Evaluating and Adapting Multimodal LLMs for Graphical Perception in Color-Encoded Scalar Field Visualization

(Supplemental Material)

Category: Research

Paper Type: evaluation

Abstract—This supplemental material provides comprehensive experimental details and results that complement our main paper. We present: (1) responses data from our empirical study and workshop on human perception workflow, including participants' questionnaire responses after the study and workshop summarizations; (2) detailed comparative analysis between our fine-tuned models and Reda *et al.*'s human study results; and (3) complete quantitative results across different colormaps and spatial frequencies for both value identification and gradient comparison tasks. Our full experimental datasets, implementation code, and trained models are available at https://osf.io/y4pgm/?view_only=be060a5816bf4edbaaf66a695d57dee0.

1 Human Perception Workflow

This section presents questionnaire responses from our user study and summarizations from the workshop. Both records are available in our OSF repository under: `Files/Program Codes and Supplementary/questionnaire or workshop`

1.1 Questionnaire Responses

Table 1 presents responses from participants describing their perception workflow after completing the value identification task (Task 1) and gradient comparison (Task 2).

1.2 Workshop Results

Table 2 presents summaries from the workshop where participants reflected on their task completion strategies. The workshop discussions were guided by four key questions:

- **Key Steps**: What steps did you mentally take when completing the two tasks?
- **Key Focusing Aspects**: What aspects did you focus on when viewing the scalar field visualization?
- **Prior Knowledge**: What prior knowledge do you think helped you with these tasks?
- **Difficulties**: What challenges did you encounter, such as difficulty recognizing colors or completing certain tasks, and why?

Our study initially included a third task–pattern perception–to match the experimental design in [1]. However, during our study we found that many participants reported this task is ***very difficult***: multiple groups (G2, G3, G4, and G5) specifically highlighted Task 3 as being particularly challenging. Participants struggled with the cognitive load of switching between different perceptual modes—precise value localization and comparative gradient assessment—within the same task. Given these consistent reports of difficulty and the fact that ***Task 3 can be conceptualized as a combination of Task 1 and Task 2***, we chose to focus our analysis and model evaluation on the two fundamental tasks in the main paper. This allowed us to establish fundamental human perception workflows and evaluate MLLM capabilities on the core perceptual elements before moving forward to more complex combined tasks.

2 Quantitative Comparisons for Fine-tuned MLLMs

This section presents detailed performance comparisons between our fine-tuned models and the human perception results reported by Reda *et al.* [1].

2.1 Task 1: Value Identification

Figure 1 illustrates the performance of our InternVL2-8B model (fine-tuned using the baseline prompting strategy) compared to Reda *et al.*'s original human experiment results [1] on Task 1. Both results exhibit a common trend: error rates increase with higher spatial frequencies, indicating that both humans and MLLMs find it more challenging to identify specific values in more complex scalar fields.

2.1.1 Task 1 Fine-tuning Approach

For Task 1, we developed a specialized fine-tuning approach to enhance the model's ability to identify coordinates corresponding to specific values in scalar field visualizations. Rather than using standard cross-entropy loss alone, we implemented a multi-component loss function that specifically addresses the spatial nature of coordinate prediction:

- **Position-aware Loss**: We added a position loss component that evaluates the Euclidean distance between predicted coordinates and ground truth positions. This encourages the model to predict coordinates that are spatially close to the correct values rather than just syntactically similar responses.
- **Spatial Accuracy Weighting**: The loss function assigns higher weights to spatial accuracy (40% weight) compared to response structure (10% weight), while maintaining cross-entropy as the primary component (50% weight). This balance helps the model focus on accurate coordinate prediction while maintaining coherent text generation.
- **Coordinate Extraction**: We implemented automatic coordinate extraction from both model outputs and ground truth text using pattern matching, allowing the model to learn correct formatting for coordinate responses (as either parentheses or bracket notation).

Table 1: Participant descriptions of their completion strategies. These responses highlight the common patterns in how participants engaged with the scalar field visualizations.

