InterChart: Breadth-to-Depth Language Model Cognition Across Linked Visuals

Anonymous ACL submission

Abstract

Existing benchmarks in chart-based visual question answering (VQA) often fail to evaluate visual cognitive load variations in vision-language models (VLMs) and lack structured multi-visual context reasoning. We introduce *InterChart*, a novel benchmark designed to assess multi-visual context reasoning across varying levels of cognitive complexity. *InterChart* comprises 5,214 carefully crafted QA pairs spanning 983 multi-chart visual contexts, structured into three distinct sets of breadth-to-depth cognitive context load. The dataset covers a spectrum of approaches and reasoning tasks, including decomposition, numerical analysis, entity inference and more. We conduct a comprehensive baseline evaluation across multiple VLMs, exploring different prompting strategies and a chart-to-table multi-table paradigm. Our results underscore the importance of structured cognitive decomposition in enhancing chart-based reasoning and highlight critical gaps in existing VLM capabilities.

1 Introduction

As vision capabilities in Large Language Models (LLMs) advance, tasks and benchmarks related to visual question answering (VQA) and reasoning have garnered significant attention effectively gauging performance for real-world vision tasks. An emerging context for such tasks are *charts*. Charts are a common method for representing numerically varying information across diverse fields such as scientific experiments, data analysis, business reports, and time-varying visualizations. Unlike naturally occurring images, charts have a fixed format of representation and require reasoning to interpret.

Numerous chart benchmarks have been proposed to enhance the understanding and reasoning capabilities of multi-modal large language models (MLLMs) over charts, including those by Masry et al. (2022), Methani et al. (2020a), Kafle et al. (2018), Davila et al. (2021), Li and Tajbakhsh (2023) and Kantharaj et al. (2022). Such data is prevalent in real-world scenarios, including academic papers and analytical reports, making the ability to understand and reason over charts an essential task for MLLMs. Numerous studies have explored decompositions in various modalities, including graphs (Miao et al., 2021; Jin et al., 2024), tables and premises (Ye et al., 2023b,a), and multi-hop questions (Deng et al., 2022; Prasad et al., 2024; Methani et al., 2020b; Huang et al., 2023). A key insight from these works is that the representations generated from a complex modality often fail to capture all the individual components required to reason effectively about the questions posed on them.

Cognitive Load: Cognitive load refers to the mental effort required to process and understand context, determined by the working memory resources being utilized. In cognitive psychology, this concept is central to understanding decision-making and task performance, particularly in scenarios where excessive cognitive load can lead to errors or inefficiencies. Xu et al. (2024) tries to evaluate erratic behavior of LLMs and jailbreak tendencies through overloading however, in the context reasoning for VLMs, cognitive load is an essential but under-explored factor. We hypothesize that varying levels of cognitive load, when structured instructionally, can influence VLM outcomes. This aligns with John Sweller's foundational theory on cognitive load (Sweller, 1988), which posits that instructional design can mitigate cognitive load in learners—a principle we extend to evaluating VLM reasoning over complex visual scenarios.

While many benchmarks aim to evaluate models using real-world chart "contexts," they often fail to establish clear boundaries - both a floor and a ceiling, for chart-based vision-language model (VLM) performance, even though an estimate of real-world performance can be gauged we still do not effec-

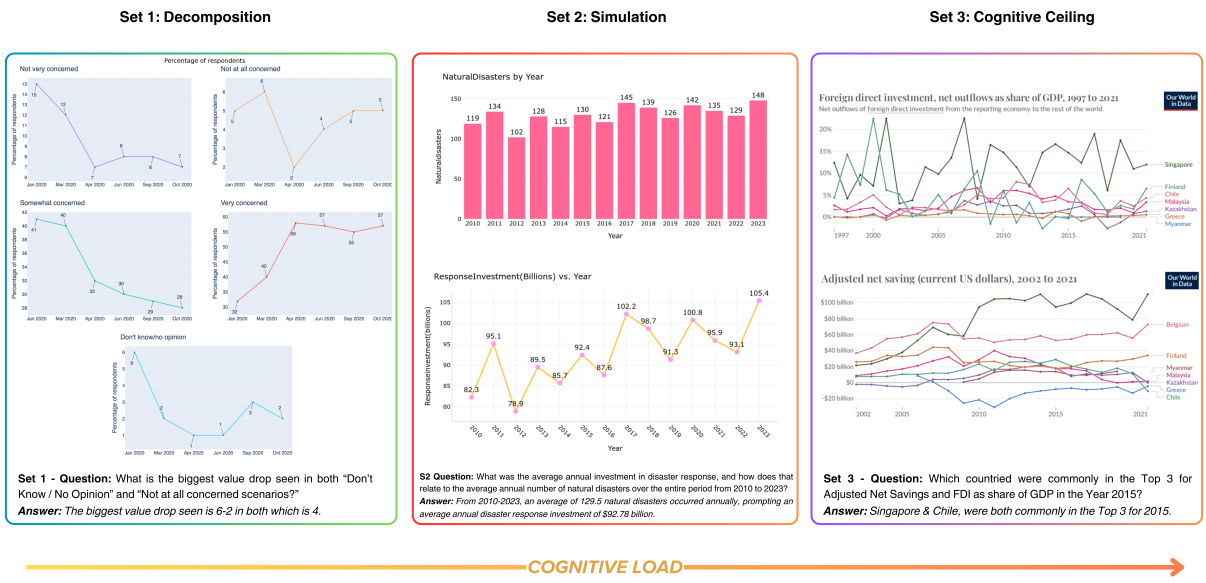

Figure 1: Illustrative examples from our InterChart Resource's Sets 1, 2 and 3. The Set 1 sample is a decomposed version of a chart similar to a single one shown in Set 3.

tively understand how differences in cognitive load in visual contexts passed to the model affect performance in complex chart scenarios. One key question remains: *To what extent can a language model reason over complex visual scenarios, and does decomposing them reduce cognitive load and improve performance?*

