INTERCHART: Benchmarking Visual Reasoning Across Decomposed and Distributed Chart Information

Anonymous ACL submission

Abstract

We introduce INTERCHART, a diagnostic benchmark that evaluates how well vision-language models (VLMs) reason across multiple related charts, a task central to real-world applications such as scientific reporting, financial analysis, and public policy dashboards. Unlike prior benchmarks focusing on isolated, visually uniform charts, INTERCHART challenges models with diverse question types ranging from entity inference and trend correlation to numerical estimation and abstract multi-step reasoning grounded in 2–3 thematically or structurally related charts. We organize the benchmark into three tiers of increasing difficulty: (1) factual reasoning over individual charts, (2) integrative analysis across synthetically aligned chart sets, and (3) semantic inference over visually complex, real-world chart pairs. Our evaluation of state-of-the-art open- and closed-source VLMs reveals consistent and steep accuracy declines as chart complexity increases. We find that models perform better when we decompose multi-entity charts into simpler visual units, underscoring their struggles with cross-chart integration. By exposing these systematic limitations, INTERCHART provides a rigorous framework for advancing multimodal reasoning in complex, multi-visual environments.

1 Introduction

Real-world settings such as scientific publications, business reports, and journalism dashboards rarely communicate data through a single chart. Instead, insight often emerges from comparing or synthesizing information across multiple visualizations. These charts may differ in type, styling, or even semantic framing, yet they jointly convey trends, correlations, and complex relationships. For humans, reasoning across such heterogeneous visual inputs is intuitive. However, vision-language models (VLMs) remain a significant challenge. While recent VLMs have shown strong performance on single-chart visual question answering (VQA) tasks (Masry et al., 2022; Methani et al., 2020), they perform inconsistently to aggregate information across multiple charts. Existing benchmarks (Li and Tajbakhsh, 2023; Kantharaj et al., 2022) have begun exploring multi-chart reasoning, but they often rely on simplified scenarios, synthetic data, static chart styles, or limited visual variation. Consequently, these datasets fail to capture key challenges in real-world chart reasoning: visual inconsistency, semantic misalignment, temporal discontinuity, and multi-step aggregation. Moreover, their evaluation metrics typically depend on string matching, which inadequately reflects semantic understanding.

We introduce INTERCHART, a diagnostic benchmark designed to probe how well VLMs can reason across multiple charts with increasing levels of complexity. Unlike prior datasets, INTERCHART spans both synthetic and real-world charts, and introduces a structured tiering system to evaluate performance under controlled and unconstrained conditions. It targets a range of reasoning abilities—from simple fact extraction to multi-step, cross-domain inference—allowing researchers to disentangle visual parsing errors from reasoning failures. INTERCHART is organized into three structured subsets, each targeting a different level of reasoning complexity. The first tier, *DECAF* (Decomposed Elementary Charts with Answerable Facts), consists of single-variable charts decomposed from compound figures. This subset emphasizes direct factual and comparative reasoning in simplified visual contexts. The second tier, *SPECTRA* (Synthetic Plots for Event-based Correlated Trend Reasoning and Analysis), introduces synthetic chart pairs that share a common axis but differ in style. They test a model's ability to reason about related quantities such as position and velocity by requiring it to perform trend cor-

relation and event-based interpretation. The third and most advanced tier, *STORM* (Sequential Temporal Reasoning Over Real-world Multi-domain charts), comprises visually complex and semantically diverse real-world chart pairs. These require models to engage in multi-step inference, align mismatched semantics, and synthesize information across domains and temporal sequences.

To ensure reliable assessment, we propose a novel LLM-assisted evaluation pipeline. Instead of relying solely on an exact string match, we employ multiple LLMs as semantic judges and aggregate their decisions through majority voting. It enables evaluators to assess paraphrased answers, numeric approximations, and equivalent units flexibly, producing more robust performance estimates.
We summarize our contributions as follows:

1. We present **INTERCHART**, the first multi-tier benchmark for multi-chart VQA, spanning decomposed, synthetic, and real-world chart contexts.
2. We design structured reasoning tasks to benchmark on various closed and open-source VLMs across three visual tiers, capturing localized and cross-visual dependencies, including trend correlation and temporal abstraction.
3. We propose an LLM-assisted semantic evaluation framework that improves alignment with human judgment and enables fine-grained error analysis.

2 The INTERCHART Benchmark

We introduce INTERCHART to systematically evaluate how reasoning difficulty, chart diversity, and visual complexity affect performance in vision-language models (VLMs). The benchmark contains 5,214 validated question-answer (QA) pairs divided into three subsets: *DECAF*, *SPECTRA*, and *STORM*. These subsets represent distinct levels of real-world chart interpretation difficulty. Appendix 6 summarizes the benchmark construction and annotation workflow for all three subsets.

2.1 DECAF - Decomposed Elementary Charts with Answerable Facts

The *DECAF* subset establishes a foundation for evaluating baseline chart understanding. It includes both real and synthetic charts that represent single variables with minimal visual clutter. The QA tasks focus on factual lookup, comparisons, and parallel reasoning across clearly presented data.

Chart Construction We selected compound charts from ChartQA (Masry et al., 2022), ChartLlama (Han et al., 2023), ChartInfo (Davila et al., 2025), and DVQA (Kafle et al., 2018), ensuring diverse sources of real-world chart styles and semantics. These charts span common types such as vertical and horizontal bar plots, line charts, box plots, dot plots, and heatmaps, covering a wide spectrum of visual encodings frequently used in analytical documents. To support reasoning at a granular level, we aimed to isolate atomic facts from multi-variable visuals. When necessary, we used DePlot (Liu et al., 2023) to regenerate missing tables from raw chart images, ensuring data fidelity and completeness. We then employed a custom decomposition script that extracted individual rows from these tables, aligned them with chart legends and axis labels, and rendered simplified single-variable charts using Plotly. This transformation allowed us to break down dense compound visuals into interpretable units, promoting focused reasoning over elementary visual elements. This resulted in 355 compound charts and 1,188 decomposed charts.

QA Generation We employed a SQL-based sampling strategy to generate table slices. We then used deterministic query templates and Gemini 1.5 pro to create natural language QA pairs, including both chart- and table-derived prompts. A filtering process reduced over 36,000 pairs to 5,800 candidates, followed by manual review to finalize 2,809 QA pairs. Table 1 details the chart types, sources, and QA generation methods in *DECAF*.

2.2 SPECTRA - Synthetic Plots for Event-based Correlated Trend Reasoning and Analysis

The *SPECTRA* subset evaluates a model's ability to integrate distributed information across visually distinct but thematically aligned synthetic charts. These scenarios simulate real-world reasoning, such as interpreting relationships between variables that evolve over time or across regions.

