General Comments

This document outlines the revisions made in our July 2025 resubmission in response to the February 2025 ACL ARR submission (ID: 8197). Reviewers and area chairs raised concerns about theoretical vagueness, inconsistent terminology, opaque data creation processes, and weak methodological grounding. In this revision, we substantially restructured the framing, benchmark design, and analysis pipeline. We eliminated the informal "cognitive load" framing, introduced a measurable complexity model, standardized terminology, and provided full transparency for annotation, evaluation, and prompting protocols. Below, we describe the major changes with references to relevant sections and tables in the updated version.

Replaced the Vague "Cognitive Load" Framing with an Operational Design. In the original submission, we used the term "cognitive load" as a proxy for complexity without defining it theoretically or operationalizing it in our dataset design. This framing lacked rigor and interpretability. In the revised version, we abandoned this terminology entirely and instead structured the benchmark around three explicitly measurable dimensions of task complexity:

- **Visual Density:** number of visual elements and their compactness (e.g., crowdedness of charts);
- **Chart Integration:** whether the question requires reasoning over one chart or synthesizing information from multiple sources;
- **Reasoning Complexity:** the type and number of logical operations required to answer a question (e.g., lookup vs. correlation over time).

This new framework is introduced in Section 2 (lines 151–193) and aligns with the design of our three benchmark subsets: DECAF (low-density single-chart lookup), SPECTRA (paired synthetic charts requiring symbolic integration), and STORM (real-world, high-density charts with multi-hop logic). This change directly addresses concerns from Reviewer **giXd**, Reviewer **5dzc**, and the AC's meta-review regarding unclear theoretical motivation.

Defined Research Questions and Aligned Them with Evaluation. The original version buried research questions within the narrative, making it difficult to trace how they motivated the experimental setup. In the revised version, we explicitly define four research questions (RQ1–RQ4) in the Introduction (lines 100–126). These correspond to concrete evaluation setups:

- **RQ1:** Does decomposition help? → Tested using DECAF results in Table 6 (Section 5.1).
- **RQ2:** How does complexity impact VLM accuracy? → Analyzed using overall trend plots in Figure 3 and Table 6 across all subsets (Section 5.1).
- **RQ3:** Can prompting strategies improve robustness? → Investigated in Table 7 (Section 5.3).
- **RQ4:** Do structured table representations improve performance? → Evaluated in Section 4.2 via baseline comparison.

This structure enhances clarity and addresses the request from Reviewer **5dzc** for stronger hypothesis–experiment alignment.

Standardized Naming and Terminology. The previous version used inconsistent or ambiguous phrases (e.g., "compound charts," "cognitive ceiling," "simulated charts"). In the revision, we introduced standardized subset names that appear in all figures, tables, and evaluation references:

- **DECAF:** Decomposed Elementary Charts with Answerable Facts
- **SPECTRA:** Synthetic Plots for Event-based Correlated Trend Reasoning and Analysis
- **STORM:** Sequential Temporal Reasoning Over Real-world Multi-domain Charts

This update (Section 2, lines 194–210 and throughout Section 4) directly addresses Reviewer **mniS**'s concern about unclear terminology.

Fully Documented Annotation and Generation Pipelines. The February submission vaguely mentioned "LLM filtering" and "human review." In the new version, we clearly describe generation, validation, and arbitration protocols for each subset:

- **DECAF:** QA pairs were generated from decomposed real charts and filtered using Gemini 1.5 Pro (Team 2024) for logical answerability, clarity, and correctness (Appendix C.1).
- **SPECTRA:** Prompt chaining was used with a Python-enabled LLM agent to evaluate numerical correctness (Appendix C.2). This approach is inspired by recent tool-use LLMs like InternVL-2 (Chen et al. 2024).
- **STORM:** Annotators wrote and verified questions over real-world charts. We recruited 11 graduate-level annotators (NLP/CV background) and used a double-pass verification process (Appendix C.3).

We added annotation templates, acceptance criteria, and protocols in Appendix C. Agreement scores are reported in Table 5 (Cohen's $\kappa = 0.71$, Jaccard index = 94.75

Introduced Qualitative Error Analyses. Section 5.4 now provides detailed examples and error clusters. These include:

- Gemini 1.5 Pro hallucinating time trends in SPECTRA;
- Qwen2-VL (Yang et al. 2024) misclassifying chart types in STORM;
- Correct DECAF responses with interpretable chain-of-thought outputs.

These qualitative cases enhance interpretability and fulfill Reviewer **mniS**'s request for deeper model diagnosis.

Added Prompting Protocols and Dual-Evaluation Setup. In Section 5.3 and Appendix A.7, we explain prompt design choices (Zero-Shot, CoT, Few-Shot-CoTD) and model evaluation protocols. We used a dual-model evaluation strategy involving Gemini Flash and Qwen2.5 (Yang et al. 2024), with 88

Added Comparative Benchmark Table. Section 2 (Table 1) now compares InterChart against ChartQA (Masry et al. 2022), PlotQA (Methani et al. 2020), and MultiChartQA (**?**). InterChart supports:

- Decomposition (via DECAF);
- Executable reasoning (via SPECTRA);
- Real-world chart density (via STORM).

This directly addresses Reviewer **giXd**'s question on novelty vs. MultiChartQA.

Expanded Tables and Error Breakdown Visuals. Section 5.5 and Appendix A.8 include:

- Table 8: Chart-type accuracy (bar, line, box, scatter);
- Table 9: Reasoning-type accuracy (correlation, lookup, estimation);
- Table 10: Error categories (hallucination, grounding failure, arithmetic).

These detailed breakdowns address Reviewer **mniS**'s call for improved interpretability.

Improved Structure and Writing. We restructured Sections 2 and 4 for logical coherence. Captions, references, and table cross-links were cleaned and standardized throughout. This improves readability and directly addresses Reviewer **yu5k**'s concerns.

Reproducibility and Release Plan. We now include:

- Prompt templates for Gemini, Qwen2-VL, GPT-4o Mini (OpenAI 2024), MiniCPM (Hu et al. 2024), InternVL-2 (Chen et al. 2024), and Idefics3 (Laurençon et al. 2024);
- Intermediate LLM generations (Appendix B);
- Filtering and arbitration scripts (Appendix C).

Code and dataset will be publicly released.

Reviewer and AC Coverage. We explicitly address the following:

- Reviewer **mniS**: terminology inconsistencies, missing error examples, unclear annotation processes;
- Reviewer **giXd**: vague framing, lack of novelty;
- Reviewer **yu5k**: need for transparency, readability, and clearer pipeline documentation;
- Reviewer **5dzc**: weak research question structure and missing prompting strategy explanation;
- Meta-review (AC): lack of clarity in framing and evaluation, reproducibility concerns—all addressed through major revisions to framing (Section 2), evaluation design (Sections 4–5), and reproducibility (Appendices A.6–C).

References

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 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*, 24185–24198.

Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; Zhang, X.; Thai, Z. L.; Zhang, K.; Wang, C.; Yao, Y.; Zhao, C.; Zhou, J.; Cai, J.; Zhai, Z.; Ding, N.; Jia, C.; Zeng, G.; Li, D.; Liu, Z.; and Sun, M. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Laurençon, H.; Marafioti, A.; Sanh, V.; and Tronchon, L. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Masry, A.; Long, D. X.; Tan, J. Q.; Joty, S.; and Hoque, E. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

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

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

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

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2406.04852*.