MLLM Perception of Color-Encoded Scalar Fields Reveals Model-Dependent Inversion of Chain-of-Thought

Minyi Liu, Yue Zhao, Xiaoyang Song, Shufan Qian, Yulong Bian, and Qiong Zeng

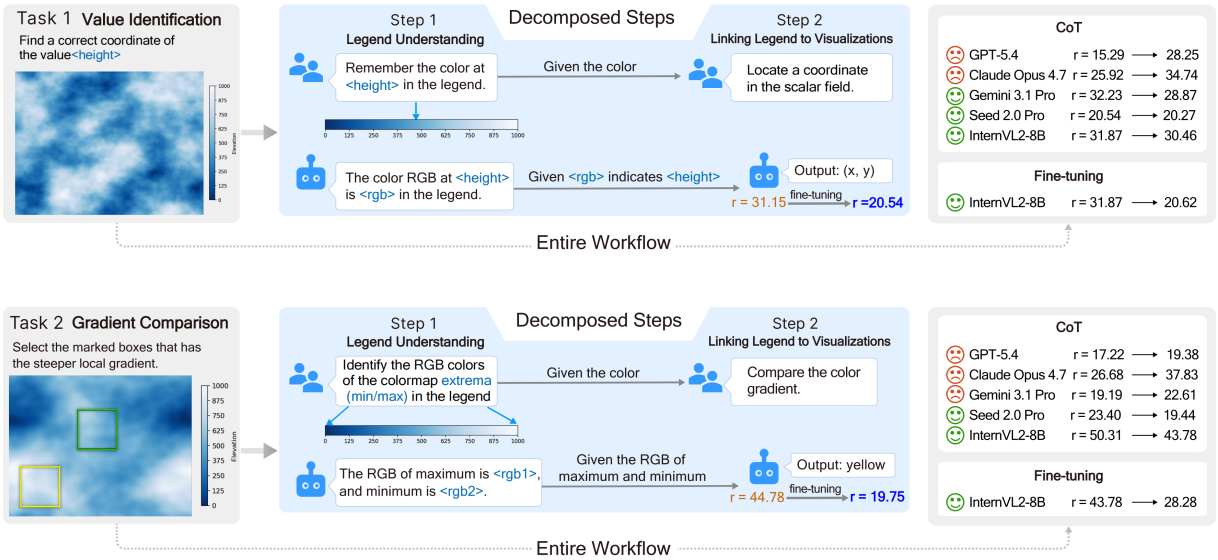

Fig. 1: **Overview of our methodology** for evaluating and adapting MLLMs on two graphical perception tasks in scalar field visualization–Task 1 (value identification, top) and Task 2 (gradient comparison, bottom). Building on the two-step human perception workflow (*legend understanding* and *linking legend to visualizations*) we evaluate and adapt MLLMs through different prompting strategies and model adaptations, revealing a model-dependent interaction between prompting and model capacity (with r indicating error rates).

Abstract— Color-encoded scalar field visualizations are fundamental to scientific data analysis, yet their interpretation depends critically on both data characteristics and colormap design. While multimodal large language models (MLLMs) have emerged as promising tools for visualization tasks, their ability to comprehend and interpret scalar field visualizations remains largely unexplored. We investigate how effectively MLLMs perceive these visualizations without direct access to the underlying data values, focusing on two graphical perception tasks: value identification and gradient comparison. Building on the well-established two-step **human perception workflow**—legend understanding and linking legend to visualization—we develop a structured prompting framework for **MLLM perception workflows** and implement it through both decomposed-step and chain-of-thought (CoT) prompting strategies across five state-of-the-art MLLMs, alongside fine-tuning techniques for model adaptation. Our results reveal a *model-dependent* interaction between prompting and model capacity: CoT prompting benefits weaker models (e.g., reducing Task 2 error by up to 21.5%) but *degrades* stronger ones (relative error increases of up to 84.76% on Task 1 for the strongest models we tested), while fine-tuning yields consistent improvements (up to a 43.80% relative error reduction across both tasks). Our work bridges human and machine perception workflows, demonstrating the promising potential of leveraging MLLMs to understand graphical perception while highlighting the need for *model-aware* prompting strategies in visualization perception.