Chart Construction We created structured tables with shared axes to emulate real-world analyses (e.g., linking urban green space with happiness), ensuring that each table reflected plausible entity relationships across dimensions such as time,

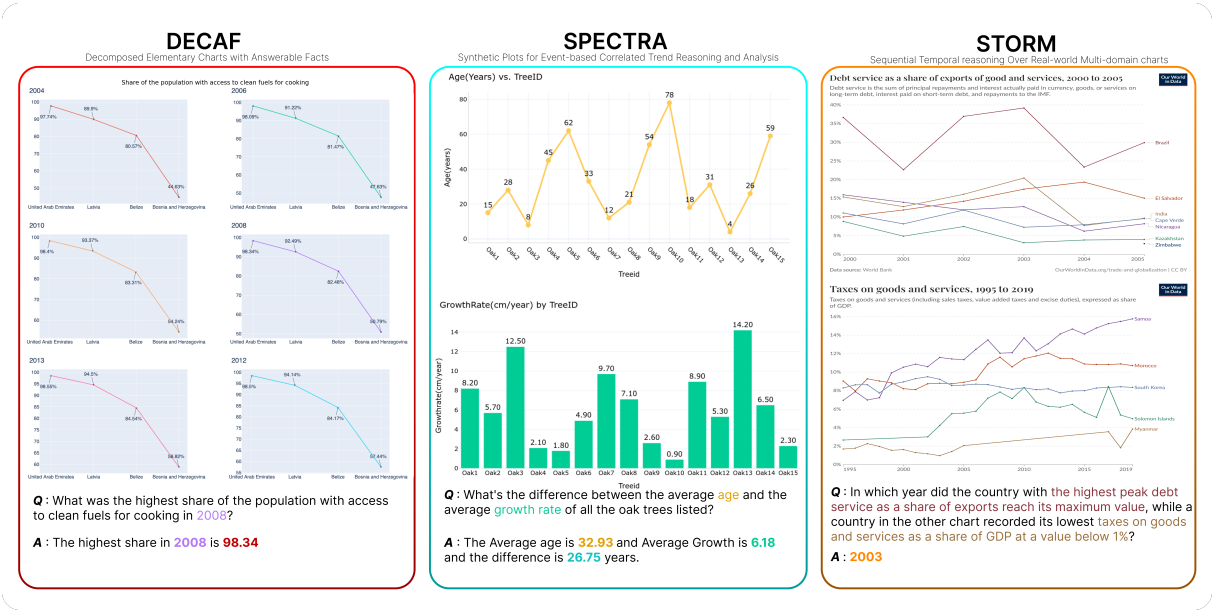

Figure 1: Illustrative examples from our INTERCHART benchmark: DECAF, SPECTRA, and STORM. The DECAF example shows a decomposed version of a chart similar to one found in STORM.

DECAF Distributions			
Chart Type		**Original Chart Sources**	
Line	22	*ChartQA*	153
Horizontal Bar	52	*DVQA*	70
Vertical Bar	149	*ChartInfo*	27
Box Plot	58	*ChartLlama*	105
Heat Map	37		
Dot	37		
QA Generation Methods		**Total**	
Original QA	665	**QA Pairs**	**2,809**
Table-LLM	1,467	**Original Charts**	**355**
Table-SQL-LLM	677	**Decomposed Charts**	**1,188**

Table 1: Summary of chart types, sources, QA generation, and totals for *DECAF*.

geography, or category. These base tables served as input to a two-step synthetic chart construction pipeline. First, we used Gemini 1.5 Pro to generate tabular data with natural variability across rows and columns, guided by template-based prompt scaffolds that preserved semantic consistency while allowing domain shifts (e.g., GDP vs. life expectancy). Second, the structured tables were rendered into visually diverse charts using a human-in-the-loop chart generation module. This included manual oversight to ensure balanced axis scales, legend consistency, and type diversity (e.g., bar-line overlays, multi-axis scales). The resulting charts preserved shared axes across pairs, promoting alignment in subsequent QA tasks. Through this pipeline, we generated synthetic yet realistic chart combinations that encouraged event-based correlation and cross-variable reasoning.

QA Generation We prompted the model to generate questions targeting *low-level reasoning*, such as computing totals or averages; *trend analysis*, including directional inferences and value predictions; and *scenario-based inference*, such as multi-condition comparisons. We used a Python-enabled LLM agent to validate answers through intermediate computation before converting outputs into natural language. After validation, the *SPECTRA* subset contains 1,717 QA pairs across 333 visual context sets and 870 unique charts. Table 2 provides detailed distributions.

SPECTRA & STORM Distribution			
SPECTRA		**STORM**	
Correlated	1,481	*Range Estimation*	198
Independent	245	*Abstract Numerical*	275
		Entity Inference	295
Totals			
QA Pairs	**1,717**	**QA Pairs**	**768**
Context Sets	**333**	**Original Charts**	**324**
Unique Charts	**870**	**Unique Images**	**648**

Table 2: Distribution of question types and overall counts in *SPECTRA* and *STORM*.

2.3 STORM - Sequential Temporal reasoning Over Real-world Multi-domain charts

The *STORM* subset probes the upper limits of current VLM capabilities. It contains complex real-world line chart pairs with diverse styles and domains. These chart combinations reflect realistic analysis settings such as economic reports, environmental trends, and public health dashboards.

Chart Collection We crawled charts and associated metadata from the Our World in Data repository. Using semantic cues and metadata attributes, we applied a semantic pairing module to group charts into coherent visual contexts that share related entities across time. The pairing process identified candidate chart pairs with aligned topics or axes, such as GDP and healthcare spending over the same time period. Each candidate pair was manually reviewed to ensure contextual relevance and analytical coherence. The chart construction pipeline followed the *STORM* algorithmic design outlined in Appendix - Algorithm 3, incorporating structured metadata extraction, entity alignment, and refinement steps to yield 324 validated chart sets comprising 648 distinct images.

QA Curation We used Gemini 2.5 Pro to generate candidate QA pairs grounded in both the chart images and their metadata. The QA generation process focused on multi-step reasoning that spans both charts in a pair, including contextual range estimation, numerical comparisons, temporal trend evaluation, and entity-based inference. Human annotators refined the generated QA pairs to ensure clarity, correctness, and depth of reasoning. Each pair was reviewed, categorized, and finalized through a collaborative validation loop, as described in Algorithm 3. The resulting *STORM* subset includes 768 QA pairs across the verified chart sets. Table 2 summarizes the distribution of question types and chart contexts.

2.4 INTERCHART Verification

We implemented a multi-stage verification pipeline that combined automated filtering and human validation to ensure the quality of INTERCHART.

We first used LLM-based acceptability checks to remove ambiguous or malformed QA pairs. Next, a team of 6 graduate-level annotators manually reviewed each item in DECAF and SPECTRA, ensuring correctness and diversity. Two graduate-level annotators independently verified every QA pair of STORM, with arbitration used to resolve disagreements.

	QA Samples	DECAF	SPECTRA
Pre	13,000	5,800	4,800
Post	5,214	2,809	1,717
% Drop	59.9%	51.6%	64.2%

Table 3: INTERCHART human filtering statistics showing QA sample counts before and after manual verification for subsets *DECAF* and *SPECTRA*.

Table 3 shows filtering statistics for the *DECAF* and *SPECTRA* subsets, revealing retention rates after manual curation. Table 4 shows the inter-annotator agreement for the *STORM* subset, measured using Cohen' Kappa. We achieved a agreement score of 70.63%, reflecting consistent annotations for complex multi-chart reasoning.

	QA Samples	Cohen's κ	Jaccard Index
Overall	768	70.63%	94.75%

Table 4: Overall inter-annotator agreement (Cohen's κ) for the STORM annotated subsets.

Final Dataset Overview: INTERCHART includes **5,214 validated QA pairs** across **1,012 multi-chart contexts** and **2,706 unique chart images**. These examples span diverse reasoning types, visual structures, and real-world complexities, making INTERCHART a comprehensive diagnostic resource for evaluating multi-chart visual question answering.

3 Experiments

We benchmark visual reasoning on INTERCHART using a diverse set of vision-language models (VLMs) and multiple input strategies. Our experiments address four core questions: (1) Does chart decomposition improve accuracy? (2) How does visual complexity affect multi-chart reasoning? (3) Can prompt engineering enhance performance? (4) Do structured tables offer an advantage over direct visual inputs?

