Evaluating Task-based Effectiveness of MLLMs on Charts

Yifan Wu[1]*, **Lutao Yan**[2]*, **Yuyu Luo**[1], **Yunhai Wang**[3], **Nan Tang**[1]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]South China University of Technology
[3]Renmin University of China
{evanwu50020, lutaoyan666, cloudseawang}@gmail.com,
{yuyuluo, nantang}@hkust-gz.edu.cn,

Abstract

The ability to automatically interpret and extract insights from charts can support various applications, including assisting individuals with low vision in capturing insights from charts. With advancements in multimodal large language models (MLLMs), particularly GPT-4V, there is potential to meet this demand. However, existing evaluations mainly focus on *high-level* chart understanding tasks, such as chart captioning, which overlook the *low-level* data analysis tasks (e.g., characterize distribution) that humans encounter in daily life. In this paper, we explore a forward-thinking question: Is GPT-4V effective at low-level data analysis tasks on charts? To this end, we first curate a large-scale dataset, named ChartInsights, consisting of 89,388 quartets (chart, task, question, answer) and covering 10 widely-used low-level data analysis tasks on 7 chart types. Firstly, we conduct systematic evaluations to understand the capabilities and limitations of 18 advanced MLLMs , which include 12 open-source models and 6 closed-source models. Starting with a standard textual prompt approach, the average accuracy rate across the 18 MLLMs is 36.17%. Among all the models, GPT-4V achieves the highest accuracy, reaching 56.13%. To understand the limitations of multimodal large models in low-level data analysis tasks, we have designed various experiments to conduct an in-depth test of GPT-4V's capabilities. We further investigate how visual modifications to charts, such as altering visual elements (*e.g.* changing color schemes) and introducing perturbations (*e.g.* adding image noise), affect GPT-4V's performance. Secondly, we present 12 experimental findings. These findings suggest GPT-4V's potential to revolutionize interaction with charts and uncover the gap between human analytic needs and GPT-4V's capabilities. Thirdly, we propose a novel textual prompt strategy, named Chain-of-Charts, tailored for low-level analysis tasks, which boosts model performance by 24.36%, resulting in an accuracy of 80.49%. Furthermore, by incorporating a *visual prompt* strategy that directs GPT-4V's attention to question-relevant visual elements, we further improve accuracy to 83.83%. Our study not only sheds light on the capabilities and limitations of GPT-4V in low-level data analysis tasks but also offers valuable insights for future research.

*Equal Contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

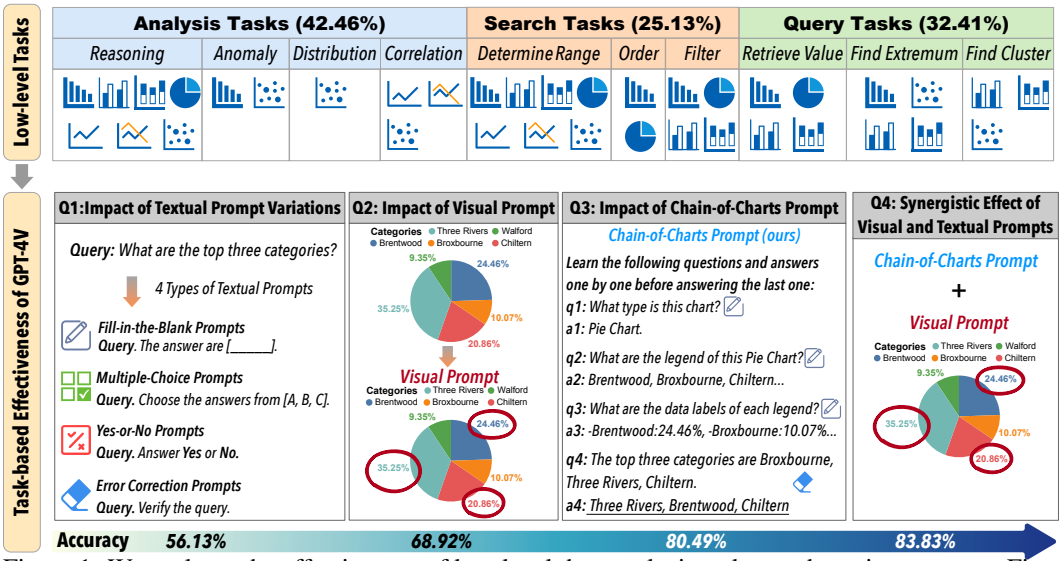


Figure 1: We evaluate the effectiveness of low-level data analysis tasks on charts in two steps. First, we develop a large-scale dataset tailored for 10 low-level analysis tasks, covering 7 widely-used chart types (top part of the figure). Second, we conduct four main experiments to study the capabilities of GPT-4V on low-level analysis tasks across different scenarios (bottom part of the figure).

Visualization can effectively convey data insights. However, due to the abundance of information a visualization may convey, it is not easy for users to accurately extract the desired information [1, 2]. Therefore, it becomes crucial to automatically help users pinpoint this information [3, 4, 5] based on their needs [6], a process known as chart question answering, or ChartQA for short.

High-Level and Low-Level ChartQA Tasks. ChartQA can generally be categorized into two types: **high-level** tasks and **low-level** tasks. High-level tasks typically refer to broader questions such as chart captioning, chart-to-text conversion, etc., while low-level tasks involve more specific inquiries, such as identifying correlations, detecting anomalies, and so forth [3, 6, 7].

Traditionally, ChartQA has been a challenging problem due to the limited capabilities in natural language understanding and the high complexity of chart reasoning. However, recent advancements in multi-modal large language models (**MLLMs**) have made it possible for users to interact with systems using natural language to extract specific information from data across various modalities. This progress has illuminated new possibilities for ChartQA on different levels of tasks.

Prior Art: High-Level Tasks using MLLMs. Recently, several studies [8, 9, 10, 11, 12, 13, 14] have explored the capabilities of MLLMs in performing high-level ChartQA tasks. The findings reveal that state-of-the-art MLLMs, such as GPT-4V, have demonstrated promising results in addressing high-level tasks. These studies have also outlined a range of intriguing future research directions in this field.

Our Focus: Low-Level Tasks using MLLMs. While the above efforts have concentrated on high-level ChartQA tasks, the effectiveness of MLLMs for low-level data analysis tasks remains underexplored. In this paper, we aim to systematically investigate the capabilities of GPT-4V in addressing 10 low-level data analysis tasks [3, 6]. Our study seeks to answer the following critical questions, shedding light on the potential of MLLMs in performing detailed, granular analyses.

- **Q1:** Impact of Textual Prompt Variations.What is the impact of different textual prompts on GPT-4V 's output accuracy? This question aims to assess the baseline performance and capabilities of GPT-4V in different low-level tasks.
- **Q2:** Impact of Visual Variations and Visual Prompts: How do different visual prompts, such as alterations in color schemes, layout configurations (e.g., aspect ratio), and image quality, affect the performance of GPT-4V in low-level tasks?

- **Q3:** Impact of Chain-of-Thoughts. Can we enhance basic textual prompts in Q1 with a chain-of-thoughts like approach?
- **Q4:** Synergistic Effect of Visual and Textual Prompts: Can the combination of visual and textual prompts lead to enhanced performance in low-level ChartQA tasks with GPT-4V? This question explores the potential for achieving better results by integrating both types of prompts.

Contributions. We make the following notable contribution.

(1) ChartInsights Dataset. We curate a large-scale dataset, ChartInsights for evaluating the low-level data analysis tasks on charts. This dataset features a diverse array of visual variants, visual and textual prompts, and comprehensive metadata, which can be used to investigate the performance of MLLMs in the low-level ChartQA task from different scenarios.

(2) Setting Benchmarks. Our study establishes benchmarks by evaluating GPT-4V, a state-of-the-art MLLM, on 10 low-level ChartQA tasks. This benchmarking effort provides valuable insights into the current capabilities of MLLMs in processing and analyzing chart information.

(3) New Experimental Findings. We conduct a thorough analysis of the experimental results and identify 12 key findings. These insights emphasize the critical role of visual prompts, chart elements, and image quality in successfully performing low-level ChartQA tasks.

(4) Chain-of-Charts. We introduce the *Chain-of-Charts* strategy, a new textual prompt designed to enhance the reasoning capabilities of GPT-4V in the context of ChartQA. The Chain-of-Charts leverages a series of interconnected question-answer pairs to guide the model.

