ChartAB: A Benchmark for Chart Grounding & Dense Alignment

Anonymous ACL submission

Abstract

Charts play important roles in visualization, reasoning, and communication in data analysis and idea exchange between humans. However, vision-language models (VLMs) still lack accurate understanding of the details and struggle to extract fine-grained structural information from charts. Such limitations in chart grounding also hinder their capability to compare multiple charts and reason about their difference. In this paper, we develop a novel "**ChartA**lign **B**enchmark (ChartAB)" to provide a full-spectrum evaluation of VLMs in chart grounding tasks, i.e., extracting tabular data, allocating visualization elements, and recognizing various attributes from charts of diverse types and complexities. We develop a JSON template to facilitate the calculation of evaluation metrics specifically designed for each grounding task. By applying a novel two-stage inference workflow, the benchmark can further evaluate VLMs' capability of aligning and comparing elements/attributes in two charts. Our analysis of evaluations on several recent VLMs sheds novel insights on their perception biases, weaknesses, robustness, and hallucinations in chart understanding. These observations expose the fine-grained discrepancies among VLMs in chart understanding tasks and indicate specific skills that need to be strengthened in existing VLMs.

1 Introduction

Recent large multimodal models (LMMs) such as vision-language models (VLMs) have achieved remarkable breakthroughs in aligning vision modality with language models, so challenging language-level reasoning can be performed on visual input signals, opening the possibility of various applications that naturally depend on interactions between the two modalities. One critical class of applications is chart understanding and reasoning, which has broad applications in finance, data science, mass media, biology and other scientific discoveries, and where ideas and information are exchanged through visualizations. In these applications, measuring the numerical values in charts, comparisons between visual elements (e.g., bars or curves), correspondence between colors/numbers/names/markers, and recognition of attributes are critical skills for downstream tasks. Most of them require accurate grounding of the structured details in charts. Moreover, dense alignment of elements in multiple charts is also a widely demanded skill in practical scenarios. These raise new open challenges to VLMs.

Instead of focusing on charts, existing VLMs have been mainly pretrained and finetuned on natural images and common questions/instructions that are not fully compatible with chart understanding tasks. Unlike perceiving objects' shape, pose, and semantic meanings in natural images, accurate measurements and comparison of geometric/graphic components, understanding of their structure and layout, and manipulation of their positions and rich texts are more important to the perception and reasoning with chart images. However, it is usually challenging for VLMs to gain these capabilities, leading to hallucinations and misinterpretations in chart-centric tasks.

Despite recent growing interest in chart-related tasks, the VLMs and benchmarks specifically designed for charts usually focus on simple QA tasks that cannot comprehensively assess the capabilities of VLMs in grounding and understanding components in charts for more general-purpose tasks. Moreover, the alignment of layout and components between multiple charts has not been explored in previous works. Hence, there is still a lack of benchmarks focusing on evaluating the above critical skills on grounding and dense alignment of charts.

In this paper, we take the first step towards evaluating and analyzing general-purpose VLMs on chart grounding and multi-chart dense alignment.

We split the information to be grounded in a chart into two categories: (1) the visualized data, and (2) the chart attributes (e.g., colors, style, legend, sizes, positions) specifying the visualization design, components, and layout. The grounding task is defined as extracting the data table and the attributes from each chart image, while the dense alignment is to find the difference between two charts. These two tasks are the upstream tasks or critical subroutines in various chart-centric applications.

To this end, we develop a comprehensive benchmark using pairs of similar charts to evaluate models' performance on the two tasks regarding each type of information in the two categories. To create a pair of similar charts, we perturb an existing chart by randomly changing (1) one or a few data cells in the data table, and/or (2) an attribute in the script used to generate the original chart. To maximize the potential of VLMs and evaluate their full capability, we propose a multi-stage information extraction and QA pipeline, in which the VLMs are first queried with a grounding task of specified information in each chart, followed by comparing the grounding results for the two charts. It harnesses specified JSON templates to guide the grounding and alignment of different types of information. We further propose several novel evaluation metrics to address the symmetry and ambiguity of various types of information, facilitating the quantitative comparison among different VLMs.

Our analysis reveals the weaknesses of existing VLMs in chart perception and understanding for dense grounding and alignment. The observed mistakes reflect their biases and hallucinations on some chart components, providing several critical insights to improve VLMs. The evaluation results also reflect how the differences between models, chart types, and queried data/attributes affect the benchmarking performance. We further evaluate the robustness of VLMs in accurately extracting data from charts under different design choices of visualizations, e.g., before and after changing the chart type or colors. **Our contributions and advantages** are summarized below:

- We introduce the first comprehensive benchmark "`ChartAB`" to systematically evaluate VLMs' capabilities on dense grounding and multi-chart alignment of visualized data and attributes of components in chart images.
- We propose a holistic evaluation suite, including a multi-stage pipeline converting charts into JSON files with specific templates for tasks regarding data/attributes, and a rating scheme of the grounding/alignment performance based on VLMs' answers.
- Our evaluation and analysis of existing VLMs expose their weaknesses in fine-grained understanding of charts, hallucinations, and their vision encoders' biases in perceiving critical features/structures of charts.
- We evaluate the robustness of chart grounding and alignment under perturbations of chart attributes. It provides novel insights for the design of high-quality charts.

2 Related Work

VLMs for Charts. Early methods like FigureQA (Kahou et al., 2017) and PlotQA (Methani et al., 2020) focused on traditional architecture and rule-based reasoning. Subsequent methods (DePlot (Liu et al., 2022a), MatCha (Liu et al., 2022b), StructChart (Xia et al., 2023)) worked on module-based augmentation for efficient grounding of chart-data and plot-code for downstream applications. Recent methods focus on an integrated multi-task paradigm. ChartAssistant (Meng et al., 2024) utilizes mixed visual encoding and augmented pre-training for robust multi-task abilities. ChartVLM (Xia et al., 2024) applies a difficulty-based cascading decoding mechanism to augment the model's reasoning abilities using intermediate representations. Increasingly, general-purpose VLMs have shown remarkable abilities in chart cognition and reasoning.