Index Terms—Continuous colormap, scalar field visualization, graphical perception, multi-modal large language models

1 Introduction

Color-encoded scalar field visualization, which maps scalar values to colors using colormapping techniques, is essential for analyzing scientific data [20, 21, 34]. Effective colormap designs enable users to accurately interpret visual information, such as identifying values and comparing variations, while poorly designed colormaps can lead to misinterpretation [37, 39]. For instance, while the *rainbow* colormap supports accurate value identification [24, 35], its perceptual drawbacks, such as misleading contours caused by abrupt hue transitions [27], limit its effectiveness for pattern perception [24]. This highlights the importance of understanding graphical perception in colormaps–*how viewers perceive and decode information from color*-encoded visualizations–as a foundation for designing effective colormaps [28, 39].

Existing research has explored graphical perception in colormap design, as well as its interaction with data distribution and analytical tasks, primarily through empirical studies [7, 18, 25, 33]. These guidelines provide valuable insights for manually selecting and designing colormaps and even inform the development of automatic colormap optimization algorithms [32, 37] when well-defined mathematical metrics are available. Despite these advancements, domain experts and designers still struggle to *predict graphical perception under varying*

- *Minyi Liu, Xiaoyang Song, Shufan Qian, and Yulong Bian are with Shandong University. E-mail: LiuMinyi.Chuck@gmail.com; songxiaoyang222@gmail.com; qiansf.13@gmail.com; bianyulong@sdu.edu.cn.*
- *Yue Zhao is with Shandong Second Medical University. E-mail: zhaoyue@sdsmu.edu.cn.*
- *Qiong Zeng, the corresponding author, is with Shandong University and Shandong Provincial Key Laboratory of Computing-Network Integration. Email: qiong.zn@sdu.edu.cn.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

conditions. The major challenge is that prior studies often focus on a limited set of colormaps and data types, typically under controlled conditions with idealized users (e.g., those with normal color vision) and fixed tasks [24]. Consequently, it is difficult to generalize these findings to complex, real-world scenarios where user variability, dynamic tasks, and diverse data distributions come into play. This raises an **open question**: (**Q***) *Can computational models be leveraged to understand viewers' graphical perception in color-encoded visualizations?*

As MLLMs are increasingly integrated into visualization and data analysis workflows—from automated chart captioning and insight generation to interactive colormap recommendation—understanding their perceptual capabilities and limitations in reading scientific visualizations has become a practical necessity. Multimodal large language models (MLLMs) have shown their abilities in semantic analysis, spatial information understanding, as well as textual and visual reasoning [4, 13, 31, 36], which hold potential for advancing graphical perception tasks in color-encoded visualizations. For example, prior work has demonstrated their applicability in visualization design and generation through multimodal interactions between text prompts and visual feedback [6, 11], proving their ability to understand design principles. Recent studies have begun evaluating graphical perception of MLLMs, primarily focusing on discrete chart types (such as bar chart, area chart, and histogram) [10, 30, 38]. MLLMs have strong capabilities in interpreting conventional charts (e.g., line chart), but are limited in data-dense visualizations (e.g., bubble charts) [23]. Zhang *et al.* [38] further identified key limitations in state-of-the-art MLLMs for graphical perception, including *poor generalization across chart types*, *weak understanding of fundamental visual elements*, and *difficulties in referencing values within charts*. Compared to discrete charts, color-encoded visualizations present additional challenges–requiring MLLMs to perceive color transitions, smooth variations, and spatial coherence–capabilities that remain unexplored. Moreover, MLLMs evolve rapidly across generations, and whether prompting strategies that simulate human perception workflows benefit all models uniformly—or interact with model capability in non-trivial ways—remains an open question.

Motivated by these gaps, we refine the open question into two specific, **executable research questions**: (**Q1**) How well do state-of-the-art MLLMs perceive scalar field visualizations without direct access to numerical scalar values, following the processes that viewers perceive and decode information? (**Q2**) How do structured prompting and adaptation strategies–guided by human perception workflows–interact with model capability across diverse MLLMs? To enable systematic evaluation, we focus on two fundamental and well-established graphical perception tasks in color-encoded visualizations [24, 33]: *value identification* (**Task 1**, locating scalar values) and *gradient comparison* (**Task 2**, assessing changes across regions).

To answer these research questions, we adopt the well-established two-step *human perception workflow* for decoding color-encoded visualizations [14, 24, 29]–*legend understanding* and *linking legend to visualization*–as evaluation scaffolding. Guided by this workflow, we introduce a *structured prompting framework* that simulates human perception workflow, enabling MLLMs to reason step-by-step in the way that humans identify specific values or compare changes across regions. We then apply the structured prompting framework to design two prompting strategies: a *decomposed-step* prompting that breaks the workflow into two reasoning steps, and a *chain-of-thought* (CoT) prompting that simulates the entire workflow. We compare both against a baseline strategy that directly poses task-related queries, e.g., finding the coordinate of a specific value. Additionally, we explore fine-tuning techniques to adapt MLLMs to graphical perception tasks by leveraging ground-truth annotations.

