Benchmarking Multimodal Large Language Models for Scientific Visualization Literacy

Patrick Phuoc Do* Chau M. Ta† Chaoli Wang‡

University of Notre Dame

ABSTRACT

Multimodal large language models (MLLMs) are increasingly used to interpret visualizations, yet current evaluations remain largely chart-centric and provide limited evidence of understanding of scientific visualization (SciVis). We benchmark six MLLMs on the scientific visualization literacy assessment test, a standardized SciVis literacy assessment comprising 49 items based on 18 scientific visualizations and illustrations, spanning 8 techniques and 11 task types. We evaluate three closed-source and three open-source models under a closed-world protocol and compare their performance using data from 485 human participants. Results show that current MLLMs do not exhibit uniform SciVis literacy. Gemini is the strongest model overall, exceeding the human mean across the evaluated subsets, whereas the open-source models remain below the human baseline. Performance is highly uneven across techniques and tasks: models perform best on scientific illustration, search, and spatial understanding, but struggle on texture-based and integration-based visualizations and on quantitative estimation. Error analysis reveals recurring failures in fine-grained quantitative estimation, flow-direction interpretation, and grounded encoding interpretation. These findings position SciVis literacy as a necessary benchmark dimension for evaluating multimodal AI systems. Our code and model outputs are publicly available at https://github.com/patdmp/mllm-scivis-lit-benchmark.

Keywords: Scientific visualization, visualization literacy, multimodal large language model

1 INTRODUCTION

Scientific visualization (SciVis) has long served as an important medium for exploring, analyzing, and communicating complex scientific data, including simulation output, medical imaging, climate data, and other spatially grounded scientific phenomena [19, 25]. In contrast to information visualization (InfoVis), SciVis typically represents data with intrinsic spatial or physical structure. This distinction is closely tied to the notions of physically based data and given spatialization, which differentiate much of SciVis from InfoVis [31]. In addition, when interpreting SciVis, viewers need to reason about spatial structure, multivariate relationships, and time-varying or physically meaningful patterns encoded in the visual representation [19, 31].

Meanwhile, multimodal large language models (MLLMs) have rapidly advanced in their ability to process images alongside text and to support visualization interpretation, scientific communication, and visual analytics workflows [1, 2, 34]. However, existing evidence suggests that their performance on visualization tasks remains uneven, and current evaluations have focused primarily on InfoVis rather than SciVis [4, 16]. As a result, whether current MLLMs possess *SciVis literacy* remains an open question [32].

To address this gap, we benchmark MLLMs using the scientific visualization literacy assessment test (SVLAT) [11], a recently introduced instrument designed specifically to measure how well non-experts read, understand, and interpret scientific visualizations and illustrations. Because SVLAT spans multiple SciVis techniques and tasks under a closed-world design, it provides a suitable benchmark for evaluating SciVis understanding in MLLMs.

Using SVLAT, our goal is to examine whether current MLLMs demonstrate measurable SciVis literacy and, if so, where their strengths and weaknesses lie across techniques and task categories. More broadly, we position SciVis literacy as a missing benchmark dimension for multimodal AI systems. By doing so, we aim to contribute a clearer evaluation target for future work on trustworthy model assessment, model-assisted scientific communication, and human–AI interaction around scientific visual content.

2 RELATED WORK

Visualization literacy assessment. Visualization literacy refers to a viewer's ability to read, understand, and interpret data visualizations [5, 6, 23]. Boy et al. [7] introduced a principled approach based on item response theory (IRT) to assess visualization literacy, showing that literacy can be modeled as a latent ability rather than treated solely as raw accuracy. Building on this psychometric perspective, Lee et al. [21] developed VLAT, a standardized assessment for non-expert users that measures literacy across 12 visualization types and 8 visualization tasks. CALVI [14] extended the notion of visualization literacy beyond routine reading by focusing on critical thinking about misleading or erroneous visualizations. To improve practicality and reduce administrative time, later work introduced shorter, more specialized instruments. For instance, Mini-VLAT [28] condensed VLAT into a brief but effective short-form test while preserving much of its measurement utility. Adaptive versions of VLAT and CALVI [9] reduced testing burden while retaining reliability and validity. These studies show that visualization literacy is a mature topic for assessment within InfoVis.

SciVis literacy. Compared with InfoVis, SciVis introduces additional interpretive demands. Scientific visualizations often encode data with inherent spatial and physical structure, requiring users to reason about geometry, depth, topology, motion, and relationships among scalar, vector, or tensor fields. Prior work on SciVis tasks has emphasized the importance of task taxonomies tailored to scientific data analysis, especially for volume data and spatial reasoning [20]. These characteristics make SciVis literacy qualitatively different from InfoVis literacy in the context of conventional charts and graphs. Despite the importance of SciVis in scientific analysis and communication, standardized assessment in this area has remained largely absent until recently. SVLAT [11] is a recently proposed psychometrically grounded instrument for measuring SciVis literacy. SVLAT is designed around 8 canonical SciVis techniques and 11 interpretation tasks, enabling literacy to be measured in a way that better reflects the demands of scientific visual representations. Because SVLAT is standardized, validated, and explicitly constructed for scientific visual material, it provides an appropriate foundation for benchmarking MLLMs on SciVis literacy rather than

*e-mail: mdo23@nd.edu
†e-mail: cta@nd.edu
‡e-mail: chaoli.wang@nd.edu

relying on chart-oriented proxies.