The task-specificity in chart-specific VLMs from instruction-tuned datasets make them infeasible for general or newer tasks. The strong performing general purpose VLMs with task flexibility are hence evaluated in our benchmark experiments.

Chart Understanding Benchmarks. are intended for tasks like question answering (PlotQA (Methani et al., 2020), ChartQA (Masry et al., 2022)), summarization (Chart-to-text (Kantharaj et al., 2022b)), explanation-generation (OpenCQA (Kantharaj et al., 2022a)). Multi-task benchmarks including ChartLlama (Han et al., 2023), ChartX (Xia et al., 2024) rely on GPT-4 (Achiam et al., 2023) driven data curation, and agglomerating of modality information for downstream cognition tasks. Recent works specifically focus on expanding QA scope to overcome increased saturation by

VLMs, for example CharXiv (Wang et al., 2024b) focuses on charts in research papers, SciGraphQA (Li and Tajbakhsh, 2023) evaluates multi-turn QA, ChartQAPro (Masry et al., 2025) includes diverse visualizations such as dashboards, infographs, and flexible questions (hypothetical, unanswerable). T

he QA driven benchmarks limit model's ability to question-specific encodings and fail to evaluate understanding of finer-level chart details.

Visual Grounding has been extensively utilized for augmenting dense-level abilities. DePlot (Liu et al., 2022a) trained transformer for image-to-csv generation utilizing novel table comparison for evaluation. MatCha (Liu et al., 2022b) developed image to data-table & plot-code derendering for subsequent downstream tasks. Beyond charts: Grounded-SAM model (Ren et al., 2024) utilizes Grounding-DINO (Liu et al., 2024) for enhanced dense-level open-set object tracking. BLIP-2 (Li et al., 2023) is extensively integrated with VLMs for VQA related tasks. LlaVa-grounded (Zhang et al., 2024) has enabled detailed text description of multi-object natural images utilzing image-text grounding for instruction tuning.

The above works showcase inference augmentation with grounding to expand model capabilities especially for finer-level tasks requiring precision in values.

Multi-Image Reasoning. Multiple benchmarks have been developed on evaluating VLM's multi-image reasoning. MMMU's (Yue et al., 2024) encompasses interleaved examples with multi-images mainly from medical, cartoon, art and technical domains. MUIRBench's (Wang et al., 2024a) multi-chart based diagram QnA questions are focsed on coarse-level understanding. MMIR's (Zhao et al., 2024) chart understanding section is centered on cross-modal alignment i.e. chart-image & plotting-code correctness matching. MileBench's (Song et al., 2024) semantic understanding tasks contain text-rich images attending to text extraction and understanding in domain of image-OCR, documents and slides.

Current multi-image reasoning paradigm's chart understanding centers on traditional image-based-QA, image-to-code, image-to-OCR, interleaved text-image tasks missing evaluation of finer-level understanding of chart's plot attributes and data.

3 ChartAB: Chart Grounding and Alignment Benchmark

We present ChartAlignBench: the first dataset for evaluating dense-level alignment in charts across following 3 tasks: *Data Alignment*, *Plot-Attribute Alignment*, *Robustness*. The 3 alignment tasks consist of $\sim$ 3,600, $\sim$ 2,000, $\sim$ 3,300 instances respectively. For Data Alignment & Plot-Attribute Alignment, each instance consists of pair of chart-images diverging in finer-level chart-data & plot-attribute respectively. For Robustness, each instance contains 5 pairs of chart-images, each pair with identical chart-data divergence but variation in a plot-attribute across the 5 pairs.

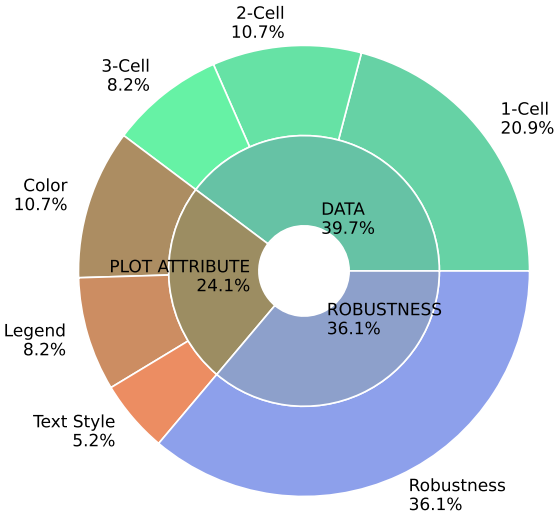


Figure 1: Statistics of tasks in ChartAB.

3.1 Grounding of Single Chart

VLM's *multi-modal capabilities* are tend to be constrained due to cross-modal bottlenecks.

Dense-level understanding is constrained due to multiple reasons. The vision-encoder driven global embedding for image doesn't include object-level representations. The prompt dependent task-learning of the models lack robustness against variation in prompt variations making them unsuitable for generalization and diverse application.

Multi-image reasoning also shows weakness. Each image is evaluated as separate entity and lacks any sequence-level comparison between images. The use of inter-leaved learning for evaluation of twined text & image inputs shows attention bias towards a specific image. Model's prompt-text understanding suffers from position-reference ambiguity leading to hallucinations and mixed-up references due to absence of spatial anchoring.

Grounding of chart-image to textual form enables extraction of rich semantic meaning for performing subsequent dense-alignment for the chart image-pairs. The tokenization of different visual components into textual format allows element-wise representation and correspondence ability between specific chart objects. And prevents any attention bias or prompt sensitive alterations in model outputs. The approach shows cognitive parallels to human understanding of charts: first parsing their structure (i.e. axes, legends, marks), and then mapping to semantic information.

