VisHintPrompt: A Visual Hint Prompting Strategy for Numerical Inference from Charts

Fengling Zheng, Yongle Peng, Zi Rong, Chenyun Cai, Yumeng He, Dekun Qian, Zhiguang Zhou, Yigang Wang, and Yong Wang

***Abstract*—Numerical inference from charts underpins many visualization tasks, including chart question answering and chart redesign. Although many chart-specific models have been developed for numerical extraction, their reliance on the chart types and visual encodings represented in training data limits their generalizability. Modern Multimodal Large Language Models (MLLMs) possess competitive general-purpose visual understanding, offering the potential for numerical inference without additional training. Yet how to reliably elicit such fine-grained numerical inference from pretrained MLLMs via visual prompting remains largely unexplored. To address this gap, we propose *VisHintPrompt*, a scaffolded visual hint prompting strategy grounded in the explicit extraction of axis priors from both Cartesian and polar charts. *VisHintPrompt* elicits the latent numerical inference abilities of MLLMs by progressively narrowing the search region. It supplies structured visual cues that guide the model toward accurate quantitative interpretation. The strategy consists of three coordinated components: *Axis-aware Grid Enhancement*, which extracts fine-grained axis priors and further exposes them as structure-aligned grids; *Iterative Visual Feedback*, which overlays alignment cues derived from intermediate predictions to enhance the next prediction; and *Progressive Zoom-in Refinement*, which enlarges the region of interest and applies progressive grid densification to enhance local detail for precise value estimation. *VisHintPrompt* is visual-centric and applies to both Cartesian charts (e.g., bar, scatter) and polar charts (e.g., pie, radar). Experiments across diverse chart types on both proprietary and open-source MLLMs demonstrate consistent improvements over standard prompting. Quantitative and qualitative analyses, along with ablation studies, validate the effectiveness and essential role of each component. These results underscore the broad applicability of *VisHintPrompt* in enabling reliable and more accurate numerical inference from charts.**

***Index Terms*—Chart understanding, multimodal large language models, visual prompting, data inference, and information visualization.**

I. Introduction

NUMERICAL inference from charts is a fundamental capability that can benefit a wide range of visualization tasks, including chart question answering [1]–[3] and chart redesign [4]. It requires accurately mapping visual encodings, such as positions, lengths, or angles, to their corresponding numerical values. Prior research primarily addressed chart numerical inference through chart-specific models designed for particular chart types or visual encodings [5]–[7]. While these models can achieve competitive performance within their targeted settings, their effectiveness often relies on assumptions about chart structure and encoding styles, limiting their ability to generalize to unseen or structurally diverse charts.

Recent advances in Multimodal Large Language Models (MLLMs) have shown strong general-purpose visual and linguistic understanding capabilities, which motivates growing efforts to adapt MLLMs to chart-related numerical inference tasks through task-specific training, including specialized dataset construction and fine-tuning or instruction tuning [6], [8]. Representative approaches include Deplot [1], and ChartSketcher [9]. While these methods have shown promising performances on chart understanding benchmarks, they rely on substantial task-specific data and training to modify model behavior, often incurring high computational and data collection costs.

An alternative and potentially more lightweight approach is to enhance MLLMs' chart numerical inference through visual prompting. A recent evaluation study, ChartInsights [10]demonstrates that explicitly augmenting charts with visual cues can significantly enhance models' ability to analyze data and perform reasoning, without additional training. Beyond chart-specific tasks, a growing body of work in general computer vision has explored visual prompting strategies for object detection, keypoint localization, and spatial reasoning. Methods such as DetToolChain [11], RedCircle [12], and SCAFFOLD [13] provide converging evidence that explicitly externalizing spatial priors through image manipulation, such as zoom-in operations and coordinate-reading cues (e.g., overlaying rulers, compasses, or dot matrices), can effectively guide MLLMs' attention and improve spatial localization during inference. However, accurate chart numerical inference fundamentally requires establishing explicit correspondences between visual coordinates and numerical values under axis constraints. While existing visual prompting strategies have shown effectiveness in extracting explicitly annotated values and improving spatial localization, it remains unclear whether and how visual prompting alone can systematically enable pretrained MLLMs to perform fine-grained numerical inference, where values must be inferred from visual encodings via precise alignment with axis scales. This unresolved question motivates the need for a systematic and generic visual prompting strategy for chart numerical inference.

Fengling Zheng is with Hangzhou Dianzi University and Nanyang Technological University. A part of this work was done when she was a visiting student supervised by Yong Wang at Nanyang Technological University. E-mail: {fenglingzheng}@hdu.edu.cn.

Yongle Peng, Zi Rong, Chenyun Cai, Dekun Qian, Zhiguang Zhou, and Yigang Wang are with Hangzhou Dianzi University, Hangzhou, China. E-mail: {22330236, rongzizi, 23280129, zhgzhou, yigang.wang}@hdu.edu.cn.

Yumeng He is with the Viterbi School of Engineering, University of Southern California, United States of America. E-mail: heyumeng@usc.edu.

Yong Wang is with Nanyang Technological University, Singapore. E-mail: yong-wang@ntu.edu.sg.

(Corresponding authors: Yong Wang and Zhiguang Zhou.)

To fill this gap, we propose *VisHintPrompt*, a scaffolded visual prompting strategy designed to support fine-grained numerical inference from charts using pretrained MLLMs, without requiring any additional training. The core idea of *VisHintPrompt* is to explicitly expose axis-aware spatial structure by introducing and progressively refining grid-based visual references, which serve as a foundation for structured visual hints to guide a more accurate numerical inference (Fig. 1). Our experiments show that explicitly extracting and externalizing *axis priors*, i.e., axis-related structural information such as axis orientation, tick locations, and value scales, provides an effective geometric foundation for guiding fine-grained numerical inference. Specifically, *VisHintPrompt* consists of three coordinated components: *(1) Axis-Aware Grid Enhancement*, which introduces structure-aligned Cartesian or polar grids to provide a stable reference for mapping visual positions to numerical values; *(2) Iterative Visual Feedback*, which overlays alignment cues derived from intermediate predictions to iteratively correct coarse localization errors; and *(3) Progressive Zoom-in Refinement*, which selectively enlarges regions of interest and applies progressive grid densification to introduce finer axis-aligned references, amplifying subtle mark-axis relationships that are critical for precise value estimation. These components form a coherent strategy that incrementally refines spatial focus and numerical grounding. Experiments across diverse chart types on both proprietary and open-source MLLMs show consistent performance gains over standard prompting. Quantitative and qualitative analyses, together with component-wise ablation studies, further validate the complementary roles of each component. Our main contributions are summarized as follows:

- **A scaffolded visual hint prompting strategy (*VisHintPrompt*).** We propose *VisHintPrompt*, a structured visual prompting framework that integrates Axis-aware Grid Enhancement, Iterative Visual Feedback, and Progressive Zoom-in Refinement. By progressively narrowing the model's effective search space, *VisHintPrompt* enables more precise numerical inference from charts with pretrained MLLMs, without additional training.
- **A comprehensive evaluation of *VisHintPrompt* on numerical inference across different MLLMs and chart types.** We conduct systematic experiments spanning three proprietary and three open-source MLLMs across nine chart types, covering both Cartesian and polar encodings. Our evaluation integrates quantitative results, qualitative analysis, and component-wise ablation studies. The results demonstrate consistent improvements over standard prompting and confirm the functional role of each component.

II. Related Work

In this section, we review three lines of research most relevant to our work: (i) traditional chart information extraction approaches that reconstruct structured data from chart images; (ii) chart understanding enhancement strategies for MLLMs from two perspective: model-centric approaches that enhance chart understanding via pretraining or instruction tuning of Multimodal Large Language Models (MLLMs), and visual-centric prompting strategies that guide MLLMs without changing model parameters. [Yong: The overview here is inconsistent with the subsequent subtitles below.]

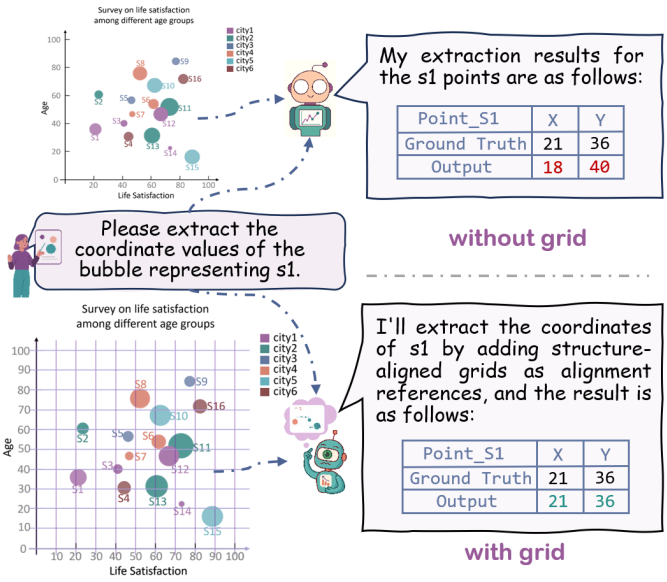

Fig. 1. An illustrative example of scatterplot data inference using an MLLM with or without visual hints (i.e., the grid here). Black text represents the ground truth, red text highlights large numerical extraction errors, and green text indicates accurate extraction with minimal deviation from the ground truth. Grid hints substantially improve extraction performance.

