8. Appendix

This appendix presents additional experimental results, visualizations, and detailed descriptions of data construction, providing a more comprehensive view of the models' performance and analyses.

8.1. Experimental Setup

We conduct all experiments on a single NVIDIA H20 GPU. The sampling temperature of each model is set to 0 to ensure deterministic and reproducible outputs. For the maximum token generation limits, we adopt each model's default settings.

For the evaluation of model responses, we use GPT-5 with the temperature fixed at 0 to assess the correctness of model-generated answers. Specifically, for each sample, GPT-5 is provided with the question, the reference answer, and the model's generated response, and is instructed to determine whether the output is correct according to our evaluation criteria, following the prompt template shown in Table 5.

8.2. Image Variants Generation

This subsection details the generation of several image variants designed to evaluate the model's robustness under various visual perturbations. Each variant introduces a distinct form of distortion or noise, allowing for a comprehensive assessment of the model's ability to accurately extract relevant information despite changes in image quality or the introduction of irrelevant elements.

- *Blurred Variant:* This variant is generated by applying a Gaussian blur to the image using OpenCV's `cv2.GaussianBlur`, with a default kernel size of 5 and a standard deviation of 0. The blur reduces image sharpness while preserving the overall structure of the chart, aiming to evaluate the model's robustness to visual degradation and its ability to extract information under softened or less distinct visual conditions.
- *Noisy Variant:* This variant is generated by adding random integer noise to the image pixels, with values ranging in $[-50, 50]$, using NumPy. The added noise simulates real-world visual disturbances, aiming to evaluate the model's robustness in extracting key information under realistic noisy conditions.
- *Watermarked Variant:* This variant is created by overlaying randomly generated watermark text onto the image using PIL's `ImageDraw` library. By default, two text strings—each 10 to 20 characters long—are rendered in dark gray and placed at random positions on the image. These watermarks act as irrelevant visual distractions, allowing us to test the model's ability to focus on the main content despite the presence of such noise.
- *Annotation-Removed Variant:* In this variant, numeric annotations such as data labels and value markers are removed from the chart while retaining the original structure, including axes, bars or lines, and overall layout. By omitting explicit numerical information, this version tests the model's ability to interpret the chart based purely on visual cues such as shape, relative position, and trend patterns.

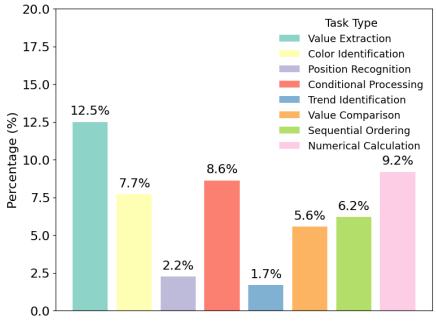


Figure 4. Average proportion of each sub-task type across the 200 main tasks.

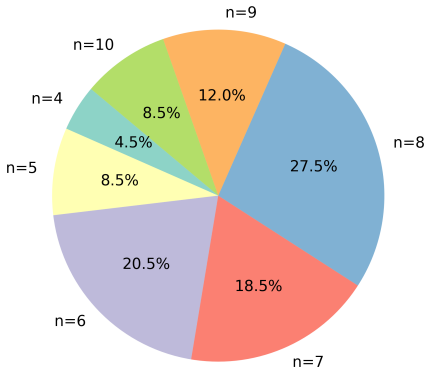


Figure 5. Distribution of the 200 main tasks by their number of sub-task QAs (n = sub-task count).

8.3. Dataset Statistics

Our dataset covers 20 chart types, including 11 single-series charts: single-series line chart (SLC), single-series area chart (SAC), single-series bar chart (SBC), single-series horizontal bar chart (SHBC), pie chart (PIC), rose chart (ROC), ring chart (RIC), tree map (TRM), single-series radar chart (SRC), funnel chart (FUC), and scatter chart (SCC); and 9 multi-series charts: multi-series line chart (MLC), stacked area chart (STAC), multi-series bar chart (MBC), multi-series horizontal bar chart (MHBC), stacked bar chart (STBC), stacked horizontal bar chart (STHBC), multi-series radar chart (MRC), box plot (BOP), and candlestick chart (CAC).