VLMs We evaluate both closed- and open-source VLMs. **Closed-source models** include Google Gemini 1.5 Pro (Team, 2024) and OpenAI GPT-4o Mini (OpenAI, 2024). **Open-source models** include Qwen2-VL-7B-Instruct (Yang et al., 2024b), MiniCPM-V-2_6 (Hu et al., 2024),

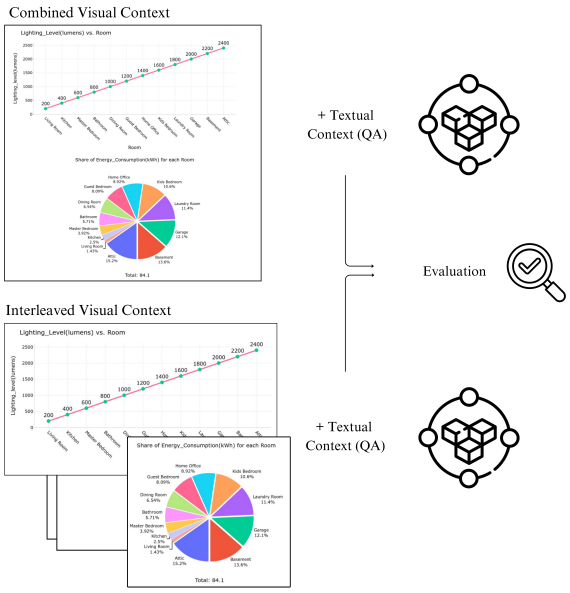

Figure 2: Visual input formats in INTERCHART: **Combined** (stitched multi-chart image) vs. **Interleaved** (separate sequential chart images).

InternVL-2-8B (Chen et al., 2024), and Idefics3-8B-LLaMA3 (Laurençon et al., 2024). We also include DePlot (Liu et al., 2023) and Chart-to-Text (Kantharaj et al., 2022) to assess reasoning over structured outputs.

3.1 Evaluation Pipelines

We compare two reasoning pathways: direct chart-based VQA and a chart-to-table pipeline using intermediate structured representations.

Direct Chart Question Answering We test two visual formats: (i) **Combined**, where charts are stitched into a unified image, and (ii) **Interleaved**, where charts are passed sequentially. For DECAF, we also evaluate original compound charts to quantify gains from simplification.

Prompting styles include **Zero-Shot**, **Zero-Shot CoT** (stepwise reasoning), and **Few-Shot with Directives** (Tannert et al.), which gives structured step-level guidance. Due to input size limits, InternVL and Idefics3 are excluded from interleaved inputs.

Table as Intermediate Representation This setup evaluates whether structured conversion aids reasoning. It includes: (1) *Chart-to-Table Conversion*, where models extract metadata and tables from images, and (2) *Table-Based QA*, where models answer using these tables via CoT prompts. We compare Gemini 1.5 Pro, Qwen2-VL, and MiniCPM. To address DePlot's title extraction issues, we augment it using Gemini title generation, yielding an improved hybrid we term **DePlot++**. This isolates the benefit of structure vs. visual inputs under matched prompts.

Evaluation Strategy We use LLM-based semantic judges to score answers beyond exact string matching, supporting paraphrases, numerics, and unit variations if reasoning is correct. Evaluators include **Gemini 1.5 Flash (8B)** (Team, 2024), **Phi 4** (Abdin et al., 2024), and **Qwen2.5-7B-Instruct** (Yang et al., 2024a). Each receives the question, reference answer, and model output, and returns a binary correctness score along with its reasoning. Final scores use majority voting.

To validate the majority voting agreement, we benchmarked 10,000 sampled responses. In over **78.67%** of cases, all three evaluators agreed on a common answer. Per-model breakdowns appear in Appendix 6.

4 Results and Analysis

We analyze performance on INTERCHART across visual input formats, prompting strategies, and subset difficulty levels by answering targeted questions that highlight emerging trends, model strengths, and failure modes. Tables 5 through 9 summarize these results.

4.1 Performance across Chart Subsets

Do Interleaved Charts Help Models Perform Better than Combined Charts? Not consistently. As shown in Table 5, interleaving charts sometimes improves performance but often leads to minimal or negative changes. For example, Gemini-1.5 Pro improves slightly in STORM from 34.8% to 36.0% but drops from 65.2% to 64.7% in DECAF. Qwen2-VL decreases in DECAF (50.2% to 49.3%) and SPECTRA (32.8% to 32.9%). MiniCPM improves modestly in STORM (21.5% to 25.2%). These results suggest interleaving may help with visual clutter in complex charts but does not offer consistent benefits across all subsets.

Does Decomposing Charts Improve Model Accuracy? Yes. As shown in Table 6, converting charts into structured tables improves accuracy in many cases. Gemini-1.5 Pro achieves 69.9% accuracy using structured DECAF tables, outperforming both DePlot (54.3%) and C2T (43.8%). DePlot++ further improves performance to 63.2% by enhancing title and metadata alignment. Qwen2-VL and MiniCPM also benefit modestly, though

Model	Zero-Shot				Zero-Shot CoT				Few-Shot CoT$_D$			
	Net	DECAF	SPECTRA	STORM	Net	DECAF	SPECTRA	STORM	Net	DECAF	SPECTRA	STORM
Combined Visual Context Image												
GPT-4o-mini	44.8	59.3	45.6	29.7	48.5	68.3	47.9	29.4	48.8	68.6	47.2	30.6
Gemini-1.5-Pro	**53.0**	**65.2**	**59.1**	34.8	**55.0**	**71.6**	**58.5**	34.9	**56.3**	**73.9**	**61.5**	**33.7**
Qwen2-VL-7B	37.3	50.2	32.8	28.9	41.8	59.9	37.3	28.4	40.4	56.3	37.0	27.9
MiniCPM-V-2_6	34.3	52.2	32.4	21.5	35.3	52.7	31.9	21.3	32.4	48.7	30.1	18.6
InternVL-2-8B	30.4	40.0	26.6	24.8	32.3	45.2	28.2	23.6	31.6	46.3	27.3	21.2
Idefics3-8B-Llama3	23.2	39.3	19.4	11.1	23.8	38.8	19.6	13.1	25.9	35.7	25.1	17.1
Mean	37.2	51.0	36.0	25.1	39.5	56.1	37.2	25.1	39.2	55.0	38.0	24.9
Interleaved Visual Context												
GPT-4o-mini	41.9	44.4	50.0	31.5	44.5	51.5	50.3	31.9	44.4	51.7	50.4	31.1
Gemini-1.5-Pro	52.7	64.7	57.4	**36.0**	54.1	68.1	57.8	**36.4**	54.2	70.3	59.6	32.9
Qwen2-VL-7B	37.0	49.3	32.9	28.9	39.4	52.8	38.7	26.7	36.1	47.9	35.2	25.2
MiniCPM-V-2_6	37.1	49.3	36.8	25.2	36.6	49.6	36.2	24.2	35.5	48.1	35.1	23.5
Mean	42.2	51.9	44.3	30.4	43.7	55.5	45.8	29.8	42.6	54.5	45.1	28.2

Table 5: Accuracies using our evaluation method with majority voting of evaluators on all models and prompting strategies. Results are grouped by visual context format (top: Combined, bottom: Interleaved), and broken down by set type (DECAF, SPECTRA, STORM) and strategy (Zero-Shot, Zero-Shot CoT, Few-Shot CoT with Directives). Net scores refer to the mean score of the model across different subsets.

their scores remain lower (50.1% and 33.8%, respectively). These results suggest that SQL-based decomposition paired with table-driven reasoning can improve clarity and support more accurate inference compared to image-only inputs.

Why Do Models Perform Poorly on Real-World Multi-Chart Tasks? As seen in Table 5, accuracy drops sharply in the STORM subset. Gemini-1.5 Pro falls to 34.8%, Qwen2-VL to 28.9%, and MiniCPM-V-2_6 to 21.5%. These real-world chart pairs demand semantic alignment and temporal synthesis. Table 9 shows abstract numerical reasoning is hardest (15.6%), followed by range estimation (33.4%) and entity inference (39.1%). These declines reflect the challenge of integrating misaligned metadata, irregular axes, and domain-specific trends across diverse visual styles.