A. Traditional Chart Information Extraction Approaches

For decades, researchers have devoted efforts to extracting structural information from chart images to recover their underlying structured data, thereby supporting downstream tasks such as chart understanding and analysis. Early approaches typically adopted multi-stage parsing pipelines. The core idea was to first detect chart elements using image processing and OCR techniques, and then reconstruct values based on geometric or semantic relationships. Since different chart types are defined by distinct graphical primitives, these methods often relied on predefined rules to locate components through color continuity and edge features, followed by graphical mark extraction [4]. For example, *ReVision* [4] employed edge detection and pattern recognition techniques to separate axes, data marks, and legends, thereby recovering data from bar and pie chart images. *ChartSense* [14] leveraged a mixed-initiative design that combined image processing with guided user interaction to achieve fast and accurate data extraction. *ChartKG* [15] integrated object recognition, OCR, and rule-based parsing of graphical marks, and organized the recovered data in the form of a knowledge graph. Although such approaches provide interpretability, they depend on complicated pipelines, where errors can be amplified across multiple stages, ultimately limiting robustness. With the advent of deep learning, research has shifted toward end-to-end frameworks for chart data extraction, aiming to directly predict structured outputs from chart images. *Scatteract* [16] automatically recovered data points from scatter plots. *DVQA* [17] proposed a

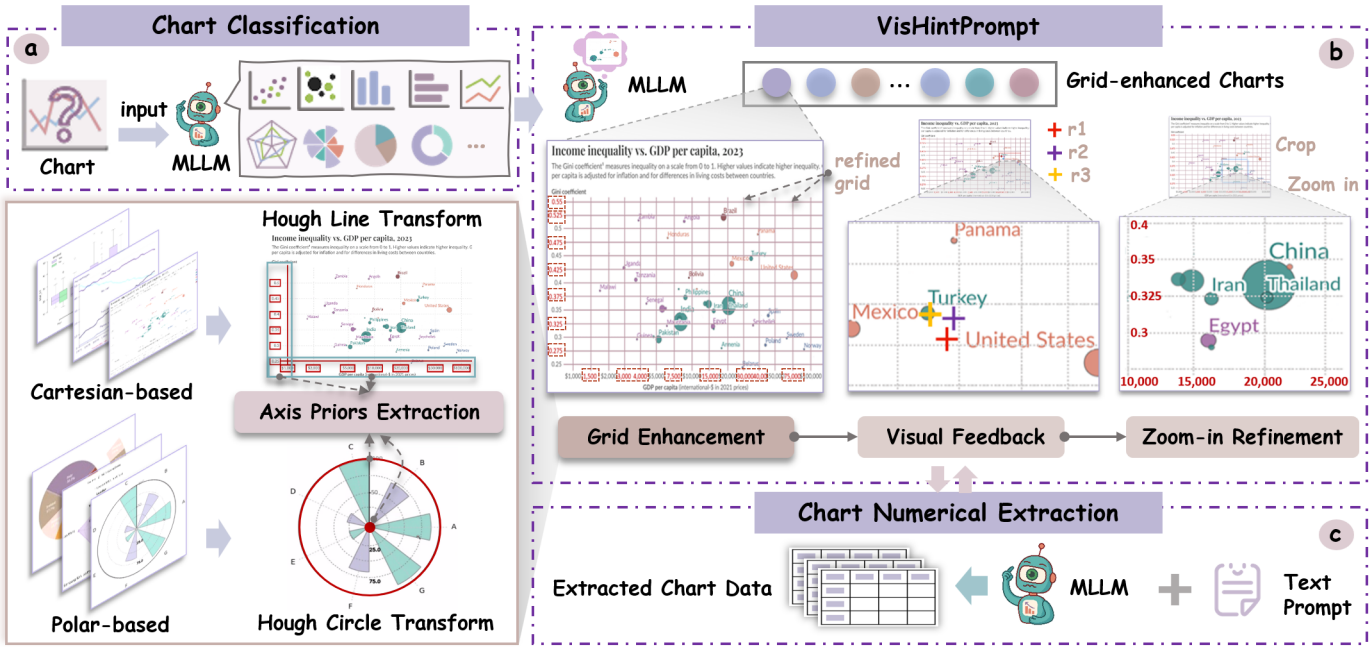

Fig. 2. An overview of *VisHintPrompt*, a visual hint prompting strategy for numerical inference from charts. The method operates in three stages: (1) *Axis-aware Grid Enhancement*, which exposes fine-grained axis priors by overlaying grids aligned with the chart's underlying structure; (2) *Iterative Visual Feedback*, which introduces model-informed alignment cues to refine intermediate predictions; and (3) *Progressive Zoom-in Refinement*, which enlarges the relevant region to provide high-resolution visual detail for fine-grained and reliable numerical inference across both Cartesian and polar charts.

deep dual-network model to parse values directly from bar chart images. *ChartOCR* [5] combined deep learning with rule-based methods, extracting key points of chart elements to support data extraction from three common chart types. However, these algorithms were often tailored to specific chart types and only covered a limited set of formats (e.g., line, bar, and pie charts), showing clear limitations in handling the diversity and complexity of real-world charts. More recent studies have introduced end-to-end Vision–Language Models [6], [18]–[20], which learn mappings from chart images to structured tables via vision encoders and autoregressive decoders. By leveraging weak supervision or self-supervision for cross-modal alignment, these methods aim to achieve direct chart understanding from images. However, in the absence of explicit numeric annotations or under diverse chart styles, their ability to extract visual structural information and to generalize remains limited, which will be addressed in this work.

B. Chart Understanding Enhancement Strategies for MLLMs

[Yong: Fengling, pls check if the change is aligned with your intent, and revise the paragraphs below.]

The emergence of MLLMs such as GPT-4o, Gemini, and LLaVA has further transformed the paradigm of chart understanding, with end-to-end reasoning becoming increasingly mainstream. A variety of strategies have been proposed to improve their performance in chart data inference, which can be broadly categorized into *model-centric enhancements* and *visual-centric prompting strategies*.

[Yong: For the following paragraphs, 1) the introduction of existing work can be condensed; 2) the difference between prior work and our approach is a bit redundant and can be shortened.]

1) Model-centric Enhancements: Model-centric studies primarily improve performance through large-scale pretraining or instruction tuning on chart-specific datasets [21]–[25]. For example, *ChartLlama* [26] fine-tuned LLaMA with an instruction-tuning corpus generated by GPT-4, achieving strong performance. *ChartAst* [27] and *MMC* [28] pretrained models on chart–table pairs to align visual and tabular modalities. Yang et al. [29] constructed a high-quality dataset with diverse visual details to strengthen fine-grained understanding. *TinyChart* [30] introduced a lightweight 3B-parameter model and proposed a parameter-free visual instruction merging module to reduce input sequence length, alleviating the computational burden of high-resolution chart images. Nevertheless, many methods rely on charts synthesized from underlying data tables [31], thereby neglecting the ability of models to directly extract values from chart images themselves. To address this limitation, Zeng et al. [8] proposed Visualization-Referenced Instruction Tuning, which guides training by unfreezing vision encoders and introducing hybrid-resolution adaptive strategies to enhance fine-grained perception. *Onechart* [32] incorporated auxiliary tokens to direct model attention to critical components. *CHOPINLLM* [33] integrated original data alignment during pretraining, and during fine-tuning first extracted chart data before answering questions, thereby capturing both structural and numerical properties of chart images and significantly improving performance on unlabeled charts. Other approaches have advanced in complementary directions. *ChartGemma* [34] trained on instruction-tuning data generated directly from chart images, without requiring access to underlying data tables, thereby capturing both global trends and local visual details. *ChartMoE* [35] employed a mixture-of-experts architecture, enabling more accurate data extraction and attribute

parsing under multi-task alignment and instruction tuning. *Chart-r1* [36] leverages a group relative policy optimization strategy, employing relatively mild reward signals to calibrate numerical sensitivity during reinforcement fine-tuning. While these approaches have achieved remarkable advances, they either incur high training and deployment costs or are intrinsically proprietary and impossible to conduct model training.

2) Visual-centric Prompting Strategies: Visual-centric prompting strategies, such as multimodal chain-of-thought and visual prompting [11], focus on decomposing tasks into subproblems and guiding models to emulate human-like reasoning in chart interpretation. For instance, *ChartThinker* [37] first proposed chain-of-thought reasoning for charts, but mainly emphasized textual logic rather than visual parsing. Subsequent studies increasingly incorporate explicit visual guidance to complement textual reasoning. *ChartInsights* [10] introduced Chain-of-Charts prompting, which combined textual and visual cues to direct models' attention to specific elements for data interpretation. *VisualCoT* [38] incorporated visualization-driven chain-of-thought, employing cropping mechanisms to focus on key regions and enhance local visual understanding. Similarly, *DetToolChain* [11] presented a detection-based prompting toolkit that guided MLLMs to attend to regional information through staged visual hints, integrating coordinate measurement and self-verification to improve consistency across iterations. *VProChart* [39] drew inspiration from human visual alignment principles, modeling spatial proximity, color consistency, and crosshair alignment rules among visual elements to significantly boost numerical inference accuracy. These visual-centric prompting strategies enhance MLLMs in a lightweight yet effective manner by injecting structured visual cues into the input, without modifying model parameters. However, most of them are restricted to local regions or specific visual scenes. [Yong: Are our approaches also for local regions?]

III. VisHintPrompt

We propose *VisHintPrompt*, a visual-hint-based prompting strategy to elicit MLLMs' abilities of numerical inference from charts. It introduces three coordinated visual hint prompting components (Fig. 2(b)), *(1) Axis-aware Grid Enhancement*, *(2) Iterative Visual Feedback*, and *(3) Progressive Zoom-in Refinement*, which progressively narrow the search region and substantially improve numerical accuracy and robustness across diverse chart types. *VisHintPrompt* begins with chart type classification, which determines the appropriate axis prior extraction procedure and the corresponding visual prompting strategy. We consider two coordinate families: Cartesian (bar, line, scatter, and bubble charts) and polar (pie, donut, radar, and rose charts).

A. Axis-aware Grid Enhancement

1) ***Extraction of Axis Priors from Chart Images****:* Axis-prior extraction (Fig. 2(a)) aims to obtain high-precision axis information that is readily available from chart images using simple, off-the-shelf image processing algorithms. Specifically, the extracted axis priors across these two coordinate systems comprise the geometric layout (x–y axes or circular frame, axis positions, circle center, and radii) and the associated numeric structure (tick locations, tick labels, and scales), providing a high-precision reference for subsequent visual-hint prompting.

a) Axis priors in cartesian coordinates: For Cartesian charts, we extract axis priors describing the x–y axis geometry and numeric scale.

Extraction of x–y axes. We begin by detecting candidate horizontal and vertical axis lines using the classical Hough Transform, which is robust for identifying long linear geometric structures. Among the detected lines, we select the pair that satisfies the characteristic property of Cartesian coordinates—one dominant horizontal line and one dominant vertical line that are orthogonal to each other. To additionally ensure robustness across charts with different axis placements, we employ an MLLM to identify the layout of the chart and verify the correct axis pair. This combination of geometric detection and layout validation reliably yields the final axis extraction for Cartesian charts.