Figure 4 presents the distribution of sub-task types across the entire dataset. *Value Extraction* is the most frequent sub-task type (12.5%), followed by *Numerical Calculation*

Table 5. Prompt template for evaluating the correctness of model-generated answers.

You are a QA consistency evaluator. Determine whether the given Response matches the provided Answer based on the Question.

Task:

1. Judge whether the Response is consistent with the Answer.

2. Consistency criteria:

• If the Answer is numeric, consider the Response consistent if it is within ±5% of the Answer; differences in units should be ignored.

• If the Answer is textual, an entity, or non-numeric, the Response must fully preserve the meaning and information of the Answer.

• Minor wording differences that do not affect meaning are allowed.

3. Only output a single word:

• YES for consistent

• NO for inconsistent

4. Do not provide any explanations.

Input:

Question: {question}

Answer: {answer}

Response: {response}

Your output:

(9.2%) and *Conditional Processing* (8.6%). In addition, *Trend Identification* appears least frequently (1.7%), as it is only included in line charts.

Figure 5 shows the distribution of sub-task counts per main question. Most main tasks consist of 6–8 sub-questions: 27.5% contain 8 steps, 20.5% contain 6 steps, and 18.5% contain 7 steps. This controlled step range maintains a balance between reasoning complexity and interpretability.

8.4. Illustrative Example of the Dataset

Figure 6 presents a representative example from our dataset. It showcases the original input chart alongside four variants (Blurred Image, Noisy Image, Image with Watermark, and Image without Labels) to highlight the dataset's focus on robustness.

The original chart displays **"Sports Participation by Type"** as a percentage. The figure also details the associated question–answer pairs, demonstrating the range of tasks included in our benchmark, from basic data retrieval to complex, multi-step numerical reasoning.

8.4.1. Complex Reasoning QA

The benchmark includes complex questions that require decomposing the task into multiple atomic steps.

> **Q:** "What is the percentage difference between the sport with the highest participation and the combined percentage of the two sports with the lowest participation?" Answer: 30.

8.4.2. Decomposed Atomic QA Pairs

This complex question is typically solved by sequentially executing several simpler, atomic tasks, which are also included in the dataset and serve as an explanation for the complex task. The following examples illustrate the primary atomic task categories:

- *Value Extraction* (Q1, Q5): Retrieving the specific data points from the chart.

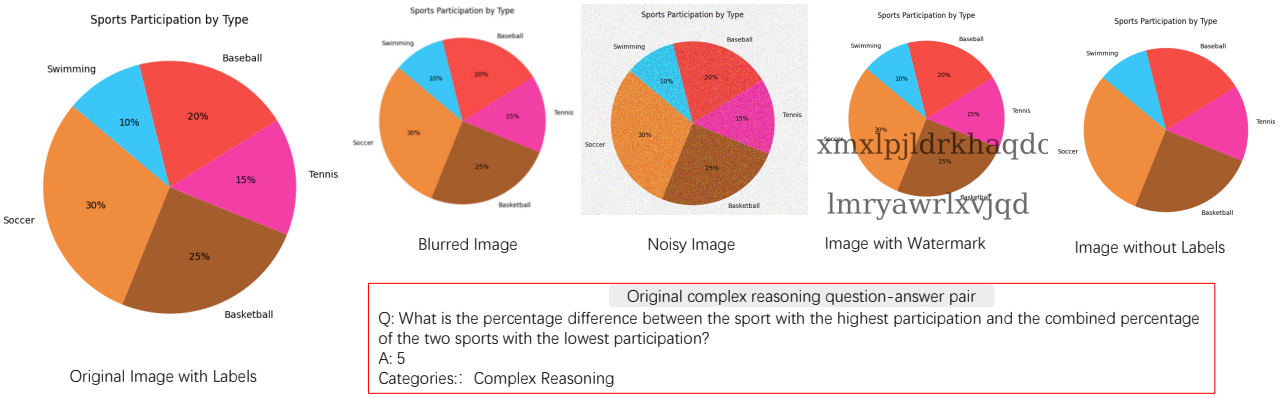
  > Q1: "What are the participation percentages for each sport?" Answer: ["Soccer": 30, "Basketball": 25, "Tennis": 15, "Baseball": 20, "Swimming": 10]. Q5: "What is the participation percentage of Score?" Answer: 30.