3.2 Dense Alignment between Two Charts

The fundamental model inference for the dense-alignment evaluation involves comparing 2 similar looking charts-images, which differ in (1) *data points* OR (2) *plot attributes*. The task is structured as pair comparison instead of single chart-grounding as we want to evaluate end-to-end model ability of identifying finer-level differences between the charts. It is intended to mirror human-analysis of chart data, which focuses on comparative reasoning. A visualization practitioner while developing charts also tries to iterate across designs by applying finer changes each time (e.g. changing a color, text size). In a task involving identification of pair-wise changes allows dense-alignment labels to be clearly and consistently defined, and provide rich supervision for model learning.

Following are the tasks for evaluating model ability to detect dense-level alignment to analyze distinctive chart perception and reasoning abilities.

3.2.1 Data Alignment

The task evaluates data alignment in image-pairs, i.e. difference in values of cells in the data-table which is visualized by the charts. The finer-level cell-changes involves performing (1) *1-cell change*, (2) *2-cell change*, (3) *3-cell change* between the chart images. The task aims to analyze model's ability to perceive change in visual encoding property (e.g. position, shape, size) in the chart image, and ability to map it to the specific cell i.e. row & column header in data-table modality. Along with measure of the cell-change utilizing the visual components of image describing scale and values.

3.2.2 Plot-Attribute Alignment

The task evaluates plot-attribute alignment in image-pairs, i.e. difference in values of attributes which are part of the plotting-design. We assess the capability through three alignment tasks:- (1) *color alignment*, (2) *legend alignment*, (3) *text-style alignment*. The plot-alignment task aims to analyze model's ability to perceive finer-level visual design changes.

The details are ediscussed in the Appendix (A.1).

3.2.3 Robustness

The task evaluates robustness of data-alignment against variation of plot attributes namely: colors, legend, text-style. A set of multiple chart-pairs with same pair-wise data value difference but variation in a plot attribute is utilized to infer impact on data-alignment ability of model from changes in the attribute values across the set. The task aims to analyze model's ability to consistently deliver accurate data alignment under variations of chart plot's design. Robust model is expected to generate unchanged response despite the plot changes while variation in model responses can be utilized to quantify the susceptibility to plot design changes. Hence providing measure of data-alignment robustness against plot attribute.

Figure 2: Multi-step approach of *o4-mini* reasoning model on color-alignment evaluation example.

3.3 A Two-Stage Evaluation Pipeline

The two-stage approach fundamentally envisions the dense-alignment task as decomposable into sub-tasks utilizing the visual-to-text grounding to perform finer-level analysis. The task decomposition enables splitting complex finer-level reasoning into smaller steps for efficient element-wise comparisons and handling model biases.

It is inspired from the multi-step approach used in SOTA reasoning models. Fig. 2 shows an example color-alignment evaluation for the o4-mini model. The model's thought-stream shows initial grounding of an entity's color from each of two chart images respectively, followed by dense-alignment on the grounded color information. This multi-step approach of the model validates our task decomposition approach and its ability for efficient multi-image dense alignment.

Fig. 3 describes the data-alignment pipeline. *First-stage* results in an intermediate text with semantic information on data of the chart image. The interpretable nature and element-wise representation enables subsequent reasoning for dense-level alignment. *Second-stage* involves VLM reasoning by applying discriminative comparison on the grounded results from first-stage as input information, for the specific dense-alignment task.

(1) Pipeline details and (2) Essentiality of second-stage are discussed in the Appendix (A.3).

3.4 Dataset Curation

We apply *perturbation* based approach to generate ChatAlignBench dataset along with ground-truth, from the *ChartX dataset* as source dataset. Details are discussed in Appendix (A.2).

3.5 Evaluation metric

All alignment evaluation scores are normalized to $[0, 10]$**.** We average scores across the chart-type (e.g. bar-chart, line-chart) for all chart-pairs corresponding to them. We prepared the evaluation criteria for effectively differentiating model performance across tasks, and quantify performance aspects for chart data and plot attributes part of dense alignment.

The alignment evaluation metrics are discussed in the Appendix (A.4).

We also performed grounding analysis to evaluate the grounding stage ability of the models. Grounding being for elements of single image, we calculate an image's grounding score and average it across all the chart images. Grounding evaluation: (1)legend grounding (discussed in Fig. 8) & (2) text-style (discussed in Fig. 10), we simply apply categorical correctness. (3) Color grounding (discussed in 9) is evaluated using L2 distance:

$$\mathcal{G}_{\text{color}} = \sqrt{(r-\hat{r})^2 + (g-\hat{g})^2 + (b-\hat{b})^2} \quad (1)$$

4 Experiments

4.1 Experimental Setup

Models We evaluate diverse open source models: Phi-3.5 vision-instruct (Abdin et al., 2024), InternVL-2.5 8B (Chen et al., 2024), LLaVA-1.6 Mistral 7B (Liu et al., 2023), QWEN-2.5 VL 8B (Bai et al., 2025). And GPT-4o (Hurst et al., 2024) as the proprietary model.

4.2 Ablations

We performed ablation experiments to vigorously compare differing approaches to our 2-stage approach.

The ablation experiments aimed to thoroughly compare single-stage based alignment approaches for performing multi-image reasoning vis-a-vis our two-stage approach. The ablation techniques:-

(1) *stitched-charts* inference: The chart-pair images are vertically concatenated resulting in a single image of stitched chart-pairs which undergo single-stage inference.

(2) *multi-image* inference: The VLM inputs multiple images, and contextualizes output based on the input images with aim of better understanding across of finer-level alignment in multi-image reasoning.