Alignment of tick values to pixel positions. Tick marks are first scanned along each extracted axis as short line segments that are perpendicular to the axis line. Because this scanning process may occasionally be affected by background noise or decorative elements, we analyze the distribution of intervals between adjacent detected ticks and identify the interval that appears most frequently as the true tick spacing. Using this dominant spacing, we refine the tick locations and correct potential outliers, which produces a more reliable set of tick positions. For each refined tick, we apply OCR to obtain the corresponding numeric label and pair it with its pixel coordinate. Through this refinement procedure, the method produces accurate pixel–value pairs that support stable and precise value interpolation in subsequent stages of our prompting framework.

b) Axis priors in polar coordinates: For polar charts, we extract axis priors that capture both the polar geometric layout and the two forms of numeric encoding in polar coordinates: angular encoding in pie and donut charts, and the joint angular–polar encoding used in radar charts.

Extraction of circle center and radius. To obtain the geometric parameters of polar charts, we detect the chart center and its outermost circular boundary. The image is converted to grayscale, smoothed with a Gaussian filter, and processed with a Canny operator to enhance structural edges. The resulting edge map is analyzed using the Hough Circle Transform (CHT), which is robust to noise and partial edge loss and returns a set of candidate circles. Because polar charts often contain multiple concentric rings, we select the circle with the largest radius and adopt its center (x_c, y_c) and radius $r_{\max}$ as the geometric parameters of the chart.

Alignment of polar values to pixel radii. For pie and donut charts, the polar coordinate structure is fully determined by the center and the outer boundary, since these charts encode values solely through angular spans. In contrast, radar charts and rose diagrams encode data through polar distances, which requires an additional alignment between pixel radii and numeric values. To obtain this polar mapping, we detect two reference circles using CHT and extract the annular region between their

radii r_1 and r_2. This region contains the numeric tick labels corresponding to the two grid circles. The annular patch is provided to the MLLM, which reads the labels and yields the paired observations (r_1, V_1) and (r_2, V_2). With the shared center (x_c, y_c) and these two radius–value pairs, we construct a linear mapping from pixel radius to numeric values. This mapping defines the chart's radial metric and completes the polar-axis representation for subsequent prompting components.

c) Unified axis-prior representation: To enable a unified prompting design, we store the extracted axis priors for Cartesian and polar charts in a shared representation that encodes valid value ranges, axis orientation, and pixel-to-value mappings. These priors are then used to construct grid-based visual hints, prediction overlays, and progressive zoom-in refinement in *VisHintPrompt*.

*2) **Axis-Aware Grid Enhancement**:* Building on the extracted axis priors, this component overlays geometry-aligned reference grids as explicit visual cues for MLLM. Depending on the chart type, these grids take the form of orthogonal grids, concentric rings, or angular polar divisions, corresponding to different geometric encoding families (Fig. 2(b)). By embedding axis-guided visual scaffolds into the chart, the perceptual distance between data marks and nearby reference boundaries is reduced, enabling value inference within a more localized and interpretable neighborhood.

a) Constructing orthogonal grid hint: This form of prompting applies to charts embedded in Cartesian coordinates, such as bar charts, line charts, scatterplots, and bubble charts. For value-encoded axes, we generate a fine-grained orthogonal grid by subdividing the original tick intervals into visually manageable spans, approximately 50 pixels in width, as shown in Fig. 4(Scatter). These subdivisions maintain integer tick values whenever possible to avoid excessive floating-point annotations. For categorical axes, only label-aligned guide lines are inserted. Scatterplots and bubble charts adopt the same subdivision strategy symmetrically along both dimensions to preserve the underlying value mapping.

b) Constructing concentric ring hint: Charts such as radar and rose plots encode data magnitudes along polar directions. For these chart families, we construct a hierarchy of concentric rings by subdividing the polar axis, thereby generating finer reference levels around the original polar ticks, as illustrated in Fig. 5(Rose). These rings preserve the polar geometry while supplying clearer reference distances for the model to interpret polar magnitudes.

c) Constructing polar-division hint: Donut and pie charts encode proportional information through angular spans. For these charts, we construct polar-division prompting by inserting angular divisions aligned with the chart's outer radius, as shown in Fig. 4(Donut). These divisions are optionally annotated with angle values, enabling the model to directly read start and end angles for proportion estimation in a geometrically explicit manner.

Together, these geometry-specific prompting structures form a unified axis-guided reference framework that extends across Cartesian and polar chart families. By providing clear, geometry-aligned visual anchors, this stage substantially enhances the model's ability to localize, interpret, and quantify underlying chart encodings.

B. Iterative Visual Feedback

With the axis priors already extracted, this stage introduces an explicit form of visual feedback by overlaying the model's previous prediction onto the original chart. The key idea is to render the predicted data position back into the chart space so that the model can visually compare its estimate against the true geometric layout. This prediction-overlay mechanism provides a direct and interpretable hint that constrains subsequent inference and reduces ambiguity in the model's internal representation of the chart.

Given a prediction p_t obtained from the previous iteration, we project p_t back to chart coordinates using the extracted axis mapping and generate a visual overlay that marks the predicted location. This feedback overlay is defined as

$$F_t = \mathcal{O}(I, p_t, P_{\text{axis}}), \tag{1}$$

where $\mathcal{O}$ denotes the overlay operator and P_{axis} denotes the axis priors. The operator renders a chart-type–specific visual cue (e.g., marker, line, or angular indicator) at the predicted position. The resulting feedback-augmented visualization F_t is then fed into the MLLM for the next iteration.

This visual feedback serves two complementary functions. First, it makes the model's own uncertainty explicit by exposing the discrepancy between the predicted and the actual data location in the chart. Second, it creates a stable and geometry-aligned visual anchor that guides the model to re-evaluate its earlier prediction and adjust it toward the correct value. As a result, the prediction-overlay feedback forms a self-correcting loop that enhances the model's ability to align numerical inference with the chart's spatial encoding.

C. Progressive Zoom-in Refinement

To further elevate numerical inference accuracy, we introduce a *Self-evolving Visual Refinement* mechanism that performs progressively localized zoom-in reasoning based on axis priors anchored in the preceding stage. These axis-aware geometric priors provide a calibrated coordinate frame that supports a more disciplined and increasingly discriminative localization process.

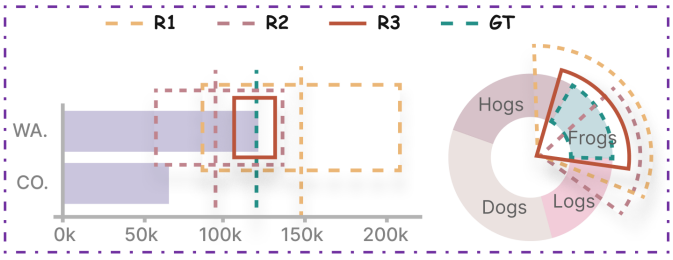

Fig. 3. An illustration of the Progressive Zoom-in Refinement mechanism on a horizontal bar chart and a donut chart. Successive regions (R1, R2, R3) show how the model iteratively crops closer to the target region, enabling increasingly precise numerical inference.

Each refinement iteration treats the model's previous prediction as an endogenous cue, effectively a self-generated supervisory signal that guides the construction of tighter visual constraints through localized cropping, magnification, and grid densification. We formalize these operations through an iterative region-refinement operator. Let p_t denote the model's numerical prediction at iteration t. The refined visual region is defined as

$$R_t = \mathscr{G}\big(\mathscr{M}(\mathscr{C}(I, p_t))\big), \tag{2}$$

where $\mathscr{C}(I, p_t)$ crops a region centered at the current predicted location p_t, $\mathscr{M}(\cdot)$ scales this crop, and $\mathscr{G}(\cdot)$ overlays a denser grid based on the axis priors.

Given the refined region R_t, the model updates its prediction through

$$p_{t+1} = f_\theta(R_t), \tag{3}$$

where f_θ denotes the multimodal model.

To ensure correctness, a self-verification criterion is introduced to check whether the cropped region still encloses the target chart element:

$$\mathscr{V}(R_t) = \begin{cases} 1, & \text{if the target element lies within } R_t, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

If $\mathscr{V}(R_t) = 0$, the cropping region is adaptively expanded and re-centered before refinement continues. Formally, the system updates

$$R_t \leftarrow \mathscr{C}\big(I, p_t; \alpha \cdot s_t\big), \tag{5}$$

where s_t is the current crop scale and $\alpha > 1$ controls the expansion ratio. Only when $\mathscr{V}(R_t) = 1$ does the algorithm proceed to compute p_{t+1}.

The complete refinement sequence forms a perceptual narrowing trajectory

$$p_0 \to p_1 \to \cdots \to p_T, \tag{6}$$

which empirically converges to a high-confidence estimate p^* that minimizes prediction error:

$$p^* = \arg\min_p \|p - p_{\text{true}}\|. \tag{7}$$

In practice, we run at most $T = 3$ refinement iterations, or stop earlier if the predicted value changes by less than ε. Through the interplay between self-evolving localization and verification-driven correction, this refinement stage produces progressively tighter value estimates and steers the model toward precise and reliable numerical predictions.

IV. Experiments

A. Experimental Setup

1) Evaluation Datasets: To rigorously evaluate *VisHint-Prompt*, we construct a unified evaluation dataset spanning two coordinate families: Cartesian charts, including vertical and horizontal bar, line, scatter, and bubble charts, and polar charts, including pie, donut, radar, and rose charts. Each chart is paired with pixel-accurate ground truth values and pixel-level representations of axis priors, enabling joint evaluation of numerical value extraction and axis-aware understanding. Existing benchmarks such as ChartQA [2], PlotQA [3], and ExcelChart400K [5] are mainly designed for chart question answering or chart OCR tasks. For our task setting, differences in three aspects make these benchmarks difficult to use: limited chart type coverage, the presence of explicit numeric labels within charts, and the lack of ground truth for both underlying data values and pixel-level axis priors needed to evaluate numerical inference across diverse chart families.