We evaluate five state-of-the-art MLLMs on the two tasks under all three prompting strategies, plus task-specific fine-tuning, and uncover a *model-dependent inversion* in chain-of-thought reasoning: CoT improves accuracy on some models (e.g., reducing Task 2 error by up to 21.5%) while degrading it on others (relative error increases of up to 85% on Task 1). Fine-tuning, by contrast, yields consistent gains across both tasks (up to 43.8%), bringing an open-source 8B model close to or above several proprietary zero-shot results. The inversion echoes recent reports of CoT-induced failures in broader LLM tasks [16, 17], and motivates *capability-aware* prompting strategies for visualization.

Our main contributions are listed as follows:

- We design a structured prompting framework grounded in the two-step human perception workflow [14, 24, 29], with decomposed-step and CoT variants that allow us to localize where in the perception process MLLMs succeed or fail.
- We benchmark prompting strategies and task-specific fine-tuning on five state-of-the-art MLLMs and report a *model-dependent inversion* in CoT prompting that has not been documented in prior visualization-perception evaluations.
- We release our datasets, prompts, code, and fine-tuned models at `https://osf.io/y4pgm/?view_only=dadc9ff8b62d4a27a2a0d852ccde30bc` for future research on graphical perception of MLLMs in scalar field visualization.

2 Related Work

We review three lines of prior work. **(i) Human graphical perception in colormaps.** Research on visualization comprehension has established a structured perception workflow when viewers interpret visual information. Trickett and Trafton [29] proposed a comprehensive model of graph comprehension emphasizing spatial cognition; Lee *et al.* [14] developed a grounded model of how people make sense of unfamiliar visualizations through iterative exploration. For scalar fields specifically, Reda *et al.* [24] identified *value identification* and *gradient comparison* as two fundamental tasks and showed how performance varies with spatial frequency and colormap design [7, 25, 26]. We *adopt* the resulting two-step workflow—*legend understanding* and *linking legend to visualization*—as evaluation scaffolding in Sec. 4. **(ii) MLLMs in visualization perception.** Recent studies evaluate MLLMs on graphical perception, primarily for discrete charts (e.g., bar, line, area) [9, 10, 23, 30, 38], revealing limitations in fundamental visual elements and value referencing [38]. Continuous color-encoded scalar fields, which require perceiving smooth color transitions and spatial coherence, remain unexplored. **(iii) Adapting LLMs through prompting and fine-tuning.** Chain-of-thought (CoT) prompting [5, 19, 36] and parameter-efficient fine-tuning such as LoRA [4, 12] have been widely used to improve LLM reasoning. However, recent work shows that CoT can in fact *degrade* performance when models already internalize effective reasoning strategies [16, 17]—a phenomenon whose implications for visualization perception have not been examined.

Our work bridges these three lines: using Reda's tasks [24] as a controlled testbed, we systematically evaluate how prompting and fine-tuning interact with model capability in scalar field perception.

3 MLLM Perception Workflow

Following the two-step human perception workflow described in Sec. 2, we design the methodology for evaluating and adapting MLLMs on the two graphical perception tasks, termed the *MLLM perception workflow*. We first propose a structured prompting framework with reusable components, then implement it through decomposed-step and chain-of-thought (CoT) prompting strategies, and finally describe our fine-tuning protocol for model adaptation.

3.1 Structured Prompting Framework

We set a unified context ("You are an image analysis assistant, and your task is to complete the following steps") and developed reusable components that can be combined depending on the task: ***image provision*** (image dimensions and the scalar-field/legend split), ***structure explanation*** (the legend region is excluded from the field), ***region explanation*** (two square boxes for **Task 2**), ***legend observation*** for ***Legend Understanding*** (colormap meaning, value range, and the colors representing the extrema), ***coordinate system clarification*** (origin at the lower-left corner), and ***task guidance*** for ***Linking Legend to Visualization*** (locating a coordinate matching a target value for **Task 1**; comparing the rate of color change between two boxes for **Task 2**). For brevity, the parenthetical descriptions only sketch each component;

full definitions, including the exact image <width>/<height> placeholders, RGB <rgb> and value-range parameters <color_range>, and the complete prompt templates for both tasks, are provided in the supplementary material. These components are flexibly combined to instantiate the prompting strategies below. Separating Step 1 from Step 2 lets us attribute errors to either color interpretation or spatial localization (Sec. 5).

