Seeing Through the Overlap:

Evaluating and Adapting Multimodal LLMs for Graphical Perception in Color-Encoded Scalar Field Visualization (Supplementary Material)

Appendices:

We provide additional experimental details and results shown as below. We have also made our experimental data, code, and models available at https://osf.io/y4pgm/?view_only=dadc9ff8b62d4a27a2a0d852ccde30bc

A. Appendix A presents partial results from the user study.
B. Appendix B presents the Task 1 results of Reda's method and our model.
C. Appendix C presents the Task 2 results of Reda's method and our model.

A. HUMAN PERCEPTION WORKFLOW

We collected task results from human participants along with think-aloud protocols, which were recorded as CSV files.

(1) Partial Results from Task 1/2:

A	B	C	D	E
username	experiment_index	image_id	csv_file	true_values
P1	5	gray1	./static/Data/data1.csv	[0.968323554431438, 0.7452261039318275, 0.4090125512171292, 0.152874113208949, 0.0187049190492216]
P2	6	gray2	./static/Data/data2.csv	[0.9456387803793644, 0.6156818744479161, 0.44087587882887513, 0.15682850349623478, 0.06852732803393705]
P3	3	hot4	./static/Data/data4.csv	[0.988825265926391 3, 0.7628710841887123, 0.4688670869007887, 0.341530945923688, 0.09624059397142275]
P4	3	hot4	./static/Data/data4.csv	[0.9420796675789277, 0.7729329443087757, 0.4856570801777266, 0.19875581411322, 0.030249558235488545]
P5	6	gray2	./static/Data/data2.csv	[0.9386310324062205, 0.6769160711810509, 0.43629430206406, 0.2825344575780691, 0.05632074542461912]
P6	4	hot5	./static/Data/data5.csv	[0.9056475714994979, 0.6831094367437702, 0.5155963042941022, 0.3146413877908389, 0.08611709389888748]
P7	5	gray1	./static/Data/data1.csv	[0.9539216756804093, 0.8489206602487536, 0.430187064761321, 0.29147998624405047, 0.002959737646948046 4]
P8	3	hot4	./static/Data/data4.csv	[0.9746726887056579, 0.8580585624483673, 0.4194245773236388, 0.22918317461755763, 0.16780661427697732]
P9	9	gray5	./static/Data/data5.csv	[0.940614752142621 8, 0.6588352623552841, 0.4603956857152232, 0.1796073449779013, 0.17485892455656477]
P10	6	gray2	./static/Data/data2.csv	[0.863184324051263, 0.6541464988944006, 0.4041547859872596, 0.23041296064177405, 0.0299930188090062]

(2) Question: How did you finish the Task 1 and Task 2? Please describe your think-aloud protocols:

Example of think-aloud protocols for Task 2:

"First, observe which bounding box contains a greater variety of colors. Generally, the one with more color variations tends to have a steeper gradient.

If the two boxes contain a similar variety of colors, connect the highest and lowest value points within each box (assuming the maximum and minimum values in both boxes are comparable). The box with the shorter line segment indicates a steeper gradient.

The first two steps are based on prior intuitive judgments. If it is still difficult to estimate, locate the region within each box where color changes are most concentrated. Then, connect two points with the same difference in value (the larger the difference, the better) and compare the lengths of the resulting line segments. The box with the shorter segment has the steeper gradient."

A	B	C	D	E	
No	Submit Time	Name	Age	Major	How did you complete Test Task 1? Please describe your thought process step-by-step (1, 2, 3...).
2	2025/3/8 16:53:5	P1	20	Computer Science	First look at the coordinates, then find points from light to dark in the image.
3	2025/3/8 16:56:3	P2	18	Computer Science	1. Compare the overall brightness of colors; 2. Compare with surrounding color blocks; 3. Confirm the selected area;
4	2025/3/8 16:57:5	P3	19	Biology Science	1. Observe the overall image, choose the lightest point; 2. Select a small area range and set a linear area from light
5	2025/3/8 17:01:4	P4	19	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 b
6	2025/3/8 17:03:1	P5	19	Biology Science	1. Identify easily distinguishable extreme points (e.g., 1000 and 0); 2. Segment and find critical values like 750 close
7	2025/3/8 17:04:3	P6	21	Biology Science	Except for the first and last points which are easiest to judge by brightness, other points are at the transition betweer
8	2025/3/8 17:05:3	P7	19	Computer Science	1. I first looked at the standard, identified that the scale marks indicated finding five colors from light to dark 2. Comp
9	2025/3/8 17:05:4	P8	19	Computer Science	1. First find the brightest (white) areas; 2. For other points, use a gradual approximation method because boundaries
10	2025/3/8 17:06:2	P9	19	Biology Science	1. I saw the 'color-number' correspondence chart on the right, found 1000 as the darkest, 0 as the lightest. 2. Easily