As shown in Table I, the dataset contains 450 charts (50 per type), and each chart includes several visual targets such as bars, points, or sectors. In total, this yields over 4,000 point-level prediction instances, providing sufficient statistical stability for per-type evaluation while keeping the computational cost of multi-round LLM inference tractable. To approximate the visual and structural diversity of real-world graphics, we introduce controlled variations in canvas size, image resolution, color palettes, font families and sizes, label layouts, object counts, and chart-specific properties such as bar orientation, line and marker styles, bubble label positions, sector spacing, ring widths, and radar dimensions. All charts are programmatically rendered within a unified pipeline; the "Real" and "Synthetic" labels in Table I refer only to the provenance of the underlying numeric data rather than to visual rendering. Real charts are generated from real-world numeric tables, whereas synthetic charts rely on randomly sampled distributions with controlled statistical patterns. This design ensures consistent visual rendering while preserving semantic variability across data sources. Table I summarizes the chart categories, real or synthetic composition, and key style variations incorporated in our dataset. [Yong: 1. Do we plan to publish the dataset? It will be good if we can; 2. It should be "Table I", NOT "table I". Please revise it accordingly.]

TABLE I
Overview of Chart Types and Data Composition

Category	Type	Total	Real	Syn.	Variations
Cartesian	Vertical Bar	50	25	25	1. Single/Multi bars
	Horizontal Bar	50	25	25	1. Single/Multi bars
	Line	50	25	25	1. Vertex shape; 2. Line style
	Scatter	50	25	25	1. Vertex shape; 2. Labels
	Bubble	50	25	25	1. Label position
Polar	Pie	50	25	25	—
	Rose	50	0	50	1. Sector spacing
	Radar	50	0	50	1. Dimensions; 2. Fill
	Donut	50	0	50	1. Ring width

2) Evaluation Metrics: Following the prior chart understanding research [2], [3], [6], we adopt a metric called *range-normalized error* to enable consistent comparison across various charts with different value scales. The evaluation is conducted hierarchically, starting from *point-wise* errors to aggregated *chart-level* errors as well as *type-level* errors.

Point-wise Range-normalized Error (RNE). For a chart c with ground truth values $\{v_i\}$ and predictions $\{\hat{v}_i\}$, the range-

normalized point-wise error is

$$e_{c,i}^{\mathrm{RNE}} = \frac{|\hat{v}_i - v_i|}{v_{\max,c} - v_{\min,c}}. \tag{8}$$

This normalization maps all errors into the $[0,1]$ interval, enabling fair cross-chart comparison and aligning with prior work that evaluates chart value prediction in normalized numeric space [3], [40].

Point-wise Relative Error (RE). To complement range normalization with a scale-sensitive metric, we also report the relative error with respect to the ground-truth value:

$$e_{c,i}^{\mathrm{RE}} = \frac{|\hat{v}_i - v_i|}{\max(|v_i|, \varepsilon)}, \tag{9}$$

where ε is a small constant to avoid division by zero. This metric captures the relative deviation from the true value and is particularly informative when charts contain values with different magnitudes.

Chart-level Error. For a given point-wise error metric $m \in \{\mathrm{RNE}, \mathrm{RE}\}$, the chart-level error is obtained by averaging the point-level errors:

$$E_c^{(m)} = \frac{1}{K_c} \sum_{i=1}^{K_c} e_{c,i}^{(m)}, \tag{10}$$

where K_c denotes the number of data points in chart c. This produces a single error measure per chart that reflects overall numerical consistency under metric m.

Type-level Error. Since our evaluation is conducted independently for each chart type, the final score for a chart type is computed, for each metric m, by averaging the chart-level errors over all charts of that type (e.g., bar chart and scatterplot) in our testing dataset:

$$E_{\mathrm{type}}^{(m)} = \frac{1}{C_{\mathrm{type}}} \sum_{c \in \mathrm{type}} E_c^{(m)}, \tag{11}$$

where C_{type} denotes the number of charts of that type. We report type-specific aggregates following recent unified chart understanding models [6], [26] to avoid cross-type mixing.

Overall, this point–chart–type hierarchy, combined with both range-normalized (RNE) and relative-to-ground-truth (RE) errors, provides a scale-invariant, type-consistent, and literature-aligned evaluation of numerical inference performance for *VisHintPrompt*.

3) Model Selection: To evaluate the model-agnostic generality of *VisHintPrompt*, we apply our prompting framework to a diverse set of MLLMs spanning both closed-source and open-source systems. This design allows us to examine whether structured visual hints consistently enhance numeric inference ability across heterogeneous architectures without any fine-tuning or model-specific adaptation.

a) Closed-source MLLMs: We evaluate *VisHintPrompt* across three closed-source MLLMs with advanced visual–language capabilities.

- **GPT-4o** — a general-purpose vision–language model known for its robust perception and reasoning.
- **Gemini 2.0 Flash** — an efficient and high-throughput model with competitive visual understanding.
- **Gemini 2.5 Flash** — an upgraded version with improved fine-grained spatial reasoning.

b) Open-source MLLMs: We further evaluate three representative open-source models covering a wide range of parameter scales and architectural families:

- **InternVL3-78B** — a large, high-capacity open-source VLM with strong visual recognition.
- **Pixtral-12B-2409** — a transformer-based multimodal model with competitive image understanding.
- **GLM-4.6V-Flash** — a lightweight model suitable for evaluating low-resource settings.

c) Rationale: Our focus is to examine how much numeric extraction accuracy can be gained *without* any training, finetuning, or synthetic chart supervision. Modern MLLMs already possess latent spatial and geometric reasoning capabilities, and *VisHintPrompt* provides structured chart-specific visual hints that enable these capabilities to be more effectively activated and utilized.

d) Exclusion of chart-specific and OCR systems: Chart-specific QA models (e.g., PlotQA, ChartQA, UniChart) rely on synthetic training and support only a narrow set of chart families, making them incompatible with our full-range numeric inference setting.

Overall, this model selection covers a comprehensive spectrum of MLLMs and provides a rigorous testbed for assessing the generality and effectiveness of *VisHintPrompt*.

4) Implementation Details: For closed-source models, we access GPT-4o, Gemini-2.0-Flash, and Gemini-2.5-Flash via their official APIs; GPT-4o uses deterministic decoding (temperature $= 0$), while the Gemini models follow their default inference settings. For the open-source model Pixtral-12B-2409, inference is performed locally on a BI-V150 GPU server using the official released implementation, without any parameter modification. In contrast, InternVL3-78B and GLM-4.6V-Flash are accessed via their official APIs under default inference settings. All prompts follow fixed templates for each elicitation stage, and each model prediction is parsed through a strict JSON schema with one retry if a malformed response is encountered. We did not perform any prompt tuning or model-specific adaptation beyond these fixed templates.

V. Results and Analysis

A. Quantitative Evaluation

This section presents the quantitative performance of MLLMs under our proposed *VisHintPrompt*. The evaluation covers more than 4,000 point-level prediction instances across Cartesian and polar chart families. Unless otherwise specified, all reported errors are type-level Range-normalized Error (RNE) and Relative Error (RE) as defined in Section IV, obtained by averaging point-wise errors over charts within each chart type and reported as percentages. Importantly, none of the benchmark charts include explicit numeric labels on the visual marks, precluding trivial solutions based on OCR. Consequently, all predictions must rely on visual inference from geometric cues such as bar height, scatter-point position, angular extent, and polar distance.

*1) **Overall Performance**:* As shown in Table II, averaged over five MLLMs and nine chart types, *VisHintPrompt* leads to overall reductions in numerical error (RNE and RE) compared

TABLE II
MAIN EXPERIMENTAL RESULTS ACROSS FIVE MLLMS ON NINE CHART TYPES. WE REPORT RANGE-NORMALIZED ERROR (RNE) AND RELATIVE ERROR (RE; LOWER IS BETTER) FOR BOTH THE BASELINE (ORIGINAL CHARTS) AND OUR VISHINTPROMPT METHOD.

MLLMs	Setting	Scatter		Bubble		v−Bar		h−Bar		Line		Radar		Rose		Pie		Donut		Avg.	
		RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE	RNE	RE
GPT-4o	Baseline	6.72	4.69	6.63	25.34	5.7	14.22	9.05	49.44	3.84	9.77	25.54	55.71	16.58	22.64	**3.72**	**25.72**	**2.64**	**30.93**	8.94	26.5
	VisHintPrompt	**3.44**	**2.23**	**3.38**	**10.90**	**3.03**	**6.91**	**1.94**	**9.22**	**2.13**	**4.11**	**11.92**	**32.14**	**9.99**	**13.4**	12.52	43.81	7.94	67.15	**6.25**	**21.1**
Gemini 2.0 Flash	Baseline	3.08	13.43	3.03	12.79	**0.67**	**0.26**	5.43	14.03	**0.31**	**1.02**	20.63	50.21	18.9	25.94	3.8	23.53	2.81	27.84	6.52	18.78
	VisHintPrompt	**1.38**	**4.30**	**1.41**	**4.06**	1.65	0.64	**1.65**	**4.56**	0.81	1.66	**6.93**	**21.77**	**10.0**	**13.57**	**3.03**	**15.63**	**1.75**	**16.92**	**3.18**	**9.23**
Gemini 2.5 Flash	Baseline	3.19	9.44	3.31	10.37	2.21	5.49	3.41	37.02	**0.54**	**1.84**	27.46	52.0	12.23	16.83	2.92	23.53	2.44	21.56	6.41	19.79
	VisHintPrompt	**2.62**	**7.20**	**3.10**	**7.89**	**0.95**	**2.78**	**0.77**	**5.42**	1.84	3.56	**6.63**	**21.6**	**7.16**	**10.45**	**1.89**	**9.07**	**1.47**	**13.38**	**2.94**	**9.04**
InternVL3-78B	Baseline	3.30	2.19	5.04	12.47	**1.27**	4.35	5.39	16.69	**0.83**	**2.07**	16.25	37.0	11.73	16.97	**3.62**	25.72	**2.28**	29.89	5.52	16.37
	VisHintPrompt	**1.45**	**1.01**	**3.79**	**9.95**	1.40	**4.19**	**0.64**	**5.43**	1.77	3.79	**6.78**	**15.85**	**8.97**	**13.09**	4.94	**18.73**	3.02	**26.95**	**3.64**	**11.0**
Pixtral-12B-2409	Baseline	7.38	4.86	6.69	18.92	**2.03**	**3.27**	**6.98**	43.43	**2.62**	7.38	33.15	72.36	34.57	46.68	**6.13**	55.04	9.73	84.60	12.14	37.4
	VisHintPrompt	**4.73**	**3.22**	**6.36**	**9.32**	15.26	24.6	9.28	**36.29**	2.95	**6.99**	**15.56**	**44.2**	**11.89**	**17.16**	13.31	**50.28**	**6.15**	**44.82**	**9.50**	**26.32**
All MLLMs	Baseline	4.73	6.92	4.94	15.98	**2.38**	**5.52**	6.05	32.12	**1.63**	4.42	24.61	53.46	18.8	25.81	**4.04**	30.71	**3.98**	38.96	7.91	23.77
	VisHintPrompt	**2.72**	**3.59**	**3.61**	**8.42**	4.46	7.82	**2.86**	**12.18**	1.9	**4.02**	**9.56**	**27.11**	**9.6**	**13.53**	7.14	**27.5**	4.07	**33.84**	**5.1**	**15.34**

to the baseline. These reductions are observed across all five evaluated MLLMs. Specifically, across all model–chart type combinations, type-level RNE and RE decrease in 32 of the 45 cases, with closed-source MLLMs achieve improvements on at least 7 of the 9 chart types, while open-source MLLMs also improve on at least 5 of the 9.