3.2 Prompting Strategies

Decomposed-step prompting explicitly separates the two perceptual steps to identify strengths and weaknesses of MLLMs. *Step 1 (Legend Understanding)* prompts the model with ***image provision*** and ***legend observation*** and asks it to output the RGB color corresponding to the target value (**Task 1**) or the RGB values of the colormap extrema (**Task 2**). *Step 2 (Linking Legend to Visualization)* provides the reference RGB and asks the model to locate the matching coordinate in the field (**Task 1**) or to determine which of the two marked boxes has the steeper color gradient (**Task 2**).

Chain-of-thought prompting [36] merges the two steps into a single prompt that systematically guides the model through (1) image provision and structure explanation, (2) legend observation encouraging attention to the overall color mapping, (3) coordinate system clarification, and (4) the explicit task statement (locate the value-bearing coordinate for **Task 1**; identify the steeper-gradient box for **Task 2**).

Baseline prompting serves as a reference: for **Task 1** we instruct the model with "Identify a coordinate of a point with the value <elevation> in the scalar field image," and for **Task 2** we provide a definition of terrain steepness and ask which marked box is steeper.

3.3 Model Adaptation via Fine-tuning

We adapt MLLMs by fine-tuning on paired scalar-field visualizations and ground-truth answers. Since Step 1 does not directly contribute to the final answer, we fine-tune with Step 2, CoT, and baseline prompts. Following Reda *et al.* [24], we generate scalar fields at five spatial frequencies and nine colormaps. For **Task 1**, each field is queried with 21 distinct value-identification questions, with 30% random crops for augmentation, yielding 9,450 pairs. For **Task 2**, 9,000 gradient-comparison pairs are generated.

Loss function. The standard cross-entropy (CE) loss is well-suited for the binary classification of **Task 2**. For **Task 1**, predictions in the form of (x, y) coordinates require spatial-distance awareness, so we use $\mathscr{L}_{\text{Task1}} = \mathscr{L}_{\text{CE}} + \lambda \log(1 + d(\hat{p}, \mathscr{G}))$, where $\hat{p}$ is the predicted coordinate, $\mathscr{G}$ is the set of ground-truth points, $d(\hat{p}, \mathscr{G}) = \min_{p \in \mathscr{G}} \|\hat{p} - p\|_2$, and λ balances the two terms. The log term prevents disproportionately large penalties for distant predictions.

Implementation. We fine-tune InternVL2-8B [3, 4] with LoRA. Hyperparameter details are provided in the supplementary material.

4 Experimental Results

4.1 Experimental Setup

Dataset. Following Reda *et al.* [24], we generated scalar fields at five spatial frequencies using Perlin noise, each rendered with nine colormaps, producing 45 visualizations for Task 1 (value identification) and 180 for Task 2 (gradient comparison). All visualizations were rendered at 820×630 pixels and paired with their colorbar legends and ground-truth annotations.

Models. We evaluated five state-of-the-art MLLMs spanning proprietary and open-source, lightweight and large-scale: *GPT-5.4* [22], *Claude Opus 4.7* [1], *Gemini 3.1 Pro* [8], *Seed 2.0 Pro* [2], and *InternVL2-8B* [3, 4]. All models were tested in a zero-shot setting; fine-tuning was performed on InternVL2-8B. All proprietary models were queried with *temperature* $= 0$ (GPT-5.4 used reasoning effort `medium`); we used `max_tokens` $= 20$ and `top_p` $= 1$.

Metrics. For **Task 1**, we use the absolute error between the predicted value v_1 and the requested value v_2, $|v_1 - v_2| \times 100\%$, with values normalized to $[0, 1]$. For decomposed Step 1, where the model outputs