Do Models Generalize Well from Synthetic to Real-World Chart Distributions? No. Table 5 shows a consistent drop in performance from SPECTRA to STORM across all models. Gemini-1.5 Pro declines from 59.1% in SPECTRA to 34.8% in STORM. Qwen2-VL drops from 32.8% to 28.9%, and MiniCPM-V-2_6 from 32.4% to 21.5%. These results suggest that while models handle synthetic trend-based reasoning to some extent, they struggle to transfer those skills to real-world chart pairs that involve domain shifts, visual diversity, and temporal reasoning.

4.2 Effect of VLMs

Why Does Gemini-1.5 Pro Outperform Other Models? Gemini-1.5 Pro consistently leads across all subsets and prompting strategies. As shown in Table 5, it scores 65.2% in DECAF, 59.1% in SPECTRA, and 34.8% in STORM—well ahead of all other models. GPT-4o-mini is the next best, but lags in STORM (29.7%). Open-source models like Qwen2 and MiniCPM perform reasonably in DECAF but decline sharply on harder subsets. Gemini's strength likely stems from its training on structured inputs and strong instruction-following capabilities.

How Do Open-Source Models Compare Across Subsets? Open-source models perform well in DECAF but struggle in SPECTRA and STORM. Qwen2-VL-7B drops from 50.2% in DECAF to 32.8% in SPECTRA and 28.9% in STORM. MiniCPM-V-2_6 shows a similar decline: 52.2% $\rightarrow$ 32.4% $\rightarrow$ 21.5%. InternVL and Idefics3 perform lower across all subsets, particularly in STORM. These trends point to challenges in generalization, especially when models face domain shifts and complex temporal reasoning.

4.3 Effect of Strategies

Which Prompting Strategies Work Best Across Subsets? Few-Shot CoT$_D$ generally yields the highest accuracy across models and subsets. Table 5 shows Gemini-1.5 Pro improves from 65.2%

Model	DECAF	SPECTRA	STORM	$DECAF_o$
C2T	43.8	46.3	14.7	62.6
Gemini-1.5-Pro	**69.9**	**68.1**	**29.5**	**76.0**
Deplot	54.3	57.9	22.2	63.8
Deplot++	63.2	58.1	23.6	61.9
MiniCPM-V-2_6	33.8	22.1	12.2	35.6
Qwen2-VL-7B	50.1	34.3	18.4	52.4

Table 6: Accuracies from the chart-to-table prompting and rendering strategies for *DECAF*, *SPECTRA*, *STORM*, and *DECAF* compound charts: $DECAF_o$.

(Zero-Shot) to 71.6% (Zero-Shot CoT), and further to 73.9% using Few-Shot CoT_D in DECAF. Qwen2-VL follows a similar pattern, improving from 50.2% to 59.9%, before dropping slightly to 56.3%. While MiniCPM sees minor gains with CoT, it drops slightly under Few-Shot CoT_D. Overall, structured prompting helps most in DECAF and SPECTRA, but offers limited advantage in STORM due to its high complexity.

Does Chain-of-Thought (CoT) Consistently Help? Mostly in simpler subsets. Table 5 shows that CoT improves performance in DECAF and SPECTRA but offers limited benefit in STORM. For example, Gemini-1.5 Pro jumps from 65.2% to 71.6% in DECAF and from 59.1% to 58.5% in SPECTRA. Qwen2-VL improves from 50.2% to 59.9% in DECAF, and MiniCPM sees only a marginal gain (52.2% to 52.7%). In STORM, scores remain largely unchanged or even decline slightly, indicating that verbal reasoning alone cannot compensate for high visual and semantic complexity.

4.4 Effect of Intermediate Representation

How Do Different Table Extraction Methods Compare? DePlot++ consistently outperforms DePlot in DECAF and SPECTRA. As shown in Table 6, DePlot++ achieves 63.2% in DECAF and 58.1% in SPECTRA, compared to 54.3% and 57.9% with DePlot.

DECAF *Chart Type*	Mean	Best
DECAF-Decomposition		
Line	39.66	57.76
Horizontal Bar	50.95	73.36
Vertical Bar	56.17	78.63
Box Plot	**64.3**	**84.23**
Heat Map	55.36	81.35
Dot	58.24	78.63

Table 7: Distribution of Accuracies for Chart Decomposition Approach for *DECAF*.

SPECTRA *Question Category*	Mean	Best
DECAF-Decomposition		
Correlated	39.49	67.43
Independent	**43.22**	**73.47**

Table 8: Distribution of Accuracies for Question Categorization Approach for *SPECTRA*.

This improvement reflects better title and axis alignment, which helps structured models parse tabular input more accurately. The gains are modest but consistent, affirming the importance of clean preprocessing and metadata fidelity.

When Do Structured Tables Hurt Performance Instead of Helping? In STORM. As shown in Tables 6 and 5, structured representations often degrade accuracy on complex real-world charts. Gemini-1.5 Pro drops from 34.8% with visual inputs to 29.5% using tables. C2T performs even worse at 14.7%. These trends suggest that tables cannot capture semantic and temporal alignment across axes, which are critical for accurate reasoning in real-world multi-chart settings.

4.5 Effect of Chart Types, Question Category, and Reasoning Type

Which Chart Types Are Easier or Harder in DECAF? According to Table 7, box plots (64.3%) and dot plots (58.24%) are the easiest for models to interpret, followed by vertical bars (56.17%). Line charts (39.66%) and horizontal bars (50.95%) yield lower accuracy, likely due to visual ambiguity in axis orientation and overlapping labels. These results suggest that models perform best when the chart layout is clean and the data encoding is visually distinct.

Which Question Types Are Easier in SPECTRA? Table 8 shows that independent questions achieve higher accuracy (43.22%) than correlated ones (39.49%).

STORM	**Interleaved**		**Combined**	
Reasoning Type	**Mean**	**Best**	**Mean**	**Best**
Abstract Numerical	13.6	23.7	**15.6**	**25.5**
Entity Inference	**42.1**	**51.3**	39.1	50.9
Range Estimation	31.2	**52.3**	**33.4**	47.5

Table 9: Distribution of accuracies for reasoning type categorization in *STORM*, comparing interleaved and combined visual formats.

This suggests that isolating variables in SPECTRA makes reasoning easier for models, while correlated questions introduce multi-step dependencies across charts that are harder to track and align.

How Consistent Are VLMs Across Chart Types? Model performance varies significantly across chart types. Table 7 shows accuracies ranging from 39.66% for line charts to 64.3% for box plots. This variation suggests VLMs lack consistent chart-type generalization and are sensitive to layout complexity, axis orientation, and label density. Even high-performing models like Gemini show dips on dense or ambiguous formats, highlighting the need for chart-aware visual parsing.

How Do Reasoning Types Impact Performance in STORM? As shown in Table 9, reasoning type has a clear impact on accuracy in STORM. Entity inference yields the highest mean accuracy (42.1% interleaved), followed by range estimation (33.4%), and abstract numerical reasoning is lowest (13.6–15.6%). Interleaved visual formats offer modest gains for entity and range tasks but have limited effect on abstract numerical reasoning, where semantic alignment and aggregation across charts remain key challenges.

5 Comparison with Related Work

Understanding visualizations through natural language has long been a goal in multimodal AI. Early chart-based VQA datasets such as FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022), and ChartLlama (Han et al., 2023) introduced benchmarks over synthetic or real-world plots, focusing on factual or reasoning questions in isolated visual contexts. Recent efforts like ChartInfo (Davila et al., 2024) and SciGraphQA (Li and Tajbakhsh, 2023) extended this by incorporating structured data such as tables and graphs. However, these datasets center on single-chart scenarios and do not evaluate a model's reasoning ability across multiple, semantically related charts. Complementary work on multi-hop (Deng et al., 2022) and graph-based QA (Jin et al., 2024) has demonstrated that decomposing complex inputs into smaller units improves reasoning and interpretability.