Closed-source MLLMs achieve the lowest post-prompting type-level RNE and RE. For example, for Gemini-2.5-Flash, the mean type-level RNE across chart types drops from 6.41% to 2.94% and the mean type-level RE from 19.79% to 9.04%, reflecting its stronger visual–language integration. Open-source models, in contrast, often exhibit larger *relative* reductions in error. For a representative open-source model, Pixtral-12B-2409, the mean type-level RNE across chart types drops from 12.14% to 9.5% and the mean type-level RE from 37.4% to 26.32%, suggesting that scaffolded visual hints can partially compensate for weaker geometric priors rather than merely amplifying existing strengths. The fact that these gains appear across heterogeneous architectures indicates that the proposed strategy behaves largely model-agnostically.

Overall, *VisHintPrompt* reduces numerical error across five MLLMs and nine chart types, with lower chart-averaged errors observed for each open- and closed-source MLLM. Taken together, these results confirm the effectiveness of *VisHintPrompt* in improving numerical inference from charts across diverse chart types and heterogeneous MLLMs.

*2) **Performance Across Chart Types**:* We group the nine chart types into three categories for focused analysis: charts with high baseline accuracy (vertical bar and line charts), charts with visually complex structures (horizontal bar, scatter, bubble, radar, and rose charts), and charts involving proportion-based encodings, for which *VisHintPrompt* introduces a grid-based angle estimation strategy for proportion inference (pie and donut charts).

a) Charts with high baseline accuracy: **Highly prevalent vertical bar and line charts with dense, grid-aligned structures exhibit the strongest baseline performance across both closed- and open-source MLLMs, resulting in limited room for improvement under *VisHintPrompt*.** Most cases that do not show clear improvements arise from this group. For example, in one of the strongest closed-source models, Gemini-2.0-Flash, baseline errors on these chart types are already extremely low, with RNE below 0.67% and 0.31% and RE below 0.26% and 1.02% for vertical bar and line charts, respectively. Under such near-saturated conditions, *VisHintPrompt* yields only marginal changes, and minor fluctuations may occasionally appear. Model-specific differences can nevertheless be observed. While Pixtral-12B-2409 shows limited gains on vertical bar charts, GPT-4o and Gemini-2.5 Flash still benefit from *VisHintPrompt* due to their relatively higher baseline errors. In one specific case, Pixtral-12B-2409 on vertical bar charts, *VisHintPrompt* may introduce slight degradations, which can be explained by near-saturated baseline performance and sensitivity to additional visual cues.

b) Structurally Demanding Chart Families: **Structurally complex chart types, icluding horizontal bar, scatter, bubble, radar, and rose charts, exhibit substantially higher baseline errors, yet consistently benefit from *VisHint-Prompt*, with pronounced error reductions observed across all evaluated open- and closed-source MLLMs.**

When averaged across MLLMs, all five structurally demanding chart types exhibit approximately 50% relative reductions in both RNE and RE compared to their baseline errors. At the chart–MLLM combination level, *VisHintPrompt* typically reduces type-level RNE by approximately 2%–6% and RE by 5%–10% for most models. More pronounced improvements are observed for polar charts. On radar charts, type-level RNE across models decreases from approximately 25%–70% to 6%–27%, with a median reduction of around 8%. Similarly, on rose charts, RNE is reduced from roughly 10%–30% to 7%–10%, again with median reductions close to 8%. Comparable trends are observed for scatter and bubble charts. Horizontal bar charts exhibit the largest gains. For example, two closed-source MLLMs (Gemini 2.0 Flash and Gemini 2.5 Flash) and one open-source MLLM (InternVL3-78B) reduce RE from approximately 10%–40% at baseline to around 5% under *VisHintPrompt*. Taken together, these results provide strong empirical evidence that *VisHintPrompt* effectively enhances numerical inference under challenging visual structures where baseline model performance is limited.

c) Proportion-Based Charts with Angle Estimation: **Donut charts** depend on accurate estimation of angular

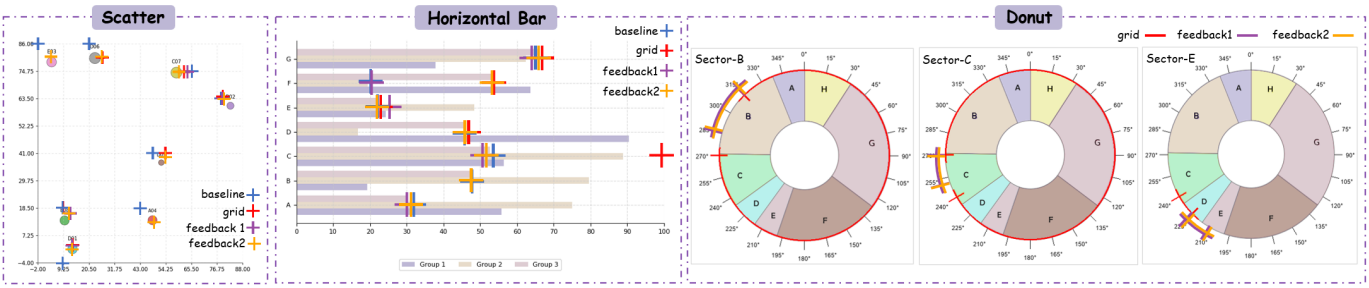

Fig. 4. Qualitative examples illustrating grid enhancement and iterative visual feedback (Case 1) on three chart types: scatterplot, horizontal bar, and donut. Colored crosshairs denote predictions from different stages: baseline (blue), grid-aware enhancement (red), and the first and second feedback rounds (purple and yellow). Grid enhancement reduces coarse localization ambiguity, while iterative feedback further corrects large residual deviations and stabilizes predictions across chart types.

boundaries, where small localization errors directly translate into value errors.

*3) **Summary**:* Taken together, the type-level results show that *VisHintPrompt* delivers its greatest benefits on chart families that impose complex geometric reasoning, while still offering modest but reliable refinements on visually simpler chart types. The resulting performance distribution delineates the visual structures that current multimodal models naturally handle well and those for which targeted visual prompting remains essential.

TABLE III
QUALITATIVE COMPARISON ON A SCATTERPLOT, A HORIZONTAL BAR CHART, AND A DONUT. FOR EACH CHART TYPE, WE REPORT THE GROUND-TRUTH VALUES ALONGSIDE THE BASELINE MLLM PREDICTIONS AND THE OUTPUTS OF VISHINTPROMPT.

Scatterplot									
Setting \ **Point Name**		D01	D02	E03	A04	D05	D06	C07	E08
Ground Truth	X	13.22	82.65	3.90	48.22	52.20	22.95	58.90	9.77
	Y	1.66	60.55	78.39	13.03	37.04	80.23	74.14	13.39
Baseline	X	9.25	79.75	-2.00	43.00	48.75	20.50	65.50	**9.25**
	Y	-4.00	63.50	86.00	18.50	41.00	86.00	**74.75**	18.50
VisHintPrompt	X	**13.52**	**82.45**	**4.27**	**47.98**	**54.25**	**24.06**	**59.23**	**10.75**
	Y	**3.23**	**60.82**	**78.97**	**11.94**	**36.98**	**80.00**	**74.75**	**13.47**

Bar Chart (Group 3)								
Setting	A	B	C	D	E	F	G	
Ground Truth	29.97	47.33	52.55	45.71	22.02	52.95	65.86	
Baseline	32.00	**47.00**	52.00	**46.00**	**22.00**	20.00	65.00	
VisHintPrompt	**30.00**	**47.00**	**53.00**	**46.00**	**22.00**	**53.00**	**66.00**	

Donut								
Setting	A	B	C	D	E	F	G	H
Ground Truth	0.060	0.190	0.100	0.050	0.050	0.200	0.260	0.090
Baseline	0.083	0.160	0.120	**0.063**	**0.055**	0.250	**0.250**	0.120
VisHintPrompt	**0.056**	**0.192**	**0.106**	0.067	0.083	**0.200**	0.233	**0.111**

B. Qualitative Analysis

To illustrate how *VisHintPrompt* operates and how its components contribute to numerical inference, we present a qualitative analysis with two representative cases. Case 1 examines the effect of axis-aware grid enhancement and iterative visual feedback, showing how structured visual scaffolds progressively reduce localization ambiguity and stabilize value estimation. Case 2 focuses on the zoom-in refinement mechanism, illustrating how progressive amplification enables increasingly fine-grained value discrimination.

*1) **Case1. Grid-aware Enhancement and Iterative Visual Feedback**:* We examine the effects of grid augmentation and iterative visual feedback using three representative chart types: bubble charts, bar charts, and donut charts. As shown in the Fig. 4, predictions from baseline, grid-aware enhancement and feedback rounds are encoded using colored crosshairs, following the same visual convention as in the actual iterative feedback process. Specifically, blue denotes the baseline prediction, red corresponds to predictions made on grid-aware enhancement images, and purple and yellow indicate predictions obtained after the first and second rounds of visual feedback, respectively. Overall, compared to the baseline, grid-aware enhancement improves prediction quality in most cases across both bubble and bar chart types. A closer inspection reveals that for the bubble chart, all examined data points benefit from grid enhancement relative to the baseline. For the horizontal bar chart, the introduction of grids also yields positive effects for more than half of the data points. [Yong: How should we find the correspondence between the findings here and what is shown in Figure 5? It is not that clear to me.]