- *Value Comparison* (Q2, Q4): Identifying the maximum or minimum values and their corresponding categories.

  > Q2: "What are the two sports with the lowest participation percentages?" Answer: Tennis, Swimming. Q4: "Which sport has the highest participation percentage?" Answer: Soccer.

- *Numerical Calculation* (Q3, Q6): Performing arithmetic operations on the extracted values.

  > Q3: "What is the combined participation percentage of Tennis and Swimming?" Answer: 25.
  > Q6: "What is the difference between the participation percentage of Soccer and the com-



Original complex reasoning question-answer pair

Q: What is the percentage difference between the sport with the highest participation and the combined percentage of the two sports with the lowest participation?
A: 5
Categories:: Complex Reasoning

Decomposed atomic question-answer pair

Q1: What are the participation percentages for each sport?
A1: {"Soccer": 30, "Basketball": 25, "Tennis": 15, "Baseball": 20, "Swimming": 10}
Categories:: Value Extraction
Pre_questions: [-1]

Q2: What are the two sports with the lowest participation percentages?
A2: ["Tennis", "Swimming"]
Categories:: Value Comparison
Pre_questions: [Q1]

Q3: What is the combined participation percentage of Tennis and Swimming?
A3: 25
Categories:: Numerical Calculation
Pre_questions: [Q1, Q2]

Q4: Which sport has the highest participation percentage?
A4: Soccer
Categories:: Value Comparison
Pre_questions: [Q1]

Q5: What is the participation percentage of Soccer?
A5: 30
Categories:: Value Extraction
Pre_questions: [Q1, Q4]

Q6: What is the difference between the participation percentage of Soccer and the combined percentage of Tennis and Swimming?
A6: 5
Categories:: Numerical Calculation
Pre_questions: [Q3, Q5]

Figure 6. The benchmark example in our dataset

bined percentage of Tennis and Swimming?" Answer: 5.

This structured approach—with clear *Categories* (e.g., Value Extraction, Numerical Calculation) and explicit *Pre-questions* dependencies—ensures that models are benchmarked not only on the final answer accuracy but also on their ability to perform the necessary intermediate steps.

8.5. Analysis Across Chart Types

To further understand how chart structure affects model performance, we analyze accuracy across different chart types by treating each question as an independent evaluation instance. Tables 6 and 7 report model accuracy on single-series and multi-series chart categories, respectively.

Overall, models perform consistently better on single-series charts than on multi-series charts, indicating that reasoning complexity increases significantly when multiple data series are involved. Among single-series charts, bar-based charts (SBC and SHBC) and scatter chart (SCC) achieve the highest accuracies across most models. For example, Gemini-2.0-flash reaches 99.49% accuracy on SBC and 96.59% on SCC. These chart types provide clear visual separation between data elements, making value extraction and comparison relatively straightforward. In contrast, models struggle more with radar charts (SRC) and rose charts (ROC). These chart types require interpreting values along radial axes, which may complicate spatial reasoning and value estimation. Even strong models such as Gemini-2.0-flash show noticeable performance drops on SRC compared to standard bar or line charts.

The gap becomes more pronounced in multi-series charts. Performance remains relatively strong on multi-series bar charts (MBC) and multi-series horizontal bar charts (MHBC), which preserve clear categorical separation between series. However, models show substantial degradation on more complex chart types such as multi-series radar charts (MRC), box plots (BOP), and candlestick charts (CAC).