We also share all our data and code to facilitate further research: `https://anonymous.4open.science/r/ChartInsights-D43E`

2 Related Work

2.1 Low-Level Analysis Tasks on Charts

Visualization charts offer numerous insights that aid users in performing data analysis tasks. Low-level data analysis tasks typically involve activities requiring direct interpretation and processing of specific visual elements within a chart, such as data retrieval, outlier identification, and correlation determination [3, 7, 15]. Amar et al. [3] identified ten low-level tasks, highlighting the real-world activities users undertake with visualization tools to understand their data. Subsequently, Saket et al. [6] evaluated the effectiveness of five basic charts across ten low-level analysis tasks using two datasets through a crowdsourced experiment. In this paper, we aim to evaluate how effectively GPT-4V can interpret charts by using these ten low-level data analysis tasks as a framework.

2.2 Multimodal Large Language Models

The field of Multimodal Large Language Models (MLLMs) is experiencing rapid advancements, with efforts concentrated on developing artificial intelligence systems capable of processing and producing multi-modal content, including text, images, videos, and more. Early research such as CLIP [16] demonstrated the effective combination of visual and linguistic information through contrastive learning, while subsequent work like DALL-E [17] further showcased the potential of Transformer [18] architecture in generating images that match text descriptions. Building on these foundational successes, the research community has ventured into refining these models for diverse multi-modal applications, employing strategies like fine-tuning and prompt-based learning. For example, VisualGPT [19] and BLIP [20] have been adapted for Visual Question Answering (VQA) tasks, significantly enhancing their multi-modal task performance. Concurrently, the development of various benchmarks [21, 22, 23, 24, 25], including MME [26], has been crucial. These benchmarks provide a wide array of tasks and datasets, facilitating a comprehensive evaluation of MLLMs' abilities across different contexts. In this paper, we try to harness the off-the-shelfMLLMs for low-level data analysis tasks on charts.

2.3 MLLMs for Chart Question Answering

With the advancements in MLLMs, such as GPT-4V, it becomes increasingly promising to automatically comprehend charts and extract insights according to user queries [1, 2, 4, 27]. This process is known as chart question answering, *i.e.* ChartQA for short. Recent research efforts have focused on understanding the capabilities of MLLMs in performing ChartQA tasks. These studies can be categorized into two groups: evaluation studies and the construction of datasets for ChartQA.

Evaluating MLLMs on ChartQA Tasks. Several recent studies [8, 9, 10, 11, 12, 13, 14] have attempted to leverage the capabilities of MLLMs to perform high-level ChartQA tasks such as chart captioning and chart-to-text. For example, Huang et al. evaluated the capabilities of representative MLLMs, such as GPT-4V and Bard (*i.e.* Gemini) [28], on chart captioning tasks. Their findings indicated that GPT-4V faces challenges in generating captions that accurately reflect the factual information presented in charts. Moreover, these studies have highlighted various promising directions for future research in this field. Diverging from the emphasis on high-level tasks in previous works, our research uniquely targets *low-level* ChartQA tasks [3, 6].

ChartQA Datasets. In the last decade, several ChartQA datasets have been presented [2, 8, 11, 12, 13, 29, 30, 31, 32], as shown in Section 3 Table 1. For example, ChartBench [11] includes 2.1K charts for four types of ChartQA tasks. However, a gap remains evident in the landscape of existing ChartQA datasets: none are tailored to comprehensively evaluate the 10 *low-level* tasks identified as critical to the ChartQA task. Moreover, to conduct more customized evaluations, such as modifying the visual elements or adding a visual prompt, we need access to the metadata (*e.g.* the underlying data) of the charts, not just the chart images. Therefore, we curate a large-scale dataset ChartInsights, which consists of a total of 89,388 quartets, each including a chart, a specified task, a corresponding query, and its answer.

3 ChartInsights for Low-level Analysis on Charts

In this section, we will first discuss the design goals for curating datasets for low-level tasks (Section 3.1). We will then provide details of constructing ChartInsights (Section 3.2). We will close this section by elaborating on the characteristics of ChartInsights (Section 3.3).

3.1 Design Goals

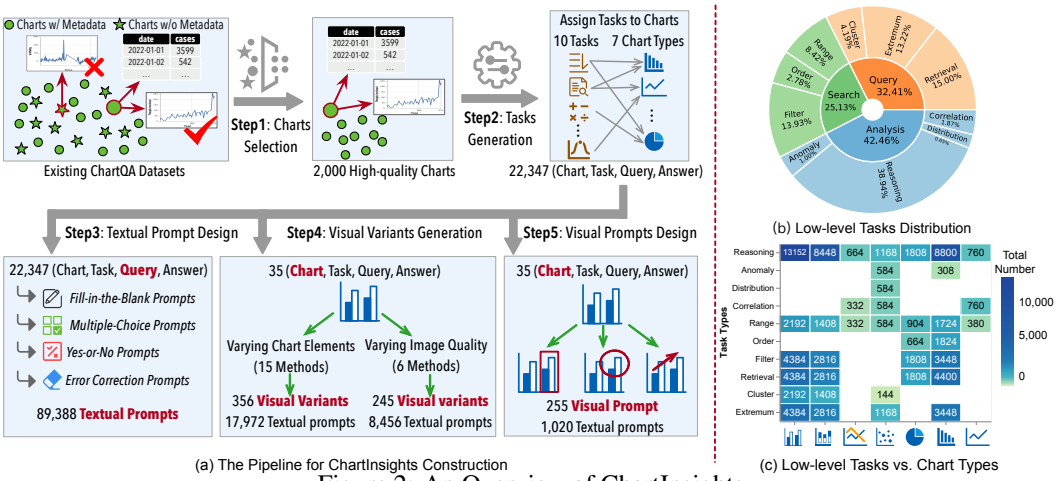
G1: Supporting Low-level Data Analysis Tasks. Our first goal is to facilitate the support of 10 low-level data analysis tasks [3, 6]. This focus addresses a critical gap in existing ChartQA datasets, which often overlook the granularity required to fully understand and interact with the data presented in charts.

G2: Evaluating Visual and Textual Variants on Charts. We highlight the critical role of visual variants (*e.g.* color, size, shape) in data visualizations, which are key to conveying and interpreting information effectively. Despite their importance, these variants are often neglected in existing ChartQA datasets and evaluations. Our goal is to address this issue by incorporating a diverse array of visual variants, including varying chart elements, image quality, and visual prompts. In addition, we also want to investigate the impact of different textual prompts on the low-level analysis task.

G3: Making Metadata Available. The third goal tackles the prevalent issue of inaccessible data and metadata in current ChartQA datasets. By offering comprehensive access to each chart's metadata, such as the source data, chart type, and visual element specifics (like color schemes and labels), our dataset enhances analytical depth into chart design's impact on ChartQA performance.

3.2 ChartInsights Construction

To fulfill our three design goals, our construction process begins with the collection of charts with metadata from existing datasets. Next, we meticulously assign specific low-level data analysis tasks to appropriate chart types. Lastly, we develop diverse textual prompt strategies, along with visual

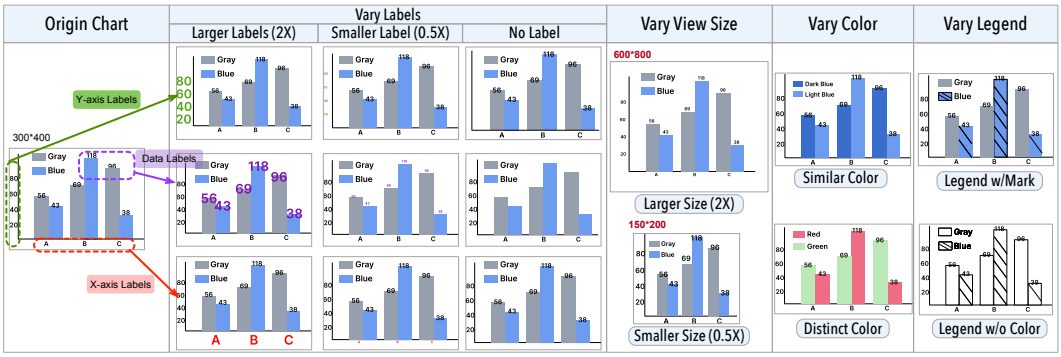


(a) The Pipeline for ChartInsights Construction

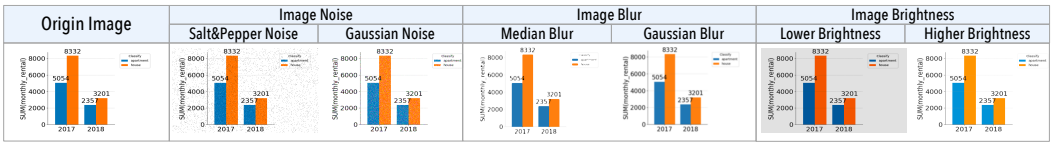
(b) Low-level Tasks Distribution

(c) Low-level Tasks vs. Chart Types

Figure 2: An Overview of ChartInsights



(a) Vary Chart Elements



(b) Vary Image Quality

Figure 3: Vary Visual Elements on Charts. (a) We vary chart labels, view size, color, and legend in a total of 15 ways. (b) We alter the image quality by adding noise, applying blur, and adjusting brightness.

variants and prompts, tailored to each chart. Note that we save all metadata during the construction process, which can make the users customize their dataset based on ChartInsights easily.