Building on the grid-enhanced predictions, we examine the effect of iterative visual feedback across different chart types. In most cases, feedback contributes positively by progressively adjusting predictions toward the ground truth positions. Across chart types, this effect is primarily reflected in the correction of large localization errors that persist after grid enhancement. For the bubble chart, a representative example is data point *E08*, where feedback effectively corrects a substantial deviation that remains after grid enhancement. [Yong: Pls check my Chinese comments.] For the bar chart, the *C, Group 3* data point [Yong: where is "the *C, Group 3* data point" in Figure 5??? Are you referring to the bar?] illustrates a similar pattern: grid enhancement alone was insufficient to correct a large deviation, while after one round of feedback, the prediction returned to a reasonable error range. [Yong: How do we know that from Figure 5??] For donut charts, feedback not only mitigates large angular errors but also reduces the likelihood of confusing start and end angles introduced at the grid-aware

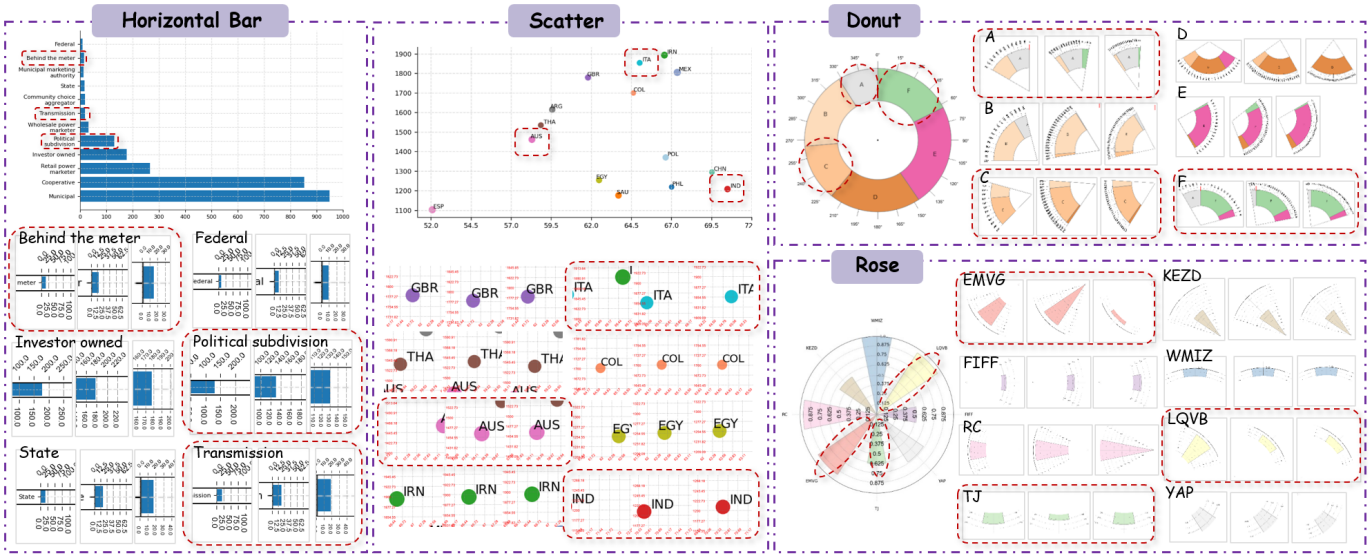

Fig. 5. Qualitative examples of our *VisHintPrompt* on four chart families—h-bar, scatter, donut, and rose. Each dashed red box highlights the iterative zoom-in process guided by the feedback prediction, illustrating how the model progressively converges to the correct region in the final refinement stage.

enhancement stage, as illustrated by data point *B*. [Yong: How do we know that from Figure 5?? Where should we look at?]

At the same time, feedback does not uniformly lead to monotonic error reduction for all data points. For instance, for the data point *C, Group 3* in the bar chart, while feedback corrects a large initial deviation, the second feedback round results in a slightly larger error relative to the first. This observation is consistent with our ablation results, indicating that feedback alone provides limited accuracy gains. Nevertheless, by improving localization reliability, feedback enables subsequent zoom-in refinement to further improve overall prediction accuracy.

Importantly, a comparison of the point-wise ground truth coordinates, baseline predictions, and *VisHintPrompt* predictions reported in Table III shows that, despite non-monotonic behavior at intermediate feedback rounds, the final predictions achieved by *VisHintPrompt* are substantially more accurate than the baseline across the examined data points.

[Yong: Please revise Case 2 by checking and addressing similar issues existing in Case 1.]

*2) **Case2. Progressively Zoom-in Refinement**:* Our zoom-in refinement mechanism is centered on the model's current prediction: when the target mark lies near the geometric center of the cropped grid, the prediction is likely accurate. For donut charts, whose sectors represent discrete angular partitions, we further widen the tolerance band such that symmetric left and right neighboring sectors provide a visual cue of convergence. Fig. 5 summarizes these refinement trajectories showing how predictions evolve across iterations and table IV lists the ground-truth values together with the baseline and *VisHintPrompt* predictions for the same representative points.

Scatterplot (Cartesian): Scatterplots provide the clearest setting for analyzing geometric convergence: a correct prediction results in the corresponding scatter point lying at the crop center. Across the dataset, the *VisHintPrompt* pipeline reliably produces such centered configurations, consistent with the quantitative gains reported in Table IV. A closer examination reveals that nearly all points undergo meaningful refinement. For labels *AUS*, *ITA*, and *IND*, the feedback predictions still exhibit substantial deviations from the ground truth, visible as points far from the crop center and similar in magnitude to the baseline errors. With successive amplifier stages, however, the crops progressively tighten around the correct coordinates and the predictions shift inward. These trajectories highlight the pipeline's ability to correct coarse localization errors and to stabilize fine-grained spatial reasoning on Cartesian layouts.

Donut chart (polar with angular value mapping): For donut charts, accurate inference requires identifying the correct angular segment. In our framework, a successful prediction is reflected by the target sector appearing centrally in the crop and being symmetrically bounded by its neighboring sectors. As with scatterplots, *VisHintPrompt* consistently achieves this centered configuration after refinement. Specifically, labels *A*, *C*, and *F* illustrate typical correction patterns: although the initial angular estimates are noticeably misaligned, the visual feedback and zoom-in refinement stages incrementally steer the prediction toward the correct boundaries. Ultimately, the target sectors appear symmetrically framed in the final crop. This behavior demonstrates the efficacy of scaffolded visual hints for fine-grained angular reasoning and confirms the robustness of the approach on polar charts with angular value encoding.

Rose chart (polar with polar value mapping): Rose charts combine angular segmentation with polarly varying magnitudes, making them considerably more challenging than scatterplot or donut charts. Baseline predictions often deviate substantially from the ground truth, revealing the intrinsic difficulty of interpreting irregular polar geometry. Despite these challenges, *VisHintPrompt* still produces clear improvements. Labels *EMVG*, *TJ*, and *LQVB* again show strong refinement, with predictions moving from large angular and polar mismatches toward the crop center over iterations. Notably,

TABLE IV
QUALITATIVE COMPARISON ON AN H-BAR CHART, A SCATTERPLOT, A DONUT, AND A ROSE CHART. FOR EACH CHART TYPE, WE REPORT THE GROUND TRUTH COORDINATES OR NORMALIZED VALUES FOR REPRESENTATIVE POINTS ALONGSIDE THE BASELINE MLLM PREDICTIONS AND THE OUTPUTS OF *VisHintPrompt*, ILLUSTRATING THE REDUCTION IN NUMERICAL ERROR. [YONG: PLS CHECK MY CHINESE COMMENTS.]

H-bar Chart

Setting	Municipal	Cooperative	Retail power marketer	Investor owned	Political subdivision	Wholesale power marketer	Transmission	Community choice aggregator	State	Municipal marketing authority	Behind the meter	Federal
Ground Truth	950	853	266	178	131	32	20	19	17	14	15	9
Baseline	945	**850**	**270**	188	**130**	40	25	25	25	20	25	20
VisHintPrompt	**950**	**850**	**270**	**180**	**130**	**31**	**19.5**	**19**	**20**	**12.5**	**17.5**	**10**

Scatterplot

Point Name / **Setting**		GBR	ESP	AUS	PHL	POL	ITA	ARG	CHN	EGY	MEX	THA	SAU	IND	COL	IRN
Ground Truth	X	61.78	52.08	58.29	66.99	66.63	65.01	59.55	69.51	62.48	67.34	58.85	63.68	70.48	64.62	66.54
	Y	1779.98	1104.30	1462.91	1220.70	1371.76	1854.68	1616.10	1297.37	1255.41	1806.33	1536.34	1177.09	1209.53	1701.17	1894.95
Baseline	X	62.00	**52.00**	57.30	66.74	66.50	65.10	**59.50**	**68.50**	62.00	67.00	**59.01**	63.10	70.97	64.50	67.00
	Y	1740.00	**1100.00**	1475.76	**1222.73**	1333.85	1857.60	**1619.70**	**1300.00**	1278.79	1750.00	1524.62	1140.00	1160.00	**1700.00**	1874.24
VisHintPrompt	X	**61.70**	**52.00**	**58.39**	**67.00**	**66.72**	**64.92**	59.67	**69.50**	**62.44**	**67.42**	59.22	**63.66**	**70.47**	**64.64**	**66.67**
	Y	**1772.73**	**1100.00**	**1466.09**	1224.24	**1377.27**	**1854.55**	1627.27	1290.00	**1263.64**	**1800.00**	**1545.45**	**1186.36**	**1211.37**	**1700.00**	**1900.00**

Donut Chart							**Rose Chart**							
Setting	A	B	C	D	E	F	FIFF	LQVB	WMIZ	KEZD	RC	EMVG	TJ	YAP
Ground Truth	0.090	0.180	0.130	0.200	0.230	0.180	0.52	0.95	1.00	0.58	0.96	0.92	0.63	0.73
Baseline	0.167	0.167	0.167	**0.170**	0.170	0.170	0.40	0.80	**1.00**	0.82	0.60	0.66	0.30	0.30
VisHintPrompt	**0.097**	**0.175**	**0.125**	0.167	**0.233**	**0.172**	**0.54**	**0.91**	0.97	**0.50**	**0.81**	**0.90**	**0.59**	**0.53**

in example *EMVG*, the cropping window does not shrink monotonically in the second round. Instead, after detecting that the refined window no longer contains the target sector, the verification mechanism triggers an adaptive enlargement of the crop. This ensures that subsequent refinement begins from a valid region, enabling recovery even from severe initial errors. This case highlights the extensibility and stability of the pipeline when applied to more complex polar chart structures.