The ablation experiments analyzed Phi-3.5 model's performance on data-alignment task. As shown in table. 1, the single-stage approach fared poorly compared to out two-stage approach reaffirming the two-stage approach. Multi-image inference showed the weakest performance. Despite increasing training efforts towards improved VLM training, the models still face issues in reasoning ability on fine-grained tasks. Stitched-charts approach showed better results than multi-image, however they too underperformed vis-a-vis our two-stage approach. The comparatively stronger image self-attention capabilities seem to augment multi-image by utilzing the stitched connection. However the better prevailing capabilities of two-stage approach capture the gain of grounding generation. The VLM's multi-modal understanding though improving still suffers from finer-level nuances missed by information loss in image-encoding and cross-attention mechanisms.

4.3 Key Findings and Analysis

Finding 1

VLMs' dense grounding and alignment of data/color information are not satisfying on complex charts.

Compared to simpler and more common charts, e.g., bar/line charts and numbered bar/line charts, dense grounding/alignment on complex charts such as 3D/box/radar/rose/multi-axes charts with more components and irregular layouts is more challenging to most VLMs. Despite the similar alignment performance for *legend* (Fig. 6) and *text-style*

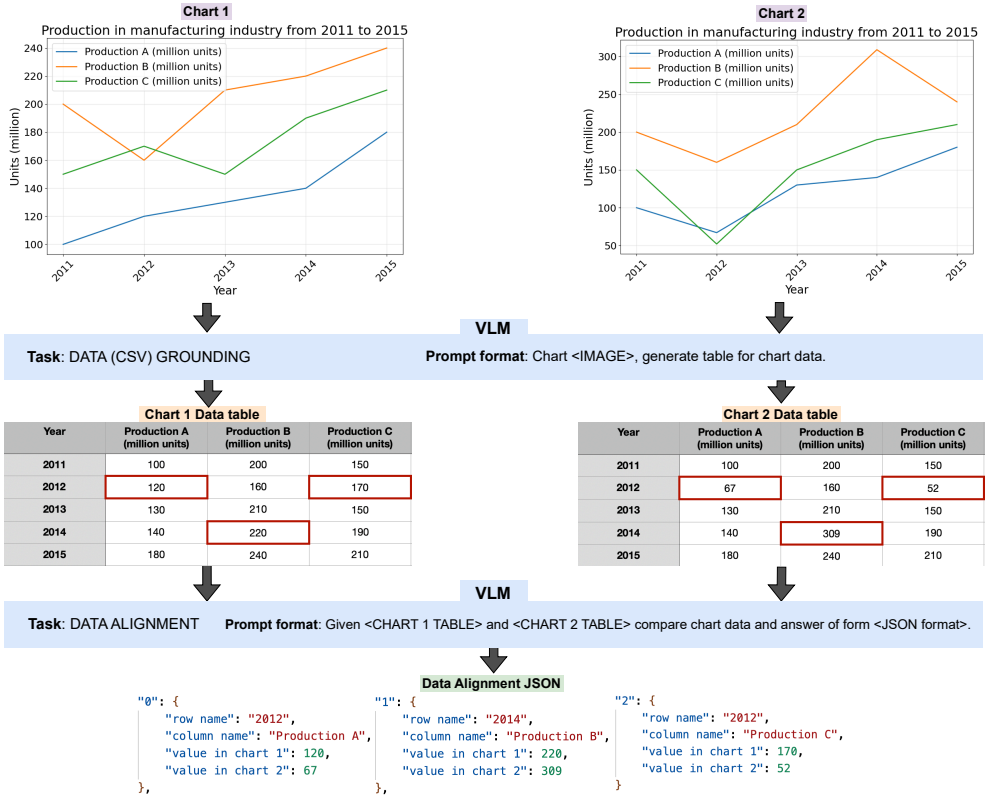


Figure 3: **Two-Stage Evaluation Pipeline of data alignment in ChartAB.** The first stage focuses on grounding the data visualized in each chart to a table, while the second stage requires the VLMs to find the difference between the two charts' tables and output a JSON file listing the different cells in the two tables. Evaluation of other attributes adopts similar multi-stage pipelines, with details in the Appendix.

Type	Approach	Bar	Bar #	3D Bar	Line	Line #	Radar	Rose	Box	Multi-axes
1-stage	Multi-chart	4.8	7.4	4.7	3.3	4.7	4.9	3.1	3.2	3.3
	Stitched-chart	5.0	4.8	3.0	4.5	3.5	3.0	2.7	2.8	3.2
2-stage	Ours	6.5	8.3	4.1	6.1	6.3	3.8	3.4	2.9	3.5

Table 1: **Ablation study of 1-stage vs. 2-stage evaluations** on data alignment (1-cell change) task. Mean scores across nine chart types show that our 2-stage evaluation reflects VLMs' greatest potential on chart alignment.

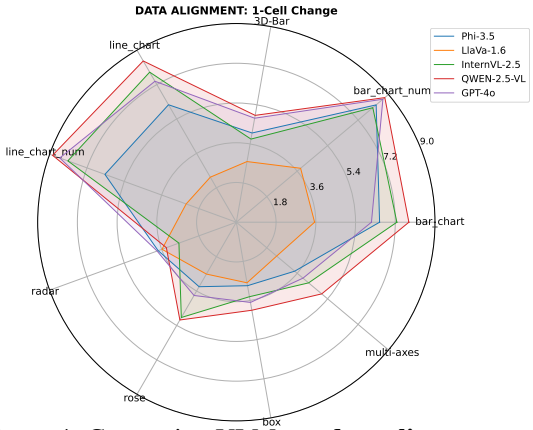


Figure 4: Comparing VLMs on **data alignment** tasks when two charts' data tables differ in only **one cell**. Llava-1.6 is worse than most other VLMs. QWEN-2.5-VL outperforms GPT-4o on most chart types. Related discussion can be found below Finding 1.

