Evaluating VLMs as Accessibility Bridges for Process Visualizations

Kazi Tasnim Zinat, Saad Mohammad Abrar, Sharmila Duppala, Saimadhav Naga Sakhamuri, Zhicheng Liu
University of Maryland College Park
{kzintas, sabrar, sduppala, leozcliu}@umd.edu

Abstract

Vision-Language Models (VLMs) show promise as accessibility tools that can transform data visualizations into natural language descriptions for users who have difficulties interpreting the visuals. Yet, their ability to interpret process visualizations remains underexplored, as such visualizations are largely absent from existing VQA benchmarks. In this work, we systematically evaluate five VLMs on more than 100 expert-generated QA pairs across 21 process visualizations spanning three types of visual structures. Results show that while frontier models like GPT-4.1 and Gemini 2.0 perform well, open-source models struggle—especially with aggregation and negation tasks. The figures and QA dtasets are available on GitHub.

1. Introduction

Data visualizations serve as powerful tools for understanding and communicating complex information, but remain largely inaccessible to people with visual impairments. Vision-Large Models (VLMs) present promising opportunities to bridge this accessibility gap by automatically generating descriptive text annotations and answering specific questions on visualizations. For visually impaired users, such capabilities could transform previously inaccessible visual information into comprehensible insights delivered through assistive technologies.

Although recent advances in VLM-based chart interpretation have shown promise, there has been little research on whether and how modern VLMs can interpret process visualizations effectively. Unlike statistical charts that summarize quantitative data, process visualizations depict time-ordered workflows and causal relationships. They are widely used in domains like healthcare, manufacturing, education, and cyber security to support the analysis and illustration of sequential activities. Despite the widespread use of process visualizations, most visual question answering (VQA) benchmarks remain limited to common statistical chart types [4–6].

To address this gap, we presents a systematic evaluation of five VLMs—two proprietary and three open-source—across three types of process visualizations and four task types. The results reveal both notable strengths and critical limitations in their ability to interpret process-oriented diagrams.

2. Methodology

2.1. Dataset Description

Our evaluation focuses on 21 process visualizations of mined frequent patterns from two event sequence datasets from healthcare and sports domains [7], spanning three types of visual structures (Fig. 1): *tree-based* branching patterns, where each node has a single parent and potentially multiple children [3], *graph-based* patterns allowing nodes to have multiple incoming edges [2], and *linear sequence clusters* that group similar sequences into representative patterns [1].

We manually generated 144 multiple-choice questions (6-8 questions per visualization), each with four answer options. Ground-truth answers were annotated by one expert and cross-validated by another to ensure accuracy. The questions were categorized into four task types: *Value Extraction* (identify numerical values associated with nodes or edges), *Sequential Reasoning* (determine antecedent–sequela relationships between events), *Value Aggregation* (compute totals across multiple nodes or edges), and *Negative Cases* (identify values or relationships absent from the visualization).

2.2. Experiment Setup

We evaluated five VLMs (Gemini 2.0, GPT 4.1, Gemma 3, Pixtral and Qwen 2.5 VL) using a consistent zero-shot prompting strategy. For each type of process visualization, the models received a tailored prompt explaining how to interpret the diagram, along with explicit instructions to reason step-by-step before selecting an answer. Each model was given only the visualization and the question—without access to underlying data or exact values—reflecting real-world conditions and offering a more challenging, realistic test of VLMs as accessibility tools.

Performance was measured by accuracy across task types and visualization styles. In addition to quantitative