In breadth-focused scenarios, a complex chart is decomposed into multiple independent charts, thereby reducing information density and cognitive load. In contrast, depth-focused scenarios retain highly information-dense data, challenging the VLM's capacity for reasoning over tightly interlinked visual and textual elements. To test this setting, we propose a VQA task where the context provided for QA consists of multiple charts, which seems ideal for evaluating reasoning capabilities. We segment our approach into levels: **breadth-level decomposition and depth analysis**. A complex multi-entity compound chart is broken down into simpler, single-entity charts in breadth-level decomposition. Increasing breadth essentially refers to expanding the token context size passed to the model, allowing it to process multiple simpler charts simultaneously. On the other hand, depth evaluation involves presenting pairs of complex compound charts with high information density, challenging the model to reason over intricate and tightly interconnected data. We scale this gradient by testing across **three distinct *Inter-Chart* VQA Sets**, providing a comprehensive evaluation of both breadth and depth scenarios linked via a true multi-chart set. We then evaluate baselines across a spectrum of models, a myriad of approaches including directional CoT prompting, Chart-to-Table paradigms and more. More details about the *InterChart* Dataset are in section 2. All data and approaches will be made public.

2 Proposed *InterChart* Resource

In this section, we provide a detailed overview of the construction process for our comprehensive *InterChart* Benchmark. The benchmark is divided into three distinct sets: **Set 1 (*S1*)**, **Set 2 (*S2*)**, and **Set 3 (*S3*)**. ***S1*** is decomposition focused, ***S2*** mimics real world multichart contexts through simulated AI Table Generation and ***S3*** which is our hard set that measures visual context ceiling. We detail our dataset creation, combining raw data collection, multi-step processing, and comprehensive human annotation.

S1: Compound Chart Decomposition

This set focuses on decomposing complex, multi-entity compound charts into their corresponding single-entity charts, followed by relevant question generation.

Chart Creation: We utilize established datasets such as ChartQA (Masry et al., 2022), ChartLlama (Han et al., 2023), ChartInfo (Davila et al., 2025) and DVQA (Kafle et al., 2018) filtering for multi-

S1 Distributions	Count
Chart Type:	
Line	22
Horizontal Bar	52
Vertical Bar	149
Box Plot	58
Heat Map	37
Dot	37
Original Chart Sources:	
ChartQA	153
DVQA	70
ChartInfo	27
ChartLlama	105
QA Generation Methods:	
Original QA	665
Table-LLM	1,467
Table-SQL-LLM	677
Total QA Pairs	2,809
Total Original Charts	355
Total Decomposed Charts	1,188

Table 1: Summary of Chart Data and QA Pairs for *S1*.

S2 Distributions	Count
Question Types:	
Correlated	1,481
Independent	245
Total QA Pairs	1,717
Unique Context Sets	333
Total Unique Charts	870

Table 2: Summary of Chart Data and QA Pairs for *S2*

entity charts, including hbar, vbar, scatter plots, box plots, and line charts. For charts with existing tables, we use them directly; otherwise, we generate tables via DePlot (Liu et al., 2023). A custom script iteratively parses each chart's table, decomposing data into individual entities, mapping them to legends and axes, and rendering decomposed charts using the Plotly library. The final dataset consists of 355 complex charts, decomposed into 2809 single-entity charts.

QA Generation: To ensure quality and scalability, we implement a "generate and filter" pipeline inspired by prior work (Han et al., 2023; Singh et al., 2024).

Generate: Constrained SQL sampling of linked data points within a chart's table forms the basis for definitive SQL queries. These queries, pre-templated with *WHERE* conditions, produce objective answers, replicating multi-row and multi-column reasoning. The SQL query, selected data points, derived answer, and chart context are fed into Gemini-1.5 (Vertex), prompting naturalized QA pair generation. We also prompt the Gemini model to create questions directly from the table and use a subset of the original questions as well. However, this approach may introduce entropy and noise, the filtration steps deals with this.

Filter: The initial method generated 36,000+ QA pairs across 6,200+ charts. These pairs, along with tables, were re-evaluated via an LLM acceptability test, reducing the set to 5,800. A final human review refined the dataset to 2,809 high-quality QA pairs, optimizing naturalness and minimizing entropy.

S3 Distributions	Count
Question Types:	
Range Estimation	270
Abstract Numerical Analysis	254
Entity Inference	164
Total QA Pairs	688
Unique Context Sets	295
Total Unique Images	590

Table 3: Summary of Question Types and Counts for *S3*

S2: Synthetic Simulation

This set evaluates multi-chart reasoning in scenarios where information is distributed across related visualizations rather than a single chart. For this we craft all charts from LLM generated context tables through a human-in-the-loop process. For example, understanding urban living conditions may require analyzing one chart depicting city-to-green-space ratios and another showing happiness indices.

Chart Creation: We first generate structured entity relationships to simulate diverse real world situations through Gemini 1.5 Pro (Vertex). This is followed by a table creation process by the using the same model, ensuring that they are linked through one common axis and focusing on creating realistic data incorporating noise as well. These tables are then converted into charts through a human-in-the-loop process, ensuring readability, accuracy, and diversity in visualization types.

QA Generation: We use Gemini 1.5 Pro to generate questions requiring direct data extraction, cal-

culations, and counting operations (e.g., calculating averages, counting occurrences under conditions) as well as questions demanding commonsense reasoning, trend identification, and extrapolation based on the data (e.g., situation-based scenarios, trend analysis, predictive inferences). We then generate accurate and human-readable answers using the tables and questions generated. We use a prompt-chaining approach with an LLM agent equipped with a Python REPL tool to generate accurate answers, followed by another step to convert the generated answers to natural language.

Filter: The dataset undergoes rigorous human validation to ensure correctness, clarity, and relevance, with low-quality and unsuitable entries removed. This meticulous human verification process guarantees the accuracy and reasoning integrity of the final dataset. After filtration we are left with 1,717 QA pairs and 333 context pairs.