MultiChartQA (Zhu et al., 2025) takes a step toward multi-chart reasoning through synthetic chart triplets and four structured tasks: direct, parallel, comparative, and sequential. While it offers controlled diagnostics, the benchmark uses uniformly styled charts with fixed layouts and semantics. It does not assess model performance under visual diversity, semantic drift, or layout complexity, which are standard features in real-world chart collections.

INTERCHART addresses these gaps with a broader diagnostic lens. It introduces three subsets *DECAF*, *SPECTRA*, and *STORM* spanning single-chart to real-world multi-chart reasoning under increasing difficulty and diversity. Unlike prior benchmarks, it combines synthetic and real-world charts to evaluate robustness to visual heterogeneity and abstraction. Additionally, it incorporates an LLM-based evaluation framework that assesses semantic correctness beyond string overlap. INTERCHART thus serves both as a benchmark for evaluating performance and a diagnostic framework for identifying where current models fail in complex, multi-chart reasoning scenarios.

6 Conclusion and Future Directions

We introduced INTERCHART, a diagnostic benchmark for evaluating vision-language models (VLMs) on multi-chart visual reasoning. Structured across three progressively complex subsets *DECAF*, *SPECTRA*, and *STORM*. INTERCHART enables detailed analysis of model behavior under controlled visual transformations. Our findings show that while current VLMs perform well on simplified, decomposed visuals, their accuracy drops significantly when required to integrate or infer across visually complex, semantically misaligned chart sets. Rather than treating VQA as a binary success metric, INTERCHART provides a controlled setting to explore *why* models succeed or fail by varying presentation while holding semantic content constant. This enables diagnostic analysis of model robustness, attention mechanisms, and failure modes—offering insights relevant to model design, training strategies, and interface development.

In future work, we plan to expand INTERCHART beyond traditional charts to include infographics, annotated scientific plots, and hybrid layouts. We also aim to explore multilingual question sets and incorporate neuro-symbolic or retrieval-augmented approaches to support structured abstraction and cross-domain transfer. These directions can advance model transparency, scalability, and applicability in real-world decision-support settings.

Limitations

INTERCHART offers a flexible diagnostic framework but comes with limitations. First, our evaluations rely entirely on zero- and few-shot prompting due to resource constraints. This setup does not capture the full potential of models that might benefit from fine-tuning on chart-specific data. Second, all questions and visual content are English-only, which limits multilingual applicability. Additionally, the current version does not support spatial reasoning tasks such as bounding box grounding or region referencing. While we plan to add fine-grained annotations and structured parsing outputs in future versions, this study focuses solely on answer-level reasoning. Several potential extensions—such as dynamic chart distillation, symbolic chart indexing, or JSON-based parsing supervision—remain conceptual due to scope limitations. Despite these constraints, INTERCHART lays a foundation for expanding multimodal evaluation toward structured, visual-first tasks. Future extensions could include layout-aware fine-tuning pipelines, grounded CoT prompting, and multimodal summarization agents tailored for multi-chart analytics.

Ethics Statement

This work adheres to ethical standards in data collection, annotation, and reproducibility. All visual data used in INTERCHART originate from publicly available or synthetically generated sources under permissible licenses. No sensitive or personally identifiable information is included. Annotations were conducted by graduate-level volunteers based in the United States and India, all of whom provided informed consent. To promote transparency and reproducibility, we will publicly release the full dataset, evaluation scripts, prompt templates, and annotation guidelines. All filtering heuristics and design decisions have been carefully documented to facilitate future research and benchmarking efforts. We also employed AI tools, including large language models, to assist with aspects of the project such as prompt development and explanatory text generation. All AI-generated outputs were reviewed and refined by human authors to ensure accuracy and clarity. Overall, this project reflects our commitment to data privacy, transparency, annotator welfare, and the responsible integration of AI tools throughout the research process.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. Technical Report arXiv:2412.08905, Microsoft Research. V1.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Kenny Davila, Rupak Lazarus, Fei Xu, Nicole Rodríguez Alcántara, Srirangaraj Setlur, Venu Govindaraju, Ajoy Mondal, and C. V. Jawahar. 2025. Chart-info 2024: A dataset for chart analysis and recognition. In *Pattern Recognition*, pages 297–315. Springer Nature Switzerland.

Kenny Davila, Rupak Lazarus, Fei Xu, Nicole Rodríguez Alcántara, Srirangaraj Setlur, Venu Govindaraju, Ajoy Mondal, and CV Jawahar. 2024. Chart-info 2024: A dataset for chart analysis and recognition. In *International Conference on Pattern Recognition*, pages 297–315. Springer.

Zhenyun Deng, Yonghua Zhu, Qianqian Qi, Michael Witbrock, and Patricia Riddle. 2022. Explicit graph reasoning fusing knowledge and contextual information for multi-hop question answering. In *Proceedings of the 2nd Workshop on Deep Learning on Graphs for Natural Language Processing (DLG4NLP 2022)*, pages 71–80, Seattle, Washington. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2311.16483*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Tang, Suhang Wang, Yu Meng, and Jiawei Han. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 163–184, Bangkok, Thailand. Association for Computational Linguistics.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland. Association for Computational Linguistics.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2310.04949*.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, Toronto, Canada. Association for Computational Linguistics.

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*.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence.

Simon Tannert, Marcelo Feighelstein, Jasmina Bogojeska, and Joseph Shtok. Flowchartqa. `https://document-intelligence.github.io/DI-2022/files/di-2022_final_11.pdf`.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2406.04852*.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11341–11359, Albuquerque, New Mexico. Association for Computational Linguistics.

Appendix A: Prompt Templates

Zero-Shot Prompt

Zero-Shot Prompt

Your task is to answer the question based on the given {img_word}. Your final answer to the question should strictly be in the format `"Final Answer:" <final_answer>`.
Question: {question}

Zero-Shot Chain-of-Thought Prompt

Zero-Shot Chain-of-Thought Prompt

Your task is to answer the question based on the given {img_word}. Your final answer to the question should strictly be in the format `"Final Answer:" <final_answer>`. Let's work this out in a step by step way to be sure we have the right answer.
Question: {question}

Data Extraction Prompt

Data Extraction Prompt

Your task is to extract all data from the chart image provided. Make sure to include the chart's title. Output the data in a structured format. Ensure every data point is accurately captured and represented. Be meticulous and do not omit any information.
Think step by step. Identify the chart type to extract data accordingly.

Table-Based Question Answering Prompt

Table-Based QA Prompt

You are tasked with answering a specific question. The answer must be derived solely from information provided, which is extracted from image(s) of chart(s). This information will include the data extracted from the chart, including the chart title. Your final answer to the question should strictly be in the format `"Final Answer:" <final_answer>`. Let's work this out in a step-by-step way to be sure we have the right answer.
Data extracted from charts: {tables}
Question: {question}

Chart Title Extraction Prompt

Chart Title Extraction Prompt

Your task is to extract the main title of the chart image. The main title is typically located at the top of the chart, above the chart area itself, and describes the overall subject of the chart. The title usually describes what data is being presented, the time period, or the geographic location, if applicable.
If the chart does not have a discernible main title, your response should be `"Title: None"`. Otherwise, your response should be in the format `"Title: <title>"`.