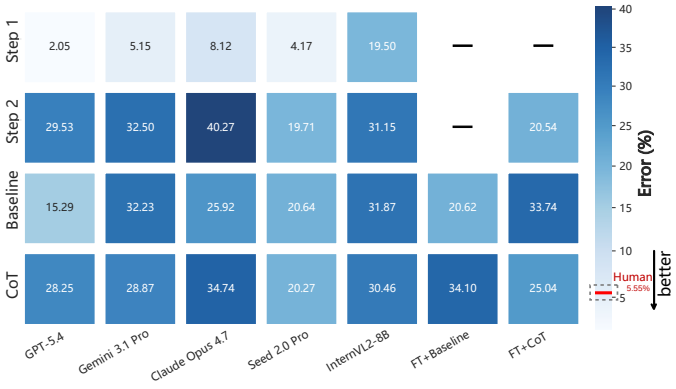

Fig. 2: Error values (%) for **Task 1** across models and prompting strategies. Rows: Step 1 (*legend understanding*), Step 2 (*linking legend to visualization*), Baseline, and CoT. Columns: models; FT+Baseline and FT+CoT denote InternVL2-8B fine-tuned with Baseline and CoT prompts, respectively. Lower (lighter) is better; dashes indicate configurations not evaluated. The red tick on the colorbar marks the human baseline (5.55%), derived from the data of Reda *et al.* [24] following our evaluation protocol on the same data-generation procedure.

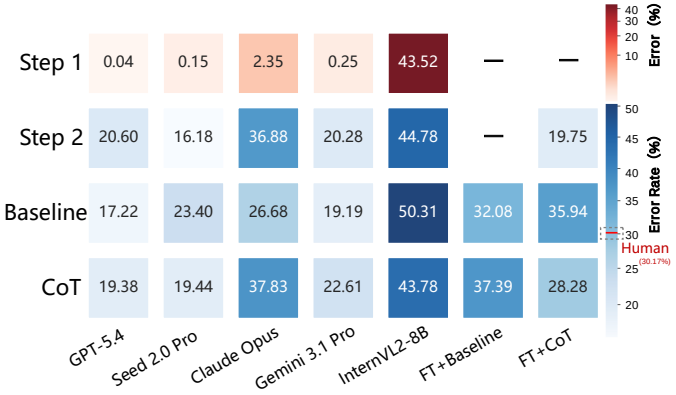

Fig. 3: Error rates (%) for **Task 2** across models and prompting strategies (lower is better). The red tick on the colorbar marks the human baseline (30.17%), derived from the data of Reda *et al.* [24] following our evaluation protocol; several top models outperform humans on this task.

an RGB, we map the closest color on the colormap in CIELAB2000 back to its scalar value before computing the error. For **Task 2**, we report the error rate $r = 1 - n_1/n_2$ over n_2=1,800 tested visualizations (180 images, each queried 10 times under deterministic decoding to capture API-side variability).

4.2 Task 1: Value Identification

Fig. 2 summarizes performance across models and prompting strategies.

Decomposed steps. Proprietary models excel at *Step 1 (Legend Understanding)* (GPT-5.4: 2.05%, Seed 2.0 Pro: 4.17%, Gemini: 5.15%, Claude: 8.12%), while InternVL2-8B lags (19.50%). For *Step 2 (Linking Legend to Visualization)*, errors are larger across all models—Seed 2.0 Pro leads at 19.71%, followed by GPT-5.4 (29.53%), Gemini (32.50%), and Claude Opus 4.7 (40.27%); InternVL2-8B reaches 31.15%. This consistent Step 1-to-Step 2 gap confirms that spatial localization—not color recognition—is the primary bottleneck.

Entire workflow. The effect of CoT prompting varies across models. Seed 2.0 Pro improves slightly from 20.64% (baseline) to 20.27% (CoT), and Gemini improves from 32.23% to 28.87%. In contrast, **GPT-5.4 degrades from** 15.29% **to** 28.25% **(a relative degradation of** 85%) and **Claude Opus 4.7 degrades from** 25.92% **to** 34.74%. CoT thus improves accuracy on some models while reducing it on others, revealing a model-dependent interaction between explicit step-by-step instructions and each model's spatial-reasoning ability.