ID	Major	Task 1 Completion Strategy	Task 2 Completion Strategy
P1	Computer Science	First look at the coordinates, then find points from light to dark in the image.	When seeing dark colors in the middle of a square, I feel a greater difference from the surrounding colors and choose that square. If the gradient is smooth, it looks softer.
P2	Computer Science	1. Compare the overall brightness of colors; 2. Compare with surrounding color blocks; 3. Confirm the selected area; 4. Repeat comparisons to complete selection.	Roughly compare the degree of change from the outside to the center; separately find and compare the path with the most drastic change in the square.
P3	Biology Science	1. Observe the overall image, choose the lightest point; 2. Select a small area range and set a linear area from light to dark; 3. Observe the legend on the right and carefully compare colors; 4. Select the point.	First observe the color at the boundary (0.5) on the right color bar; roughly observe the variety of colors inside the square; the more color types, the steeper the gradient.
P4	Computer Science	1. Look at the legend on the right to find corresponding colors; 2. Find matching colors in the scalar field on the left by comparing with surrounding colors; 3. Repeat to select five points.	First look at the color scale on the right, noting that the colors change from purple to blue to light green to red; in the square, first find red and purple. If both exist, the change is more drastic. If not, find red and blue, or red and light green, and gradually narrow the range.
P5	Biology Science	1. Identify easily distinguishable extreme points (e.g., 1000 and 0); 2. Segment and find critical values like 750 close to 1000, then find nearby points.	Find the extreme colors in the square and compare them with the scale on the right, recording the value differences; compare the differences in extreme color values between two squares; if the differences are small, compare the mid-range color gradients by recording the differences between intermediate and extreme colors.
P6	Biology Science	Except for the first and last points which are easiest to judge by brightness, other points are at the transition between lighter and darker areas; find points matching the legend at the boundary.	First identify the highest and lowest colors in the scale contained in the square to determine the range of color change—the larger the range, the steeper the gradient. When two squares have similar color change ranges, compare the distribution areas of the two end colors—more distribution at the ends and less in the middle indicates a steeper gradient.
P7	Computer Science	1. I first looked at the standard, identified that the scale marks indicated finding five colors from light to dark; 2. Compared colors roughly mapping them to white, yellow, orange, red, black; 3. Located the easiest white area first, then found matching yellows and so on.	First, observe the overall color proportion in the square. If most of the square is a single color, the gradient will not be steep. If two squares have similar color proportions, check which one has a richer variety of colors—more colors mean a steeper gradient. Compare the biggest color contrast within a square—the larger the contrast between warm and cool colors, the steeper the gradient.
P8	Computer Science	1. First find the brightest (white) areas; 2. For other points, use a gradual approximation method because boundaries are not obvious; compare across boundaries and finalize.	First find the maximum and minimum values in the square; judge the size of the boundary between different color regions and whether the color change is sharp. Use the analogy of topographic maps: green as the baseline, red as peaks, blue as valleys, and infer the slope.
P9	Biology Science	1. I saw the 'color-number' correspondence chart on the right, found 1000 as the darkest, 0 as the lightest; 2. Easily identified the regions with these extreme values first. 3. Moved slightly from 1000 to find value close to 750 for point; 2. 4. Found midpoints by feeling when no clear reference.	I mainly judge the color range within the square. For example, in task 10, the darkest color value of the black square is close to 0, and the brightest is close to 0.99, while the brightest value in the white square is only around 0.8. Thus, I believe the black square has a steeper color gradient. My judgment relies on instant memory and visual perception of the color scale.
P10	Biology Science	1. Identifying the whitest part is easy; 2. For 0.75 points, yellow in color, light gray in grayscale; 3. 0.5 is orange/mid-gray; 4. 0.25 similarly by color comparison; 5. 0 and 1 are easy to find as pure white and black.	Based on the following judgment criteria: The greater the color span, the steeper the gradient. A color transition from red to blue has a steeper gradient than from red to green or purple to green. Based on color extremities. If bright red and deep purple appear in one image but not the other, the one reaching extremities has a steeper gradient. Compare small areas within the square. If a small region shows an abrupt color change, such as from red to blue in a small area, the gradient is steeper.
P11	Computer Science	1. Find extreme points 0 and 1.0; 2. Other points: find distinct colors like orange and yellow in the color image; in grayscale, locate similar brightness regions near edges.	In most cases, the two squares have obvious differences (similar upper limits but different lower limits: one blue, one purple; or vice versa). For extreme colors (purple-red), the difference is not very noticeable (I basically can't tell).
P12	Biology Science	First observe the color order, realize it goes from light to dark; then based on the legend identify corresponding areas; locate the lightest white area, then sequentially find matching colors until the darkest.	Find several color changes according to the color scale; compare the general color distribution between the black and white squares; match these colors to intervals on the right; subtract the interval values and compare—larger differences indicate steeper color changes.
P13	Computer Science	1. Find the brightest spot; 2. Find darker areas near the brightest spot; 3. Find a third point with tone clearly different from the first and last; 4. Locate transitional colors; 5. Find the darkest point.	Prioritize based on tone change; when the color gradients seem similar, the square with more dark areas usually has a steeper gradient; if the proportion of dark areas is similar, the square with dark areas closer to the center has a steeper gradient.
P14	Computer Science	1. Find the highest value point in areas with clear color gradients; 2. Find points near boundaries at 750, 500, 250; 3. Find the lowest value point.	Divide the square into sections based on color, identifying the areas corresponding to the maximum and minimum values on the scale; compare these maximum and minimum regions to determine which square has a larger color difference.
P15	Computer Science	Two steps: 1. Match numeric values with colors on the right legend; 2. Find matching colors on the left scalar field, continuously comparing selected points to standard legend.	First observe which square has more color types—the one with more colors generally has a steeper gradient. If both squares have roughly the same variety, connect the highest and lowest values (assuming similar differences) and compare the line lengths—the shorter the line, the steeper the gradient. If still unsure, find the region where color changes are most concentrated, connect points with the same difference (larger differences preferred), and compare line lengths.
P16	Computer Science	1. The experiment requires identifying five distinct colors; 1 and 5 are the extreme colors, easy to find; 2. Since color picker tools aren't allowed, compare target colors with surroundings manually.	Gradient judgment is about judging the slope: select two fixed colors, find the corresponding points, and measure the distance.
P17	Computer Science	1. Find intervals corresponding to 1000 and 0; 2. Find relatively uniform transition regions; if available, pick three points close to expectation; otherwise, find regions with large contrast range.	Observe the color scale to determine values corresponding to colors (typically 0, 0.25, 0.5, 0.75, 1); check if the square contains colors close to 0 and 1; if found, continue; if not, determine the approximate range. Compare the maximum and minimum values in the images—if one extreme is close, base the judgment on the other.
P18	Biology Science	1. Roughly scan the image, pick the brightest spot for 1000; 2. Find two steps darker color for 750; 3. Match corresponding colors using legend; 4. Find darkest area for 250; 5. Pick the darkest spot.	First look at the upper and lower limits of colors within the two squares; if the upper and lower limits are similar, compare the colors at the opposite ends of the scale; if not, use the color scale as a reference.