(3) Workshop:

(i) What steps did you mentally take when completing the two tasks?

(ii) What aspects did you focus on when viewing the scalar field visualization?

(iii) What prior knowledge do you think helped you with these tasks?

(iv) What challenges did you encounter, such as difficulty recognizing colors or completing certain tasks, and why?

Here is an example of answer for (i):

"Five Steps:

(1) Read the question carefully to understand the task requirements.

(2) Identify the key elements of the question.

(3) Extract these elements from the visualization.

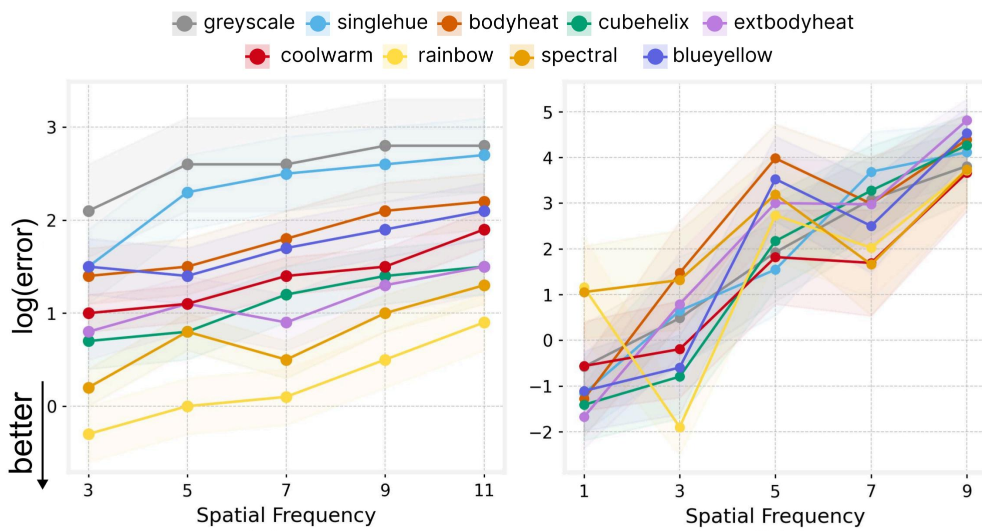
(4) Quantify them using the color legend.

(5) Compare trends or numerical values to make a decision."

Submission Time	In the experiment, what are the key steps you took to complete the task? How many steps did you generally take? Which steps? (Required)
3/8 2025 17:19	Three steps: Compare the overall color blocks, compare with the surrounding color blocks, and confirm.
3/8 2025 17:21	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 a
3/8 2025 17:22	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
3/8 2025 17:22	First observe the legend, for example, check its highest and lowest points, determine which values are higher or lower, and estimate the color of the intermediate point; secc

B. Fig. 5 (5.2.2 Quantitative Comparisons for Fine-tuned MLLMs):

Figure 5 illustrates the performance of our InternVL-8B model, fine-tuned using the Baseline prompt, and Reda's original experiment on Task 1. Both results exhibit a trend of increasing error with higher spatial frequencies.