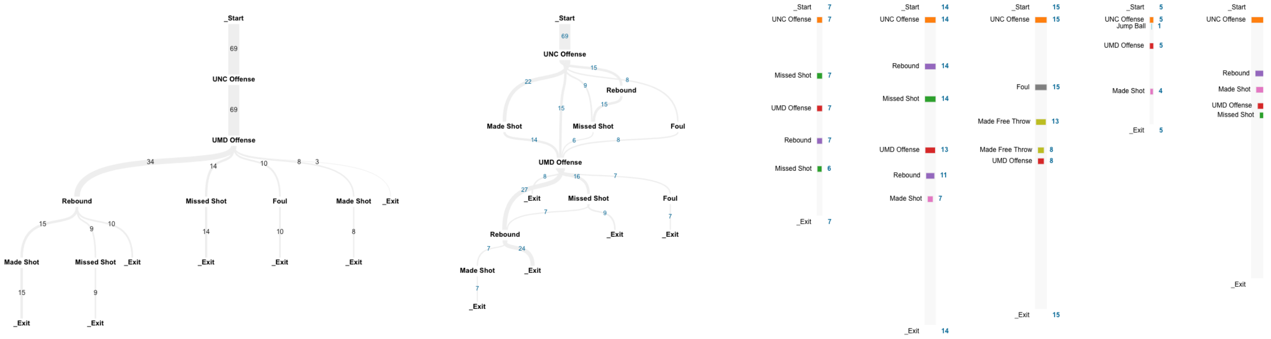

Figure 1. Example process visualizations across three types of visual structures: (left) *tree-based*, (center) *graph-based*, and (right) *linear sequence clusters* summarizing event flows. These charts depict basketball play sequences and are used to evaluate VLMs' capabilities in interpreting complex sequential structures.

Task	#Q	Gemini 2.0	GPT-4.1	Gemma 3 (27B)	Pixtral (12B)	Qwen 2.5 VL (7B)
Value Extraction	16	**100.00%**	*93.75%*	87.50%	62.50%	62.50%
Sequential Reasoning	51	*90.20%*	**94.12%**	88.24%	66.67%	86.27%
Value Aggregation	54	*74.07%*	**75.93%**	61.11%	44.44%	20.37%
Negative Cases	23	*78.26%*	**82.61%**	21.74%	21.74%	17.39%
Total	144	*83.33%*	**85.41%**	67.36%	50.69%	47.92%

Table 1. Table shows the accuracy of five VLMs across four task types on node-link event sequence visualizations. Higher values indicate better performance. **Bold** and *italic* entries denote the best and second-best performers, respectively. GPT-4.1 and Gemini 2.0 consistently outperform open-source models.

metrics, we conducted a qualitative analysis of the reasoning traces to identify common errors and barriers to effective interpretation, especially in sequential reasoning tasks.

3. Results

Table 1 shows that proprietary models—GPT-4.1 and Gemini 2.0—consistently outperformed others across all tasks.

Gemini 2.0 achieved perfect accuracy on *Value Extraction*, while GPT-4.1 excelled in both *Sequential Reasoning* and *Value Aggregation*, the latter being the most challenging task overall. *Negative Cases* showed the widest performance gap: GPT-4.1 led with 82.61%, while open-source models scored below 22%, highlighting a key weakness in detecting absent information.

Although all models generally struggled with *Value Aggregation*, especially on information-dense visuals, we observed exceptions that challenge overall performance trends. For example, in the graph-based visualization shown in Figure 1, models were asked: *"How many rebounds were made after UMD Offense?"*—a task requiring aggregation of values ($27 + 7 = 34$) across multiple paths. Surprisingly, proprietary models like GPT-4.1 and Gemini 2.0 failed to compute the correct total, while open-source models such as Gemma-3 (27B) and Pixtral (12B) were successful.

4. Conclusion

Improving VLM capabilities for interpreting process visualizations is crucial for inclusive information access. This work presents the first systematic evaluation of their performance on such visualizations. While we identify important strengths and limitations, further research is needed to uncover deeper insights for effective integration into accessibility-focused applications.

References

[1] Yuanzhe Chen, Panpan Xu, and Liu Ren. Sequence synopsis: Optimize visual summary of temporal event data. *IEEE transactions on visualization and computer graphics*, 24(1):45–55, 2017. 1

[2] Mengdie Hu, Krist Wongsuphasawat, and John Stasko. Visualizing social media content with sententree. *IEEE transactions on visualization and computer graphics*, 23(1):621–630, 2016. 1

[3] Zhicheng Liu, Bernard Kerr, Mira Dontcheva, Justin Grover, Matthew Hoffman, and Alan Wilson. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, pages 527–538. Wiley Online Library, 2017. 1

[4] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[5] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020.

[6] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023. 1

[7] Kazi Tasnim Zinat, Jinhua Yang, Arjun Gandhi, Nistha Mitra, and Zhicheng Liu. A comparative evaluation of visual summarization techniques for event sequences. *Computer Graphics Forum*, 42(3):173–185, 2023. 1