MLLMs and visualization understanding. Work on machine understanding of visualized data initially developed through benchmark datasets for chart and figure question answering. Early datasets such as FigureQA [18] and DVQA [17] focused on synthetic scientific-style figures and bar charts, respectively, establishing controlled testbeds for visual reasoning over plotted data. PlotQA [26] moved closer to realistic scientific plots by introducing open-vocabulary and numerical answers, while ChartQA [24] expanded the task to charts requiring both visual and logical reasoning. More recent benchmarks have further stressed the limitations of current MLLMs. CharXiv [33] evaluated realistic charts collected from scientific papers and showed that prior benchmarks can overestimate model capability, and MultiChartQA [35] studied multi-chart reasoning, where models must integrate information across several related visualizations. Beyond chart-only settings, SPIQA [29] evaluated multimodal question answering over figures and tables embedded in scientific papers, extending the problem toward document-grounded scientific figure understanding.

Within the visualization community, recent work has begun to study MLLMs through the lens of visualization literacy itself. Bendeck and Stasko [4] evaluated GPT-4 on a suite of visualization-literacy-related tasks and found a mixed pattern of strengths and weaknesses: the model could identify higher-level trends and extreme values, but it struggled with precise value retrieval, color discrimination, and consistency. Hong et al. [16] then examined whether LLMs truly possess visualization literacy by testing them on modified VLAT-style visualizations designed to probe generalization and reliance on visual evidence rather than prior knowledge. Their findings suggest that model performance can degrade substantially under benign visual modifications, raising concerns about robustness and genuine visual grounding. Das et al. [10] further showed that structured prompting can substantially improve performance on VLAT-style tasks, highlighting that measured capability depends not only on the model itself but also on the evaluation protocol. Complementing these studies, Varona et al. [13] analyzed visualization-literacy barriers in MLLMs by qualitatively coding erroneous responses, providing a taxonomy of failure modes that helps explain why models struggle even when benchmark accuracy appears competitive. Dong and Crisan [12] further probed the internal reasoning behavior of vision-language models for chart understanding, arguing that answer correctness alone is insufficient for characterizing visualization literacy in models.

Most existing datasets, evaluations, and prompting studies focus on InfoVis-style charts or chart-like figures rather than SciVis-native representations. Together, these studies show that evaluation of MLLMs on visualization tasks remains largely chart-centric, leaving SciVis-native assessment comparatively underexplored.

3 Methodology

Assessment instrument. We evaluate MLLMs using SVLAT [11], a standardized assessment of SciVis literacy. SVLAT consists of 49 items based on 18 scientific visualizations and illustrations, including static images and animations, spanning 8 SciVis techniques and 11 task types. It follows a closed-world design in which each item is intended to be answerable using only the provided visualization and caption, making it suitable for assessing SciVis literacy in MLLMs while limiting reliance on external knowledge.

Model selection and configuration. We evaluate both closed-source and open-source MLLMs. The closed-source models are GPT-5.4 (GPT) [27], Claude-Opus-4.6 (Claude) [3], and Gemini-3.1-Pro-Preview (Gemini) [15], all accessed via the OpenRouter API for consistent deployment. For animation items, GPT and Claude were evaluated using frame extraction (1 frame per second), as their APIs did not support direct video input at the time of evaluation. The open-source models are Qwen3.5-9B (Qwen) [30], InternVL3.5-8B (InternVL) [8], and LLaVA-OneVision-1.5-8B-Instruct (LLaVA-OneVision) [22]. All models were evaluated with *temperature* set to 0 and *max_tokens* set to 300 to improve reproducibility and consistency across models and runs. Setting the *temperature* to 0 makes decoding as deterministic as possible, thereby reducing sampling variability in both answer selection and rationale generation. We set *max_tokens* to 300 to ensure sufficient space for the required JSON output and brief explanation while avoiding unnecessarily long responses. Because some backend nondeterminism may remain, we repeated each evaluation 10 times and report the averaged results.

Prompt design. We use a single instruction prompt for all models to standardize the evaluation setting. The prompt frames the model as a SciVis expert and explicitly constrains it to use only the provided visualization and caption. It also requires the model to return a structured JSON object containing the selected option, the full answer text, and a brief rationale. To discourage unsupported guessing, the prompt includes a `"Not sure"` option when the answer cannot be determined confidently from the given evidence. The prompt used in our experiments is shown in Figure 1. This prompt is designed to align with the closed-world principle of SVLAT and to make model outputs easier to parse automatically.