(a) (b)

(a) Reda's result for Task 1

You can find this dataset on OSF: **Files/program codes and supplementary/visualization/LineChart Dataset/data1.csv**

Mean value:

Colormap/Frequency	3	5	7	9	11
blueyellow	1.557562	1.457273	1.688821	1.982881	2.16987
bodyheat	1.475544	1.500168	1.857633	2.157204	2.212857
coolwarm	1.056333	1.169798	1.395913	1.449663	1.903913
cubehelix	0.819745	1.093812	1.183603	1.441475	1.838876
extbodyheat	0.663722	1.218771	0.901003	1.495629	1.488034
greyscale	2.075869	2.551534	2.572152	2.778305	2.741026
rainbow	-0.37664	-0.05495	0.114717	0.495496	0.885991
singlehue	1.506048	2.322828	2.443536	2.487061	2.662049
spectral	0.234703	0.822706	0.540161	1.047867	1.337277

Confidence interval:

Lower:

Colormap/Frequency	3	5	7	9	11
blueyellow	1.294081	1.195107	1.422741	1.768182	1.950408
bodyheat	1.207772	1.24028	1.578843	1.90142	1.973689
coolwarm	0.753196	0.874484	1.12271	1.166186	1.675466
cubehelix	0.59177	0.859155	0.923373	1.196606	1.596691
extbodyheat	0.423608	0.972627	0.634974	1.271377	1.229809
greyscale	1.763083	2.264463	2.262908	2.519511	2.494019
rainbow	-0.61646	-0.31632	-0.14698	0.235894	0.645352
singlehue	1.188471	2.017388	2.157814	2.205176	2.408484

spectral	-0.04349	0.54521	0.292795	0.79833	1.107459

Upper:

Colormap/Frequenc	3	5	7	9	11
blueyellow	1.821043	1.719439	1.9549	2.197581	2.389332
bodyheat	1.743316	1.760056	2.136423	2.412988	2.452025
coolwarm	1.35947	1.465111	1.669115	1.733139	2.13236
cubehelix	1.047721	1.328468	1.443832	1.686344	2.081062
extbodyheat	0.903837	1.464914	1.167031	1.719881	1.746259
greyscale	2.388655	2.838605	2.881395	3.037098	2.988034
rainbow	-0.13682	0.206427	0.376415	0.755099	1.12663
singlehue	1.823625	2.628267	2.729259	2.768946	2.915613
spectral	0.512897	1.100202	0.787526	1.297403	1.567095

(b) Our model's result for Task 1

You can find this dataset on OSF: **Files/program codes and supplementary/visualization/LineChart Dataset/8b_baseGT_baseline.xlsx**

Model: InternVL-8B fine-tuned using the Baseline prompt.

Prompt:

f"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∈(0, 820), y∈(0, 630). Please ensure the output **strictly** follows this format (x, y)."

Mean value:

Colormap/Frequency	1	3	5	7	9
blueyellow	-1.1081	-0.5948	3.5268	2.5012	4.5294
bodyheat	-1.2788	1.4737	3.98	2.9868	4.3904
coolwarm	-0.5618	-0.193	1.8199	1.6937	3.6658
cubehelix	-1.4103	-0.7937	2.1704	3.2747	4.2668
extbodyheat	-1.6749	0.7922	3.0021	2.9739	4.8121
greyscale	-0.5787	0.4853	1.9249	3.0815	3.8088
rainbow	1.1578	-1.8977	2.7349	2.0266	3.7226
singlehue	-1.1639	0.6467	1.5477	3.6871	4.1147
spectral	1.0553	1.3175	3.1909	1.6587	3.7367

Confidence interval:

Lower:

Colormap/Frequency	1	3	5	7	9
blueyellow	-1.956983416	-1.592876553	2.621473983	1.505393321	3.979198395
bodyheat	-2.082093598	0.380038079	3.255306694	2.006733439	3.853826768
coolwarm	-1.5125727	-1.222285084	0.819230855	0.563373077	2.843096386