Another notable observation is that stacked charts (STAC, STBC, STHBC) introduce additional reasoning difficulty because values must be interpreted relative to cumulative totals or stacked segments. This often requires models to first identify individual components before aggregating them, increasing the likelihood of reasoning errors.

Overall, these results suggest that current multimodal models are more reliable on chart types with explicit, visually separated data elements, while they struggle with charts that encode information through radial layouts, distributions, or cumulative structures. Improving robustness on these complex chart types remains an important direction for future multimodal reasoning research.

Table 6. Model accuracy (%) across single-series chart types.

Model	SLC	SAC	SBC	SHBC	PIC	ROC	RIC	TRM	SRC	FUC	SCC
General-Purpose MLLMs											
Gemini-2.0-flash	91.81	92.96	99.49	98.78	82.96	78.56	84.64	82.91	74.35	86.77	96.59
InternVL2.5-8B	51.00	41.89	79.42	75.71	70.01	48.62	58.34	69.33	50.51	75.23	61.03
MiniCPM-o-2_6-8B	51.19	54.62	75.52	72.28	65.77	42.23	63.66	64.14	36.04	77.98	57.23
Qwen2.5-VL-7B	75.56	86.52	92.02	87.94	79.31	67.88	79.69	73.16	57.28	82.26	84.90
Janus-Pro-7B	42.79	35.43	56.28	62.54	42.83	21.36	44.69	32.08	12.50	57.67	26.96
Deepseek-VL-7B	27.00	28.84	29.94	49.21	31.05	17.20	20.37	22.61	14.84	35.05	14.94
Phi-4-Multimodal-5.6B	61.57	68.14	77.60	73.76	71.61	44.93	71.55	67.58	53.58	75.88	67.28
Qwen2.5-VL-3B	67.60	70.01	78.75	80.22	67.53	55.87	66.98	63.72	41.80	78.47	70.82
InternVL2.5-2B	37.14	33.09	64.78	56.72	43.35	28.87	38.71	57.46	23.56	57.77	48.30
Chart-Specific MLLMs											
ChartMoE-8B	51.87	46.92	76.68	69.56	53.19	26.36	57.23	53.69	25.29	68.59	62.41
TinyChart-3B	27.41	39.74	54.53	39.77	29.01	17.60	21.73	25.07	16.90	48.23	35.19
ChartGemma-2.4B	40.71	44.16	50.07	53.07	30.43	14.87	37.72	37.92	13.56	42.01	36.00

Table 7. Model accuracy (%) across multi-series chart types.

Model	MLC	STAC	MBC	MHBC	STBC	STHBC	MRC	BOP	CAC
General-Purpose MLLMs									
Gemini-2.0-flash	85.59	77.12	93.81	87.69	84.05	78.97	44.91	71.94	66.36
InternVL2.5-8B	49.43	42.59	77.12	57.99	57.06	54.40	29.97	24.47	20.29
MiniCPM-o-2_6-8B	46.92	39.21	53.02	33.34	46.48	45.44	26.44	26.55	26.89
Qwen2.5-VL-7B	69.34	48.62	90.46	85.30	65.80	70.30	38.40	51.85	53.32
Janus-Pro-7B	23.12	26.99	24.99	13.52	18.71	19.03	15.86	19.16	15.32
Deepseek-VL-7B	15.56	23.85	20.49	10.15	13.53	15.91	9.67	11.04	8.00
Phi-4-Multimodal-5.6B	52.54	55.39	70.37	46.46	61.61	48.98	49.09	47.73	24.75
Qwen2.5-VL-3B	57.00	41.74	76.22	67.68	62.54	63.57	29.68	43.82	38.25
InternVL2.5-2B	27.76	30.57	43.19	32.04	33.96	35.74	20.89	11.75	13.36
Chart-Specific MLLMs									
ChartMoE-8B	41.92	39.38	49.16	42.64	29.10	41.35	15.33	24.69	19.50
TinyChart-3B	25.73	24.04	39.92	31.81	29.60	26.24	16.87	10.04	6.32
ChartGemma-2.4B	16.71	15.83	24.40	18.96	22.27	20.01	13.75	10.00	13.21