Evaluation protocol. We evaluate each model on SVLAT one question at a time. For each item, we run the model 10 times using different random seeds and compute the average score across runs. This repeated-run protocol helps reduce the impact of residual sampling variability or backend nondeterminism. In addition to the selected answer, we record the model's rationale for each response. These explanations are used to examine whether the model's reasoning is grounded in the provided visualization and caption rather than unsupported prior knowledge or speculation. Although the rationale is not used directly for scoring accuracy, it provides qualitative evidence about how the model arrives at its answers and whether incorrect responses stem from misinterpretation of visual evidence, overconfident guessing, or reliance on external knowledge. We also analyze performance by SciVis techniques and task categories to identify systematic strengths and weaknesses. To contextualize MLLM capabilities, we compare accuracy against the human performance baseline from the original SVLAT study [11], established via an online tryout with 485 non-expert participants (246 females; ages 18–65; education ranging from high school to postgraduate).

Table 1: Overall accuracy (mean ± std) per model split by item format. Image = static image items; Animation = animation items; All = combined. For GPT-5.4 and Claude-Opus-4.6, animation items were evaluated via frame extraction.

	Accuracy (%)		
Model	Image	Animation	All
GPT-5.4	75.7 ± 5.5	73.6 ± 10.4	75.1 ± 4.9
Claude-Opus-4.6	81.7 ± 5.8	59.3 ± 11.7	75.3 ± 5.5
Gemini-3.1-Pro-Preview	90.9 ± 4.3	82.9 ± 8.5	88.6 ± 3.9
Qwen3.5-9B	69.4 ± 6.0	67.1 ± 9.4	68.8 ± 5.0
InternVL3.5-8B	60.9 ± 6.7	72.9 ± 9.9	64.3 ± 5.6
LLaVA-OneVision-1.5-8B-Instruct	60.3 ± 6.9	73.6 ± 9.7	64.1 ± 5.8
Human	76.2 ± 2.4	73.9 ± 4.7	75.6 ± 2.2

4 Results

4.1 Overall Performance

Table 1 reports the overall performance of the six evaluated MLLMs across 10 runs and of 485 human participants on SVLAT. Gemini is the strongest model by a wide margin, achieving 90.9% on image items, 82.9% on animation items, and 88.6% overall, exceeding human performance of 76.4%, 74.5%, and 75.9%, respectively. The three open-weight models trail humans in combined accuracy, with Qwen performing best (68.8% overall), followed by InternVL (64.3%) and LLaVA-OneVision (64.1%). On animation items, Gemini remains the top-performing model. Via frame extraction, GPT matches human-level accuracy (73.6% vs. 73.9%), while Claude scores the lowest of all models and humans at 59.3%. Among

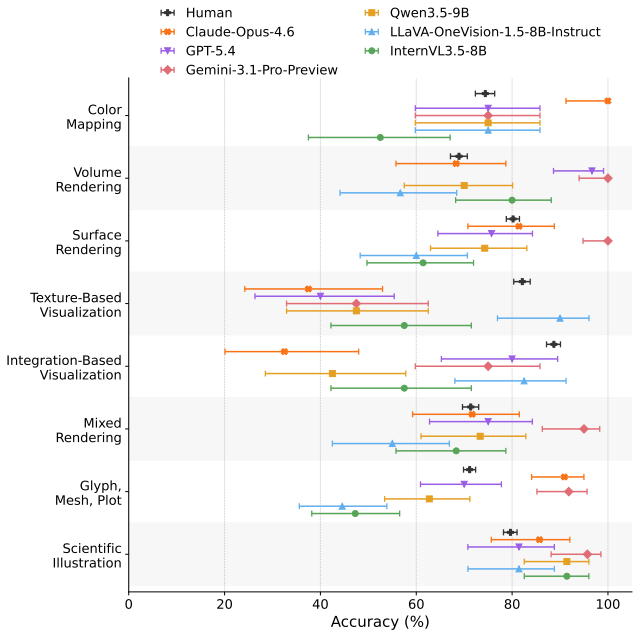


Figure 1: Model and human performance across different techniques in SVLAT. Each dot shows the mean accuracy, with the bars indicating 95% confidence intervals.

the open-source models, LLaVA-OneVision (73.6%) and InternVL (72.9%) come close to the human level, and both perform better on animation items than on static image items, whereas Qwen remains lower at 67.1%. MLLMs exhibit greater performance heterogeneity across item types (SD: 3.9–11.7) than humans (SD: 2.2–4.7).

4.2 Performance by Techniques

Human and model performance differ substantially across techniques, as shown in Figure 1. Humans exhibit the most balanced pattern, with mean accuracy ranging from 68.9% to 88.8% across all techniques, whereas the MLLMs show much larger fluctuations. Among the evaluated models, Gemini is the strongest and broadest-performing system. It achieves 100.0% on Surface Rendering and Volume Rendering, 95.0% on Mixed Rendering, 95.7% on Scientific Illustration, and 91.8% on Glyph, Mesh, Plot, exceeding human performance on the same techniques. Closed-source models consistently outperform open-source models across Surface Rendering, Mixed Rendering, and Glyph, Mesh, Plot, with the gap being most pronounced on the latter.