Fine-tuning. Fine-tuning InternVL2-8B with baseline prompting

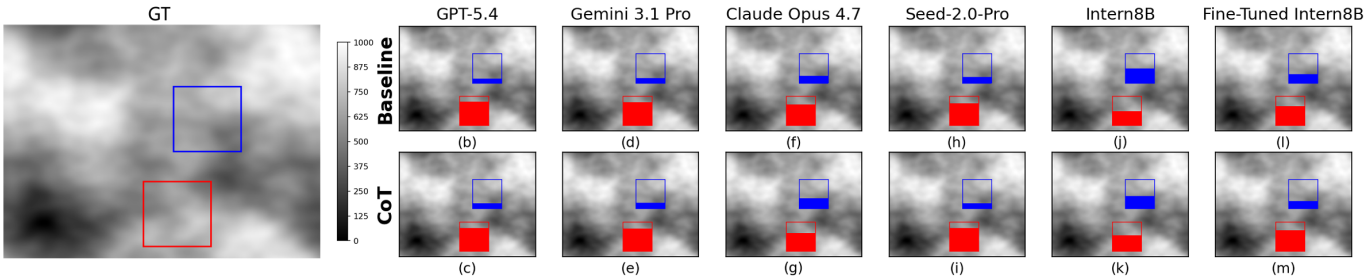

Fig. 4: Model selections for **Task 2** (a: example image with two gradient boxes). Each cell encodes the proportion of trials in which a model selected the box of that color. The fine-tuned InternVL2-8B (rightmost columns) most consistently matches the ground truth.

(FT+Baseline) reaches 20.62%, a 35.3% reduction from the non-fine-tuned baseline (31.87%). FT+CoT reaches 25.04%, still outperforming all zero-shot proprietary models except GPT-5.4 baseline. Decomposed-step fine-tuning further reduces Step 2 error to 20.54%, matching or beating the proprietary models on this task.

4.3 Task 2: Gradient Comparison

Fig. 3 reveals patterns analogous to **Task 1**, with the inversion appearing on a subset of models.

Decomposed steps. Proprietary models achieve near-perfect *Step 1* legend interpretation (GPT-5.4: 0.04%, Gemini: 0.25%, Claude: 2.35%, Seed 2.0 Pro: 0.15%); InternVL2-8B lags at 43.52%. For *Step 2*, Seed 2.0 Pro leads (16.18%), followed by Gemini (20.28%), GPT-5.4 (20.60%), and Claude Opus 4.7 (36.88%).

Entire workflow. Seed 2.0 Pro benefits the most from structured prompting: baseline error 23.40%, CoT 19.44%, decomposed-step 16.18%, indicating a monotonic improvement aligned with the human perception workflow. By contrast, GPT-5.4 degrades from baseline (17.22%) to CoT (19.38%); Gemini from 19.19% to 22.61%; and Claude from 26.68% to 37.83%. Task 2 reproduces the Task 1 split along the same model boundaries.

Fine-tuning. FT+Baseline reaches 32.08% and FT+CoT reaches 28.28%, corresponding to relative error reductions of 36.2% and 43.8% over the non-fine-tuned baseline (50.31%). We compare against Reda *et al.* [24]: while human accuracy is highest at high spatial frequencies, MLLM accuracy is highest at the lowest frequency and decreases as frequency rises, suggesting that MLLMs do not yet replicate human-like sensitivity to gradient sharpness. Per-frequency results are provided in the supplementary material.

Qualitative selection. Fig. 4 shows model selection of the steeper-gradient box. Under baseline prompting, GPT-5.4 and Gemini largely select correctly, while Seed, Claude, and InternVL2-8B frequently misidentify the region. CoT improves Seed 2.0 Pro's consistency but reduces Claude's reliability, echoing the model-dependent pattern. The fine-tuned InternVL2-8B (FT+CoT) achieves near-perfect alignment with the ground truth, matching proprietary models.

5 Discussion

Model-dependent inversion of CoT. Across both tasks, structured prompting affects different models in qualitatively different ways. It helps Seed 2.0 Pro, Gemini, and InternVL2-8B, where Seed 2.0 Pro's Task 2 error drops from 23.4% to 16.18% under decomposed-step prompting. On GPT-5.4 and Claude Opus 4.7 the same prompts have the opposite effect (GPT-5.4's Task 1 error rises 84.76%). One possible explanation is that some MLLMs have already internalized effective implicit reasoning, and forcing explicit verbalization disrupts those internal pathways. This pattern echoes recent reports of CoT-induced failures in overthinking [17] and instruction-following [16] tasks, where explicit reasoning can divert attention from task-relevant visual signals.

Fine-tuning is a robust, model-agnostic adaptation. Independent of the prompting inversion, fine-tuning InternVL2-8B yields gains across both tasks—a 43.8% relative error reduction on Task 2 (from 50.31% to 28.28%) and a Task 1 reduction from 31.87% to 20.62%, and the fine-tuned 8B model surpasses several proprietary zero-shot results. This indicates that adaptation can substantially close the gap between open- and closed-source models when prompt-side adaptation alone is insufficient or counterproductive.