Table 2: Key task completion steps reported by participants during the workshop. These responses highlight how participants systematically approached the perception tasks.

Group	Key Steps	Key Focusing Aspects	Prior Knowledge	Difficulties
G1	Three steps: Compare the overall color blocks, compare with the surrounding color blocks, and confirm.	Focus on the legend and surrounding color blocks. Distinguish colors and observe the target area in relation to its surroundings.	Able to distinguish between the target area and its surroundings.	Task 2
G2	Five steps: 1. Read the question and understand the requirements; 2. Identify the key elements of the question; 3. Extract these elements from the image; 4. Quantify them using the color legend; 5. Compare trends or numerical values to make a decision.	Key considerations include the relationship between colors and numerical values. Focus on the information conveyed by different colors and the legend.	Necessary knowledge: Analyze the problem and summarize the thought process.	Task 3
G3	1. Find extreme values and differences; 2. Locate areas with large gradients and peak values; 3. When difficult to distinguish, use details to perform weighted averaging, and estimate intermediate values.	Identify extreme points; 2. Identify points with major changes.	It is necessary to convert colors into numerical values.	Task 3
G4	First observe the legend, check its highest and lowest points, determine which values are higher or lower, and estimate the color of the intermediate point; second, look for areas in the visualization that match those colors.	Use the legend and scalar field diagram for reference.	Distinguish colors clearly and accurately interpret the numerical values on the legend.	Task 3
G5	Experiment 1: Read the questions, compare the data, and find the color blocks from light to dark; Experiment 2: Read the question, first see if there is a large area of color blocks, find the two colors with the biggest differences, and look at the proportion of each color, the sudden change of color will make the observer think that the color steepness is greater;	Pay attention to color thresholds and the range of variation.	Judging colors accurately is essential.	Task 3