S3: Visual Context Ceiling

This set shifts focus from decomposition to assessing the performance ceiling of Visual Language Models (VLMs) in high-complexity contexts. It features dense, multi-entity compound charts, requiring retrieval not just within but across chart pairs. Rather than measuring gains from decomposition, this dataset evaluates inference limits under extreme contextual and visual complexity.

Chart Creation: We curate chart pairs from Our World in Data's line chart repository using a metadata-driven semantic pairing algorithm, followed by manual refinement. Each pair contains complex, interrelated charts sharing common entities, ensuring relational coherence. This shared-entity structure acts as a *keying* mechanism, anchoring relationships across contexts and enabling robust visual grounding assessments.

QA Generation: A team of five independent annotators crafted inferential QA pairs, ensuring relevance and challenge. Questions fall into three categories:

1. Contextual Range Estimation: Evaluating value ranges across both charts, testing contextual reasoning.

2. Abstract Numerical Analysis: Requiring arithmetic and logical deductions from data points.

3. Entity Inference: Identifying trends and patterns across entities to prompt meaningful conclusions.

This is further filtered by another independent human verification team. The final S3 set includes 295 chart pairs and 688 corresponding QA pairs.

The final *InterChart* Resource contains **5,214** QA pairs over **983** Visual Context Sets and **2,648** Individual Chart Images. Tables 1, 2, and 3 show a summary of the internal distribution for all sets. Deeper processes for all sets are outlined in Algorithms 1,2 and 3 in the Appendix along with flowcharts in Figures 4, 5, and 6. Table 4 shows pre and post filtration stats for all sets.

	# QA Samples	# S1	# S2	# S3
Pre	13,000	5,800	4800	2,400
Post	5,214	2,809	1,717	688
% drop	59.9%	51.6%	64.2%	71.3%

Table 4: InterChart Human Filtering stats pre and post human verification and filtration for QA pairs sets *S1*, *S2* and *S3*

.

3 Experimentation

This section details the experimental setup, including the models, prompting strategies, and evaluation methodology used to assess the performance of various multimodal models on chart-based reasoning tasks. We address the following research questions through our experiments:

RQ1. How does multi-entity chart decomposition impact the reasoning performance of VLMs?

RQ2. To what extent does cognitive load variation affect VLMs' ability to process multi-chart contexts?

RQ3. How do different prompting strategies influence model accuracy on multi-chart question answering?

RQ4. How does multi-chart to multi-table paradigms differ in accuracies? Is visual overload tougher on the model than data-based overload?

3.1 Models

We evaluated diverse state-of-the-art multimodal models, including closed-source (Google Gemini 1.5 Pro (Vertex) and OpenAI's GPT-4o mini (OpenAI, 2024) via API) and open-source options. Our open-source selection included Qwen2-VL-7B-Instruct (Yang et al., 2024), MiniCPM-V-2_6 (Hu et al., 2024), InternVL-2-8B (Chen et al., 2025), and Idefics3-8B-Llama3 (Laurençon et al., 2024) (a Llama3-based vision-language model). This allows comparison of open-source and closed-source performance on chart reasoning. We also

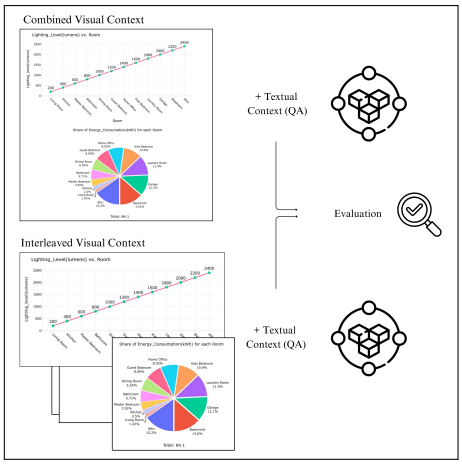

Figure 2: Visual Representation for Combined Visual Context and Interleaved Context

examined chart-to-table specialized models, DePlot (Liu et al., 2023) and Chart-to-Text (Kantharaj et al., 2022), to assess the utility of intermediate table representations.

3.2 Baselines and Methodologies

We established baselines to evaluate chart understanding, categorized into Chart Question Answering and Chart-to-Table Question Answering.

3.2.1 Chart Question Answering

This section evaluates models' ability to answer questions from charts, using two image input formats (illustrated in Figure 6):

1. **Combined Image:** Multiple charts combined into a single image.
2. **Interleaved Images:** Each chart as a separate image, presented sequentially.

Original Charts: For charts from set 1, we also performed experiments using corresponding charts from their original dataset

Three prompting techniques assessed reasoning capabilities:

- **Zero-Shot:** Here the model was prompted to answer through a direct question and no other hints. (Appendix C.1).
- **Zero-Shot Chain-of-Thought (COT):** Here the model was prompted for step-by-step reasoning for improved transparency and accuracy. (Appendix C.2).
- **Few-Shot with Directives:** Here the model was instructed to follow a structured approach to answer the question. The key steps focused on identifying key entities, extracting required values from charts and performing reasoning to reach the final answer. (Appendix C.6)

Note: Interleaved images were not tested for InternVL (context limits) and Idefics3 (compute constraints).

3.2.2 Chart-to-Table Question Answering

This approach uses a two-stage process: (1) converting charts to tables and (2) answering questions using the tables, aiming to improve reasoning with structured data.

1. **Data Extraction:** Models extract **all** data (including title and chart type) from chart images into a structured format (e.g., a table), prompted as in Appendix C.3.
2. **Table-Based Question Answering:** The extracted table and original question are given to the model. A zero-shot chain-of-thought prompt (Appendix C.4, similar to Section 3.2.1) requires answers based on the table, to test if structured data improves accuracy and interpretability.