As shown in Figure 2-(a), the construction of our ChartInsights consists of five steps: Candidate Charts Selection, Low-Level Tasks Generation, Textual Prompts Design, Visual Variants Generation, and Visual Prompts Design.

Step 1: Candidate Charts Selection. In order to more comprehensively evaluate the ability of MLLMs on low-level data analysis tasks, and to conduct more detailed and extended experiments, the datasets (tabular data) and visualization charts we collected need to meet the following three requirements: First, these datasets should contain the original metadata of the chart such as the underlying data for rendering, allowing us to create customized reasoning tasks based on the metadata data. Second, the charts in these datasets should contain data labels, because the lack of data labels will greatly limit the types of low-level tasks. Third, these datasets should contain both simple and complex charts so that the difficulty of the charts is reasonable.

After considering the above three aspects and the characteristics of existing datasets, we chose to extract charts and corresponding metadata from the two existing datasets, namely ChartQA [2] and nvBench [33]. Then, we get a total of 2K high-quality charts as well as their metadata as our initial dataset. The initial dataset contains a total of 7 types of charts, namely stacked bar charts, grouped bar charts, basic bar charts, line charts, grouped line charts, scatterplots, and pie charts.

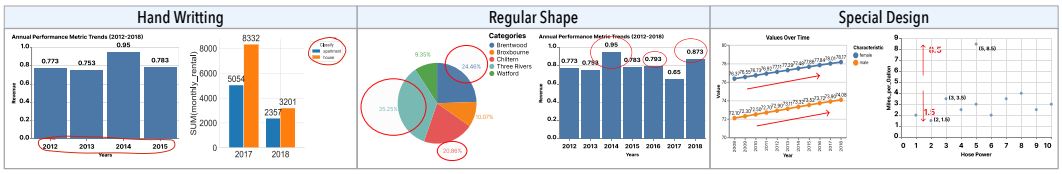


Figure 4: Three Types of Visual Prompts.

Step 2: Low-level Tasks Generation. Next, we design a set of low-level tasks for the collected charts. We follow the approach of previous works on designing low-level tasks for charts [3, 6, 7], resulting in 10 low-level tasks in this paper, as shown in top of Figure 1. We group the 10 low-level tasks into three categories, namely Analysis, Search, and Query, based on their purpose and required reasoning abilities [7].

Next, we should decide which tasks are applicable to which types of charts. We will follow the recommendations on the task-based effectiveness of humans to assign the tasks to each chart type [6]. Finally, we have 22,347 (chart, task, question, answer).

Step 3: Textual Prompts Design. In order to better explore the impact of different prompting methods on GPT-4V. We have designed 4 textual prompt methods, namely Fill-in-the-Blank, Multiple-choice, Yes-or-No, and Error Correction prompts. 1) For Fill-in-the-Blank prompt, we maintain the asking method of the initial question and set the answer format for Fill-in-the-Blank prompt; 2) For Multiple-choice prompt, we still maintain the asking method of the initial question, but at this time we will provide a list of choices for GPT-4V, which usually contains one correct answer and two wrong answers, and tells GPT-4V to choose the answer from the options; 3) For Yes-or-No prompt, we first change the initial question to a true or false question and tell GPT-4V whether it needs to be answered correctly or Wrong; and 4) For Error Correction prompt, we put the wrong answer into the original question with a certain probability and change it into a statement.

We expand the above 4 textual prompt variants on the 22,347 (chart, task, question, answer), and thus produce 89,388 (chart, task, question, answer) at the end.

Step 4: Visual Variants Generation. Visual variants (*e.g.* color, size, shape) of a chart play a key role in delivering insights, but these variants are often overlooked in existing ChartQA datasets and evaluations, and thus we aim to bridge this gap. To this end, we vary the chart elements and add image noise to vary the chart quality.

Step 4.1: Varying the Chart Elements. As shown in Figure 3(a), we change the visual elements of these charts from four aspects, namely labels, chart scale, element color, and legend. To achieve this, we sample 5 charts from each category of charts as seeds, resulting in 35 charts. For varying labels, we enlarge, reduce, and remove the x-axis, y-axis, and data labels, respectively. For varying view sizes, we enlarge and reduce the chart, respectively. For varying element color, we change the elements in the chart to the same color or a higher contrast color; For varying legend, we first add marks to different types of categories, and then delete the colors. Finally, we generate 356 visual variants for 35 charts. These 356 visual variants (charts) are associated with 17,972 textual prompts and cover 10 low-level tasks.

Step 4.2: Varying the Image Quality. We add image noise, apply image blur, and adjust the brightness to vary the chart image quality, as shown in Figure 3(b). To achieve this, we sample 5 charts from each category of charts as seeds, resulting in 35 charts. For adding image noise, we choose Gaussian noise and salt and pepper noise; For applying image blue, we use median blur and Gaussian blur; For adjusting image brightness, we choose to make the brightness of the chart higher and lower. Finally, we generate 245 visual variants for 35 charts. These 245 visual variants (charts) are associated with 8,456 textual prompts and cover 10 low-level tasks.

Step 5: Visual Prompts Design. Kong et al. [34] presented five types of graphical overlays to enhance users' capabilities in performing data analysis tasks such as extraction and comparison of numerical values. Intuitively, we want to verify whether overlays would have a positive impact on GPT-4V's performance. Therefore, we design three types of visual prompts (*i.e.* graphical overlays) for the charts.

Table 1: Comparison with Existing Datasets

Task Levels	Datasets	#-Task Types	#-Chart Types	#-Charts	#-Questions	#-Questions / #-Charts	Metadata Available?
High	DVQA [31]	3	1	300K	3.5M	11.7	✓
	ChartQA [2]	4	3	4.8K	9.6K	2	✓
Mix	FigureQA [30]	6	5	120K	1.5M	12.9	✗
	ChartLlama [12]	7	10	11K	160K	14.5	✗
	ChartBench [11]	4	11	2.1K	16.8K	8	✗
Low	**ChartInsights (ours)**	**10**	7	2K	89K	**44.5**	✓

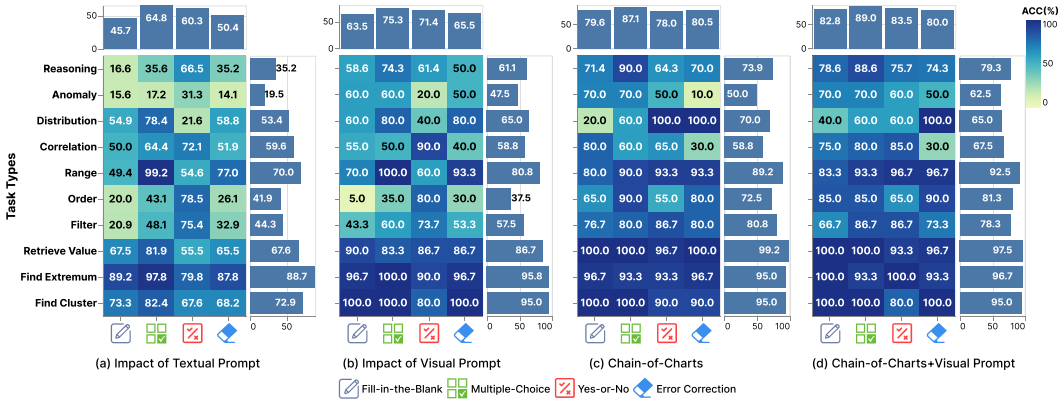


Figure 5: The Effectiveness of GPT-4V across 10 Low-level Tasks and 4 Textual Prompts

We consider three types of visual prompts, as shown in Figure 4. The first is to directly circle the content in the chart that is highly relevant to the question in handwriting, such as circling the values of the two elements mentioned in the reasoning question. The second method is regular shapes, which uses regular shapes (such as circles or rectangles) to label elements in the diagram. This makes it easier to use the size of a shape to imply the sequential relationship of elements. For example, use three circles of different sizes to correspond to the three values in the ordering task. The third way is special design. We design effective visual prompts tailored for different low-level tasks. For example, we use arrows to represent the monotonicity of the trend, for the correlation task. To generate the visual prompts, we first sample 35 charts from seven chart types, then apply various visual prompt strategies to them, resulting in 255 charts with different visual prompts. These 255 charts are associated with 1020 questions for 10 low-level tasks.