(Fig. 7) between simple vs. complex charts, the *color* alignment (Fig. 5) and *data* alignment (Fig. 4) on complex charts are much poorer than those on simple charts. The color grounding requires identifying each component's visual encoding and corresponding color, while the data grounding needs to find the mapping from visual encoding to tabular values. Hence, complex layouts with more components make these tasks more difficult. In contrast, identifying the position of legends and text styles (which both have limited options) is easier and less affected by the chart complexity.

Finding 2

Most VLMs suffer from biases when allocating the position of legends.

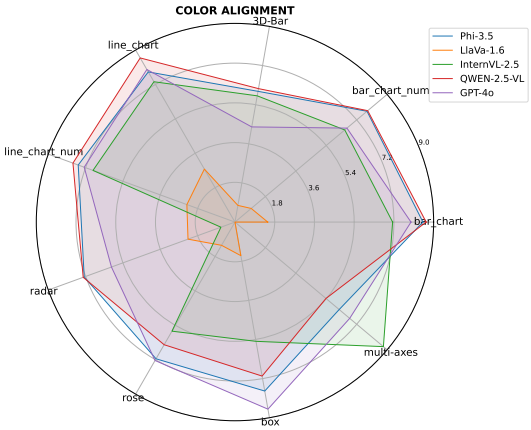


Figure 5: **Color alignment** between two charts on fine-grained visual elements (e.g., bars, lines, sector). VLMs perform better on simpler and more common charts. Related discussion can be found below Finding 1.

The grounding of the legend's position (Fig. 8) suffers from a strong bias of pretrained VLMs. The Phi-3.5 model shows the strongest prior towards the *upper-left* position. The 7-8B scale VLMs, e.g., LlaVa-1.6, Inten-VL-2.5, QWEN-2.5-VL, all show a similar level of bias but towards the *upper-right* position instead. The GPT-4o model exhibits the minimal bias among all evaluated VLMs The grounding bias strongly affects the legend alignment (Fig. 6) where Phi-3.5 performs the worst, GPT-4o has the best performance, while the other 3 models' performance is similar and between Phi-3.5 and GPT-4o.

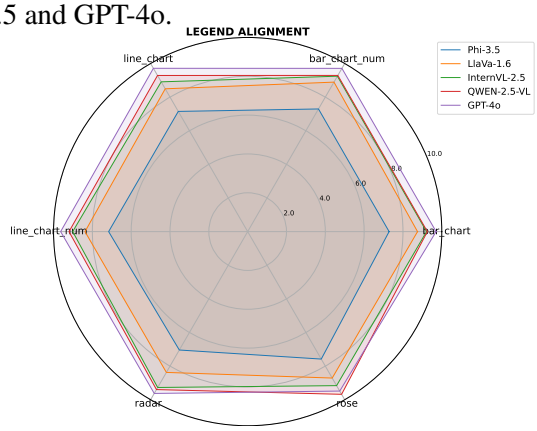


Figure 6: **Legend alignment** of legend positions between two charts. Each VLM shows similar performance across different chart types. Phi-3.5 performs the worst while GPT-4o is the best among all five VLMs. Related discussion can be found below Finding 1&2.

Finding 3

VLMs' weak color recognition ability.

As shown in Fig. 9, all models' color grounding error (L2 distance in RGB space) has a median exceeding 50. This suggests their inability to understand color shades beyond common ones, e.g., red, blue, green, etc., which exposes their weaknesses in color recognition. The lack of color understanding affects the perception of detailed differences in charts and leads to mismatches in color-related/conditioned reasoning tasks. Consequently, the VLMs' performance in color alignment tasks (Fig. 5) is consistent with that on color grounding. These results suggest improving the color understanding capability by adding more color-sensitive data or tasks in VLMs' pretraining and finetuning stages.

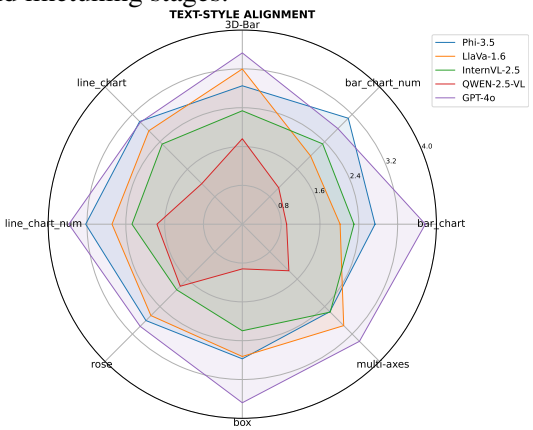


Figure 7: **Text-style (size, weight, font family) alignment.** Worst: QWEN-2.5-VL, Best: GPT-4o. Differences between chart types are not consistent across VLMs. Related discussion can be found below Finding 1&4.

Finding 4

VLMs' text-style grounding and alignment performance is poor in general, and it varies across text size, weight, and font family.

Fig. 10 shows that most VLMs fail to detect the correct text size and font family, suffering from an accuracy below 20% (except GPT-4o's performance on font family alignment). These indicate a lack of knowledge on these two text attributes. VLMs' performance on text weight ((light/normal/bold)) is much better (∼60%) and close to each other, but still not satisfying. Although LLMs can select reasonable text sizes in code generation for plots, they tend to rely on the default sizes in their priors or relative sizes to other chart components. They still lack sufficient capability to identify text sizes in chart images.