- **Penalty System**: Our loss function includes penalties for out-of-bounds predictions (coordinates outside the visualization area) and for failure to identify any coordinates when they were expected, which helps guide the model toward sensible predictions.

This customized fine-tuning approach directly addresses the spatial reasoning requirements of Task 1, leading to the improved performance seen in our experimental results.

2.1.2 Task 1 Results Comparison

The datasets for both Reda *et al.*'s results and our model's results are available in our OSF repository under: `Files/program codes and supplementary/visualization/LineChart Dataset/`.
For Task 1, we designed a straightforward prompt that clearly defines the task requirements while providing the necessary spatial context:

```
"Find the coordinate of the point with the value {value} in the scalar field image.
The final output must strictly follow this format: [x, y], representing the coordinates
of the color value {value} in the scalar field image. Each coordinate should be represented
as (x, y), where x and y are positive integers within the range: x in (0, 820), y in (0, 630).
Please ensure the output strictly follows this format (x, y)."
```

This prompt emphasizes both the task objective (locating a specific value) and the expected response format. By explicitly defining the valid coordinate range (x in (0, 820), y in (0, 630)), we guide the model toward predictions within the visualization boundaries. The prompt also clarifies that coordinates should be represented as ordered pairs, which helps standardize the model's output for consistent evaluation.

Table 3: Our fine-tuned InternVL2-8B model's results for Task 1 across different colormaps and spatial frequencies. Values in each cell represent $\mu \sim 95\%$CI for logarithmic error rates, where μ is the mean error rate.

Colormap	Freq 1	Freq 3	Freq 5	Freq 7	Freq 9
blueyellow	-1.11[-1.96, -0.26]	-0.59[-1.59, 0.40]	3.53[2.62, 4.43]	2.50[1.51, 3.50]	4.53[3.98, 5.08]
bodyheat	-1.28[-2.08, -0.48]	1.47[0.38, 2.57]	3.98[3.26, 4.70]	2.99[2.01, 3.97]	4.39[3.85, 4.93]
coolwarm	-0.56[-1.51, 0.39]	-0.19[-1.22, 0.84]	1.82[0.82, 2.82]	1.69[0.56, 2.82]	3.67[2.84, 4.49]
cubehelix	-1.41[-2.17, -0.65]	-0.79[-1.69, 0.10]	2.17[1.15, 3.19]	3.27[2.31, 4.24]	4.27[3.63, 4.91]
extbodyheat	-1.67[-2.35, -1.00]	0.79[-0.32, 1.90]	3.00[2.05, 3.95]	2.97[2.07, 3.88]	4.81[4.36, 5.26]
greyscale	-0.58[-1.48, 0.33]	0.49[-0.55, 1.52]	1.92[0.87, 2.98]	3.08[2.16, 4.00]	3.81[3.05, 4.57]
rainbow	1.16[0.09, 2.23]	-1.90[-2.48, -1.32]	2.73[1.77, 3.70]	2.03[0.95, 3.10]	3.72[2.88, 4.57]
singlehue	-1.16[-1.99, -0.34]	0.65[-0.51, 1.81]	1.55[0.57, 2.53]	3.69[2.84, 4.54]	4.11[3.50, 4.72]
spectral	1.06[0.07, 2.04]	1.32[0.26, 2.38]	3.19[2.30, 4.08]	1.66[0.57, 2.74]	3.74[2.96, 4.52]

Table 3 shows our fine-tuned InternVL2-8B model's results for Task 1. Each cell contains the lower bound of the 95% confidence interval, the mean value, and the upper bound, formatted as "lower/mean/upper". Our fine-tuned model results show relatively consistent performance across