3.3 ChartInsights Characteristics

The ChartInsights dataset contains 89,388 (chart, task, query, answer) ChartQA samples across 7 chart types for 10 low-level data analysis tasks on charts. Figure 2(b) shows the proportion of 10 low-level tasks (3 task groups) in ChartInsights. Among them, the analysis task group is the largest, accounting for 42.46%. This task group examines the reasoning power of multi-modal large models on charts. Figure 2(c) shows the distribution of 10 low-level tasks and 7 chart types. Note that because we consider the effectiveness of different chart types for different tasks, for some tasks, we did not allocate chart types in **Step 2**.

Comparison with Existing Datasets. As shown in Table 1, our dataset differs from existing ChartQA datasets [2, 29, 30, 31] in three aspects. First, our dataset covers 10 low-level data analysis tasks across 7 chart types. Second, we have designed a variety of textual prompts, visual variants, and visual prompts for each chart type and task, enabling us to perform numerous ChartQA tasks from different perspectives to evaluate GPT-4V comprehensively. Third, we make all metadata relevant to the charts available, facilitating future research.

4 Experiments

In this section, we randomly selected 20% of the data from ChartInsights as the test set. The test set contains 17,552 (charts, tasks, questions, answers) samples, involving 400 charts, covering 7 chart types and 10 low-level tasks. The basic textual prompts in the test set is shown in 1 **Q1**. We first

use the test set to evaluate 18 advanced MLLMs and find that GPT-4V has the best performance among all models. The performance of each model on 10 tasks is shown in Table 2. On this basis, we want to better explore the limitations and boundaries of MLLMs in Low-Level Analysis tasks. Taking GPT-4V as a representative research object, in addition to studying the impact of text cue changes, we will also systematically evaluate the performance of GPT-4V on the low-level ChartQA task through three steps: **Q2**: Impact of visual Prompts, **Q3**: Impact of Chain-of-Charts, and **Q4**: Synergistic Effect of Visual and Textual Prompts.

The overall result of GPT-4V is shown in Figure 5, which comprises four heatmap subfigures, labeled (a), (b), (c), and (d), each corresponding to one of the four evaluation methods discussed previously. These heatmaps visualize the performance of GPT-4V on various low-level ChartQA tasks under different prompt conditions. The progression from subfigures (a) to (d) clearly indicates the incremental benefits of incorporating visual prompts, Chain-of-Charts prompts, and their combination, culminating in the most effective approach for improving GPT-4V's performance in low-level ChartQA tasks.

In the subsequent sections, we will delve into a detailed analysis of each method: Section4.1.2 focuses on the evaluation of GPT-4V using textual prompt variations alone; Section 4.2 explores the impact of visual prompt variations on GPT-4V's performance; Section 4.3 is dedicated to the evaluation of GPT-4V with Chain-of-Charts prompts; and Section 4.4 covers the combined approach, where both textual and visual prompts are integrated.

4.1 Textual Prompt Evaluation Models on ChartInsights

Experimental Settings As discussed in Section 3.3, we use 17,552 testing samples to evaluate MLLMs. Then, we analyze the answers of MLLMs, compare them with Ground Truth, and calculate the accuracy.

4.1.1 Overall Evaluation on Various MLLMs

In Table 2 and 3, we found that among the 18 models, the performances of closed-source models are far superior to those of open-source models, and the average accuracy rate of these 18 models is calculated to be 38.25%. Among them, VisCPM [35]has the worst performance at 26.19%, while GPT-4V has the best performance at 56.13%.In all models, GPT-4V achieves its best performance in seven out of 10 tasks. We believe this is because although some open-source models may outperform current cutting-edge proprietary models such as GPT-4V or Gemini-Pro [28] on certain specific tasks after being fine-tuned on particular datasets, on more general multimodal datasets, closed-source models like GPT-4V still possess strong generalization capabilities and show clear advantages in aspects such as logical reasoning.

Finding-1: *Closed-source models exhibit far superior generalization performance in low-level analysis tasks compared to open-source models.*

4.1.2 In-depth Evaluation of Textual Prompt Variations on GPT-4V

In this set of experiments, we aim to achieve two main goals: First, to benchmark GPT-4V's performance across 10 low-level ChartQA tasks. Second, to investigate the impact of 4 types of Textual Prompt strategies on GPT-4V.

Overall Results. Figure 6 displays GPT-4V's performance across a range of chart types and task categories, highlighting a notably low overall accuracy within the Analysis task category. Specifically, GPT-4V shows its worst performance on stacked bar charts, with a mere average accuracy of 19.8%. Conversely, its strongest performance is observed in the Query task category, particularly with scatter plots, where it achieves an impressive accuracy of 89.8%.

The disparity in GPT-4V's experimental results can be attributed primarily to the nature of the tasks within each category. The Analysis task category includes a range of data analysis tasks that require complex reasoning, calculations, determination of correlations, understanding of data distributions, and identification of anomalies. In contrast, the Query task category involves simpler tasks, such as acquiring specific data values, which are inherently less complex.

Table 2: Accuracy scores of 18 MLLMs across 10 tasks on ChartInsights

Models	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Open Source MLLMs											
VisCPM-Chat-v1.1 [35]	28.4	**46.1**	33.3	51.9	23.0	6.4	25.1	15.8	32.0	29.6	26.2
BLIP2 [36]	24.8	23.4	25.0	15.1	25.3	20.2	39.8	27.8	30.3	30.1	28.3
CogVLM-17B [37]	20.3	23.1	43.6	29.6	**37.7**	10.8	9.1	37.9	56.6	26.7	29.4
OmniLMM-12B [?]	24.7	19.9	27.0	34.9	35.7	28.3	30.0	33.0	39.9	33.1	31.1
LLaVA1.5 [38]	32.4	6.3	30.9	23.1	21.7	32.7	35.6	32.6	35.8	43.5	32.2
ChartAssistant [39]	24.6	27.7	35.8	28.1	30.5	22.5	14.7	39.4	63.0	26.4	32.4
MiniCPM-v2 [40]	19.5	55.1	33.3	56.5	24.9	16.7	36.3	37.9	52.4	32.0	33.0
mPLUG-Owl2 [41]	31.0	27.0	29.4	35.3	28.4	22.5	40.3	30.9	41.1	27.3	33.3
Qwen-VL-Chat [42]	27.8	36.3	**45.1**	**55.8**	33.8	20.0	28.7	31.3	50.2	27.1	33.4
ViP-LLaVA [43]	28.8	6.6	34.8	30.3	21.9	**35.8**	**40.4**	42.2	38.3	33.8	33.8
LLaVA-NEXT [44]	**30.6**	7.4	26.5	38.0	29.5	33.3	23.4	**53.5**	59.8	**52.3**	38.5
Sphinx-v2 [45]	30.0	28.9	37.8	36.1	25.8	23.5	36.7	49.7	**66.3**	45.3	**40.2**
Closed Source MLLMs											
Qwen-VL-Plus [46]	30.8	27.3	47.1	47.1	43.0	34.6	20.7	58.7	65.5	62.5	42.6
Gemini-Pro-Vision [28]	25.6	**30.1**	45.6	58.7	75.3	32.9	30.1	60.4	80.9	55.3	48.4
ChatGLM-4V [47]	34.1	28.9	39.2	42.3	55.5	18.9	43.4	58.1	69.3	71.4	48.4
Claude3-Haiku [48]	33.0	9.0	42.7	46.2	60.4	26.2	40.0	62.3	75.1	66.8	49.5
Qwen-VL-Max [46]	28.8	25.8	**62.3**	**63.0**	66.1	40.2	38.9	67.0	79.6	66.8	51.7
GPT-4V [49]	**35.2**	19.5	53.4	59.6	**70.0**	**41.9**	**44.3**	**67.6**	**88.7**	**72.9**	**56.1**

Table 3: Accuracy scores of 18 MLLMs across 7 chart and 4 question types on ChartInsights. FB: Fill-in-the-Blank Prompt; MC: Multiple-Choice Prompt; YN: Yes-or-No Prompt; EC: Error Correction Prompt.