8.6. Supplementary Results for Section 5.2: Model Performance Across Visual Variants

As a complement to the main results presented in Section 5.2, we further evaluate four recently released MLLMs—Gemini-2.5-flash, GPT-5-mini, Qwen3-VL-8B, and InternVL3.5-8B—across five visual variants of each chart type. The complete performance scores are reported in Table 8, and the corresponding robustness metrics are summarized in Table 9.

Overall, all four models maintain strong performance on both clean and perturbed charts, demonstrating stable step-wise reasoning (with moderate ISA–CSA gaps) and consistent final-answer reliability (reflected by stable FAA–CFA gaps). Among them, GPT-5-mini achieves the best overall performance and robustness across nearly all metrics, underscoring its superior multi-step reasoning capabilities and resilience to visual perturbations.

8.7. Supplementary Results for Section 5.3: Reasoning Consistency Distribution of All Models

As an extension of the analysis presented in Section 5.3, where only three representative models were shown, we provide in Figure 7 the full distributions of step-wise accuracy (ISA) and chained accuracy (CSA) for all evaluated

Table 8. Model performance across visual variants.

Model	Original chart				Noisy version				Blurred version			
	SAp	PA	FQA	ARA	SAp	PA	FQA	ARA	SAp	PA	FQA	ARA
gemini-2.5-flash	92.76	81.79	90.50	75.00	91.98	79.30	89.50	68.00	91.28	82.52	88.00	74.50
gpt-5-mini	96.04	88.50	90.50	81.00	94.21	84.61	91.00	76.50	95.37	85.74	91.50	76.50
Qwen3-VL-8B	87.62	73.32	70.00	55.50	87.66	73.42	71.50	55.50	87.38	72.89	72.50	58.00
InternVL3.5-8B	74.66	45.93	56.50	22.50	74.63	44.05	57.50	21.50	75.49	46.84	62.00	29.00
Model	**Watermarked version**				**Anno_removed version**				**Average**			
	SAp	PA	FQA	ARA	SAp	PA	FQA	ARA	SAp	PA	FQA	ARA
gemini-2.5-flash	89.32	75.07	85.50	64.50	54.30	28.06	44.50	20.00	83.93	69.35	79.60	60.40
gpt-5-mini	94.36	83.70	90.50	76.00	89.59	77.97	86.00	70.00	93.91	84.10	89.90	76.00
Qwen3-VL-8B	86.61	69.64	74.00	56.00	57.89	35.10	46.00	21.00	81.43	64.87	66.80	49.20
InternVL3.5-8B	72.39	44.22	55.00	22.00	50.35	24.38	38.00	7.00	69.50	41.08	53.80	20.40

Table 9. Model robustness under different metrics.

Model	AR_{ISA}	AR_{CSA}	AR_{FAA}	AR_{CFA}
gemini-2.5-flash	0.1190	0.1902	0.1506	0.2433
gpt-5-mini	0.0277	0.0621	0.0083	0.0772
Qwen3-VL-8B	0.0883	0.1440	0.0571	0.1419
InternVL3.5-8B	0.0863	0.1319	0.0597	0.1167

models. These additional results exhibit the same overall pattern observed in the main text: models whose ISA values are highly concentrated at the upper end tend to preserve accuracy across the full reasoning chain, achieving stronger CSA. In contrast, models with more dispersed ISA distributions undergo larger performance degradation when transitioning from step-wise to complete-chain reasoning.