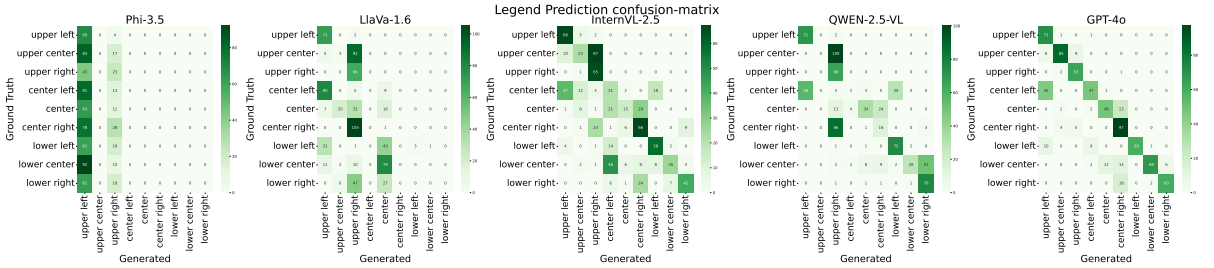


Figure 8: **Confusion matrix of legend position grounding for each VLM.** The dark non-diagonal entries highlight the fail patterns and biases of incorrectly identifying position-i as position-j. Phi-3.5 exhibits a severe bias towards *upper-left* position while GPT-4o shows the minimal bias. More discussion is provided below Finding 2.

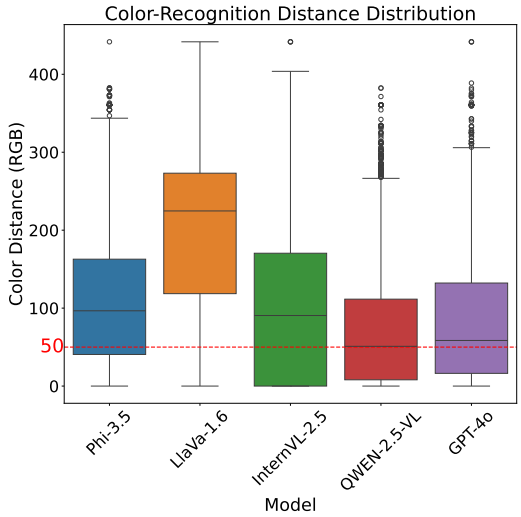


Figure 9: **Color recognition** in grounding performance measured by L2 errors in RGB space. The error distribution of each VLM is visualized by a box plot. Median of the errors for all models exceeds 50, indicating weak color recognition capability. More discussion can be found below Finding 3.

Finding 5

VLMs' weak scaling law on chart grounding and alignment tasks.

We fail to observe a clear scaling law on the evaluated models of different scales, i.e., Phi-3.5 (3B), LlaVa-1.6-Mistral (7B), Intern-VL-2.5 & QWEN-2.5-VL (8B), GPT-4o (proprietary). The Phi-3.5 model shows better or on-par alignment when compared with Llava-1.6 and Intern-VL-2.5 on all except data/text style/legend alignment (Fig. 3, 5, 7, 6). In addition, Fig. 7 shows that Qwen-2.5 is the weakest baseline in text style alignment. The current chart understanding benchmarks heavily focus on QA (as discussed in 2), but cannot fully capture the detailed information. The chart-specific VLMs are constrained by their task-specific architecture (as discussed in 2).

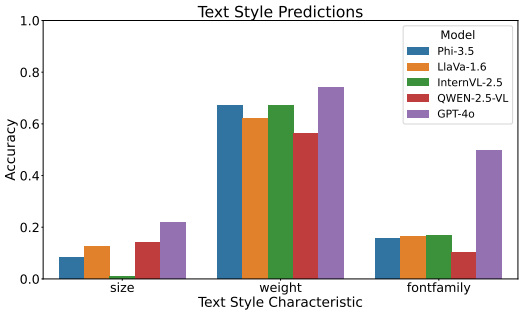


Figure 10: **Text-style grounding and alignment on size, weight, and font family.** Most VLMs suffer from a low accuracy on size and font family, indicating a lack of related knowledge in VLM training. Further discussion can be found below Finding 4.

5 Conclusion

In this work, we present ChartAB the first benchmark to comprehensively evaluate fine-grained chart grounding and multi-chart dense alignment capabilities of general-purpose vision-language models (VLMs). Through rigorous evaluations across diverse chart types and VLMs, we uncover consistent challenges faced by current models, including perceptual biases, hallucinations, and limited spatial understanding, particularly on complex and information-dense visualizations.

Our benchmark facilitates detailed assessment across dimensions such as data extraction, color and legend grounding, and robustness to visual variations. These insights expose specific areas for improvement in chart perception and reasoning, offering valuable guidance for future VLM development. The consistent superiority of our two-stage pipeline further emphasizes the necessity of grounding-based decomposition for achieving human-parallel chart understanding.

Limitations

Our work has the following limitations:-

- Model Training: We focus on zero-shot evalu-

ations for our work, and don't assess few-shot or instruction tuned performance. They may yield better performance but deflect from the problem statement of general purpose VLM's dense-level understanding.

- Real-World Chart Corpus: Various datasets contain more sophisticated real-world examples. However due to requirement of precise ground-truth for dense-alignment evaluation, we chose the ChartX dataset due to availability of plotting-code and corresponding csv data.
- Limited Task Diversity: The ChartAB focuses only on dense-alignment evaluation, missing the high-level reasoning or related dense-level downstream tasks. The work intended to perform a comprehensive evaluation of various dense alignment tasks and grounding based two-stage evaluation hence missed those aspects.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. 2022a. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022b. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Shengzhi Li and Nima Tajbakhsh. 2023. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2022a. Deplot: One-shot visual language reasoning by plot-to-table translation. *arXiv preprint arXiv:2212.10505*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2022b. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. *arXiv preprint arXiv:2504.05506*.

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

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*.

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

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, and 1 others. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, and 1 others. 2024a. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, and 1 others. 2024b. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Renqiu Xia, Bo Zhang, Haoyang Peng, Hancheng Ye, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, and Junchi Yan. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *arXiv preprint arXiv:2309.11268*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and 1 others. 2024. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

A Appendix

A.1 Dense Alignment between Two Charts: Pot-Attribute Alignment

Color Alignment evaluates alignment of encoding colors, i.e. difference in colors of visual encodings representing chart data: bars in bar chart, lines in line chart, segments/spokes in rose chart etc.

