. Navigate to [fig_numb] graph in the compound figure to answer the question.
Figure Type	**Line Chart:** Focus on the following aspects of the line chart: • Colors of different lines and their meanings • X and Y axis labels and their units • Scale and range of values • Trends and patterns in the lines **Bar Chart:** Focus on the following aspects of the bar chart: • Colors of different bars and their meanings • X and Y axis labels and their units • Scale and range of values • Height and position of bars **Box Plot:** Focus on the following aspects of the box plot: • Median line position • Box boundaries (Q1 and Q3) • Whisker extent • Outliers if present **Confusion Matrix:** Focus on the following aspects of the confusion matrix: • Row and column labels • Numerical values in each cell • Color intensity if present • Overall distribution of values **Pie Chart:** Focus on the following aspects of the pie chart: • Segments and their labels • Percentage or proportion values • Colors of different segments • Size of each segment relative to others **Others:** Focus on the following aspects of the figure: • Colors and the labels present in the figure • Any other relevant information present in the figure
Binary	This is a binary question. Answer with 'Yes' or 'No' based on [visual/textual] evidence. Respond affirmatively only if supported.
Choice	Return only the corresponding letter(s) of the correct answer(s). Only output the letter(s) corresponding to the correct choice. [answer_choices]

Table 7: Instruction Prompts for Single Prompt

Prompt Type	Prompt Content
Step 1 Base Prompt	STEP 1: INITIAL ANALYSIS Given the figure, caption, and question, analyze and answer step by step. Regularly perform self-questioning, self-verification, self-correction to check your ongoing reasoning, using connectives such as "Wait a moment", "Wait, does it seem right?" etc. Caption: [caption] Question: [question] Analyse the key visual elements (lines, shapes, colors) that address the question and analyze the relationships between elements. Then, extract the specific numerical/positional information from the figure and caption to answer the question.
Compound Images Prompt	Same as single prompt
Step 2 Base Prompt	STEP 2: COT INFERENCE Answer the question with only the raw numerical value or single word/phrase, omitting all units, context words, and explanatory text. Approximations in the scale are allowed.
Figure Type Prompt	Same as single prompt
Binary Prompt	Same as single prompt
Choice Prompt	*(For non-binary finite answer sets)*: Based on the reasoning above, match it to one or more of the provided answer options: [answer_choices] Return only the corresponding letter(s) of the correct answer(s). Do not explain your choice, do not rephrase the answer, and do not repeat the option text. Only output the letter(s) corresponding to the correct choice. If multiple letters are correct, separate them by commas without spaces (for example: B,C). If all options are correct, return A,B,C,D. Do not add anything else.

Table 8: Instruction Prompts for CoT

Chart Type	Bespoke		InternVL3		Qwen2.5-VL		Phi-4	
	Acc. (mean)	Std. Dev.	Acc. (mean)	Std. Dev.	Acc. (mean)	Std. Dev.	Acc. (mean)	Std. Dev.
line_chart	54.23	49.82	63.97	48.01	50.68	50.00	42.40	49.42
line_chart,table	42.86	49.49	85.71	34.99	42.86	49.49	57.14	49.49
tree	56.19	49.62	61.90	48.56	53.33	49.89	44.76	49.72
scatter_plot	55.71	49.67	70.00	45.83	57.14	49.49	40.00	48.99
pie_chart	67.35	46.89	73.47	44.15	67.35	46.89	44.90	49.74
architecture_diagram	67.86	46.70	76.79	42.22	55.36	49.71	28.57	45.18
box_plot	50.00	50.00	50.00	50.00	50.00	50.00	35.71	47.92
neural_networks	62.50	48.41	71.43	45.18	58.93	49.20	32.14	46.70
confusion_matrix	54.76	49.77	64.29	47.92	57.14	49.49	40.48	49.08
graph	57.14	49.97	60.71	48.84	46.43	49.87	41.07	49.20
bar_chart	53.06	49.91	69.39	46.09	51.02	49.99	40.82	49.15
histogram	35.71	47.92	71.43	45.18	35.71	47.92	50.00	50.00
venn_diagram	57.14	49.49	85.71	34.99	57.14	49.49	57.14	49.49
vector_plot	71.43	45.18	100.00	0.00	85.71	34.99	85.71	34.99
other	42.86	49.49	57.14	49.49	42.86	49.49	42.86	49.49
line_chart,bar_chart	28.57	45.18	71.43	45.18	14.29	34.99	28.57	45.18
flow_chart	85.71	34.99	85.71	34.99	71.43	45.18	42.86	49.49
tree,graph	28.57	45.18	42.86	49.49	42.86	49.49	14.29	34.99
illustrative_diagram	28.57	45.18	71.43	45.18	28.57	45.18	57.14	49.49
line_chart,scatter_plot	71.43	45.18	71.43	45.18	42.86	49.49	42.86	49.49
heat_map	57.14	49.49	71.43	45.18	28.57	45.18	57.14	49.49

Table 9: Scores for exact match across models for various chart types. Accuracy and standard deviation (both in %) are shown.

Rank	Team	R1-F1	R1-P	R1-R	RL-F1	RL-P	RL-R	BS-F1	BS-P	BS-R
1	ExpertNeurons	0.805	0.809	0.811	0.804	0.808	0.810	0.985	0.985	0.985
2	THAii_LAB	0.790	0.796	0.795	0.789	0.795	0.794	0.984	0.984	0.984
3	Coling_UniA	0.786	0.798	0.786	0.786	0.796	0.785	0.982	0.983	0.981
4	florian	0.763	0.766	0.770	0.762	0.765	0.769	0.983	0.983	0.984
5	Infyn	0.735	0.744	0.744	0.734	0.743	0.743	0.979	0.978	0.980
6	Soham Chitnis	0.706	0.719	0.705	0.705	0.719	0.704	0.980	0.982	0.979
7	psr123	0.607	0.609	0.617	0.606	0.608	0.616	0.959	0.959	0.959

Table 10: Top-7 leaderboard rankings