Consistent with the earlier findings, these extended results further confirm that early-step errors—especially samples with extremely low ISA—are the primary source of accuracy loss in multi-step reasoning. The pronounced CSA drop associated with low-ISA samples reinforces that error propagation remains the dominant bottleneck limiting model robustness across all evaluated models.

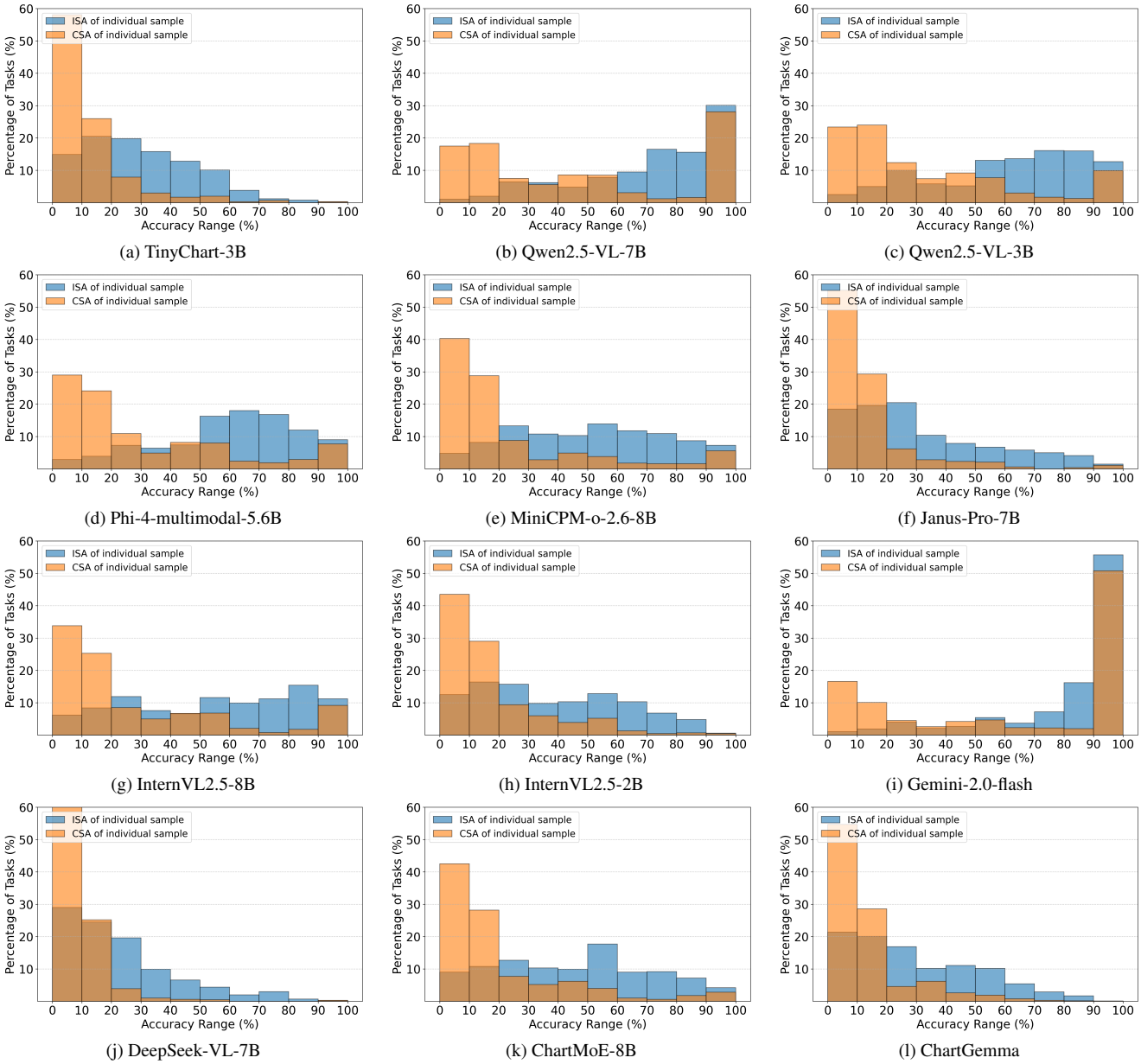
8.8. Supplementary Results for Section 5.5: Comprehensive Task Error Dependency and Recovery Analysis