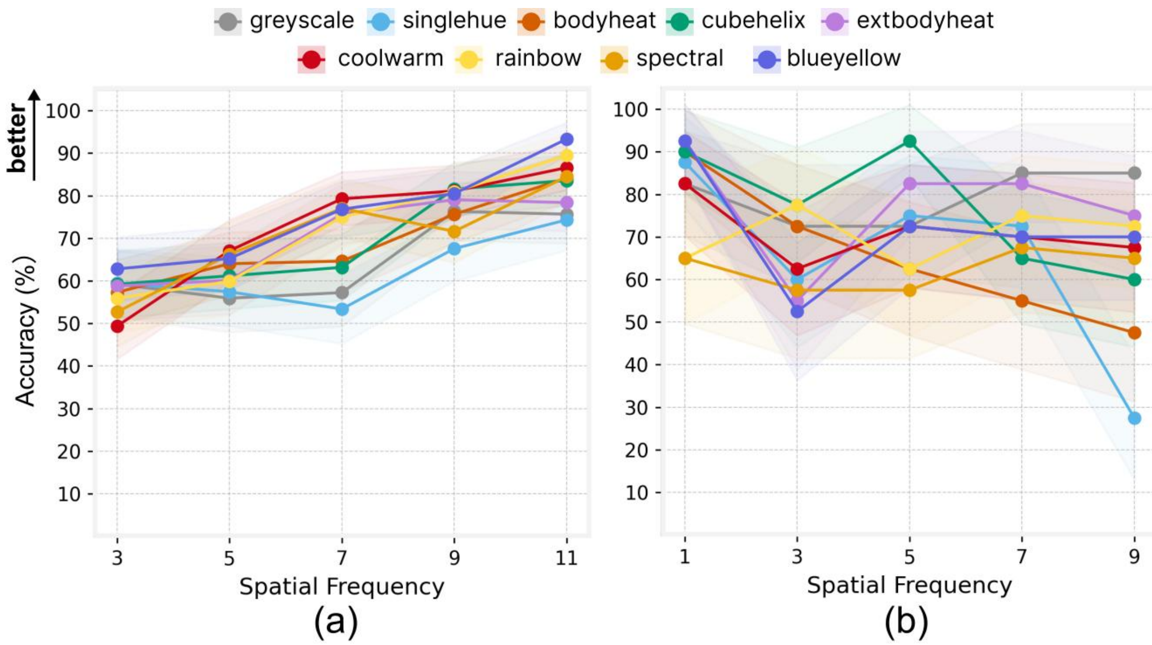
cubehelix	-2.170859216	-1.688294678	1.149943907	2.313930156	3.627366591
extbodyheat	-2.349953196	-0.320014814	2.05022858	2.065181974	4.363639284
greyscale	-1.483057418	-0.547462355	0.874504356	2.163309323	3.048733307
rainbow	0.088638354	-2.479820878	1.770451521	0.951722745	2.875216762
singlehue	-1.985277154	-0.512818594	0.569971721	2.837833763	3.504698543
spectral	0.073700392	0.258155175	2.303533449	0.574844804	2.956556864

Upper:

Colormap/Frequency	1	3	5	7	9
blueyellow	-0.259187855	0.403249641	4.432213717	3.49693547	5.079701082
bodyheat	-0.475540658	2.567408284	4.704716765	3.96686242	4.926881508
coolwarm	0.388983193	0.83621336	2.82054063	2.824083891	4.488438313
cubehelix	-0.649730187	0.100965483	3.190827555	4.23549085	4.906324113
extbodyheat	-0.999753221	1.904456168	3.954043411	3.882636511	5.260508924
greyscale	0.32567948	1.517984967	2.975343194	3.999694338	4.568932009
rainbow	2.227037163	-1.315491431	3.699313865	3.101430593	4.570000174
singlehue	-0.342533659	1.80627882	2.525453931	4.536443801	4.724649023
spectral	2.036876005	2.376943466	4.078317062	2.742583907	4.516760588

C. Fig. 8 (5.3.2 Quantitative Comparisons for Fine-tuned MLLMs)

Figure 8 illustrates the performance of our InternVL-8B model, fine-tuned using the CoT prompt and Reda's original experiment on Task 2. the overall accuracy is highest at the lowest spatial frequency, but the performance decreases as the spatial frequency increases, which are contrast to the human perception.