Implications for visualization perception evaluation. Our findings caution against "one-prompt-fits-all" benchmarks for MLLM visualization perception. Furthermore, while MLLMs can simulate aspects of the human perception workflow, they fail to reproduce fine-grained human patterns, e.g., the human error trend across spatial frequencies reported by Reda *et al.* [24] is not mirrored in our results, indicating that machine perception of color-encoded fields remains coarse. The human baselines (5.55% on Task 1; 30.17% on Task 2) quantify this gap: all models trail humans on Task 1, whereas several top models surpass human accuracy on Task 2.

Limitations and future work. Our study is limited to two fundamental tasks on synthetic scalar fields; future work should extend to higher-level analytics and more complex visualization types. We suggest three directions: *(i) Model-aware evaluation protocols:* as MLLMs increasingly enter visualization workflows, from automated chart description to colormap recommendation, future benchmarks should test models under multiple prompting strategies rather than rely on a single one, since the same prompt can flip from helpful to harmful across models. *(ii) Beyond 2D scalar fields:* extending the inversion analysis to vector fields, volume renderings, and time-varying data will clarify whether the model-dependent inversion is specific to color-encoded perception or reflects a broader property of MLLM-based visualization understanding. *(iii) Visualization-grounded reasoning:* having models inspect intermediate visual states during inference [15], e.g., re-rendering a region with an alternative colormap to verify a value, can align machine perception more closely with how human analysts iteratively explore scientific visualizations.

6 Conclusion

We evaluated five state-of-the-art MLLMs on two graphical perception tasks in color-encoded scalar field visualization. Rather than benchmarking MLLMs as black-box evaluators, our study examines how prompting and fine-tuning strategies should be matched to model capability. Prompting strategies grounded in the human two-step perception workflow help weaker models—reducing Seed 2.0 Pro's Task 2 error to 16.18%—but hurt stronger ones, with GPT-5.4 and Claude Opus 4.7 degrading by up to 85% under CoT. Fine-tuning InternVL2-8B, by contrast, yields consistent gains across both tasks (up to 43.8% relative error reduction) and surpasses several proprietary zero-shot results. All tested models fall short of human accuracy on Task 1, though several surpass humans on Task 2 [24], confirming spatial localization as a harder challenge. Together, these results motivate *model-aware* prompting for visualization. Extending beyond two fundamental tasks on synthetic scalar fields—to higher-level analytical reasoning and more complex visualization types—will require further validation. Our released datasets, prompts, and fine-tuned models provide a foundation for future work at the intersection of visualization and multimodal AI.

Acknowledgments

This study is supported by grants from the National Key R&D Program of China (No. 2021YFF0704300) and the National Natural Science Foundation of China (No. 62372271).

REFERENCES

[1] Anthropic. Claude Opus 4.7 system card. https://anthropic.com/claude-opus-4-7-system-card, Apr. 2026. 3

[2] ByteDance Seed. Seed2.0: A family of general-purpose agent models. https://seed.bytedance.com/en/seed2, Feb. 2026. 3

[3] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 3

[4] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. 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*, pp. 24185–24198, 2024. 2, 3

[5] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, Z. Sun, D. Gutfreund, and C. Gan. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 1254–1262, 2024. 2

[6] Y. Cui, L. W. Ge, Y. Ding, L. Harrison, F. Yang, and M. Kay. Promises and pitfalls: Using large language models to generate visualization items. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1094–1104, 2025. doi: 10.1109/TVCG.2024.3456309 2

[7] A. Dasgupta, J. Poco, B. Rogowitz, K. Han, E. Bertini, and C. T. Silva. The effect of color scales on climate scientists' objective and subjective performance in spatial data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics*, 26(3):1577–1591, 2020. doi: 10.1109/TVCG.2018.2876539 1, 2

[8] Google DeepMind. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, Feb. 2026. 3

[9] G. Guo, J. J. Kang, R. S. Shah, H. Pfister, and S. Varma. Understanding graphical perception in data visualization through zero-shot prompting of vision-language models. *arXiv preprint arXiv:2411.00257*, 2024. 2

[10] 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 Transactions on Visualization and Computer Graphics*, 31(10):7004–7018, 2025. doi: 10.1109/TVCG.2025.3536358 2

[11] Y. Hou, M. Yang, H. Cui, L. Wang, J. Xu, and W. Zeng. C2ideas: Supporting creative interior color design ideation with a large language model. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2024. 2

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213, 2022. 2

[14] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-a. Kang, and J. S. Yi. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):499–508, 2015. doi: 10.1109/TVCG.2015.2467195 2