Summary: Across both Cartesian and polar coordinate systems, the qualitative evidence aligns closely with the quantitative findings. When geometric cues are salient, as in scatterplots and simpler polar charts, the visual-hint prompting pipeline yields stable and often highly accurate convergence. Even in structurally complex polar charts, where baseline performance is weak, the combination of axis-aware grid enhancement, visual feedback prompting, and progressive zoom-in refinement enables the model to recover from substantial initial prediction errors. These qualitative analyses collectively demonstrate the robustness, generality, and cross-family applicability of *VisHintPrompt*.

C. Ablation Study

We next examine the contribution of each elicitation component in *VisHintPrompt* by conducting ablations on one representative chart type per coordinate family: horizontal bar, donut and radar charts. All variants use the same backbone model (Gemini-2.0-Flash) and prompt template; only the visual-hint stages are modified. We consider the following settings:

- **Baseline** uses a single forward pass of the vanilla model.
- **w/o Grid Enhancement** removes the initial axis-guided global grid. To keep local refinement meaningful, we retain the fine-grained grids drawn inside cropped regions for all variants that use local amplification. This ablation therefore isolates the additional benefit of global axis priors beyond the grids that are intrinsically coupled with progressive zoom-in refinement.
- **w/o Visual Feedback** keeps global grid prompting but disables prediction overlay iterations. The model observes the gridded chart once and directly outputs values, and the progressive zoom-in refinement stage uses this single global estimate as its starting point.
- **w/o Zoom-in Refinement** retains global grid hint and visual feedback, while performing all inference at the original resolution.
- ***VisHintPrompt*** combines all three components: Axis-aware Grid Enhancement, Iterative Visual Feedback, and Progressive Zoom-in Refinement with fine-grained grids in the cropped regions.

Table V reports RNE and RE (lower is better) for these variants on the three chart types. [Yong: Three or four???] Overall, every stage contributes to the final performance: removing any single component degrades results relative to the full pipeline, although the magnitude of the drop differs across stages and chart families. The most pronounced effect comes from disabling progressive zoom-in refinement. On donut and radar charts, the w/o progressive zoom-in refinement variant shows the largest increases in RE, indicating that progressive zoom-in refinement is the main driver of precise numerical gains. Without this stage, the model cannot reliably convert coarse geometric understanding into accurate value estimates.

TABLE V
ABLATION STUDY OF MAJOR ELICITATION STAGES. WE REPORT RANGE-NORMALIZED ERROR (RNE) AND RELATIVE ERROR (RE; LOWER IS BETTER) ON ONE REPRESENTATIVE CHART TYPE FROM EACH COORDINATE FAMILY.

Method Variant	Bar		Donut		Radar	
	RNE	RE	RNE	RE	RNE	RE
Baseline (Gemini-2.0-Flash)	5.4	14	2.88	26.86	27.1	51.7
w/o Grid Ehancement	0.25	0.88	—	—	7.5	24.1
w/o Visual Feedback	0.17	1.8	2.3	26.62	7.3	22.6
w/o Zoom-in Refinement	4.5	23	4.04	38.9	16.4	48.0
VisHintPrompt	**0.15**	**0.9**	**1.99**	**17.77**	**7.2**	**22.3**

Grid and visual feedback hints yield smaller but consistent gains compared with progressive zoom-in refinement. On horizontal bar charts, the w/o Grid Prompting variant performs comparably to the full model and even slightly better under some metrics, suggesting that for simple Cartesian layouts with strong backbone priors, the marginal benefit of a global axis grid is limited. In such cases, most remaining gains are driven by crop-based zoom-in refinement rather than global grid prompting. In contrast, for radar charts, removing global grid prompting consistently degrades performance, indicating that explicit axis priors become increasingly important as chart geometry departs from canonical Cartesian layouts and values must be inferred from polar encodings.

An interesting pattern emerges when only the feedback stage is retained. The w/o progressive zoom-in refinement variant, which includes grid prompting and visual feedback but no zoom-in refinement, can yield larger errors than the baseline on some chart types. This suggests that visual feedback by itself mainly provides coarse localization cues: it helps the model find the right region but does not substantially improve numeric precision. When visual feedback and progressive zoom-in refinement are used together, feedback supplies a stable starting point that avoids extreme mislocalizations, and the zoom-in refinement component then performs fine-grained adjustment within this region.

Taken together, these ablations show that global grid hint, visual feedback, and progressive zoom-in refinement provide complementary benefits. Grid prompting and feedback shape coarse geometric reasoning and stabilize the entry point into the refinement pipeline, while grid-coupled zoom-in refinement is crucial for translating these coarse cues into accurate numerical predictions. The full *VisHintPrompt* pipeline consistently outperforms every ablated variant across Cartesian and polar coordinate families, with the largest improvements arising when all three stages are jointly enabled. [Yong: pls check my Chinese comments.]

VI. DISCUSSION

This work demonstrates that structured visual scaffolding can substantially reshape and strengthen numerical inference from charts in MLLMs. Beyond the empirical improvements, our findings reveal how current models internalize chart structure, why the proposed elicitation stages are effective, and how these insights generalize to broader visual reasoning tasks. We discuss these implications below.

A. Insights and Broader Implications

a) Differential priors in MLLMs for chart geometry: [Yong: what are differential priors here?] Our results indicate that MLLMs exhibit a structured hierarchy of difficulty across chart types that reflects underlying geometric properties. Models show strong implicit priors for canonical Cartesian layouts, such as vertical bar, line, scatter, and bubble charts, where value–position mappings are prevalent in natural image–text data. In contrast, polar charts, including donut, rose, and radar charts, are associated with substantially weaker priors, resulting in higher baseline errors and larger performance gains under *VisHintPrompt*. These findings suggest that numerical inference in MLLMs is shaped not only by pixel-level perception but also by the compatibility between chart geometry and the model's learned inductive biases.

b) Functional decomposition of visual-hint components: The ablation study clarifies how the three elicitation components contribute complementary roles. Axis-aware global grids and visual feedback mainly support coarse spatial alignment by externalizing numeric anchors and encouraging self-correction of initial predictions. Progressive zoom-in refinement, however, provides the precision mechanism required to convert coarse localization into accurate values. Effective visual prompting therefore requires both global structural cues and targeted local refinement.

c) Externalizing structured priors as visual scaffolds: The success of *VisHintPrompt* illustrates a broader methodological principle: many visual domains [Yong: what are "many visual domains"?] rely on structured knowledge that humans naturally invoke but that MLLMs do not reliably infer from raw pixels. By rendering this structure directly in the input, through grids, overlays, and localized refinements, the prompting pipeline serves as an inference-time scaffold that enables the model to compute within a well-defined geometric frame. This principle extends naturally to engineered schematics, circuit diagrams, process flows, and other structured visual artifacts where consistent grammars are available but underrepresented in training data. Externalizing such a visual structure offers a lightweight, domain-adaptable scaffold that enhances multimodal reasoning without modifying model parameters.

B. Limitations and Future Directions

Dependence on mark count and labeling density ***[yumeng: :]*** *:* Our strategy assumes that visual elements remain distinguishable under progressive crops. Charts with extremely dense marks or overlapping labels challenge this assumption. In such cases, the amplifier stage may struggle to preserve semantic correspondence between the cropped region and the original chart. A potential solution is a two-step labeling pipeline: the model first proposes a coarse identity tag for each mark (e.g., region, item, or category), and the refinement

procedure then operates relative to these inferred identities. Preliminary experiments on scatterplots with off-center labels suggest that this approach can help maintain stable grounding even under dense or cluttered layouts.

Scalability across diverse visual styles ***[yumeng: .]*** *:* Our current pipeline was evaluated on charts that follow consistent stylistic conventions (line thickness, text contrast, mark shapes, and color palettes). Extending the method to charts with extreme stylistic diversity may require adaptive tolerance setting, dynamic grid density, or learned cropping heuristics. Incorporating style-aware modules or meta-learning mechanisms represents an important future challenge.

These limitations point to several directions for future work. Integrating *VisHintPrompt* into end-to-end chart understanding pipelines would enable evaluation on large-scale natural datasets. Extending structured visual scaffolds to other schematic domains could assess the generality of axis externalization and coarse-to-fine refinement. Finally, exploring hybrid strategies, such as reinforcement-learned zoom policies or combining prompting-based scaffolds with lightweight finetuning, may consolidate the benefits of our approach while reducing inference-time cost.

VII. Conclusion

We introduce *VisHintPrompt*, a novel visual hint prompting strategy designed to unleash the potential of MLLMs in chart data extraction tasks. Unlike traditional approaches that rely solely on textual instructions, *VisHintPrompt* integrates chain-of-thought prompting with visual hint-based multimodal strategies, enabling the model to perform numerical inference directly over grid-enhanced chart images. The strategy iteratively refines predictions through a visual feedback mechanism and leverages progressive zoom-in refinement to capture fine-grained visual details, thereby enhancing the model's perception and interpretation of target data objects. Experimental results demonstrate that *VisHintPrompt* constantly improves both interpretability and accuracy of numerical extraction in challenging scenarios, such as small-scale graphical elements. Moreover, the method achieves competitive performance that matches or surpasses recent state-of-the-art MLLM-based methods, without requiring additional training.

[Yong: where is our future work? We can either add it here or in Section VI.]

References

[1] F. Liu, J. Eisenschlos, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, W. Chen, N. Collier, and Y. Altun, "DePlot: One-shot visual language reasoning by plot-to-table translation," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 381–10 399. [Online]. Available: https://aclanthology.org/2023.findings-acl.660/

[2] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A benchmark for question answering about charts with visual and logical reasoning," in *Findings of the Association for Computational Linguistics: ACL 2022*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2263–2279. [Online]. Available: https://aclanthology.org/2022.findings-acl.177/

[3] N. Methani, P. Ganguly, M. M. Khapra, and P. Kumar, "Plotqa: Reasoning over scientific plots," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, pp. 1516–1525.