Across all techniques, Scientific Illustration is the most accessible to humans and all models. Human accuracy reaches 79.5%, while all models perform strongly, ranging from 81.4% to 95.7%. In contrast, Texture-Based Visualization and Integration-Based Visualization reveal a clearer gap between human and model performance. Humans achieve high accuracies on both techniques, at 82.0% and 88.8%, respectively. LLaVA-OneVision is the main exception on Texture-Based Visualization (90.0%), where all other models fall substantially below it. On Integration-Based Visualization, LLaVA-OneVision (82.5%), GPT (80.0%), and Gemini (75%) remain competitive with humans, whereas other models are notably lower. Per-item accuracies for MLLMs and humans, categorized by techniques, are furnished in Appendix A.

4.3 Performance by Tasks

Figure 2 shows that performance varies substantially by task. Across all categories, Search and Spatial Understanding tasks are the most accessible, whereas Quantitative Estimation–Relative Estimation (Quantitative) is the most difficult; indeed, it is the hardest task for every category, with accuracy remaining below

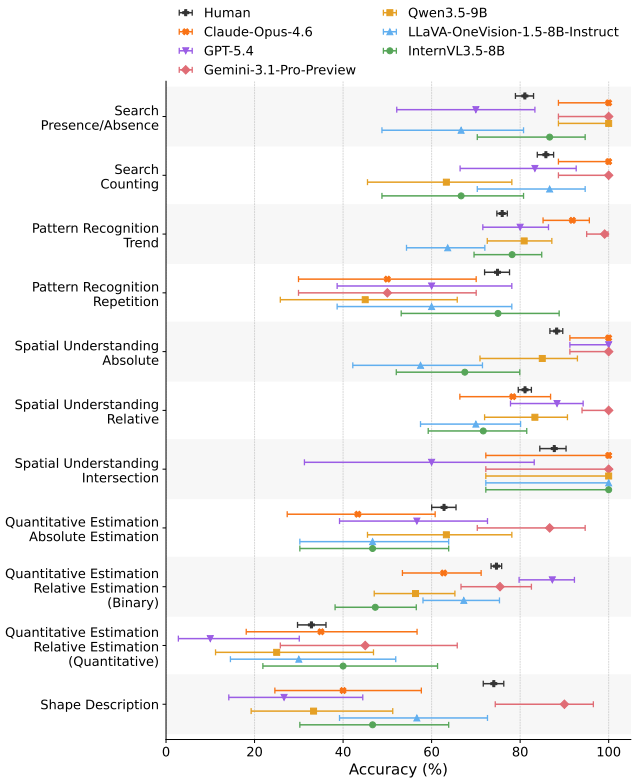


Figure 2: Model and human performance across different tasks in SVLAT. Each dot shows the mean accuracy, with the bars indicating 95% confidence intervals.

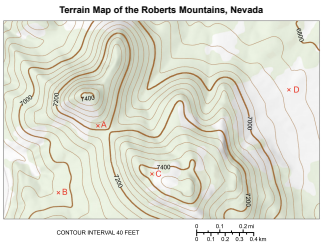
50.0% throughout. Humans perform particularly well on Spatial Understanding–Absolute (88.1%), Spatial Understanding–Intersection (87.7%), and Search–Counting (85.6%), and the models follow the same broad pattern.

Among all models, Gemini performs best, achieving the highest accuracy on nearly every task. The main exception is Pattern Recognition–Repetition and Quantitative Estimation–Relative Estimation (Quantitative), where Gemini drops to 50.0% and 45.0%, respectively. For Spatial Understanding–Absolute, all three closed-source models max out, whereas open-source models are notably lower. The open-source models are much less competitive with humans. Qwen and InternVL exceed human performance only on Search–Presence/Absence and Spatial Understanding–Intersection, while LLaVA-OneVision exceeds humans only on Spatial Understanding–Intersection. On the remaining tasks, the open-source models perform at or below the human level. It is worth noting that Quantitative Estimation–Relative (Quantitative), Spatial Understanding–Intersection, and Pattern Recognition–Repetition are each represented by only one or two items in the current benchmark, so the above findings for these categories should be treated as preliminary.

4.4 Error Analysis

The preceding results show that MLLM performance on SVLAT varies substantially across techniques, tasks, and model families. To better understand the sources of these differences, we qualitatively examine model rationales to identify recurring patterns in how models interpret, misinterpret, or fail to ground visual evidence.

Fine-grained quantitative misestimation. Models struggle with precise quantitative judgment when the answer depends on careful visual estimation rather than coarse pattern recognition. In Item 3, Gemini correctly recognized that "*the map's contour interval is 40 feet*" and located waypoint D, stating that it "*lies between the 6880 ft and 6840 ft contour lines*" (see Figure 3). However, it

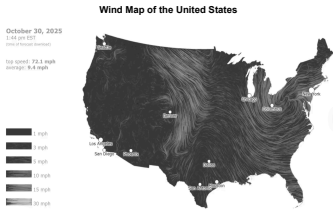


Item 3:
By approximately how many feet does the elevation of waypoint A exceed that of waypoint D?