Models	Chart Types							Question Types			
	Grouped Bar	Stacked Bar	Grouped Line	Basic Bar	Basic Line	Scatter Plot	Pie	FB	MC	YN	EC
Open Source MLLMs											
VisCPM-Chat-v1.1 [35]	24.8	21.5	24.4	30.3	25.3	**34.9**	20.4	7.4	39.8	49.3	8.3
BLIP2 [36]	32.2	31.1	24.7	34.2	15.0	18.2	8.7	3.1	46.9	57.7	5.5
CogVLM-17B [37]	26.2	24.2	**32.1**	40.1	35.3	30.2	19.6	27.6	49.7	31.3	9.2
OmniLMM-12B [?]	28.8	25.9	26.5	37.8	43.4	27.5	34.4	10.1	46.3	57.0	10.8
LLaVA1.5 [38]	31.3	30.5	22.6	35.7	40.6	26.8	36.3	9.6	41.4	67.9	9.9
ChartAssistant [39]	32.9	27.8	24.1	41.2	35.6	28.2	23.3	**30.2**	46.3	40.9	12.2
MiniCPM-v2 [40]	33.4	28.7	18.8	40.4	36.3	29.9	28.0	21.0	42.0	60.8	8.3
mPLUG-Owl2 [41]	32.4	31.8	29.2	37.0	38.4	29.2	34.3	13.2	49.9	54.6	15.5
Qwen-VL-Chat [42]	31.2	26.3	25.9	**39.3**	42.8	38.5	34.7	22.5	52.3	48.6	10.4
ViP-LLaVA [43]	32.2	31.2	16.1	39.8	34.4	28.1	39.1	7.9	40.1	**73.4**	13.7
LLaVA-NEXT [44]	37.1	33.7	18.8	50.4	**48.4**	27.4	37.1	23.1	42.2	63.2	**25.6**
Sphinx-v2 [45]	**39.4**	**35.7**	26.5	**51.2**	42.5	31.5	36.7	28.2	**56.1**	60.0	16.7
Closed Source MLLMs											
Qwen-VL-Plus [46]	36.5	35.8	26.2	57.0	32.8	44.2	42.5	32.9	59.2	54.0	24.2
Gemini-Pro-Vision [28]	45.0	43.3	35.1	57.7	43.8	47.3	51.2	40.9	55.4	43.9	53.3
ChatGLM-4V [47]	46.0	43.9	36.9	58.7	**56.3**	39.7	49.9	32.1	53.7	**72.2**	35.8
Claude3-Haiku [48]	50.7	48.1	32.1	55.6	46.9	40.4	48.3	39.4	51.2	61.8	45.9
Qwen-VL-Max [46]	47.1	46.3	38.4	66.4	43.1	47.7	48.1	**44.2**	**75.3**	48.1	39.0
GPT-4V [49]	**52.0**	**48.2**	**47.3**	**67.2**	49.4	**62.7**	**53.6**	44.1	64.4	66.4	**49.7**

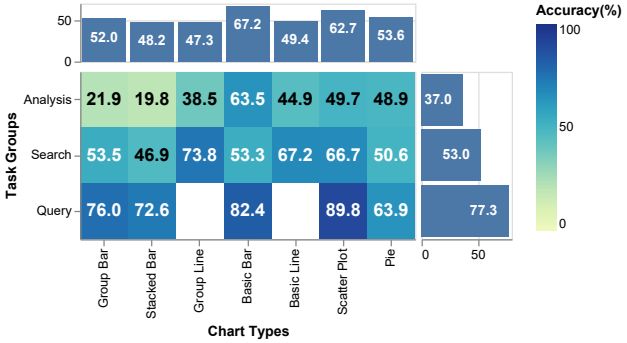


Figure 6: The Overall Accuracy: Task Groups vs. Chart Types

As depicted in the bar chart at the top of Figure 6, GPT-4V's accuracy in low-level ChartQA tasks reaches just over 60% for basic bar and scatter plots, while it hovers around 50% for similar tasks involving other chart types. Despite the dataset containing a significant number of reasoning tasks, these tasks are generally straightforward for humans. This performance gap suggests that, in the

Table 4: The Effectiveness of Textual Prompts vs. Ten Low-level Tasks

Textual Prompts	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Fill-in-the-Blank	16.59	15.62	54.9	50	49.35	20	20.85	67.47	89.24	73.3	44.05
Multiple-Choice	35.62	17.19	**78.43**	64.42	**99.22**	43.08	48.14	**81.89**	**97.85**	**82.39**	64.35
Yes-or-No	**66.46**	**31.25**	21.57	**72.12**	54.57	**78.46**	**75.42**	55.45	79.8	67.61	**66.39**
Error Correction	21.98	14.06	58.82	51.92	77.02	26.15	32.88	65.54	87.75	68.18	49.73
Overall Accuracy (%)	35.17	19.53	53.43	59.62	70.04	41.92	44.32	67.59	**88.66**	72.87	56.13

Table 5: The Average Accuracy of Different Chart Types vs.Ten Low-level Tasks ("–" means "N/A")

Chart Types	Analysis				Search			Query			Overall (%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
Grouped Bar	21.89	–	–	–	77.63	–	41.45	64.87	92.63	65	51.97
Stacked Bar	**19.79**	–	–	–	**54.41**	–	43.2	66.91	**74.08**	81.25	48.22
Grouped Line	22.02	–	–	71.43	73.81	–	–	–	–	–	47.32
Basic Bar	65.36	32.69	–	–	74.15	46.01	46.73	72.02	94.74	–	67.24
Basic Line	52.34	–	–	37.5	67.19	–	–	–	–	–	49.38
Scatter Plot	55.15	16.18	53.43	68.63	66.67	–	–	–	90.2	86.54	62.71
Pie	48.86	–	–	–	72.73	31.25	47.44	63.92	–	–	53.56

context of these tests, GPT-4V's current level of chart comprehension has not yet reached that of the average human.

Finding-2: *The overall accuracy of GPT-4V gradually decreases as the difficulty of the task increases, which is similar to human performance on chart understanding tasks, but GPT-4V cannot reach a level of chart understanding and analysis similar to that of ordinary humans.*

The Effectiveness of Textual Prompts. Table 4 presents the overall performance of GPT-4V across 10 low-level tasks with four textual prompts. Specifically, GPT-4V exhibits the highest overall accuracy with the Yes-or-No prompt, achieving 66.39%. In addition, it also performs well with the Multiple-Choice prompt, with 64.35% accuracy.

GPT-4V shows better performance on Multiple-Choice, Yes-or-No, and Error Correction prompts than the Fill-in-the-Blank prompt. The first three prompt types inherently offer candidate answers, allowing GPT-4V to select or judge, whereas Fill-in-the-Blank demands direct answer generation from GPT-4V.

Finding-3: *Structured textual prompts and candidate answers significantly enhance GPT-4V's ability to reason out correct responses.*

The Effectiveness of Chart Types. Table 5 shows the overall accuracy of GPT-4V for different chart types in 10 low-level tasks. Overall, GPT-4V has the highest performance on basic bar charts, reaching an average accuracy of 67.24%. The main reason is that the chart structure of the basic bar chart is relatively simple. Similarly, GPT-4V achieves better results on charts with simple structures such as scatter plots and pie charts. For charts with complex structures, such as stacked bar charts, grouped bar charts, and grouped line charts, the average accuracy of GPT-4V is lower than 50%. Specifically, GPT-4V achieved the worst performance in the three low-level tasks on the stacked bar chart, accounting for 50% (3/6).

Finding-4: *The complexity of the chart structures and visual elements will significantly affect the performance of GPT-4V on low-level tasks.*

4.2 The Impact of Visual Variations and Visual Prompts

Most ChartQA evaluations [8, 9, 10, 11, 12] focus on the impact of Textual Prompt, failing to consider the chart's quality and the Visual Prompts. Therefore, our research aims to explore how visual variations and visual prompts influence GPT-4V's performance.

4.2.1 Varying Chart Elements

A visualization chart can exhibit visual differences by varying its elements, as shown in Figure 3(a). Intuitively, these visual differences are likely to influence GPT-4V's performance on low-level

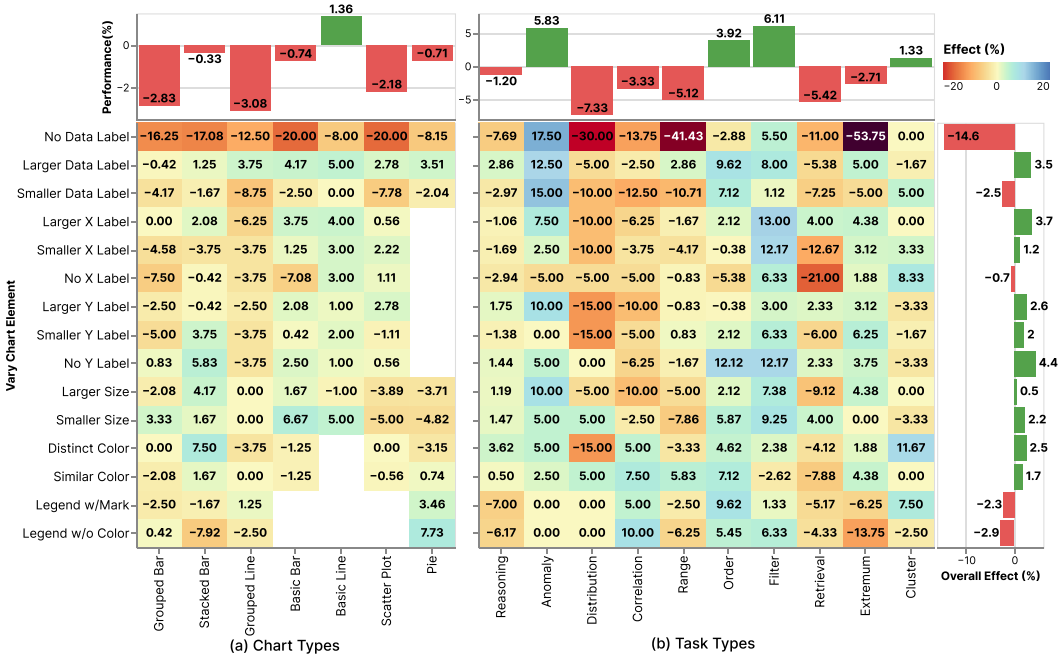


Figure 7: The Impact of Visual Variations. The heatmap shows how the performance of GPT-4V is affected by varying different chart elements.