Legend Alignment evaluates alignment of legend, i.e. difference in position of legend in the charts.

Text-Style Alignment evaluates alignment of text characteristics namely (1) *size*, (2) *weight* i.e. degree of boldness (3) *font-family* i.e. style of font applied. The text in chart corresponds to following chart sections: title, legend, axes-labels, axes-ticks.

Overall plot-alignment task aims to analyze model's ability to perceive change in visual design characteristics (e.g. visual encodings, axes, labels, legends) in the chart image, and semantic understanding to map it to specific plot attribute. And ability to precisely predict the attribute value from representation and component structure of the chart.

A.2 Dataset Curation

We used ChartX dataset (Xia et al., 2024) as source dataset for our ChartAlignBench curation. ChartX contains plotting-code and csv data-table for the chart with extremely high level of precision thus offering the flexibility for performing finer-level changes along with ground-truth generation capabilities. It contains diverse chart types of varying complexities, and chart data from multiple domains. Hence enabling analysis across charts of varying difficulties.

We utilize *perturbations* for generating fine-grained variations for given chart thus helping build dense-alignment pairs. Chart's plotting-code is perturbed for precise data or plot-attribute changes based on rigorous formatting check using regex-based search and replace, resulting in chart image generation from code execution.

The csv availability and plot-attribute information enable accurate ground-truth generation. Generated pairs for data alignment and plot-attribute alignment include randomly assigned changes, and

robustness sets include diverse plot attribute values for meticulous and unbiased evaluation.

We selected 9 diverse chart-types with ability to apply to perform chart-data and plot-attribute perturbations: (1) *simple charts*: bar chart, bar-numbered chart, line chart, line-numbered chart, (2) *complex charts*: 3D chart, box chart, radar chart, rose chart, multi-axes chart.

A.3 A Two-Stage Evaluation Pipeline: Details & Essentiality of second-stage

We utilize natural-language based instructions for zero-shot inference to enable simple execution with minimal task specific nuances for strong generalization across various models.

VLM outputs follow *JSON based formatting* due to precise nature of the key-value structure which is essential for element specific information serialization for finer-analysis, along with flexibility for variations in completion of grounding and fine-grained analysis. The alignment JSON contains finer-level attributes for which the charts differ, and the values for corresponding attribute in the two charts. E.g. for data-alignment (as shown in Fig. 3) the finer-level attributes changed between the charts i.e. cells are identified by their row & column header, along with its values in the chart-pairs, i.e. value in chart-1 & value in chart-2 respectively.

Second-stage forms essential part of evaluation pipeline. Analyzing dense-alignment ability requires performing end-to-end evaluation of VLMs. Grounding determines the chart information, and impacts the subsequent finer-level analysis. However correct grounding doesn't imply correct alignment. The VLM needs to make semantic correspondence between chart elements in the grounding result which is non-uniform and differs for each VLM. Moreover the hallucinating nature of VLMs make grounding output susceptible to ambiguities and vagueness, in which case the additional second-stage reasoning on the grounding result helps build a better overall understanding of VLM capabilities. Second-stage also allows utilization of additional contextual information (e.g. Chain-of-Thoughts) for the alignment task. Ultimately we analyze VLM's dense-alignment ability the way humans do looking at overall understanding, and at semantic shifts not captured by grounding.

A.4 Evaluation metric: Alignment

Alignment evaluation is done by calculating similarity of VLM's evaluation response JSON vis-a-vis the ground-truth anchor. The JSON encompasses finer-level *constituents* (e.g. bars of bar chart with color-difference in color-alignment task) which differ between the chart-pairs along with their specific value, and are evaluated for their correctness.

A.4.1 Plot-Attribute Alignment

For Plot-attribute alignment score, the accuracy for each N constituent is calculated for the chart-pair (chart-1 & chart-2), and averaged for all constituents to get the score. The Accuracy A_i is calculated based on the alignment task, contrasting the evaluation response value with the ground-truth value.

$$\text{Score} = 10 \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{A}_i \left(chart_1 \right) + \mathcal{A}_i \left(chart_2 \right) \right) \tag{2}$$

Legend Accuracy: The legend position accuracy using the manhattan distance, the position associated with the 3 by 3 grid:

$$\mathcal{A}^{\text{legend}} = 1 - \frac{1}{5} \text{Manhattan}(\text{position}, \hat{\text{position}}) \tag{3}$$

Color Accuracy: The color accuracy is calculated using L1 distance:

$$\mathcal{A}^{\text{color}} = 1 - \frac{1}{3} \sum_{i \in \{R,G,B\}} \frac{|intensity_i - \hat{intensity}_i|}{255} \tag{4}$$

Text Accuracy: The text alignment accuracy is calculated by correctness of size, weight, fontfamily respectively.

$$\begin{aligned} \mathcal{A}^{\text{text style}} = \\ \frac{1}{4} \sum_{i \in \{\text{title,legend,ticks,labels}\}} (0.4 \cdot \mathbb{1}[\mathbf{size}_i = \hat{\mathbf{size}}_i] \\ +0.3 \cdot \mathbb{1}[\mathbf{weight}_i = \hat{\mathbf{weight}}_i] \\ +0.3 \cdot \mathbb{1}[\mathbf{fontfamily}_i = \hat{\mathbf{fontfamily}}_i]) \end{aligned} \tag{5}$$

A.4.2 Data Alignment & Robustness

Data Alignment score calculation follows the JSON correctness discussed in evaluation metrics section. However data alignment accuracy is calculated for the combined image-pair, unlike individual image in plot-attribute. As for data alignment we also evaluate the correctness of the finer-level constituent's key (i.e. identification) which are the cell's row & column name whereas in plot-attribute alignment only constituent's value is evaluated. Data alignment scores are also averaged for all chart-pairs in a chart-type. For N being the number of cell-change between the image-pairs, data alignment score is defined as:

$$\text{Score} = 10 \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \mathcal{A}_i^{\text{cell}} \left(chart - pair \right) \right) \quad (6)$$