Gemini: "*The map's contour interval is 40 feet.* ***Waypoint A is located exactly on the thick index contour line labeled 7200,*** *giving it an elevation of 7200 ft. Waypoint D is located downhill from the 7000 ft index contour. Counting down by 40 ft intervals,* ***D lies between the 6880 ft and 6840 ft contour lines,*** *giving it an approximate elevation of 6860 ft. [...]*"

Figure 3: Example of fine-grained quantitative estimation difficulty. MLLM correctly identified the contour interval and waypoint D, but misread waypoint A as lying on the 7200 ft contour line.

misread waypoint A by claiming that it is "*located exactly on the thick index contour line labeled 7200,*" even though A lies very close to the 7200 contour line while remaining between the 7200 ft and 7240 ft contour lines. A similar problem appears in Item 41 "*What is the approximate radius of the lipid envelope?*" (refer to Appendix A). Gemini used the scale bar correctly, but estimated the diameter to be "*approximately five times the length of the scale bar,*" yielding a radius of 50 nm; Claude produced the same overestimate across all 10 runs, and GPT did so in 3 of 10 runs. By contrast, the open-source models showed no consistent bias, producing radius estimates ranging from 30 to 50 nm across runs. The correct diameter is approximately 4 times the 20 nm scale bar, giving a radius of about 40 nm. These examples suggest that the models can often identify the relevant quantitative encoding and coarse measurement strategy, a weakness reflected in the particularly low MLLM scores on Quantitative Estimation–Relative Estimation (Quantitative).



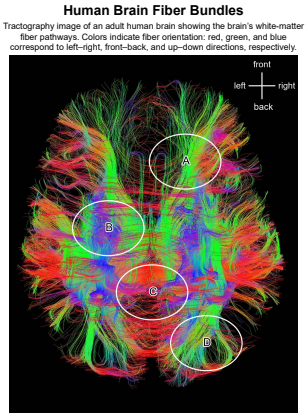
Item 20:
The winds near Chicago and Houston flow in roughly the same direction.

Gemini: "*In the visualization, the wind lines near Chicago are moving generally from* ***west to east.*** *In contrast, the wind lines near Houston are moving generally from* ***south to north.*** *Therefore, the winds in these two locations are not flowing in roughly the same direction.*"

Figure 4: Example of flow-direction interpretation error in a texture-based visualization. MLLM misread the local wind directions near Chicago and Houston, which are both roughly northwest to southeast.

Flow-direction interpretation error. Texture-based visualizations exposed difficulty in decoding local flow direction from animated vector fields. In Item 20, Gemini and Qwen incorrectly concluded that the winds near Chicago and Houston flow in different directions, reasoning that Chicago moves "*from west to east*" while Houston moves "*from south to north.*" In fact, the winds in both regions flow in roughly the same direction, approximately from northwest to southeast (see Figure 4). This pattern of difficulty extends even to models that answered correctly: InternVL identified the same direction at both cities but misread it as "*west to east,*" while LLaVA avoided any specific direction, stating only that arrows "*point in a similar direction.*" This suggests that the model misread local motion in the animated flow field, possibly by confusing streak orientation with the actual flow direction. Because most models correctly identified the hurricane's direction in Item 9 (refer to Appendix A), the difficulty appears specific to dense texture-based motion decoding rather than to directional reasoning more generally.

Encoding-mapping error during higher-level SciVis interpretation. Some failures occur when the model makes a higher-level interpretation, but grounds it in an unsupported visual mapping. In Item 35, Gemini inferred that fiber density was indicated by the opacity and solidness of the colored regions, and on that basis concluded that region C had the highest density because it appeared as a "*highly opaque, solid red mass*" (see Figure 5). Claude reached the same



Item 35:
Which region has the highest density of fiber bundles?

Gemini: "*In the provided tractography image, fiber density is visually indicated by the opacity and solidness of the colored regions. Region C contains a* ***highly opaque, solid red mass, which suggests extremely tightly packed*** *fiber bundles with very few visible individual strands. In contrast, regions A, B, and D display more visible individual fibers and* ***less uniform color, indicating lower density.*** *Therefore, Region C appears to have the highest density of fiber bundles.*"

Figure 5: Example of an encoding-mapping error during higher-level SciVis interpretation.

wrong conclusion across all 10 runs, but via a different unsupported mapping: it invoked anatomical prior knowledge, stating that region C "*corresponds to the splenium of the corpus callosum ... densely packed with fibers of multiple orientations (red, green, and blue),*" thereby grounding its density judgment in imported neuroanatomical knowledge and color variety rather than actual fiber concentration in the image. The item requires a SciVis judgment on fiber density, which can be estimated from the overall concentration and fiber overlap. Instead, both models justified their answer with unsupported cues even though the caption specifies that color encodes fiber orientation. The error, therefore, is not simply that the model made a higher-level inference, but that it grounded this inference in an unsupported encoding rather than in the actual evidence provided by the visualization and caption.