We evaluated specialized chart-to-table models using this pipeline with Gemini 1.5 Pro for question answering. We also tested **Gemini 1.5 Pro**, **Qwen2-VL-7B-Instruct**, and **MiniCPM-V-2_6**.

To address DePlot's inaccurate title extraction, we created **DePlot++**: an enhanced pipeline using Gemini 1.5 Pro to extract chart titles (Appendix C.5) before integration with Deplot's output.

3.3 Evaluation

Our methodology adopts an *"AI as an Evaluator"* approach similar to Fu et al. (2023); Lin and Chen (2023); Chiang and Lee (2023); Singh et al. (2024). We employ two evaluator models — **Gemini 1.5-Flash 8B** (Vertex), and **Qwen 2.5-7B-Instruct** (Bai et al., 2023) to assess the model-generated responses, which are compared against a *gold standard* short answer and the question. The evaluators assign a binary label to determine whether a response is correct, effectively framing the task as a "length-invariant" paraphrase detection problem for short text responses, surpassing traditional similarity metrics. Assessments rely solely on the provided information, accepting paraphrased answers with the same meaning and allowing numerical approximations with explicit assumptions. The two

Model	Zero-Shot				Zero-Shot CoT				Few-Shot CoT$_D$			
	Net	*S1*	*S2*	*S3*	*Net*	*S1*	*S2*	*S3*	*Net*	*S1*	*S2*	*S3*
Combined Visual Context Image												
GPT-4o-mini	42.6	60.9	48.5	18.6	44.7	69.8	47.2	17.9	43.9	69.4	45.5	17.6
Gemini-1.5-Pro	52.1	66.3	61.7	28.3	53.3	73.8	**62.0**	24.1	**53.1**	74.6	**62.9**	**21.9**
Qwen2-VL-7B	34.1	50.3	33.9	18.0	36.4	60.7	36.7	11.9	34.0	55.6	34.5	11.8
MiniCPM-V-2_6	32.6	53.4	34.0	10.3	32.9	53.9	33.4	11.3	29.5	50.8	27.7	9.9
InternVL-2-8B	27.1	40.3	27.8	13.1	25.0	43.4	26.2	5.5	24.1	44.3	22.4	5.5
Idefics3-8B-Llama3	22.8	38.2	19.6	10.5	22.1	38.1	18.3	9.9	23.9	33.5	27.0	11.2
Mean	35.2	51.6	37.6	16.5	35.8	56.6	37.3	13.4	34.8	54.7	36.7	13.0
Interleaved Visual Context												
GPT-4o-mini	46.0	66.1	52.2	20.3	47.5	74.0	50.9	19.0	47.6	73.0	49.8	**20.5**
Gemini-1.5-Pro	**55.7**	**74.2**	**62.9**	**30.1**	**55.4**	**75.0**	61.9	**29.4**	52.1	**76.1**	61.3	18.9
Qwen2-VL-7B	32.4	47.6	34.1	15.6	37.5	59.6	38.8	14.0	31.9	52.5	32.5	10.8
MiniCPM-V-2_6	36.4	59.1	36.6	13.4	36.0	57.1	37.2	13.7	32.5	53.3	32.2	12.1
Mean	42.7	61.8	46.5	19.9	44.2	66.4	47.2	19.0	41.1	63.7	44.0	15.6

Table 5: Baseline Accuracies using our evaluation method with Gemini-1.5 Eval Engine on All Models and Strategies broken down by Set Type Wise (*S1, S2, S3*) and Strategy wise. The highest values are highlighted.

models demonstrate a very strong Pearson's Correlation value of **0.98** for accuracies across 114 different evaluations combinations (Model+Strategy) and a strong absolute agreement of **88%** as seen in Table 9. Figure 3 validates this as well. We show results from Gemini 1.5 flash in Table 4. Results from Qwen are available in Table 10 in the Appendix.

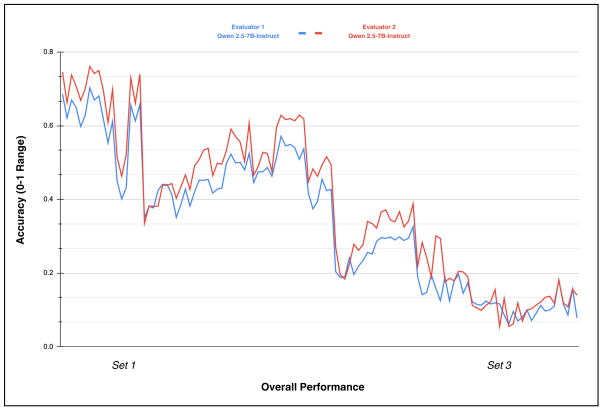

Figure 3: Visual Representation for Combined Visual Context and Interleaved Context, Qwen vs. Gemini Evaluations across 114 Evaluation Combinations

4 Results and Analysis

A New Challenging Benchmark. Our dataset benchmarks Vision-Language Models (VLMs) on fine-grained visual understanding and multi-image reasoning, emphasizing entity selection and recognition across images. Table 5 shows a significant performance gap between even the strongest VLMs and ideal scores, particularly in Set 3 (S3), highlighting limitations in reasoning capabilities for complex, real-world scenarios.

Set	Overlap
S1	89%
S2	85%
S3	89%
Total	**88%**

Table 6: Label Overlap Across Different Sets for Qwen and Gemini Eval Engines.

Model Performance Comparison. Gemini-1.5-Pro consistently outperforms other models across all strategies and visual contexts (Table 5), attributable to its strong long-context attention, data extraction, and reasoning skills. Among open-source models, Qwen2-VL-7B and MiniCPM-V-2_6 are relatively stronger, but all open-source models significantly underperform closed-source counterparts, especially on complex reasoning in S3 (often below 15% accuracy). S3 remains a consistent challenge across all models.