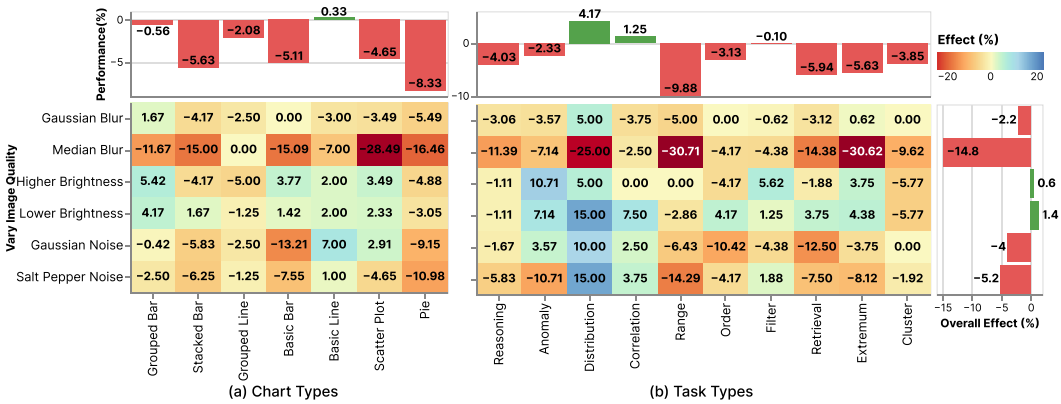


Figure 8: The Impact of Image Quality. The heatmap shows how the performance of GPT-4V is affected by varying the image quality.

ChartQA tasks. Thus, we explore how varying chart elements affect the performance of GPT-4V in this set of experiments.

Experimental Settings. The settings for varying chart elements is illustrated in Figure 3(a). As shown in Figure 2(a)**-Step 4**, we use the 356 visual variants for 35 charts as the testing samples. These 356 visual variants (charts) are associated with 17,972 textual prompts and cover 10 low-level tasks.

Overall Results. We calculate the overall accuracy of GPT-4V and compare it with the results in Tables 4 and 5 to record the change in performance. Figure 7 reports the experimental results.

Overall, the color of most areas of the heatmap in Figure 7(a) is light yellow, indicating that most chart variants have a slight negative impact on the performance of GPT-4V. *In settings without data labels*, GPT-4V's performance declines significantly, particularly across seven types of charts (Figure 7(a)). This outcome is understandable, as data labels facilitate GPT-4V's comprehension of the underlying insights conveyed by the charts. However, an interesting pattern emerges: as shown in Figure 7(b),

GPT-4V shows a performance improvement of 17.5% in anomaly detection and 5.5% in filtering tasks when data labels are absent. This suggests that data labels might, in some instances, hinder GPT-4V's ability to identify anomalies and filter values effectively.

As depicted in the heatmap of Figure 7(b), it is evident that several chart variants, such as larger x/y/data labels, positively affect GPT-4V's performance in tasks like anomaly detection, filtering, ordering, and clustering. These tasks inherently involve comparisons between elements. We hypothesize that, in many cases, alterations in chart elements may shift GPT-4V's focus towards visual comparisons rather than numerical ones, thereby enhancing its performance in these tasks.

The bar chart on the right in Figure 7 depicts the varied impacts of 15 chart variants on GPT-4V's performance. We believe that data labels play a crucial role in GPT-4V's low-level data analysis capabilities, as removing them or reducing their size tends to diminish its effectiveness. Moreover, adding marks to the legend or eliminating the legend's color negatively affects GPT-4V by introducing visual clutter and removing essential visual cues, respectively.

Finding-5: *While most chart variants slightly hinder GPT-4V's performance, especially in the absence of data labels, certain modifications like larger labels and the removal of data labels can actually improve its performance in tasks like anomaly detection and filtering, by shifting its focus to visual comparisons.*

4.2.2 Varying Image Quality

In addition to how visual modifications to chart elements alter the order and manner in which we interpret charts, the quality of chart images also plays a critical role in humans' comprehension of these visual representations. Thus, it raises a compelling question: do these factors, known to have varying degrees of negative impact on human understanding, similarly impede GPT-4V's ability to decipher charts?

Experimental Settings. The settings for varying chart elements is illustrated in Figure 3(b). We introduce six types of noise to evaluate the robustness and reliability of GPT-4V in low-level ChartQA tasks. As shown in Figure 2(a)**-Step 4**, we use the 245 visual variants for 35 charts as the testing samples. These 245 visual variants (charts) are associated with 8,456 textual prompts and cover 10 low-level tasks.

Overall Results. We calculate the overall accuracy of GPT-4V and compare it with the results in Tables 4 and 5 to record the change in performance. Figure 8 reports the experimental results. Generally, six methods of degrading image quality tend to negatively impact GPT-4V across a broad range of tasks and chart types. Among these, Median Blur stands out as the most detrimental, causing an average performance decline of 14.8%. We consider that median blurring makes numerical labels unreadable, resulting in a significant decrease in the performance of tasks directly related to numerical values.

Interestingly, adjustments to brightness—both increasing and decreasing—show a positive effect on the majority of tasks, with an average improvement of 0.6% and 1.4%, respectively.

As shown in Figure 8(b), the distribution task presents a unique case; it is adversely affected solely by Median Blur, whereas other forms of image quality manipulation tend to improve its performance.

Finding-6: *The readability of the chart after median blurring is very poor and difficult to fully understand even for humans. Thus, it's reasonable to assume that GPT-4V likewise faces difficulties in processing charts with compromised readability.*

4.2.3 The Impact of Visual Prompts

In the field of computer vision, many studies aim to enhance models' semantic extraction and entity recognition in images through the use of visual prompts [50]. Unlike general VQA (visual question answering) tasks, the ChartQA task, especially for our low-level data analysis tasks, demands greater sensitivity to detail and higher accuracy from the model. Therefore, we design different visual prompts for different low-level tasks to help GPT-4V better adapt to the requirements of different low-level tasks.

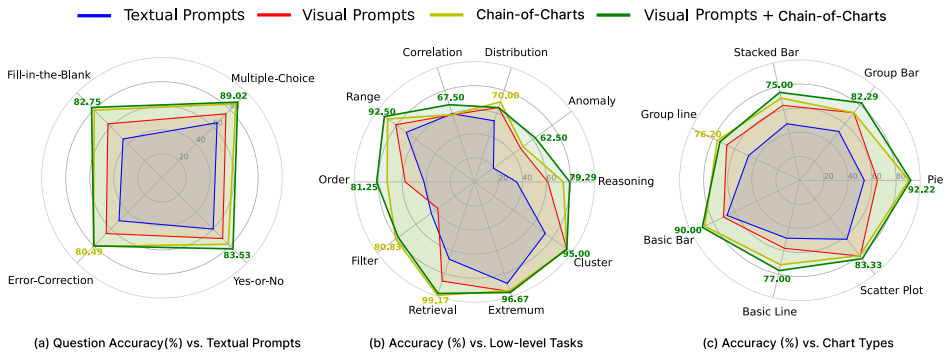


Figure 9: Comparing Different Prompting Methods.

Experimental Settings. We design three types of visual prompts, namely, hand writing, regular shape, and special design, as shown in Figure 4. As shown in Figure 2(a)**-Step 5**, we generate the 255 visual prompts for 35 charts as the testing samples. These 255 visual prompts are associated with 1,024 textual prompts and cover 10 low-level tasks.