(a) Reda's result for Task 2

You can find this dataset on OSF: **Files/program codes and supplementary/visualization/LineChart Dataset/data2.csv**

Mean value:

Colormap/Frequency	3	5	7	9	11
blueyellow	0.628049	0.652439	0.768293	0.804878	0.932927
bodyheat	0.573171	0.640244	0.646341	0.756098	0.841463
coolwarm	0.493902	0.670732	0.792683	0.810976	0.865854
cubehelix	0.592105	0.611842	0.631579	0.815789	0.835526
extbodyheat	0.587838	0.601351	0.756757	0.790541	0.783784
greyscale	0.592105	0.559211	0.572368	0.763158	0.756579
rainbow	0.559211	0.598684	0.75	0.809211	0.894737
singlehue	0.594595	0.574324	0.533784	0.675676	0.743243
spectral	0.527027	0.662162	0.77027	0.716216	0.844595

Confidence interval:

Lower:

Colormap/Frequency	3	5	7	9	11
blueyellow	0.553	0.579	0.703	0.744	0.894
bodyheat	0.497	0.566	0.572	0.69	0.785
coolwarm	0.417	0.598	0.73	0.75	0.813
cubehelix	0.513	0.533	0.554	0.753	0.776
extbodyheat	0.508	0.522	0.687	0.724	0.717

greyscale	0.513	0.479	0.493	0.695	0.688
rainbow	0.479	0.52	0.68	0.746	0.845
singlehue	0.515	0.494	0.452	0.599	0.672
spectral	0.446	0.585	0.702	0.643	0.786

Upper:

Colormap/Frequency	3	5	7	9	11
blueyellow	0.703	0.726	0.834	0.866	0.972
bodyheat	0.65	0.714	0.72	0.823	0.898
coolwarm	0.571	0.743	0.855	0.872	0.919
cubehelix	0.671	0.69	0.709	0.878	0.895
extbodyheat	0.668	0.681	0.827	0.857	0.851
greyscale	0.671	0.639	0.652	0.832	0.826
rainbow	0.639	0.677	0.82	0.872	0.944
singlehue	0.675	0.655	0.615	0.752	0.814
spectral	0.608	0.739	0.839	0.79	0.904

(b) Our model's result for Task 2

You can find this dataset on OSF: **Files/program codes and supplementary/visualization/LineChart Dataset/8b_cotGT_cot.xlsx**

Model: InternVL-8B fine-tuned using the CoT prompt.

Prompt: f"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 731 pixels in width (horizontal) and 449 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"

Mean value:

Colormap/Frequency	1	3	5	7	9
blueyellow	92.5	52.5	72.5	70	70
bodyheat	90	72.5	62.5	55	47.5

coolwarm	82.5	62.5	72.5	70	67.5
cubehelix	90	77.5	92.5	65	60
extbodyheat	92.5	55	82.5	82.5	75
greyscale	82.5	72.5	72.5	85	85
rainbow	65	77.5	62.5	75	72.5
singlehue	87.5	60	75	72.5	27.5
spectral	65	57.5	57.5	67.5	65

Confidence interval:

Lower:

Colormap/Frequency	1	3	5	7	9
blueyellow	80.13577	37.49736	57.16504	54.56998	54.56998
bodyheat	76.94822	57.16504	47.03244	39.82909	32.93547
coolwarm	68.05001	47.03244	57.16504	54.56998	52.01775
cubehelix	76.94822	62.4969	80.13577	49.50588	44.59589
extbodyheat	80.13577	39.82909	68.05001	68.05001	59.80604
greyscale	68.05001	57.16504	57.16504	70.92768	70.92768
rainbow	49.50588	62.4969	47.03244	59.80604	57.16504
singlehue	73.88788	44.59589	59.80604	57.16504	16.10802
spectral	49.50588	42.19507	42.19507	52.01775	49.50588

Upper:

Colormap/Frequency	1	3	5	7	9
blueyellow	97.4164	67.06453	83.89198	81.92515	81.92515
bodyheat	96.04205	83.89198	75.77702	69.29469	62.50264
coolwarm	91.25459	75.77702	83.89198	81.92515	79.9155
cubehelix	96.04205	87.68391	97.4164	77.86547	73.65167
extbodyheat	97.4164	69.29469	91.25459	91.25459	85.81288
greyscale	91.25459	83.89198	83.89198	92.93881	92.93881
rainbow	77.86547	87.68391	75.77702	85.81288	83.89198
singlehue	94.5405	73.65167	85.81288	83.89198	42.83496
spectral	77.86547	71.49061	71.49061	79.9155	77.86547