5 Conclusions and Future Work

Our results show that current MLLMs do not yet exhibit uniform SciVis literacy. Performance varies substantially across techniques and tasks. Across techniques, the models perform best on Scientific Illustration, while Gemini also performs strongly on Surface Rendering, Volume Rendering, and Mixed Rendering. Across tasks, their strongest results are in Search and Spatial Understanding, particularly in Presence/Absence, Counting, and Intersection. However, the open-source models remain less robust overall and show narrower, more uneven strengths. The qualitative analysis shows that they reflect recurring weaknesses in fine-grained quantitative estimation, flow-direction interpretation in texture-based visualizations, and unsupported encoding inference during higher-level SciVis interpretation. These patterns suggest that current MLLMs can often identify the general structure of a task, but still struggle when success depends on precise measurement, decoding local motion in dense dynamic fields, or grounding higher-level judgments in the actual encoding specified by the visualization and caption.

This study has several limitations. GPT and Claude were evaluated on animation items via frame extraction, as their APIs did not support direct video input during the evaluation. In addition, the evaluation is based on a single benchmark, SVLAT, and a single standardized prompt format, and the open-source comparison is limited to lightweight models. Additionally, some tasks are represented by only one or two items, which limits the reliability of per-task estimates and may not fully capture the range of difficulty within those tasks. Future work should extend this benchmark to larger and more diverse models, richer prompting and tool-use settings, and additional SciVis-native evaluation instruments. Overall, our findings show that current MLLMs are selectively capable rather than broadly SciVis literate. SciVis literacy is, therefore, a necessary benchmark dimension because scientific visualizations require forms of spatial, temporal, and encoding-grounded interpretation that are not captured by general multimodal evaluations.

Figure Credits

Figure 3: Terrain Map of the Roberts Mountains, Nevada—Image adapted from USGS National Geologic Map Database. Figure 4: Wind Map of the United States—Animation from hint.fm. Figure 5: Human Brain Fiber Bundles—Image adapted from Zeynep Saygin / MIT Koch Institute.

Acknowledgments

This research was supported in part by the U.S. National Science Foundation through grants IIS-2101696, OAC-2104158, IIS-2401144, and CCF-2550610. The authors thank the anonymous reviewers for their insightful comments.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. doi: 10.48550/arXiv.2303.08774 1

[2] R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. doi: 10.48550/arXiv.2312.11805 1

[3] Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026. 2

[4] A. Bendeck and J. Stasko. An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks. *IEEE Trans. Vis. Comput. Graph.*, 31(1):1105–1115, 2025. doi: 10.1109/TVCG.2024.3456155 1, 2

[5] S. Beschi, D. Falessi, S. Golia, and A. Locoro. Characterizing data visualization literacy for standardization: A systematic literature review. *IEEE Access*, 13:65704–65725, 2025. doi: 10.1109/ACCESS.2025.3559298 1

[6] K. Börner, A. Bueckle, and M. Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proc. Proc. Natl. Acad. Sci.*, 116(6):1857–1864, 2019. doi: 10.1073/pnas.1807180116 1

[7] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984 1

[8] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. IEEE/CVF CVPR*, pp. 24185–24198, 2024. doi: 10.48550/arXiv.2312.14238 2

[9] Y. Cui, L. W. Ge, Y. Ding, F. Yang, L. Harrison, and M. Kay. Adaptive assessment of visualization literacy. *IEEE Trans. Vis. Comput. Graph.*, 30(1):628–637, 2024. doi: 10.1109/TVCG.2023.3327165 1

[10] A. K. Das, M. Tarun, and K. Mueller. Charts-of-Thought: Enhancing LLM visualization literacy through structured data extraction. *IEEE Trans. Vis. Comput. Graph.*, 32(1):1427–1437, 2026. doi: 10.1109/TVCG.2025.3634813 2

[11] P. P. Do, K. Tang, K. Ai, and C. Wang. SVLAT: Scientific visualization literacy assessment test. *arXiv preprint arXiv:2603.19000*, 2026. doi: 10.48550/arXiv.2603.19000 1, 2

[12] L. Dong and A. Crisan. Probing the visualization literacy of vision language models: The good, the bad, and the ugly. *IEEE Trans. Vis. Comput. Graph.*, 32(1):1175–1185, 2025. doi: 10.1109/TVCG.2025.3634791 2

[13] M. Duan, Y. Jiang, M. Varona, and C. Nobre. Do MLLMs see what we see? Analyzing visualization literacy barriers in ai systems. *arXiv preprint arXiv:2601.12585*, 2026. doi: 10.48550/arXiv.2601.12585 2

[14] L. W. Ge, Y. Cui, and M. Kay. CALVI: Critical thinking assessment for literacy in visualizations. In *Proc. ACM CHI*, pp. 815:1–815:18, 2023. doi: 10.1145/3544548.3581406 1

[15] Gemini. Gemini 3.1 Pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, 2026. 2

[16] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs have visualization literacy? An evaluation on modified visualizations to test generalization in data interpretation. *IEEE Trans. Vis. Comput. Graph.*, 31(10):7004–7018, 2025. doi: 10.1109/TVCG.2025.3536358 1, 2

[17] K. Kafle, B. Price, S. Cohen, and C. Kanan. DVQA: Understanding data visualizations via question answering. In *Proc. IEEE/CVF CVPR*, pp. 5648–5656, 2018. doi: 10.1109/CVPR.2018.00592 2

[18] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio. FigureQA: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. doi: 10.48550/arXiv.1710.07300 2