Overall Results. Figure 5(b) reports the experimental results. We can see that GPT-4V equipped with visual prompts shows strong performance across 10 distinct tasks and 4 textual prompts, demonstrating its effectiveness. As shown in Figure 9(a)-(c), we also show the performance of visual prompts under different textual prompts, low-level ChartQA tasks, and chart types. Generally, visual prompts have been shown to enhance GPT-4V's performance. Notably, as depicted in Figure 9(b), GPT-4V equipped with visual prompts demonstrates great improvements in *Reasoning* and *Anomaly Detection* tasks. This indicates that visual prompts enable the model to accurately capture relevant data for analysis and reasoning.

Finding-7: *Visual prompts significantly improve GPT-4V's performance in various scenarios, such as textual prompts, low-level tasks, and chart types, highlighting the utility of visual information in aiding comprehension and reasoning.*

However, GPT-4V does not exhibit significant benefits from visual prompts in *Correlation* and *Order* tasks. These tasks often challenge GPT-4V to discern complex relationships among more than three distinct elements. In such scenarios, visual prompts may lose their specificity and lead to confusion due to the introduction of multiple new visual elements, especially in tasks like Order, where the added visual information can be misleading.

Finding-8: *Adopting dynamic visual prompt strategies tailored to specific task types is essential for optimizing performance and mitigating any potential negative impacts.*

4.3 The Impact of Chain-of-Charts

The Chain-of-Thought (CoT) prompt strategy has proven effective across various scenarios [51]. The key idea of CoT is to guide the model towards producing outputs that are more coherent and logical, by mimicking the step-by-step reasoning process humans employ to solve problems. Recently, Xu et al. [11] implemented the CoT strategy for ChartQA tasks, namely ChartCoT. The key idea is to pose a series of questions to progressively guide the model in comprehending the chart's details before it formulates an answer. However, ChartCoT struggles to ensure the accuracy of GPT-4V's responses to guiding questions, particularly with complex charts.

Chain-of-Charts Prompts. Therefore, we introduce a novel prompting strategy, termed Chain-of-Charts, which builds on the chain-of-thought approach, as shown in Figure 1-**Q3**. The core of Chain-of-Charts lies in orchestrating a sequence of questions and their corresponding answers ((q_1, a_1), (q_2, a_2),...(q_m, a_m)) to progressively guide the model towards a deeper understanding of the chart's details, thereby enhancing its ability to formulate accurate responses.

Experimental Settings. The testing samples for this set of experiments are the same as those used for evaluating visual prompts. In addition to Chain-of-Charts, we also evaluate two useful textual prompt strategies: Role-Play [52] and Tutorial.

Table 6: The Synergistic Effect of Visual and Textual Prompts (Overall Accuracy (%))

Prompts	Analysis				Search			Query			Overall(%)
	Reasoning	Anomaly	Distribution	Correlation	Range	Order	Filter	Retrieval	Extremum	Cluster	
(a) The Effectiveness of Textual Prompts											
Basic Textual Prompts	41.8	47.5	50.0	62.5	72.5	28.7	50.0	70.0	**96.7**	90.0	59.0
Role-Play	58.6	45.0	**75.0**	53.8	78.3	28.7	52.5	79.2	91.7	82.5	64.5
Tutorial	40.7	52.5	40.0	63.7	84.2	36.2	55.0	73.3	94.2	80.0	61.1
ChartCoT	48.2	**65.0**	60.0	**78.8**	83.3	26.2	80.0	72.5	94.2	90.0	67.6
Chain-of-Charts (ours)	**73.9**	50.0	70.0	58.8	**89.2**	**72.5**	**80.8**	**99.2**	95.0	**95.0**	**80.5**
(b) Synergistic Effect of Visual and Textual Prompts											
Basic Textual Prompts	61.1	47.5	65.0	58.8	80.8	37.5	57.5	86.7	95.8	95.0	68.9
Role-Play	65.7	57.5	55.0	71.3	89.2	38.8	67.5	86.7	**98.3**	90.0	73.7
Tutorial	71.4	**67.5**	65.0	**85.0**	86.7	35.0	**81.7**	88.3	95.8	92.5	78.0
ChartCoT	72.1	45.0	45.0	55.0	77.5	30.0	55.8	87.5	94.2	77.5	69.2
Chain-of-Charts (ours)	**79.3**	62.5	**65.0**	67.5	**92.5**	**81.3**	78.3	**97.5**	96.7	**95.0**	**83.8**

Shanahan et al. [52] suggested that giving large models specific roles could enhance their performance on particular tasks Inspired by this, we assigned GPT-4V the role of a visualization expert to determine whether this specialization could improve its performance. Specifically, we craft a textual prompt: "`You are an expert in ChartQA tasks with specialized skills in numerical analysis...`".

Our observations also indicate that the more detailed the prompt, the more precise and accurate GPT-4V's responses tend to be. Therefore, we created a detailed ChartQA tutorial for GPT-4V, dubbed the "Tutorial" prompt. We structure this textual prompt in a step-by-step tutoring format: "`Firstly, ... Subsequently... Next, ... Finally, summarize the information gathered.`"

Overall Results. Figure 5(c) reports the performance of GPT-4V. In comparison to Figure 5(a), it is clear that the Chain-of-Charts strategy has significantly improved GPT-4V's capabilities across 10 different tasks and four basic textual prompts.

Finding-9: *The Chain-of-Charts prompt have comprehensively improved the use of basic textual prompts in different tasks and chart types, demonstrating its effectiveness.*

Table 6(a) shows the overall accuracy of GPT-4V on 10 low-level tasks under five prompt strategies. Overall, Chain-of-Charts leads in average accuracy across all tasks with 80.49%, outperforming ChartCoT's accuracy of 67.55% by 12.94%. Specifically, Chain-of-Charts achieves the highest accuracy on five tasks including Reasoning, Determine Range, Order, Filter, and Find Cluster, with accuracy of 73.9%, 89.2%, 72.5%, 80.8%, and 95%. This method shows significant improvements, especially in Reasoning and Order tasks, where its accuracy surpasses other methods substantially: 73.9% in Reasoning, where others do not exceed 60%, and 72.5% in Order, where the rest fall below 40%. These tasks demand precise reasoning from GPT-4V, based on accurate identification of element coordinates and values. The Chain-of-Charts prompt framework effectively provides GPT-4V with the correct value and coordinate references, significantly aiding in the accurate positioning of different elements.

Finding-10: *The Chain-of-Charts prompt supplies GPT-4V with accurate chart reference information, thereby enhancing the model's comprehension and detailed reasoning of chart structures and elements.*

4.4 The Synergistic Effect of Visual and Textual Prompts

Yet, their effectiveness wanes in complex reasoning tasks such as reasoning, anomaly detection, and sorting, as shown in Figure 5(b). Conversely, the Chain-of-Charts Prompt excels in these reasoning tasks but is less effective in the Correlation task, as indicated in Figure 5(c). This observation leads us to wonder: Could combining visual prompts with Chain-of-Charts enhance GPT-4V's performance across a spectrum of low-level ChartQA tasks?

Experimental Settings. The test samples are the same as those used for evaluating visual prompts and Chain-of-Charts.

Overall Results. Figure 5(d) shows GPT-4V's accuracy following the integration of Chain-of-Charts and Visual Prompt, demonstrating a clear enhancement over the outcomes depicted in Figures 5(a), (b), and (c), which demonstrates the combined strategy's effectiveness.

Table 6 reports the performance improvements in GPT-4V after integrating various textual prompts with visual prompts. Specifically, the Chain-of-Charts with visual prompt reaches an accuracy of 83.82%, outperforming the ChartCoT with visual prompt by 14.6% and exceeding the accuracy of using Chain-of-Charts alone by 3.33%. Furthermore, this combination attained the highest accuracy in six tasks.

Finding-11: *Combining visual prompts with the Chain-of-Charts strategy significantly improves the performance, suggesting that integrating multiple types of prompts can leverage their respective strengths.*

Limitations about Visual and Textual Prompts. As depicted in Figure 5(d), the integration of Chain-of-Charts and visual prompts enables GPT-4V to outperform other settings. However, the improvement over using Chain-of-Charts alone is slight. We discuss the possible reasons behind this observation.

First, after carefully analyzing the experimental results, we discover that GPT-4V exhibits a certain degree of hallucination in chart understanding. For example, even if the calculation process is accurate, GPT-4V may provide answers that do not match any of the multiple-choice options, leading to incorrect results. This indicates that the model's accuracy is significantly affected by hallucination. Moreover, we also observe that GPT-4V might output numerical information unrelated to the chart even when explicitly recognizing values, further evidencing the hallucination phenomenon in chart reading.