Prompt Effectiveness. Table 5 shows that Zero-Shot CoT marginally outperforms Zero-Shot, contrasting with previous findings where CoT provided more substantial gains. This suggests models may be implicitly adopting step-by-step reasoning. Few-Shot CoT$_D$ doesn't consistently outperform Zero-Shot CoT, sometimes even decreasing performance

(e.g., Gemini-1.5-Pro with Combined Visual Context), possibly due to unintended biases from few-shot examples. Interleaved visual context consistently yields better results than Combined Visual Context. This also answers our third research question.

Model	*S1*	*S2*	*S3*	$S1_O$
C2T	45.9	46.0	7.1	62.7
Gemini-1.5-Pro	**70.2**	**69.8**	**15.1**	**75.2**
Deplot	57.8	58.4	8.1	63.8
Deplot++	62.8	58.7	8.4	63.3
MiniCPM-V-2_6	34.6	21.4	8.7	36.7
Qwen2-VL-7B	49.8	34.3	9.2	53.6

Table 7: Accuracies from the chart-to-table prompting and rendering strategies for *S1*, *S2*, *S3* and *S1* compound charts.

Does converting charts to tables help? Contrary to expectations, introducing a chart-to-table conversion step (Table 7) did not universally improve reasoning performance. The leading model, Gemini-1.5-Pro, saw decreased accuracy, especially in the complex S3, indicating its direct visual reasoning surpasses relying on potentially lossy tabular representations. The inconsistent results across other models and the significant performance drop in S3 highlight the crucial role of generated table quality and the potential loss of vital visual information during conversion. Thus, while explicit data extraction can be beneficial, directly processing visual input remains more effective for robust models, particularly in complex reasoning tasks. This also answers our fourth research question and highlights the need of more effective chart summarization methods to counter the information loss.

Model	ZS	CoT	FSwD
GPT-4o-mini	46.3	52.2	51.5
Gemini-1.5-Pro	**66.9**	**70.1**	**70.7**
Qwen2-VL-7B	49.0	52.8	46.6
MiniCPM-V-2_6	49.8	49.6	46.5
InternVL-2-8B	42.7	49.1	46.7
Idefics3-8B-Llama3	43.9	43.8	38.2

Table 8: Accuracy on the original single compound charts for ***S1***, comparing Zero-Shot (ZS), Zero-Shot CoT (ZSCoT), and Few-Shot with Directives (FSwD).

Comparison with Original Complex Charts (S1). To assess the influence of visual complexity, we compared model performance on the original, single compound charts (Table 8) against the modified S1 charts from Table 5, where the original complex visualizations were decomposed into multiple, simpler charts. A significant performance drop was observed for most models when faced with the original, non-decomposed charts. For each model, we can see that the scores drop by atleast 3-5% which is a significant drop in accuracy. This performance difference indicates that the models benefit from the decomposition of complex charts into simpler forms, which can be an useful method for improving chart question answering capabilities. This also answers our first research question showing that multi-entity chart decomposition can lead improved reasoning performance of VLMs.

S1 Chart Type	Mean	Best
***S1*-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 9: Distribution of Accuracies for Chart Decomposition Approach for *S1*.

Performance Variation across Chart Types (S1). Table 9 shows significant performance variation across different chart types within Set 1 (S1). Box plots proved the easiest for models (mean accuracy 64.3%, best 84.23%), likely due to their emphasis on summary statistics. Line charts were the most challenging (mean 39.66%, best 57.76%), suggesting difficulty in tracking trends and extracting precise values. Other chart types showed intermediate performance, indicating varying challenges in comparing magnitudes, identifying patterns, and interpreting spatial relationships. This highlights the crucial impact of chart type on VLM visual understanding and pinpoints areas needing further model development.

S2 Question Category	Mean	Best
***S2*-Decomposition**		
Correlated	39.49	67.43
Independent	43.22	73.47

Table 10: Distribution of Accuracies for Question Categorization Approach for *S2*.

Impact of Attending to Multiple Charts (S2). Table 10 shows that in Set 2 (S2), models performed slightly better on questions answerable from a single chart ("Independent": mean 43.22%,

best 73.47%) than those requiring correlation across multiple charts ("Correlated": mean 39.49%, best 67.43%). While the mean difference is small, the higher top performance on "Independent" questions suggests some models have greater capacity for focused single-chart analysis. The lower "Correlated" scores, even for the best model, highlight the significant challenge of multi-chart reasoning, underscoring the need for VLMs that can effectively integrate information from multiple visualizations.

***S3* Question Category**	**Mean**	**Best**
***S3*-Decomposition**		
Abstract Numerical Analysis	10.32	29.13
Entity Inference	15.34	31.09
Reasoning with Range Estimation	18.77	37.40

Table 11: Distribution of Accuracies for Question Categorization Approach for *S3*.

Challenges in Advanced Reasoning (S3). Table 11 reveals that Set 3 (S3), poses significant challenges to VLMs across all question types. Reasoning with Range Estimation achieved slightly better, though still low, scores (mean 18.77%, best 37.40%), indicating limited ability in estimation. Abstract Numerical Analysis was the most difficult (mean 10.32%, best 29.13%), highlighting a weakness in deriving non-explicit numerical insights. Entity Inference showed intermediate performance (mean 15.34%, best 31.09%), suggesting some, but not robust, capability in inferring relationships. The consistently low performance across all S3 categories underscores the need for fundamental advancements in VLM design to address these advanced reasoning challenges.

5 Comparison to Related Work

Existing ChartQA benchmarks such as Masry et al. (2022); Methani et al. (2020a); Li and Tajbakhsh (2023) are primarily designed for single-chart question answering, limiting their applicability to real-world multi-chart scenarios. However, neither address multi-chart reasoning and cognitive complexity, making them less suitable for evaluating reasoning across structured multi-visual contexts in comparison to InterChart. InterChart also introduces three cognitive complexity levels which allows for a more nuanced evaluation of how models handle varying levels of difficulty.