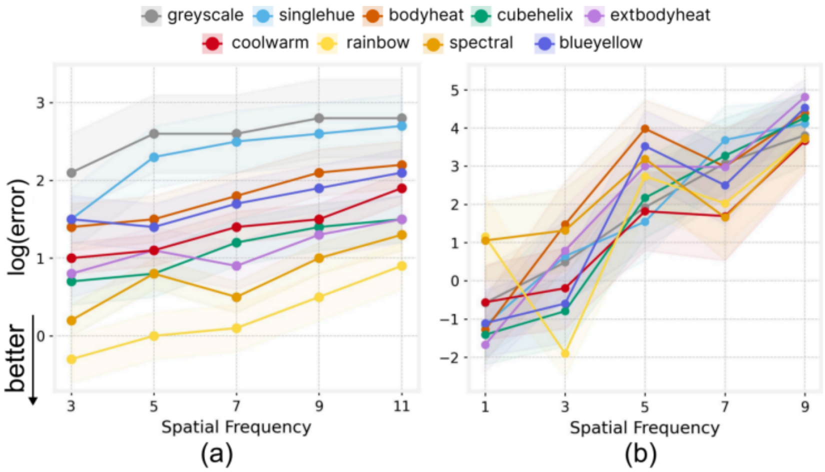

Fig. 1: Task 1 performance comparison between (a) Reda *et al.*'s human study results [1] and (b) our fine-tuned InternVL2-8B model (using baseline prompting strategy). Both results show increasing error as spatial frequency increases, though with different absolute values. The shaded regions indicate 95% confidence intervals.

different colormaps, though with some noteworthy variations. At lower frequencies (Freq 1 and Freq 3), the model often achieves negative error rates for several colormaps, suggesting a tendency toward value overestimation similar to human perception with the rainbow colormap. However, as spatial frequency increases, the error rates uniformly rise across all colormaps, demonstrating the increased difficulty of value identification in more complex scalar fields.

2.2 Task 2: Gradient Comparison

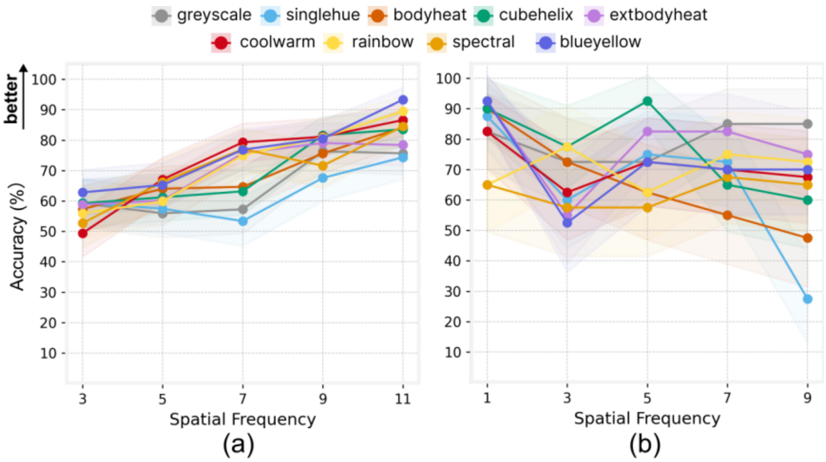

Fig. 2: Task 2 performance comparison between (a) Reda *et al.*'s human study results [1] and (b) our fine-tuned InternVL2-8B model (using CoT prompting strategy). The model shows highest accuracy at lower spatial frequencies, with performance decreasing as frequency increases—a trend that contrasts with human perception patterns. The shaded regions indicate 95% confidence intervals.

Figure 2 illustrates the performance of our InternVL2-8B model (fine-tuned using the CoT prompting strategy) alongside Reda *et al.*'s original human experiment results on Task 2. Interestingly, while our model achieves its highest accuracy at the lowest spatial frequency, its performance decreases as spatial frequency increases—a trend that contrasts with the patterns observed in human perception, highlighting fundamental differences between human and MLLM perception mechanisms.