Finding-12: *The effectiveness of using prompts to improve model performance is inherently constrained by hallucination.*

Second, there is a lack of research on developing systematic and standardized visual prompts specifically designed for enhancing ChartQA with MLLMs. In our experiment, we design visual prompts based on graphical overlay strategies [34]. We believe that crafting visual prompts that align more precisely with the model's inherent mechanisms for chart comprehension could result in significant performance improvements.

Finding-13: *Developing visual prompts tailored specifically to the characteristics of ChartQA task is a promising research direction.*

5 Lessons Learned

Effectiveness of Textual Prompts, Visual Prompts and Their Combination. Incorporating various prompt strategies, including textual and visual prompts, significantly impacts GPT-4V's accuracy. Textual prompts with structured candidate answers enhance reasoning capabilities, while visual prompts enhance the chart understanding through visual attention, particularly in anomaly detection and filtering tasks.

Importance of Chart Elements and Image Quality. Alterations in chart elements and the quality of chart images influence GPT-4V's performance. Specifically, certain modifications like larger labels or the absence of data labels can improve the model's efficiency in specific tasks by focusing its attention on visual comparisons. However, image quality degradation, especially median blurring, negatively affects the model's ability to process numerical values accurately.

GPT-4V's Strengths and Weaknesses in Low-level ChartQA Tasks. GPT-4V performs well in tasks requiring direct data retrieval and basic comparisons, showing high accuracy in Query and Search task categories. However, it faces challenges in more complex reasoning, anomaly detection, and correlation tasks, indicating a need for further optimization of prompting strategies and model training to overcome these limitations.

Potential for Future Application and Development. The experiments demonstrate a promising direction for enhancing MLLMs' performance in visual data analysis through the development of specialized prompting strategies and the careful manipulation of visual elements.

6 Limitations and Future Work

Limited Chart Types. In our experiments, we set benchmarks for the performance of GPT-4V across seven widely used chart types, providing valuable insights into the model's capabilities in chart interpretation. However, this focus inherently excludes a range of more complex chart types, such as heatmaps, radar charts, and others, which present unique analytical challenges and opportunities for data representation.

Therefore, including a more diverse chart type, especially those with complex structure and interpretation such as heat maps and radar charts, will provide a more comprehensive perspective on ChartQA for MLLM. This extension is critical for assessing the adaptability and effectiveness of MLLM in a wider range of graph interpretation tasks.

Limited Visual Prompts Design Space. Our exploration into the effectiveness of visual prompts in facilitating ChartQA tasks with GPT-4V has shown their potential to enhance model performance. Nevertheless, our investigation into the design space of visual prompts has been preliminary, lacking a comprehensive and systematic exploration of the full spectrum of visual prompt possibilities. This limitation narrows the scope of our findings and potentially overlooks more effective visual prompt strategies that could further improve the accuracy and efficiency of MLLMs in interpreting and analyzing charts.

Future research can systematically explore the design space of visual prompts tailored to various ChartQA tasks and chart types. An interesting direction is developing algorithms that can automatically generate visual prompts from given textual prompts, specific to ChartQA tasks and chart types. This would ensure that the prompts are accurately customized to improve both model interpretability and task performance.

Lacking of Considering the Data Prompts. Our approach primarily relied on chart images, neglecting the underlying data that generated these charts. This omission could hinder the model's ability to perform more complex analysis and reasoning based on the actual data points. Future work could explore integrating the underlying data as part of the prompt, potentially through multimodal inputs, to provide a richer context for the model's analyses.

Without Fine-tuning MLLMs. We only use the "off-the-shelf" GPT-4V to conduct evaluation, without considering other MLLMs because GPT-4V is known as one of the best models in the visual question-answering task. In addition, we don't perform task-specific fine-tuning because we want to benchmark GPT-4V in low-level tasks and investigate the impact of textual and visual prompts, which is orthogonal to fine-tuning the MLLMs. Future work can fine-tune MLLMs using our dataset to investigate their effectiveness.

Therefore, a promising direction is to develop a framework that includes self-correction [53], debugging, or a multi-agent approach [54]—with specialized agents for data analysis, chart understanding, and textual reasoning—that could enhance the model's accuracy and reliability in ChartQA tasks.

7 Conclusion

In this paper, we conduct a thorough evaluation to assess the performance of GPT-4V across 10 low-level ChartQA tasks. Our approach begins with the development of ChartInsights, a large-scale dataset designed to test the capabilities of MLLMs in handling various low-level ChartQA tasks across seven commonly used chart types. We benchmark GPT-4V's performance against a range of textual prompts, chart types, and specific ChartQA tasks. We further explore the impact of visual variants and visual prompts on performance, demonstrating the importance of chart quality and visual attention. To further improve the reasoning abilities of GPT-4V, we present a novel Chain-of-Charts prompt to guide the model with a series of interconnected question-answer pairs. We further enhance the effectiveness of Chain-of-Charts by integrating the visual prompts, which achieves the

best performance in the evaluation. While our findings suggest that GPT-4V has the potential to revolutionize how we interact with and interpret charts, there remains a considerable gap to bridge to meet human analytical needs fully. Future work can investigate how to incorporate the data prompts and multi-agent frameworks to further enhance the capabilities of MLLMs' effectiveness in diverse ChartQA tasks.

c

References

[1] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. Towards natural language interfaces for data visualization: A survey. IEEE Trans. Vis. Comput. Graph., 29(6):3121–3144, 2023.

[2] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In ACL (Findings), pages 2263–2279. Association for Computational Linguistics, 2022.

[3] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pages 111–117. IEEE, 2005.

[4] Zehua Zeng and Leilani Battle. A review and collation of graphical perception knowledge for visualization recommendation. In CHI, pages 820:1–820:16. ACM, 2023.

[5] Yilin Ye, Jianing Hao, Yihan Hou, Zhan Wang, Shishi Xiao, Yuyu Luo, and Wei Zeng. Generative ai for visualization: State of the art and future directions, 2024.

[6] Bahador Saket, Alex Endert, and Çagatay Demiralp. Task-based effectiveness of basic visualizations. IEEE Trans. Vis. Comput. Graph., 25(7):2505–2512, 2019.

[7] Tamara Munzner. Visualization analysis and design. CRC press, 2014.

[8] Zhi-Qi Cheng, Qi Dai, and Alexander G. Hauptmann. Chartreader: A unified framework for chart derendering and comprehension without heuristic rules. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 22202–22213, October 2023.

[9] Mingyang Zhou, Yi R. Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. Enhanced chart understanding in vision and language task via cross-modal pre-training on plot table pairs, 2023.

[10] Kung-Hsiang Huang, Mingyang Zhou, Hou Pong Chan, Yi R. Fung, Zhenhailong Wang, Lingyu Zhang, Shih-Fu Chang, and Heng Ji. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning, 2023.

[11] Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. arXiv preprint arXiv:2312.15915, 2023.

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

[13] Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. arXiv preprint arXiv:2402.12185, 2024.

[14] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning, 2023.

[15] Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. Exploring chart question answering for blind and low vision users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2023.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[17] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[19] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning, 2022.

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[21] Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception, 2024.

[22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023.

[23] Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S. Yu, and Zhijiang Guo. Do large language models know about facts?, 2023.

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024.

[25] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023.

[26] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[27] Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models, 2024.

[28] Google gemini. `https://gemini.google.com/app`. Accessed: 2024-03-23.

[29] Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), March 2020.

[30] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018.

[31] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[32] Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning, 2023.

[33] Yuyu Luo, Jiawei Tang, and Guoliang Li. nvbench: A large-scale synthesized dataset for cross-domain natural language to visualization task. arXiv preprint arXiv:2112.12926, 2021.

[34] Nicholas Kong and Maneesh Agrawala. Graphical overlays: Using layered elements to aid chart reading. IEEE Trans. Vis. Comput. Graph., 18(12):2631–2638, 2012.

[35] Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. Large multilingual models pivot zero-shot multimodal learning across languages. arXiv preprint arXiv:2308.12038, 2023.

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023.

[37] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.

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

[40] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024.

[41] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023.

[42] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

[43] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. arXiv preprint arXiv:2312.00784, 2023.

[44] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[45] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023.

[46] Alibaba qwen-vl. https://cn.aliyun.com/.

[47] zhipu chatglm-4v. https://open.bigmodel.cn/.

[48] Anthropic claude3. https://www.anthropic.com/.

[49] OpenAI. GPT-4V(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

[50] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In European Conference on Computer Vision, pages 709–727. Springer, 2022.

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS, 2022.

[52] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. Nat., 623(7987):493–498, 2023.

[53] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. arXiv preprint arXiv:2308.03188, 2023.

[54] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155, 2023.