More recent efforts such as MultiChartQA (Zhu et al., 2025) have taken important steps towards addressing the aforementioned gaps, there remains room for further refinement and expansion. InterChart features over 5000 unique QA pairs and 2500+ individual charts, providing a broader and more diverse dataset. In comparison, a portion of MultiChartQA's queries are multiple-choice or direct ChartVQA-style. While MultiChartQA serves as an important static benchmark, InterChart complements such efforts by extending the evaluation spectrum by incorporating multi-image-based QA, chart-to-table conversion, and multiple prompting strategies (zero-shot, chain-of-thought, few-shot) which enables a more comprehensive assessment of VLMs. This work also introduces a breadth-to-depth decomposition strategy to systematically structure reasoning, potentially reducing cognitive load and enhancing interpretability.

By addressing the limitations of prior benchmarks and introducing a structured evaluation methodology, InterChart establishes itself as a more scalable, generalizable, and insightful resource for evaluating chart-based reasoning in VLMs.

6 Conclusion and Future Work

In this paper, we introduced *InterChart*, a novel benchmark designed to evaluate multi-chart reasoning across varying levels of cognitive complexity. By structuring our dataset into three distinct sets, we systematically assessed the impact of cognitive load on vision-language model (VLM) performance. Our experiments demonstrate that while state-of-the-art models exhibit strong performance on simple visual contexts, their capabilities diminish significantly when faced with complex multi-visual reasoning tasks. The structured cognitive decomposition approach introduced in *InterChart* provides insights into VLM limitations, emphasizing the need for enhanced reasoning mechanisms and structured multi-modal understanding.

For future work, we aim to expand *InterChart* by incorporating additional real-world visual context datasets that are not chart specific, further increasing domain diversity. Additionally, integrating multidimesional support will help assess model performance across linguistic variations. Future studies should also explore fine-tuning methodologies and architectural innovations specifically tailored for multi-chart reasoning. By addressing these areas, we hope to further advance the field of multi-modal AI and bridge the existing gaps in chart-based reasoning tasks.

Limitations

Our work has a few notable limitations. Primarily, due to financial and computational resource constraints, we were unable to fine-tune all the models under consideration, which may have led to an under-representation of the broader capabilities of various NLP models beyond our primary focus. Additionally, the language constraints in this research, particularly the emphasis on English for generating Visual Question Answering (VQA) methods, highlight the need for greater linguistic diversity in NLP applications to enhance inclusivity and applicability. Incorporating human cognitive modeling techniques could provide deeper insights into optimizing instructional strategies for VLMs, ultimately improving their ability to handle complex structured visual data. Given the novelty of the task, it is also important to recognize that our insights may not be exhaustive, underscoring opportunities for future research. Additionally, due to certain constraints, we were not able to explore a promising avenue for improving VLM performance on combined chart images: augmenting InterChart with explicit sub-chart localization. Had resources permitted, we would have pursued a methodology wherein decomposed charts are randomly combined, with their bounding box coordinates and corresponding titles stored in a JSON format. A model, potentially such as Qwen2VL, would then be fine-tuned using LoRA to predict this JSON structure directly from the combined images. A separate tool would then leverage the predicted bounding boxes to extract the relevant sub-charts, feeding these as context for question answering. Furthermore, resource constraints prevented us from implementing a chart distillation step, where an LLM classifier would select only the necessary charts (based on titles) from a larger set to answer a given question. Several other approaches could be proposed, such as neuro-symbolic AI techniques to enhance logical and structured reasoning over multi-chart contexts, and retrieval-augmented generation (RAG) based chart retrieval methods to dynamically fetch and integrate relevant visual information. We anticipate these approaches would reduce the cognitive load on the model and hence improve model performance.

Ethics Statement

We, the authors, ensure that our research meets the highest ethical standards in both research and publication. We have carefully addressed all ethical considerations for responsible and fair use of computational linguistics methods. To help others replicate our results, we are sharing all necessary details, including code, available datasets (used according to their ethical guidelines), and other resources. This allows the research community to verify and build on our work. Our claims are backed by our experimental results. We provide detailed information on annotations, verifications, dataset splits, models, and methods used for reproducibility.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

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

Kenny Davila, Chris Tensmeyer, Sumit Shekhar, Hrituraj Singh, Srirangaraj Setlur, and Venu Govindaraju. 2021. Icpr 2020 - competition on harvesting raw tables from infographics. In *Pattern Recognition. ICPR International Workshops and Challenges*, pages 361–380, Cham. Springer International Publishing.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. 2022. Interpretable amr-based question decomposition for multi-hop question answering. In *Proceedings of the Thirty-First*

International Joint Conference on Artificial Intelligence, IJCAI-22, pages 4093–4099. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

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.

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, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies.

Xiang Huang, Sitao Cheng, Yiheng Shu, Yuheng Bao, and Yuzhong Qu. 2023. Question decomposition tree for answering complex questions over knowledge bases. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12924–12932.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng 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.

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.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

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 Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020a. Plotqa: Reasoning over scientific plots.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020b. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1516–1525.

Xupeng Miao, Nezihe Merve Gürel, Wentao Zhang, Zhichao Han, Bo Li, Wei Min, Susie Xi Rao, Hansheng Ren, Yinan Shan, Yingxia Shao, Yujie Wang, Fan Wu, Hui Xue, Yaming Yang, Zitao Zhang, Yang Zhao, Shuai Zhang, Yujing Wang, Bin Cui, and Ce Zhang. 2021. Degnn: Improving graph neural networks with graph decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1223–1233, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Gpt-4o system card.

Archiki Prasad, Alexander Koller, Mareike Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2024. Adapt: As-needed decomposition and planning with language models.

Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. 2024. FlowVQA: Mapping multimodal logic in visual question answering with flowcharts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1330–1350, Bangkok, Thailand. Association for Computational Linguistics.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Google Vertex. Gemini pro api. Accessed on Feb 4, 2024.