[19] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Trans. Vis. Comput. Graph.*, 19(3):495–513, 2013. doi: 10.1109/TVCG.2012.110 1

[20] B. Laha, D. A. Bowman, D. H. Laidlaw, and J. J. Socha. A classification of user tasks in visual analysis of volume data. In *Proc. IEEE SciVis*, pp. 1–8, 2015. doi: 10.1109/SciVis.2015.7429485 1

[21] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a visualization literacy assessment test. *IEEE Trans. Vis. Comput. Graph.*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920 1

[22] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. doi: 10.48550/arXiv.2408.03326 2

[23] A. V. Maltese, J. A. Harsh, and D. Svetina. Data visualization literacy: Investigating data interpretation along the novice–expert continuum. *J. Coll. Sci. Teach.*, 45(1):84–90, 2015. doi: 10.2505/4/jcst15_045_01_84 1

[24] A. Masry, X. L. Do, J. Q. Tan, S. Joty, and E. Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Proc. ACL Findings*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177 2

[25] B. H. McCormick, T. A. DeFanti, and M. D. Brown. Visualization in scientific computing. *Comput. Graph.*, 21(6), 1987. 1

[26] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar. PlotQA: Reasoning over scientific plots. In *Proc. IEEE/CVF WACV*, pp. 1516–1525, 2020. doi: 10.1109/WACV45572.2020.9093523 2

[27] OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026. 2

[28] S. Pandey and A. Ottley. Mini-VLAT: A short and effective measure of visualization literacy. *Comput. Graph. Forum*, 42(3):1–11, 2023. doi: 10.1111/cgf.14809 1

[29] S. Pramanick, R. Chellappa, and S. Venugopalan. SPIQA: A dataset for multimodal question answering on scientific papers. In *Proc. NeurIPS*, pp. 118807–118833, 2024. doi: 10.52202/079017-3773 2

[30] Qwen. Qwen3.5: Accelerating productivity with native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. 2

[31] M. Tory and T. Möller. Rethinking visualization: A high-level taxonomy. In *Proc. IEEE InfoVis*, pp. 151–158, 2004. doi: 10.1109/INFVIS.2004.59 1

[32] C. Wang, P. P. Do, R. S. Laramee, and A. Joshi. Scientific visualization literacy: Assessment, challenges, and opportunities. In *Proc. EuroVis (Education Papers)*, 2026. doi: 10.2312/eved.20261001 1

[33] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu et al. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Proc. NeurIPS*, pp. 113569–113697, 2024. doi: 10.52202/079017-3609 2

[34] Y. Zhao, Y. Zhang, Y. Zhang, X. Zhao, J. Wang, Z. Shao et al. LEVA: Using large language models to enhance visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 31(3):1830–1847, 2025. doi: 10.1109/TVCG.2024.3368060 1

[35] Z. Zhu, M. Jia, Z. Zhang, L. Li, and M. Jiang. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In *Proc. ACL*, pp. 11341–11359, 2025. doi: 10.18653/v1/2025.naacl-long.566 2

A Additional Details and Results

```
You are a scientific visualization expert. You will be given a scientific visualization (image or animation), its caption,
and a corresponding question.

Provide your response in the following JSON format:
{
   "choice": "<the selected answer option (e.g., 'A', 'B', 'C', 'D') or 'Not sure'>",
   "answer_value": "<the full text of the selected answer option>",
   "rationale": "<your reason for choosing this answer, in 5 sentences or fewer>"
}

Requirements:
1. Base your answer strictly on the provided visualization and caption. Do not use prior knowledge or make assumptions.
2. Keep your rationale concise, with a maximum of 5 sentences.
3. If the answer cannot be determined confidently from the provided visualization and caption, set "choice" to "Not sure"
and explain why in your rationale.
4. Return valid JSON only. Do not include markdown or any extra text outside the JSON object.
```

Figure 1: The prompt template for our MLLM literacy experiments.

Table 1: Per-item accuracies for all MLLMs and humans, categorized by techniques. I: static image, A: animation. GPT and Claude were evaluated on animation items via frame extraction.