Few-Shot with Directives Prompt

Few-Shot with Directives Prompt

Your task is to answer a question based on a given {img_word}. To ensure clarity and accuracy, you are required to break down the question into steps of extraction and reasoning. Your final answer should strictly rely on the visual information presented in the {img_word}.
Here are a few directives that you can follow to reach your answer:
Step 1: Identify Relevant Entities First, identify the key entities or data points needed to answer the given question. These could be labels, categories, values, or trends in the chart or image.
Step 2: Extract Relevant Values Extract all necessary values related to the identified entities from the image. These values might be numerical (e.g., percentages, quantities) or categorical (e.g., labels, categories).
Step 3: Reasoning and Calculation Using the extracted values, apply logical reasoning and calculations to derive the correct answer. Explicitly state the reasoning process to ensure the steps leading to the final answer are understandable and correct.
Step 4: Provide the Final Answer Based on your reasoning, provide the final answer in the following format: `Final Answer: <final_answer>`
Question: {question}

LLM-as-a-Judge Prompt

LLM-as-a-Judge Prompt

You will be given a question, the correct answer to that question (called the "Ground Truth answer"), and a student's attempt to answer the same question (called the "Student Written Answer"). Your task is to determine if the Student Written Answer is correct when compared to the Ground Truth answer.

Instructions:

- The answer should be based solely on the provided information in the question and the Ground Truth answer.
- An answer is correct if it contains the same information as the Ground Truth answer, even if phrased differently.
- Ignore minor differences in wording or phrasing that do not change the meaning.
- If the Ground Truth answer is a number, consider the Student Written Answer correct if it is approximately equal (e.g., 20.24553 vs 20.24). State assumptions clearly.
- For range-based questions, accept answers within the correct range.
- Provide a short explanation inside `<reasoning>` tags.
- Output `<answer> 1 </answer>` if correct, or `<answer> 0 </answer>` if incorrect.

Example: Question: What is the color of water? Ground Truth answer: Pink Student Answer: `Final Answer: Water is colorless.`
Response: `<reasoning> The student answer does not match the ground truth. </reasoning> <answer> 0 </answer>`
Now, answer the following: Question: {question} Ground Truth answer: {ground_truth} Student Written Answer: {student_answer}

Appendix B: Flowcharts

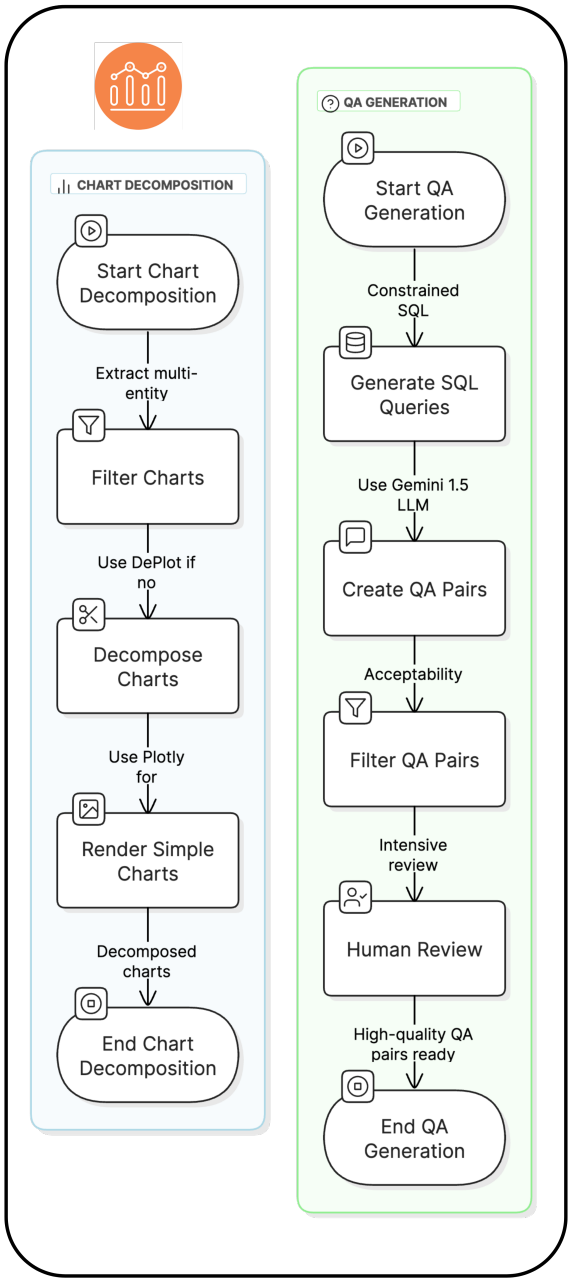

Figure 3: Pipeline for DECAF: Decomposing complex charts into simplified single-entity visuals and generating fact-based QA pairs.

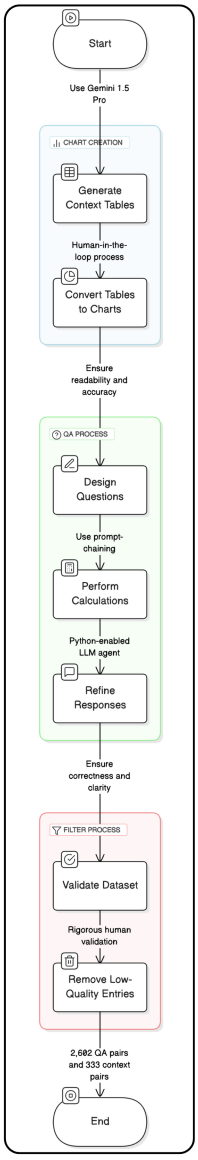

Figure 4: Pipeline for SPECTRA: Generating synthetic multi-chart contexts for correlated trend and scenario-based reasoning.

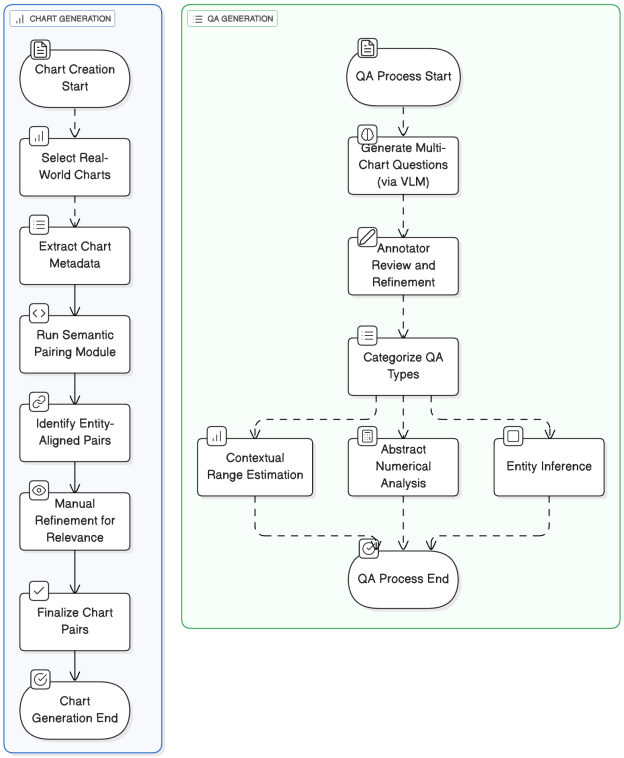

Figure 5: Pipeline for STORM: Constructing real-world chart pairs and QA for multi-step reasoning across misaligned domains.