Nan Xu, Fei Wang, Ben Zhou, Bangzheng Li, Chaowei Xiao, and Muhao Chen. 2024. Cognitive overload: Jailbreaking large language models with overloaded logical thinking. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages

3526–3548, Mexico City, Mexico. Association for Computational Linguistics.

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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023a. mplug-owl: Modularization empowers large language models with multimodality.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023b. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 174–184, New York, NY, USA. Association for Computing Machinery.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. 2025. Multichartqa: Benchmarking vision-language models on multi-chart problems.

Appendix

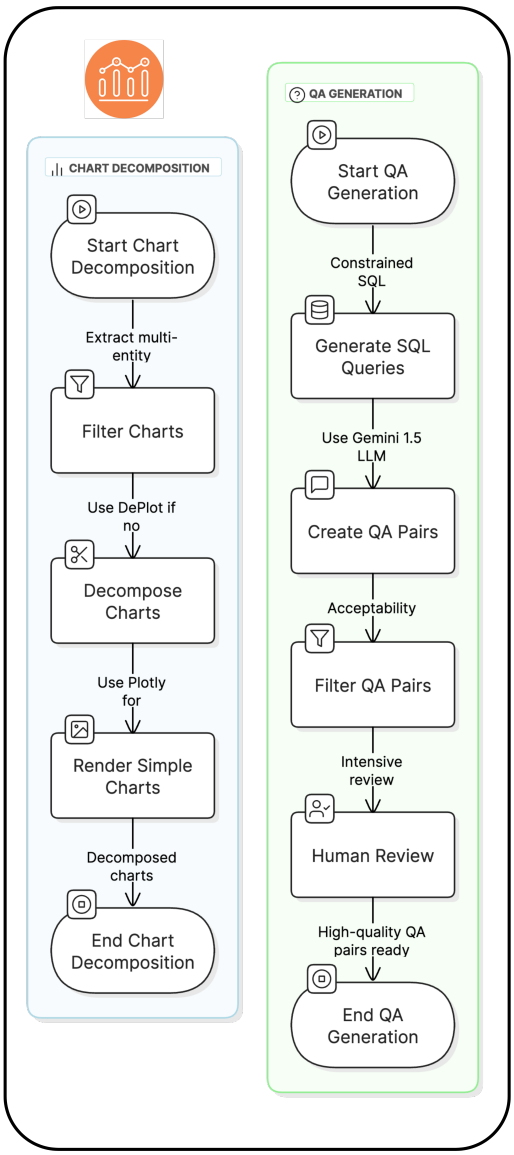

Figure 4: Flowchart for ***S1***: Constrained SQL Sampling -Multi-Entity Chart Decomposition

A Algorithms & Diagrams

Algorithm 1 ***S1*** 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 Sampling
    $select\_col \leftarrow$ 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 $C_A OP C_B$
    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

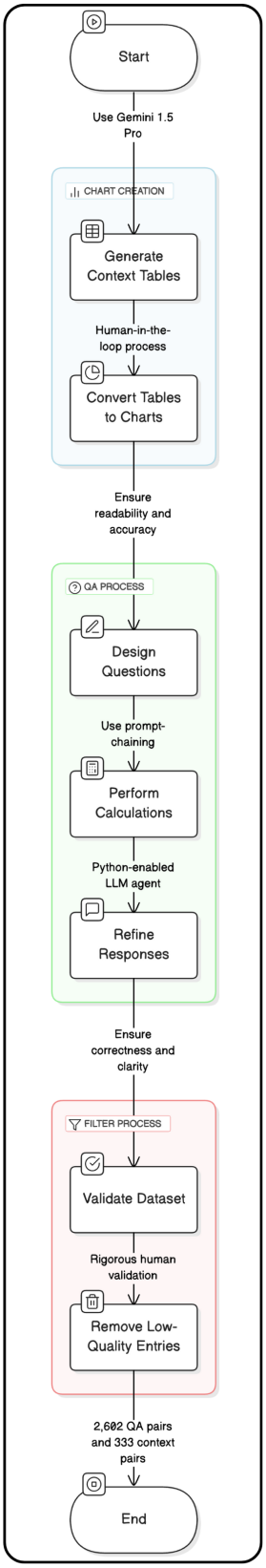

Figure 5: Flowchart for ***S2***: Synthetic Simulation - Multi-Chart Reasoning with LLM-Generated Contexts

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

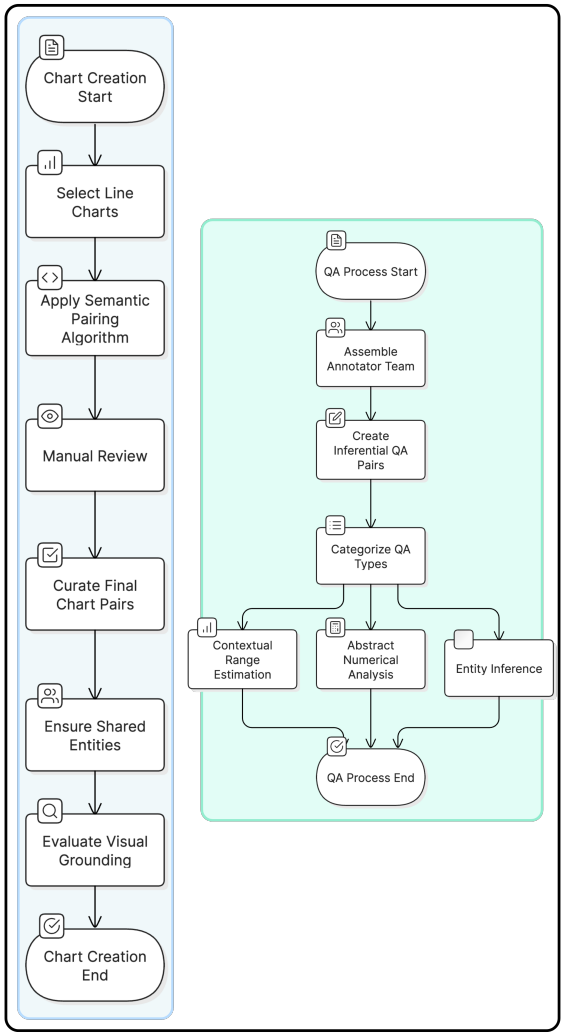

Figure 6: Flowchart for ***S3:*** Ceiling Performance - Evaluating VLM Inference Under Extreme Complexity