Technique	Item ID	Format	Task	GPT-5.4	Gemini-3.1-Pro-Preview	Claude-Opus-4.6	Qwen3.5-9B	LLaVA-OneVision-1.5-8B	InternVL3.5-8B	Human
Color Mapping	Item 13	I	Quantitative Estimation - Absolute Estimation	100.0	90.0	100.0	100.0	0.0	0.0	72.2
	Item 14	I	Quantitative Estimation - Relative Estimation (Binary)	30.0	10.0	100.0	0.0	100.0	20.0	56.5
	Item 15	I	Search - Presence/Absence	100.0	100.0	100.0	100.0	100.0	100.0	77.9
	Item 16	I	Pattern Recognition - Trend	70.0	100.0	100.0	100.0	100.0	90.0	91.2
Volume Rendering	Item 47	I	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	90.0	70.0	100.0	100.0	70.0
	Item 54	A	Pattern Recognition - Trend	80.0	100.0	10.0	50.0	0.0	80.0	51.7
	Item 55	A	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	100.0	100.0	100.0	100.0	65.6
	Item 56	A	Pattern Recognition - Trend	100.0	100.0	100.0	100.0	90.0	100.0	71.7
	Item 66	A	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	10.0	0.0	0.0	0.0	68.4
	Item 67	A	Pattern Recognition - Trend	100.0	100.0	100.0	100.0	50.0	100.0	84.7
Surface Rendering	Item 25	I	Search - Counting	50.0	100.0	100.0	20.0	80.0	0.0	89.1
	Item 27	I	Shape Description	30.0	100.0	0.0	0.0	0.0	30.0	72.0
	Item 28	I	Spatial Understanding - Intersection	60.0	100.0	100.0	100.0	100.0	100.0	87.7
	Item 31	I	Spatial Understanding - Relative	100.0	100.0	70.0	100.0	0.0	0.0	73.4
	Item 32	I	Spatial Understanding - Relative	100.0	100.0	100.0	100.0	40.0	100.0	88.4
	Item 51	A	Pattern Recognition - Trend	90.0	100.0	100.0	100.0	100.0	100.0	86.8
	Item 52	A	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	100.0	100.0	100.0	100.0	63.2
Texture-Based Visualization	Item 17	A	Pattern Recognition - Trend	100.0	100.0	100.0	100.0	90.0	100.0	82.7
	Item 18	A	Quantitative Estimation - Relative Estimation (Binary)	40.0	20.0	0.0	30.0	100.0	20.0	79.0
	Item 19	A	Shape Description	0.0	70.0	50.0	20.0	70.0	10.0	83.6
	Item 20	A	Pattern Recognition - Repetition	20.0	0.0	0.0	40.0	100.0	100.0	83.2
Integration-Based Visualization	Item 33	I	Pattern Recognition - Trend	100.0	100.0	100.0	60.0	100.0	100.0	85.8
	Item 35	I	Quantitative Estimation - Relative Estimation (Binary)	90.0	0.0	0.0	10.0	40.0	0.0	88.3
	Item 57	I	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	30.0	100.0	100.0	100.0	96.9
	Item 59	I	Spatial Understanding - Relative	30.0	100.0	0.0	0.0	90.0	30.0	83.0
Mixed Rendering	Item 60	I	Spatial Understanding - Absolute	100.0	100.0	100.0	100.0	0.0	100.0	87.1
	Item 61	I	Shape Description	50.0	100.0	70.0	80.0	100.0	100.0	66.1
	Item 62	I	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	100.0	60.0	0.0	0.0	54.1
	Item 68	A	Pattern Recognition - Trend	100.0	100.0	100.0	100.0	100.0	100.0	95.3
	Item 70	A	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	60.0	50.0	100.0	70.0	92.2
	Item 71	A	Quantitative Estimation - Relative Estimation (Quantitative)	0.0	70.0	0.0	50.0	30.0	40.0	27.0
Glyph, Mesh, Plot	Item 1	I	Quantitative Estimation - Absolute Estimation	10.0	100.0	30.0	50.0	100.0	100.0	71.2
	Item 2	I	Quantitative Estimation - Relative Estimation (Binary)	100.0	100.0	100.0	100.0	0.0	10.0	83.7
	Item 3	I	Quantitative Estimation - Relative Estimation (Quantitative)	20.0	20.0	70.0	0.0	30.0	40.0	38.8
	Item 4	I	Pattern Recognition - Trend	70.0	90.0	100.0	60.0	0.0	20.0	55.7
	Item 5	I	Pattern Recognition - Trend	60.0	100.0	100.0	70.0	70.0	40.0	61.5
	Item 7	I	Search - Presence/Absence	100.0	100.0	100.0	100.0	60.0	60.0	74.8
	Item 9	I	Spatial Understanding - Absolute	100.0	100.0	100.0	90.0	100.0	40.0	87.6
	Item 10	I	Spatial Understanding - Absolute	100.0	100.0	100.0	50.0	30.0	30.0	87.5
	Item 12	I	Pattern Recognition - Trend	10.0	100.0	100.0	50.0	0.0	30.0	66.8
	Item 22	I	Search - Counting	100.0	100.0	100.0	70.0	80.0	100.0	84.2
	Item 23	I	Pattern Recognition - Repetition	100.0	100.0	100.0	50.0	20.0	50.0	65.5
Scientific Illustration	Item 36	I	Spatial Understanding - Relative	100.0	100.0	100.0	100.0	100.0	100.0	90.0
	Item 37	I	Spatial Understanding - Relative	100.0	100.0	100.0	100.0	100.0	100.0	72.4
	Item 40	I	Spatial Understanding - Relative	100.0	100.0	100.0	100.0	90.0	100.0	79.5
	Item 41	I	Quantitative Estimation - Absolute Estimation	60.0	70.0	0.0	40.0	40.0	40.0	43.1
	Item 43	I	Search - Counting	100.0	100.0	100.0	100.0	100.0	100.0	84.3
	Item 45	I	Search - Presence/Absence	10.0	100.0	100.0	100.0	40.0	100.0	90.2
	Item 46	I	Spatial Understanding - Absolute	100.0	100.0	100.0	100.0	100.0	100.0	90.7