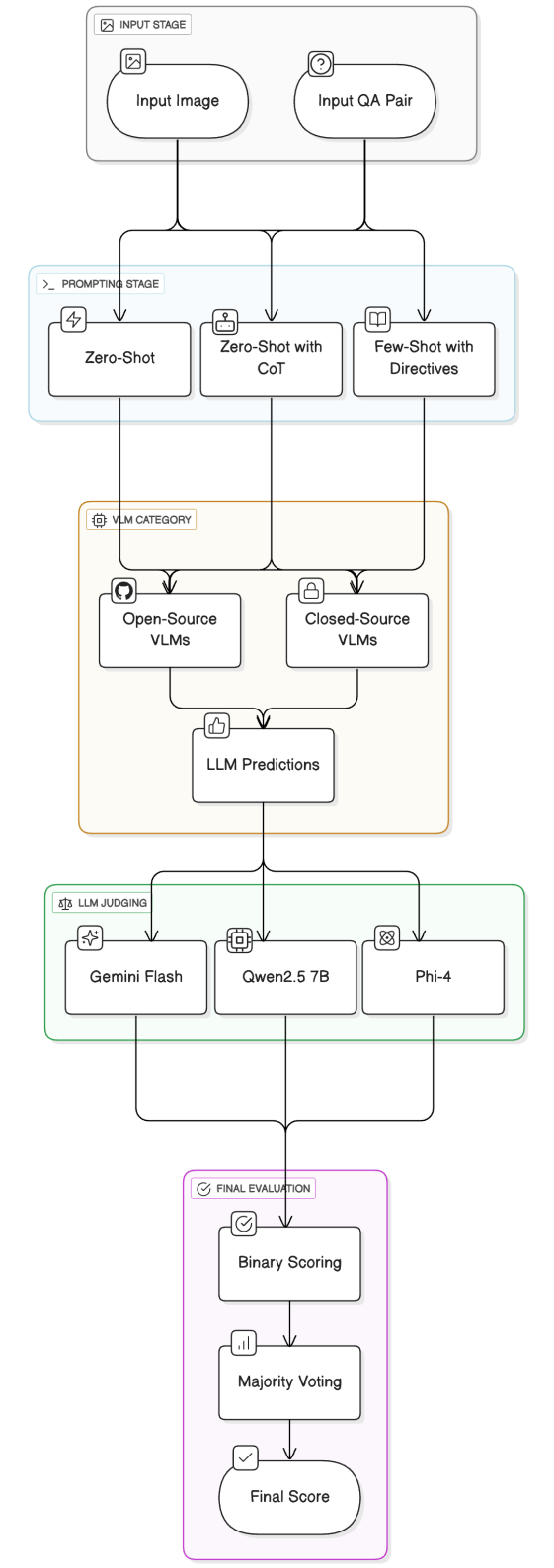

Figure 6: Evaluation pipeline overview: Combining chart-question inputs with different prompting strategies and judging model outputs via majority voting from multiple LLMs.

Appendix C: Data Generation Algorithms

Algorithm 1 ***DECAF*** Constrained SQL Sampling -Multi-Entity Chart Decomposition

Input: Table T, Level L, Operators OP_{num}, OP_{str}, FL_{ops}, STR_{ops}, C_{nj}
Output: SQL Query S
for each column C in T **do**
    Identify $C.dataType$
end for
while not ValidSQL(S, T) **do**
    Initialize empty SQL Query S ▷ Chart Decomposition via SQL
    $select\_col \leftarrow$ Random Column from T
    if $L = 1$ and Random(0,1) = 0 **then**
        Skip Selection Operation
    else
        if $select\_col$ is Numerical **then**
            Apply Numerical Operator
        else
            Apply String Operator
        end if
    end if ▷ WHERE Clause - Linked Data Points Selection
    if Random(0,1) = 1 **then**
        Choose Column C, Value V, Operator OP
        Add Condition $COPV$
    end if ▷ WHERE Clause - Multi-Row and Multi-Column Reasoning
    Extract Numeric Columns
    Choose Number of Conditions Based on L
    for each Condition **do**
        Pick Two Numeric Columns C_A, C_B
        Add Condition C_AOPC_B
    end for ▷ Combine Conditions with Conjunctions for Complex Queries
    for each Condition **do**
        Merge using C_{nj} (AND, OR)
    end for ▷ ORDER BY Clause (For L = 2)
    if $select\_col$ is Numerical and not in Conditions **then**
        Apply ORDER BY with ASC/DESC
    end if
end while
Filter by Human ▷ Ensuring Logical Consistency and Quality
return S

Algorithm 2 Synthetic Simulation - Multi-Chart Reasoning with LLM-Generated Contexts

Input: LLM Model M_{LLM}, Human Annotators A, Chart Generator G_{chart}
Output: Dataset D with Context Pairs and QA Pairs
▷ Step 1: Context Table and Chart Generation
$T_{contexts} \leftarrow \emptyset$
for each scenario S generated by M_{LLM} **do**
    Extract structured entity relationships E_S
    Construct context tables T_S based on E_S
    $T_{contexts} \leftarrow T_{contexts} \cup T_S$
end for
$C_{synthetic} \leftarrow \emptyset$
for each table T in $T_{contexts}$ **do**
    Convert T into chart C using G_{chart}
    Perform human review for accuracy and readability
    $C_{synthetic} \leftarrow C_{synthetic} \cup C$
end for
▷ Step 2: Multi-Chart QA Generation
$QA \leftarrow \emptyset$
for each related chart pair (C_1, C_2) in $C_{synthetic}$ **do**
    for each annotator a in A **do**
        Generate Questions
        Use LLM-based prompt chaining for QA refinement
    end for
end for
▷ Step 3: Dataset Filtering and Compilation
Perform Human Validation for Correctness and Clarity
Remove Low-Quality QA Pairs
$D \leftarrow \{C_{synthetic}, QA\}$
return D

Algorithm 3 STORM: Chart and QA Generation

Input: Chart Repository $\mathcal{C}$, Semantic Pairing Module P_{sem}, VLM Model M_{VLM}, Annotators A
Output: Dataset $D = \{(C_i, C_j, q, a)\}$

// Chart Generation Phase
Initialize paired chart set $\mathcal{P}_{final} \leftarrow \emptyset$
for each chart C_i in repository $\mathcal{C}$ **do**
    Extract metadata M_{C_i}
    Use P_{sem} to find matching chart C_j with aligned entities
    if valid alignment exists **then**
        Add (C_i, C_j) to candidate pairs
    end if
end for
for each pair (C_i, C_j) in candidate pairs **do**
    Manually review for relevance and coherence
    if pair is contextually valid **then**
        Add to $\mathcal{P}_{final}$
    end if
end for

// QA Generation Phase
Initialize QA set $\mathcal{Q} \leftarrow \emptyset$
for each chart pair (C_i, C_j) in $\mathcal{P}_{final}$ **do**
    Generate candidate QA pairs using M_{VLM}
    Annotators review and refine each (q, a)
    Classify QA into one of:
    - Contextual Range Estimation
    - Abstract Numerical Analysis
    - Entity Inference

    Add (C_i, C_j, q, a) to $\mathcal{Q}$
end for
return Final dataset $D \leftarrow \mathcal{Q}$

Appendix D: Model and Compute Details

Model Sizes. We evaluated a mix of closed- and open-source vision-language models (VLMs), as well as structured reasoning baselines:

- **Gemini 1.5 Pro**: 56B parameters (proprietary, estimate based on public disclosures).
- **GPT-4o-mini**: Parameter size not publicly disclosed.
- **Qwen2-VL-7B-Instruct**: 7B parameters.
- **MiniCPM-V-2_6**: 2.6B parameters.
- **InternVL-2-8B**: 8B parameters.
- **Idefics3-8B-LLaMA3**: 8B parameters.
- **DePlot (Liu et al., 2023)**: Built on encoder-decoder transformer with tabular rendering; 400M parameters.
- **Chart-to-Text (Kantharaj et al., 2022)**: Includes rule-based visual parsing + generation via T5 (220M to 3B parameters, depending on version).

Compute Infrastructure. Model inference and evaluation were performed using:

- NVIDIA A100, NVIDIA H200 GPUs on a high-memory compute cluster for open-source model inference and table-based prompting.
- Google Cloud and OpenAI APIs for Gemini 1.5 Pro and GPT-4o-mini, respectively.

Approximate Compute Budget.