2.2.1 Task 2 Results Comparison

For Task 2, we designed a more structured prompt that guides the model through the process of analyzing gradient information:

```
"You are an image analysis assistant, and your task is to complete the following steps:
1. I will provide a two-dimensional scalar field image, with dimensions of {width} pixels in width (horizontal)
and {height} pixels in height (vertical). The image is divided into two parts: the left side shows the scalar
field visualization, and the right side contains the colorbar legend.
2. The scalar field visualization contains two hollow square boxes, which color is {color_box_1} or {color_box_2}
respectively.
3. The coordinate system origin (0,0) is located at the lower-left corner, with the x-axis increasing to the right
and the y-axis growing upward.
```

```
4. Core task: Your task is to analyze the color legend and determine which of the two boxes has a steeper color
gradient. This means identifying which box has a faster rate of change in the scalar color value. You should focus
on how quickly the color changes within each box, specifically the rate of change from the highest value to the
lowest value.
5. The final output must strictly follow this format: [{color_box_1}] or [{color_box_2}], indicating the box with
the steeper color gradient. You must answer the question in English"
```

This Chain-of-Thought (CoT) prompt is more detailed than the one used for Task 1 and explicitly breaks down the gradient comparison process into steps. It first establishes the image structure and coordinate system, then defines the core task with an explanation of what "steeper gradient" means in this context. By emphasizing the need to analyze "how quickly the color changes" and "the rate of change from highest value to lowest value," the prompt encourages the model to focus on relevant gradient information rather than other visual features. The required output format is also clearly specified to ensure consistent evaluation.

Table 4 shows our fine-tuned InternVL2-8B model's results for Task 2. For the results, values represent accuracy rates as percentages. Our fine-tuned model shows the opposite trend from the Reda one. The model achieves its highest accuracy at the lowest spatial frequency (Freq 1), with performance generally declining as frequency increases. This indicates a fundamental difference in how MLLMs process gradient information compared to human visual systems. The model appears to struggle with extracting gradient information from increasingly complex spatial patterns, whereas humans seem to benefit from these additional visual cues.

Table 4: Our fine-tuned InternVL2-8B model's results for Task 2 across different colormaps and spatial frequencies. Values in each cell represent $\mu \sim 95\%$CI for accuracy rates, where μ is the mean accuracy rate (as percentages).

Colormap	Freq 1	Freq 3	Freq 5	Freq 7	Freq 9
blueyellow	92.5[80.1, 97.4]	52.5[37.5, 67.1]	72.5[57.2, 83.9]	70.0[54.6, 81.9]	70.0[54.6, 81.9]
bodyheat	90.0[76.9, 96.0]	72.5[57.2, 83.9]	62.5[47.0, 75.8]	55.0[39.8, 69.3]	47.5[32.9, 62.5]
coolwarm	82.5[68.1, 91.3]	62.5[47.0, 75.8]	72.5[57.2, 83.9]	70.0[54.6, 81.9]	67.5[52.0, 79.9]
cubehelix	90.0[76.9, 96.0]	77.5[62.5, 87.7]	92.5[80.1, 97.4]	65.0[49.5, 77.9]	60.0[44.6, 73.7]
extbodyheat	92.5[80.1, 97.4]	55.0[39.8, 69.3]	82.5[68.1, 91.3]	82.5[68.1, 91.3]	75.0[59.8, 85.8]
greyscale	82.5[68.1, 91.3]	72.5[57.2, 83.9]	72.5[57.2, 83.9]	85.0[70.9, 92.9]	85.0[70.9, 92.9]
rainbow	65.0[49.5, 77.9]	77.5[62.5, 87.7]	62.5[47.0, 75.8]	75.0[59.8, 85.8]	72.5[57.2, 83.9]
singlehue	87.5[73.9, 94.5]	60.0[44.6, 73.7]	75.0[59.8, 85.8]	72.5[57.2, 83.9]	27.5[16.1, 42.8]
spectral	65.0[49.5, 77.9]	57.5[42.2, 71.5]	57.5[42.2, 71.5]	67.5[52.0, 79.9]	65.0[49.5, 77.9]

References

[1] 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 the 2018 ACM CHI Conference on Human Factors in Computing Systems*, 2018. doi: 10.1145/3173574.3173846 1, 4