To provide a more complete view of the immediate task error dependencies discussed in Section 5.5, we extend the analysis beyond the top two preceding tasks and present the full error co-occurrence and recovery structures for Gemini-2.0-flash and Qwen2.5-VL-7B (Figure 8). While the main text reports only the most influential preceding tasks—those with the highest JointErr and RecR values—the full matrices included here reveal the complete dependency patterns across all task pairs for these two models.

The error co-occurrence matrices offer a holistic quantification of how frequently errors in each preceding task coincide with failures in each downstream task. The extended results reinforce the primary trend observed in the main text: Value Extraction (VE) consistently emerges as the dominant source of downstream errors, showing the strongest co-occurrence with CP, VC, SO, and NC. Gemini-2.0-flash and Qwen2.5-VL-7B exhibit highly similar propagation patterns, where VE errors account for the majority of immediate downstream failures, confirming its central role in triggering error cascades.

The recovery ratio matrices complement this view by quantifying the extent to which correcting each preceding task would resolve downstream errors. Consistent with the RecR trends reported in Table 4, VE demonstrates the highest recovery potential in both models, with upstream correction projected to recover 70–95% of downstream errors in tasks such as NC and VC. Other preceding tasks, such as CP and NC, show moderate recovery effects on specific downstream tasks, whereas tasks like PR and CI continue to exhibit minimal JointErr and RecR values, indicating weak immediate dependencies.

Together, these expanded analyses validate and generalize the conclusions from Section 5.5: VE is the principal immediate bottleneck in multi-step chart reasoning, simultaneously acting as the most frequent source of downstream errors and the task with the highest recovery leverage. Addressing VE errors is therefore expected to yield the greatest improvements in overall reasoning robustness for both evaluated models.



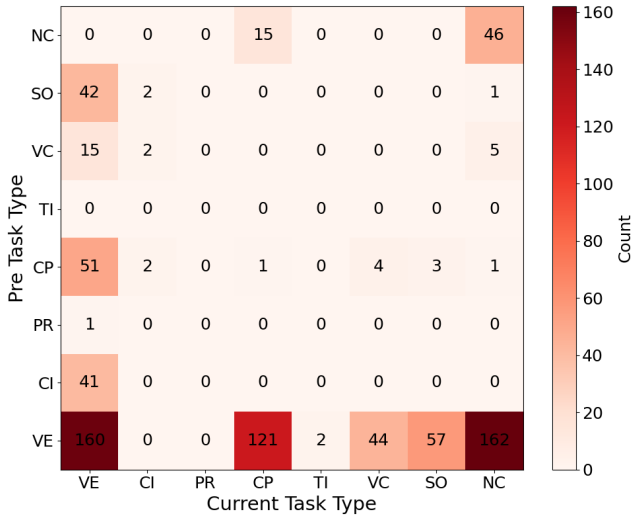
(a) TinyChart-3B (b) Qwen2.5-VL-7B (c) Qwen2.5-VL-3B