The cell accuracy A^{cell} is determined by the cell' value accuracy (for each chart), and the evaluation response's row & column similarity (for chart-pair).

$$\begin{aligned} \mathcal{A}^{\text{cell}} &= 0.3 \cdot \text{Sim}^{\text{row}} + 0.3 \cdot \text{Sim}^{\text{col}} \\ &+ 0.2 \cdot \text{A}_{\text{chart-1}} + 0.2 \cdot \text{Val}_{\text{chart-2}} \end{aligned} \quad (7)$$

The row and column name correctness is evaluated using Levenshtein distance based string comparison:

$$Sim^i = Levenshtein(i, i^) \quad (8)$$

The cell-value accuracy (for a chart) is evaluated using the percentage value difference:

$$\text{Val}_i = max \left(1 - \left(\frac{|\text{cell\_val} - \hat{\text{cell\_val}}|}{\text{cell\_value}} \right), 0 \right) \quad (9)$$

Robustness: Robustness of data alignment over variation in plot-attribute aims to evaluate model's ability to maintain consistent alignment over changing plot-attributes. The data alignment score is utilized for developing the robustness evaluation metric. For robustness, each chart has set of 5 data-alignment pairs with identical data-alignment but variation in plot-attribute values. We define μ(set) and σ(set) as the mean and standard-deviation respectively of the 5 image-pairs in the robustness set for a chart.

σ(set): It represents the deviation of 5 chart-pairs. A high value indicates of large difference between the data-alignment scores of the chart-pairs hence low robustness.

We define the Robustness metric by averaging the σ(set) for all the charts, for particular configuration: i.e. cell-change c, and the altered plot-attribute p.

$$R(c,p) = \frac{1}{N_{c,p}} \sum_{\substack{\text{cell-change}=c \\ \text{plot-attr}=p}} \sigma(\text{robustness set}) \quad (10)$$

A.5 Additional Finding & Insights

Finding 6

VLMs' data grounding and alignment are more robust to color variations than changes in legend positions and text styles.

Fig. 11 shows that robustness is the worst under text-style variations and the best under color variations. In the visualizations of data, colors are used to discretize, categorize, and measure chart constituents. As long as their colors are distinguishable, color variations will not affect the data grounding. In contrast, the text styles and legends provide critical information about the data via ticks, labels, and legend items. Moreover, changing legend position may lead to position changes and occlusion of other chart elements. Hence, their variations have a greater impact on the data grounding/alignment performance.

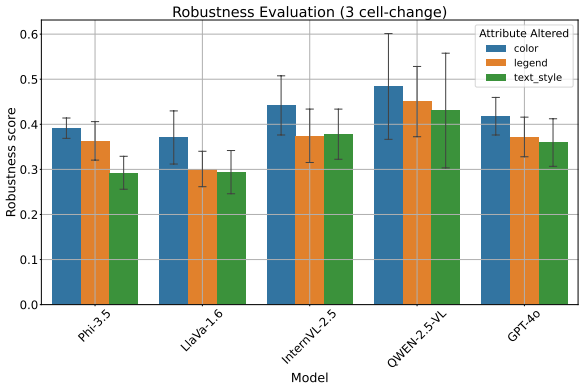


Figure 11: **VLMs' Robustness of data alignment (3-cell change) to variations in color, legend, and text-style.** VLMs show better robustness to color changes than text-style changes. QWEN-2.5-VL outperforms the other four VLMs on robustness. More discussion can be found below Finding 6.

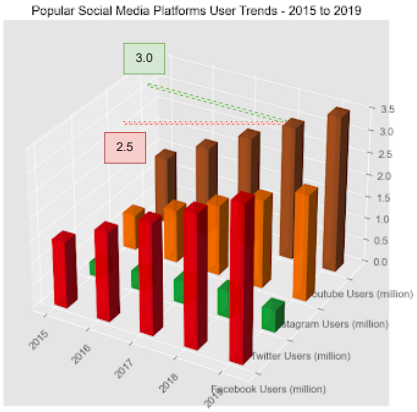
Finding 7

VLMs' spatial understanding capability affects several important chart understanding skills.

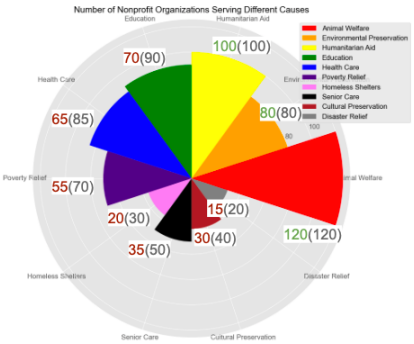
Chart understanding usually requires an accurate mapping between spatial relationships and the corresponding numerical values to be visualized.

- *Depth understanding*: Despite the high-level similarity between 3D bar charts and (2D) bar charts, as shown in Fig 4, the data alignment performance is much poorer on 3D bar charts due to the lack of depth understanding, which affects the measurement of scales and values along axes in the 3D space.

- *Text vs non-text cues*: Rose charts are extended from bar charts by allowing more polar coordinates with scale differences in radial forms. However, Fig. 12b reveals a great difference



(a) Depth estimation in 3D bar charts



(b) Text vs. non-text cues for value scaling in rose charts.

Figure 12: **VLMs' spatial understanding is poor on complex charts.** More discussion is provided below Finding 7.

between the two on data alignment performance. This is due to fewer text cues (e.g., axes ticks) in rose charts, where non-text cues such as grid lines cannot be fully leveraged.

- *Better performance on numbered charts*: numbered bar and line charts explicitly place the data values in the charts, hence facilitating VLMs to extract the data easily without precise measurements of the visual elements. Hence, as shown in Fig. 4, numbered bar/line charts usually enjoy better performance.