- **Open-source model inference**: $\sim$320 GPU-hours (covering 5,214 QA pairs $\times$ 3 prompting strategies $\times$ multiple visual formats).
- **Evaluation with LLM-as-a-Judge**: $\sim$60 GPU-hours (Gemini 1.5 Flash, Qwen2.5-7B, and Phi-4; each example judged by 3 models).
- **Chart-to-Table + Table-based QA (DePlot, DePlot++, Gemini, MiniCPM, Qwen2)**: $\sim$120 GPU-hours for rendering, metadata generation, and table-based prompting.

All experiments were implemented in Python $\geq$ 3.10 using PyTorch $\geq$ 2.0. Evaluation workflows used batch inference pipelines with structured logging, and charts were rendered or parsed using Plotly, DePlot, and in-house scripts.

Appendix E: Individual Evaluation Results

Table 10: Baseline Accuracies using our evaluation method with Gemini-1.5 Eval Engine on all models and prompting strategies. Results are grouped by visual context format (top: Combined, bottom: Interleaved), and broken down by set type (DECAF, SPECTRA, STORM) and strategy (Zero-Shot, Zero-Shot CoT, Few-Shot CoT$_D$).

Model	**Zero-Shot**				**Zero-Shot CoT**				**Few-Shot CoT$_D$**			
	Net	**DECAF**	**SPECTRA**	**STORM**	**Net**	**DECAF**	**SPECTRA**	**STORM**	**Net**	**DECAF**	**SPECTRA**	**STORM**
Combined Visual Context Image												
GPT-4o-mini	45.8	60.9	48.5	27.9	48.0	69.8	47.2	27.1	48.0	69.4	45.5	29.0
Gemini-1.5-Pro	56.3	66.3	61.7	40.8	59.3	73.8	62.0	42.2	59.1	74.6	62.9	39.9
Qwen2-VL-7B	48.7	50.3	33.8	35.2	51.0	60.7	36.6	33.9	47.8	55.6	34.5	33.3
MiniCPM-V-2_6	38.0	53.4	34.0	26.5	38.4	53.9	33.5	27.8	33.5	50.8	27.7	22.1
InternVL-2-8B	33.2	40.3	27.8	31.6	31.6	43.4	26.2	28.6	31.4	44.3	22.4	27.6
Idefics3-8B-Llama3	22.2	38.2	19.6	8.9	23.0	38.1	18.3	12.8	29.0	33.5	27.0	26.6
Mean	40.7	51.6	37.6	28.2	42.2	56.6	37.3	28.9	41.5	54.7	36.7	29.8
Interleaved Visual Context												
GPT-4o	49.3	66.1	52.2	29.7	51.8	74.0	50.9	30.6	50.6	73.0	49.8	29.0
Gemini-1.5-Pro	59.0	74.2	62.9	43.0	60.0	75.0	61.9	43.0	58.4	76.1	61.3	39.4
Qwen2-VL-7B	47.5	47.6	34.1	30.8	50.3	59.6	38.8	32.5	45.1	52.5	32.5	30.2
MiniCPM-V-2_6	41.7	59.1	36.6	29.3	41.0	57.1	37.2	28.9	38.2	53.3	32.2	29.1
Mean	49.4	61.7	46.5	33.2	50.8	66.4	47.2	33.8	48.1	63.7	43.9	31.9

Table 11: Baseline Accuracies using our evaluation method with Qwen 2.5 Eval Engine on all models and prompting strategies. Results are grouped by visual context format (top: Combined, bottom: Interleaved), and broken down by set type (S1, S2, S3) and strategy (Zero-Shot, Zero-Shot CoT, Few-Shot CoT$_D$).

Model	**Zero-Shot**				**Zero-Shot CoT**				**Few-Shot CoT$_D$**			
	Net	DECAF	SPECTRA	STORM	Net	DECAF	SPECTRA	STORM	Net	DECAF	SPECTRA	STORM
Combined Visual Context Image												
GPT-4o-mini	41.4	55.3	38.8	30.1	44.2	61.2	40.8	30.6	45.2	61.7	42.8	31.1
Gemini-1.5-Pro	51.1	66.1	54.5	32.6	51.1	67.0	54.9	31.4	52.1	68.6	57.1	30.7
Qwen2-VL-7B	33.8	48.0	29.0	24.5	35.5	52.5	29.8	24.3	34.5	50.1	29.8	23.7
MiniCPM-V-2$_6$	29.1	45.2	25.6	16.6	28.9	45.4	25.2	16.2	27.3	45.2	23.4	13.2
InternVL-2-8B	24.3	35.1	19.6	18.2	26.3	38.6	21.8	18.5	26.6	41.4	24.1	14.2
Idefics3-8B-Llama3	19.8	38.1	18.8	2.5	19.5	37.7	18.9	2.0	19.7	34.9	20.4	3.9
Mean	33.2	48.0	31.1	20.8	34.6	50.4	31.9	20.5	34.2	50.3	32.9	19.5
Interleaved Visual Context												
GPT-4o-mini	45.6	61.3	44.1	31.4	47.3	65.8	44.3	31.8	48.0	65.6	47.2	31.1
Gemini-1.5-Pro	50.0	67.0	51.0	31.9	51.6	68.1	53.8	32.9	51.3	70.3	54.1	29.5
Qwen2-VL-7B	33.5	46.3	29.5	24.7	36.4	51.4	32.6	25.1	34.1	48.7	28.8	24.7
MiniCPM-V-2$_6$	34.4	52.3	29.6	21.3	32.9	49.9	29.4	19.4	32.2	49.8	28.7	18.2
Mean	40.9	56.7	38.6	27.3	42.1	58.8	40.0	27.3	41.4	58.6	39.7	25.9

Microsoft Phi4 Evaluation Result

Model	Zero-Shot				Zero-Shot CoT				Few-Shot CoT$_D$			
	Net	**DECAF**	**SPECTRA**	**STORM**	**Net**	**DECAF**	**SPECTRA**	**STORM**	**Net**	**DECAF**	**SPECTRA**	**STORM**
Combined Visual Context Image												
GPT-4O-mini	47.5	61.8	49.4	31.3	53.3	73.8	55.8	30.5	53.0	74.8	53.2	31.0
Gemini-1.5-Pro	53.0	67.1	61.1	30.9	54.6	73.9	58.6	31.4	57.8	78.4	64.4	30.5
Qwen2-VL-7B	38.3	52.2	35.7	27.0	46.0	66.5	45.4	27.0	44.2	63.1	46.8	22.8
MiniCPM-V-2_6	38.9	57.9	37.6	21.3	38.6	58.7	37.2	20.0	36.6	50.2	39.0	20.6
InternVL-2-8B	34.1	44.6	32.5	25.0	37.9	53.6	36.6	23.7	36.9	53.4	35.5	21.9
Idefics3-8B-Llama3	27.7	41.6	19.9	21.8	28.9	40.6	21.4	24.6	27.7	38.6	27.8	16.7
Mean	39.9	54.2	39.4	26.0	43.2	61.2	42.5	26.2	42.0	59.8	44.4	23.9
Interleaved Visual Context												
GPT-4o-mini	55.1	68.5	53.6	33.8	56.0	77.6	56.6	33.5	55.8	77.7	55.6	34.1
Gemini-1.5-Pro	55.3	74.5	58.4	33.1	55.9	76.4	57.9	33.4	57.1	78.0	63.5	29.9
Qwen2-VL-7B	37.3	49.4	35.2	27.3	45.9	64.6	44.6	28.6	42.0	55.3	44.3	26.4
MiniCPM-V-2_6	45.0	66.0	44.0	25.0	43.4	64.1	42.0	24.2	44.0	63.3	44.4	24.3
Mean	48.2	64.6	47.8	29.8	50.3	70.7	50.3	29.9	49.7	68.6	51.9	28.7

Table 12: Baseline Accuracies using our evaluation method with Microsoft Phi4 Eval Engine on All Models and Strategies, broken down by Set Type (S1, S2, S3) and Strategy type (Zero-Shot, Zero-Shot CoT, Few-Shot CoT$_D$).