(d) Phi-4-multimodal-5.6B (e) MiniCPM-o-2.6-8B (f) Janus-Pro-7B

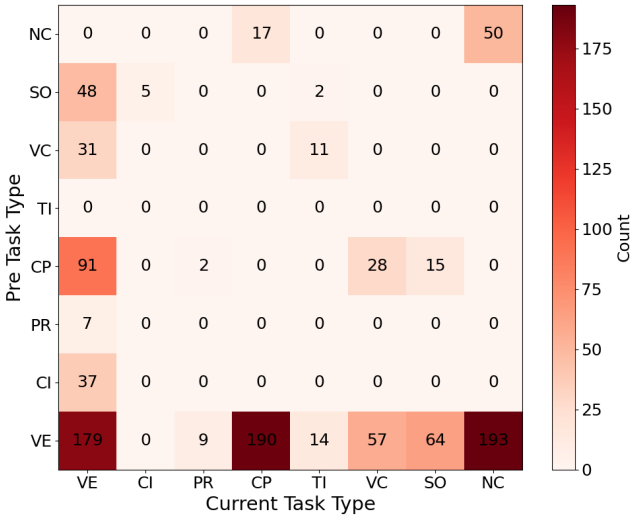
(g) InternVL2.5-8B (h) InternVL2.5-2B (i) Gemini-2.0-flash

(j) DeepSeek-VL-7B (k) ChartMoE-8B (l) ChartGemma

Figure 7. Distribution of step-wise (ISA) and chained (CSA) accuracy across individual QA samples for all evaluated models.

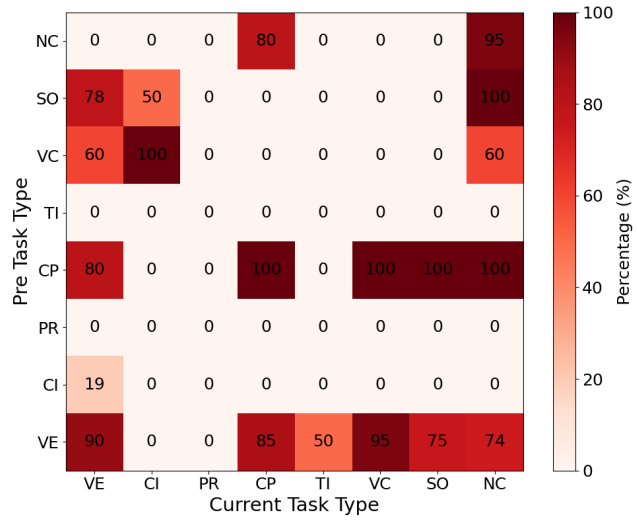


(a) Gemini-2.0-flash

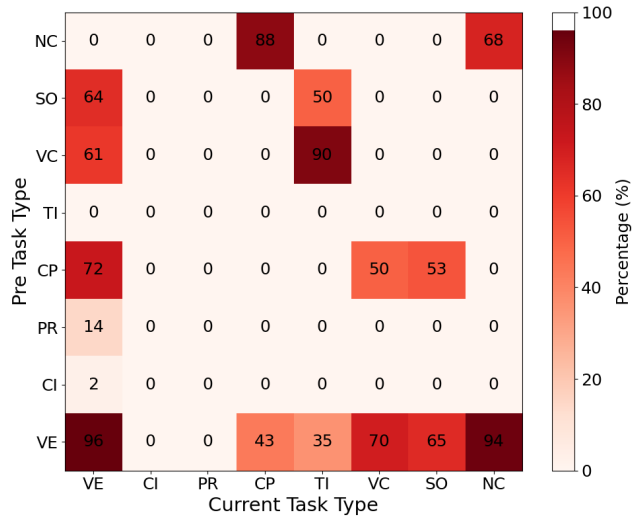


(b) Qwen2.5-VL-7B

(i) Error co-occurrence matrix: each cell shows the number of cases in which a preceding task type and a current task type are simultaneously answered incorrectly under the model's own predictions.



(c) Gemini-2.0-flash



(d) Qwen2.5-VL-7B

(ii) Recovery ratio matrix: each cell shows the proportion of previously incorrect current-task answers that become correct when the corresponding preceding-task answers are replaced with ground-truth values.

Figure 8. Task dependency analysis across models using two complementary heatmaps. (i) visualizes error co-occurrence between preceding and current task types, revealing where reasoning failures tend to originate. (ii) reports the recovery ratios obtained by correcting preceding-task answers, quantifying how much each task type contributes to downstream performance improvement. Together, these visualizations expose model-specific dependency patterns and highlight which task types exert the strongest influence on multi-step reasoning reliability.