[15] C. Li, W. Wu, H. Zhang, Y. Xia, S. Mao, L. Dong, I. Vulić, and F. Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267 of *Proceedings of Machine Learning Research*, pp. 36340–36364. PMLR, 2025. 4

[16] X. Li, Z. Yu, Z. Zhang, X. Chen, Z. Zhang, Y. Zhuang, N. Sadagopan, and A. Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in LLMs. In *Advances in Neural Information Processing Systems*, 2025. Spotlight. 2, 4

[17] R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, and T. L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. In *Proceedings of the 42nd International Conference on Machine Learning*, vol. 267, pp. 38489–38517, 2025. 2, 4

[18] Z. Meng, Y. Yang, Q. Zeng, K. Lu, L. Lu, C. Tu, F. Yang, and Y. Wang. Seeing through the overlap: The impact of color and opacity on depth order perception in visualization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3714070 1

[19] C. Mitra, B. Huang, T. Darrell, and R. Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024. 2

[20] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014. doi: 10.1201/b17511 1

[21] J. R. Nuñez, C. R. Anderton, and R. S. Renslow. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLoS One*, 13(7):1–14, 2018. doi: 10.1371/journal.pone.0199239 1

[22] OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, Mar. 2026. 3

[23] S. Pandey and A. Ottley. Benchmarking visual language models on standardized visualization literacy tests. In *Computer Graphics Forum*, vol. 44, pp. 1–12, 2025. doi: 10.1111/cgf.70137 2

[24] K. Reda, P. Nalawade, and K. Ansah-Koi. Graphical perception of continuous quantitative maps: The effects of spatial frequency and colormap design. In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems*, 2018. doi: 10.1145/3173574.3173846 1, 2, 3, 4

[25] K. Reda, A. A. Salvi, J. Gray, and M. E. Papka. Color nameability predicts inference accuracy in spatial visualizations. *Computer Graphics Forum*, 40(3):49–60, 2021. doi: 10.1111/cgf.14288 1, 2

[26] K. Reda and D. A. Szafir. Rainbow revisited: Modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 2020. doi: 10.1109/TVCG.2020.3030439 2

[27] B. E. Rogowitz and L. A. Treinish. Data visualization: The end of the rainbow. *IEEE Spectrum*, 35(12):52–59, 1998. 1

[28] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):810–819, 10 pages, Jan. 2019. doi: 10.1109/TVCG.2018.2865147 1

[29] S. B. Trickett and J. G. Trafton. Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In *International Conference on Theory and Application of Diagrams*, pp. 286–300. Springer, 2006. doi: 10.1007/11783183_38 2

[30] H. W. Wang, J. Hoffswell, S. M. Thazin Thane, V. S. Bursztyn, and C. X. Bearfield. How aligned are human chart takeaways and llm predictions? a case study on bar charts with varying layouts. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):536–546, 11 pages, 2025. doi: 10.1109/TVCG.2024.3456378 2

[31] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, and J. Dai. Visionllm: large language model is also an open-ended decoder for vision-centric tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, article no. 2688, 13 pages. Curran Associates Inc., Red Hook, NY, USA, 2023. 2

[32] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C.-W. Fu, O. Deussen, and B. Chen. Optimizing color assignment for perception of class separability in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):820–829, 2019. doi: 10.1109/TVCG.2018.2864912 1

[33] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications*, 8(5):41–49, 1988. doi: 10.1109/38.7760 1, 2

[34] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, San Francisco, CA, USA, 2004. doi: 10.1016/C2009-0-62432-6 1

[35] C. Ware, M. Stone, and D. A. Szafir. Rainbow colormaps are not all bad. *IEEE Computer Graphics and Applications*, 43(3):88–93, 2023. doi: 10.1109/MCG.2023.3246111 1

[36] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2, 3

[37] Q. Zeng, Y. Zhao, Y. Wang, J. Zhang, Y. Cao, C. Tu, I. Viola, and Y. Wang. Data-driven colormap adjustment for exploring spatial variations in scalar fields. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4902–4917, 2022. doi: 10.1109/TVCG.2021.3109014 1

[38] K. Zhang, J. Yang, J. P. Inala, C. Singh, J. Gao, Y. Su, and C. Wang. Towards understanding graphical perception in large multimodal models. *arXiv preprint arXiv:2503.10857*, 2025. 2

[39] L. Zhou and C. D. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 2016. doi: 10.1109/TVCG.2015.2489649 1