Algorithm 3 ***S3:*** Ceiling Performance - Evaluating VLM Inference Under Extreme Complexity

Input: Chart Repository C_{repo}, Semantic Pairing Algorithm S_{pair}, Annotator Team A
Output: Dataset D with Chart Pairs and QA Pairs
▷ Step 1: Chart Pairing and Preprocessing
$C_{pairs} \leftarrow \emptyset$
for each chart C in C_{repo} **do**
    Identify Metadata Attributes M_C
    Apply S_{pair} to find a semantically linked chart C' with shared entities
    if Valid Semantic Relationship Exists **then**
        Add (C, C') to C_{pairs}
    end if
end for
Perform Manual Refinement on C_{pairs} for relational coherence
▷ Step 2: Inferential QA Generation
$QA \leftarrow \emptyset$
for each (C, C') in C_{pairs} **do**
    for each annotator a in A **do**
        Generate Questions in Three Categories:
        1. Contextual Range Estimation (Value Range Evaluation)
        2. Abstract Numerical Analysis (Arithmetical & Logical Deductions)
        3. Entity Inference (Pattern Recognition Across Charts)
    end for
end for
▷ Step 3: Dataset Compilation
$D \leftarrow \{C_{pairs}, QA\}$
return D

B Additional Results

B.1 Qwen Results Table

Model	Zero-Shot				Zero-Shot CoT				Few-Shot CoT_D			
	Net	*S1*	*S2*	*S3*	*Net*	*S1*	*S2*	*S3*	*Net*	*S1*	*S2*	*S3*
Combined Visual Context Image												
GPT-4o-V	35.0	55.3	37.4	12.4	39.5	61.2	39.5	**17.9**	40.7	61.7	41.7	18.8
Gemini-1.5-Pro-V	**43.6**	**62.1**	**54.5**	14.1	**45.6**	**67.0**	**54.9**	14.8	**48.3**	**68.6**	**57.2**	**19.5**
Qwen2-VL-7B	31.7	48.0	29.0	**18.2**	31.3	52.5	29.8	11.5	30.3	50.1	29.8	11.0
MiniCPM-V-2	26.0	45.2	25.6	7.1	26.5	45.4	25.1	9.0	26.2	45.2	23.4	10.0
InternVL-2-8B	21.1	35.1	19.6	8.6	22.2	38.6	21.8	6.1	25.7	41.4	24.1	11.6
Idefics3-8B-Llama3	22.8	38.1	18.8	11.5	22.7	37.7	19.0	11.3	22.5	35.0	20.4	12.2
Interleaved Visual Context												
GPT-4o-V	39.4	61.3	42.4	14.5	41.9	65.8	42.7	**17.3**	43.6	65.6	45.5	**19.6**
Gemini-1.5-Pro-V	**44.6**	**67.0**	**50.9**	**15.8**	**44.8**	**68.1**	**53.8**	12.5	**48.0**	**70.3**	**54.1**	19.5
Qwen2-VL-7B	30.5	46.3	29.5	15.7	30.6	51.4	32.6	7.8	28.7	48.7	28.8	8.6
MiniCPM-V-2	30.5	52.3	29.6	9.7	29.8	49.9	29.4	10.0	29.9	49.8	28.7	11.2

Table 12: Qwen Baseline Accuracies using our evaluation method with Gemini-1.5 Eval Engine on All Models and Strategies broken down by Set Type Wise (*S1, S2, S3*) and Strategy wise. The highest values are highlighted.

C Prompts

C.1 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}
```

C.2 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}
```

C.3 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.
```

C.4 Table-Based Question Answering Prompt

Table-Based Question Answering 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}
```

C.5 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>'.
```

C.6 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. Think
step by step and make sure you arrive at the correct answer for the given
question.

Step 4: Provide the Final Answer
Based on your reasoning, provide the final answer in the following format:
Final Answer: <final_answer>

Here's are a few examples of reasoning using the given directives:
Example 1
Chart Provided: You are shown a chart representing the monthly sales figures
of four products (Product A, Product B, Product C, and Product D) across six
months.
Question: Which product had the highest average sales over the six months?

Model's Response:
Step 1: The relevant entities to focus on are the monthly sales figures for
Product A, Product B, Product C, and Product D.
Step 2: Extract the sales values for each product across all six months from
the chart.
Step 3: Calculate the average sales for each product by summing the sales
values across the six months and dividing by six. Compare the averages to
determine which product had the highest average sales.
Step 4: Final Answer: The product with the highest average sales is <Product X>.

Question: {question}
```

C.7 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 to the Ground Truth answer
    (e.g., if the Ground Truth answer is 20.24553 and the Student Written
    Answer is 20.24, it is correct). State these assumptions clearly in your
    reasoning.
*   For questions involving ranges,
if the model's answer falls within the ground truth range,
consider it correct.
*   Provide a brief explanation of your reasoning within `<reasoning>` tags.
*   1 means the Student Written Answer is correct. 0 means the Student Written
    Answer is incorrect.
*   State your final decision (1 or 0) within `<answer>` tags.

**Example:**

Question: "What is the color of water?"
Ground Truth answer: "Pink"
Student Written Answer: "Final Answer: Water is colorless."

Response: `<reasoning> The student answer does not match with the given ground
truth. As a result, the answer is wrong.</reasoning>`
`<answer> 0 </answer>`

**Now, answer the following:**

Question: {question}
Ground Truth answer: {ground_truth}
Student Written Answer: {student_answer}
```