[4] M. Savva, N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer, "Revision: automated classification, analysis and redesign of chart images," in *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 393–402. [Online]. Available: https://doi.org/10.1145/2047196.2047247

[5] J. Luo, Z. Li, J. Wang, and C.-Y. Lin, "Chartocr: Data extraction from charts images via a deep hybrid framework," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1916–1924.

[6] A. Masry, P. Kavehzadeh, X. L. Do, E. Hoque, and S. Joty, "UniChart: A universal vision-language pretrained model for chart comprehension and reasoning," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '23, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14 662–14 684. [Online]. Available: https://aclanthology.org/2023.emnlp-main.906/

[7] V. Cărbune, H. Mansoor, F. Liu, R. Aralikatte, G. Baechler, J. Chen, and A. Sharma, "Chart-based reasoning: Transferring capabilities from llms to vlms," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024, pp. 989–1004.

[8] X. Zeng, H. Lin, Y. Ye, and W. Zeng, "Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2024.

[9] M. Huang, L. Zhang, J. Ma, H. Lai, F. Xu, Y. Li, W. Wu, Y. Wu, and J. Liu, "Chartsketcher: Reasoning with multimodal feedback and reflection for chart understanding," *arXiv preprint arXiv:2505.19076*, 2025.

[10] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo, "ChartInsights: Evaluating multimodal large language models for low-level chart question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 174–12 200. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.710/

[11] Y. Wu, Y. Wang, S. Tang, W. Wu, T. He, W. Ouyang, P. Torr, and J. Wu, "Dettoolchain: A new prompting paradigm to unleash detection ability of mllm," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds. Cham: Springer Nature Switzerland, 2025, pp. 164–182.

[12] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 987–11 997.

[13] X. Lei, Z. Yang, X. Chen, P. Li, and Y. Liu, "Scaffolding coordinates to promote vision-language coordination in large multi-modal models," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 2886–2903. [Online]. Available: https://aclanthology.org/2025.coling-main.195/

[14] D. Jung, W. Kim, H. Song, J.-i. Hwang, B. Lee, B. Kim, and J. Seo, "Chartsense: Interactive data extraction from chart images," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 6706–6717. [Online]. Available: https://doi.org/10.1145/3025453.3025957

[15] Z. Zhou, H. Wang, Z. Zhao, F. Zheng, Y. Wang, W. Chen, and Y. Wang, "Chartkg: A knowledge-graph-based representation for chart images," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 5854–5868, 2025.

[16] M. Cliche, D. Rosenberg, D. Madeka, and C. Yee, "Scatteract: Automated extraction of data from scatter plots," in *Machine Learning and Knowledge Discovery in Databases*, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, Eds. Cham: Springer International Publishing, 2017, pp. 135–150.

[17] K. Kafle, B. Price, S. Cohen, and C. Kanan, "Dvqa: Understanding data visualizations via question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '18, June 2018.

[18] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," in *International Conference on Machine Learning*, ser. ICML '23. PMLR, 2023, pp. 18 893–18 912.

[19] F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. Eisenschlos, "MatCha: Enhancing visual language pretraining with math reasoning and chart derendering," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 756–12 770. [Online]. Available: https://aclanthology.org/2023.acl-long.714/

[20] M. Zhou, Y. Fung, L. Chen, C. Thomas, H. Ji, and S.-F. Chang, "Enhanced chart understanding via visual language pre-training on plot table pairs," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1314–1326. [Online]. Available: https://aclanthology.org/2023.findings-acl.85/

[21] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen, "Charxiv: charting gaps in realistic chart understanding in multimodal llms," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

[22] H.-K. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim, and J. Seo, "Natural language dataset generation framework for visualizations powered by large language models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642943

[23] L. Hu, D. Wang, Y. Pan, J. Yu, Y. Shao, C. Feng, and L. Nie, "Novachart: A large-scale dataset towards chart understanding and generation of multimodal large language models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, ser. MM '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 3917–3925. [Online]. Available: https://doi.org/10.1145/3664647.3680790

[24] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty, "ChartInstruct: Instruction tuning for chart comprehension and reasoning," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 10 387–10 409. [Online]. Available: https://aclanthology.org/2024.findings-acl.619/

[25] V. Ventura, L. A. Kleybolte, and A. Zarcone, "Instruction-tuned QwenChart for chart question answering," in *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*, T. Ghosal, P. Mayr, A. Singh, A. Naik, G. Rehm, D. Freitag, D. Li, S. Schimmler, and A. De Waard, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 240–251. [Online]. Available: https://aclanthology.org/2025.sdp-1.22/

[26] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang, "Chartllama: A multimodal llm for chart understanding and generation," *arXiv preprint arXiv:2311.16483*, 2023.

[27] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao, and P. Luo, "ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 7775–7803. [Online]. Available: https://aclanthology.org/2024.findings-acl.463/

[28] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu, "MMC: Advancing multimodal chart understanding with large-scale instruction tuning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1287–1310. [Online]. Available: https://aclanthology.org/2024.naacl-long.70/

[29] Y. Yang, Z. Zhang, Y. Hou, Z. Li, G. Liu, A. Payani, Y.-S. Ting, and L. Zheng, "Effective training data synthesis for improving mllm chart understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, ser. ICCV '25, 2025.

[30] L. Zhang, A. Hu, H. Xu, M. Yan, Y. Xu, Q. Jin, J. Zhang, and F. Huang, "TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1882–1898. [Online]. Available: https://aclanthology.org/2024.emnlp-main.112/

[31] F. Wang, B. Wang, X. Shu, Z. Liu, Z. Shao, C. Liu, and S. Chen, "Chartinsighter: An approach for mitigating hallucination in time-series chart summary generation with a benchmark dataset," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 6, pp. 3733–3745, 2025.

[32] J. Chen, L. Kong, H. Wei, C. Liu, Z. Ge, L. Zhao, J. Sun, C. Han, and X. Zhang, "Onechart: Purify the chart structural extraction via one auxiliary token," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 147–155.

[33] W.-C. Fan, Y.-C. Chen, M. Liu, L. Yuan, and L. Sigal, "On pre-training of multimodal language models customized for chart understanding," *arXiv preprint arXiv:2407.14506*, 2024.

[34] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty, "ChartGemma: Visual instruction-tuning for chart reasoning in the wild," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert, K. Darwish, and A. Agarwal, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 625–643. [Online]. Available: https://aclanthology.org/2025.coling-industry.54/

[35] Z. Xu, B. Qu, Y. Qi, S. Du, C. Xu, C. Yuan, and J. Guo, "Chartmoe: Mixture of diversely aligned expert connector for chart understanding," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=o5TsWTUSeF

[36] L. Chen, X. Zhao, Z. Zeng, J. Huang, Y. Zhong, and L. Ma, "Chart-r1: Chain-of-thought supervision and reinforcement for advanced chart reasoner," *arXiv preprint arXiv:2507.15509*, 2025.

[37] M. Liu, D. Chen, Y. Li, G. Fang, and Y. Shen, "ChartThinker: A contextual chain-of-thought approach to optimized chart summarization," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 3057–3074. [Online]. Available: https://aclanthology.org/2024.lrec-main.273/

[38] H. Shao, S. Qian, H. Xiao, G. Song, Z. Zong, L. Wang, Y. Liu, and H. Li, "Visual cot: advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2025.

[39] M. Huang, L. Zhang, L. Han, W. Wu, X. Zhang, and J. Liu, "Vprochart: answering chart question through visual perception alignment agent and programmatic solution reasoning," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i4.32384

[40] B. Tang, A. Boggust, and A. Satyanarayan, "Vistext: A benchmark for semantically rich chart captioning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 7268–7298.

Fengling Zheng is currently a PhD in the School of Computer Science, Hangzhou Dianzi University, and a researcher in Big Data Visualization and Human Computer Collaborative Intelligent Laboratory. Her research interests include interactive visualization, visual analytics and knowledge graph.

Yongle Peng is currently a senior undergraduate student majoring in Digital Media Technology at the School of Humanities, Arts and Digital Media, Hangzhou Dianzi University. His research interests include multimodal large language models, computer vision, and visual analytics.

Zi Rong is currently a Master's student in the School of Humanities, Arts and Digital Media, Hangzhou Dianzi University, and a researcher in Big Data Visualization and Human Computer Collaborative Intelligent Laboratory. Her research interests include chart understanding, data visualization, and human–computer interaction.

Chenyun Cai is currently an undergraduate student in Digital Media Technology at Hangzhou Dianzi University. Concurrently, he is a Research Assistant at the Lab of Big Data Visualization and Human-Computer Collaborative Intelligence, where his research focuses on knowledge graphs and interactive visualization.

Yumeng He is a Master's student in Computer Science at the University of Southern California, advised by Dr. Jernej Barbič. Her research interests lie at the intersection of computer vision, computer graphics, and robotics, with a focus on image and video understanding, 3D and 3DGS generation, physics-based simulation, real-to-sim pipelines, and embodied policy learning.

Dekun Qian is currently a Master student in the School of Humanities, Arts and Digital Media, Hangzhou Dianzi University, and a researcher in Big Data Visualization and Human Computer Collaborative Intelligent Laboratory. His research interests include human-computer interaction and digital media technology.

Zhiguang Zhou is currently a professor in School of Humanities, Arts and Digital Media, and serves as the dean of Digital Media Technology Research Institute at Hangzhou Dianzi University. His research interests include data visualization, visual analytics and knowledge graph mining. He received his Ph.D. in Computer Science from the state key Laboratory of CAD&CG in Zhejiang University.

Yigang Wang received the MS and PhD degrees in applied mathematics from Zhejiang University, Hangzhou, China. He is currently a professor with the School of Media and Design, Hangzhou Dianzi University, Hangzhou, China. His interests include image processing, computer vision, pattern recognition, and computer graphics.

Yong Wang is currently an assistant professor in the College of Computing and Data Science, Nanyang Technological University. Before that, he worked as an assistant professor at Singapore Management University from 2020 to 2024. His research interests include information visualization, visual analytics and human-AI collaboration, with an emphasis on their application to FinTech, quantum computing and online learning. He obtained his Ph.D. in Computer Science from Hong Kong University of Science and Technology. He received his B.E. and M.E. from Harbin Institute of Technology and Huazhong University of Science and Technology, respectively. For more details, please refer to http://yong-wang.org.