ChartLens: Fine-grained Visual Attribution in Charts

Manan Suri [UMD], **Puneet Mathur** [Adobe] *, **Nedim Lipka** [Adobe],
Franck Dernoncourt [Adobe], **Ryan A. Rossi** [Adobe], **Dinesh Manocha** [UMD]
[UMD] University of Maryland, College Park [Adobe] Adobe Research
manans@umd.edu, puneetm@adobe.com

Abstract

The growing capabilities of multimodal large language models (MLLMs) have advanced tasks like chart understanding. However, these models often suffer from hallucinations, where generated text sequences conflict with the provided visual data. To address this, we introduce Post-Hoc Visual Attribution for Charts, which identifies fine-grained chart elements that validate a given chart-associated response. We propose ChartLens, a novel chart attribution algorithm that uses segmentation-based techniques to identify chart objects and employs set-of-marks prompting with MLLMs for fine-grained visual attribution. Additionally, we present ChartVA-Eval, a benchmark with synthetic and real-world charts from diverse domains like finance, policy, and economics, featuring fine-grained attribution annotations. Our evaluations show that ChartLens improves fine-grained attributions by 26-66%.[1]

1 Introduction

Rapid advancements in large language models (LLMs) have revolutionized various natural language processing tasks, including understanding, generation, and reasoning (Huang and Chang, 2022; Yang et al., 2024). Building on this foundation, multimodal large language models (MLLMs), have extended these capabilities to encompass multimodal tasks like image captioning and visual question answering. However, a critical challenge faced by these models is the prevalence of hallucinations—instances where the model generates content that appears plausible but is factually incorrect or inconsistent (Zhang et al., 2023; Ye et al., 2023a; Rawte et al., 2023). In MLLMs, this issue is particularly pronounced as cross-modal inconsistencies can emerge, where generated text fails to align with provided visual inputs (Huang et al., 2024; Guan et al., 2024). To address this, attribution has emerged as a promising strategy for text-based systems, allowing models to reference external sources, thereby enhancing factual reliability. In the context of LLMs, attribution refers to the ability of a model to provide verifiable evidence, such as references or citations, that supports its generated outputs, thereby enhancing factual reliability and trustworthiness (Li et al., 2023a). Techniques such as direct generated attribution (Peskoff and Stewart, 2023; Sun et al., 2022), post-retrieval answering (Ye et al., 2023b; Li et al., 2023b), and post-hoc attribution (Huo et al., 2023; Chen et al., 2023) aim to mitigate hallucination by enabling users to trace responses back to their origins. For resolving visual hallucinations specifically, post-generation validation approaches like (Zhou et al., 2023; Lee et al., 2023b; Yin et al., 2023) have proven effective in aligning textual outputs with visual evidence, ensuring cross-modal consistency and improving the overall trustworthiness of MLLMs.

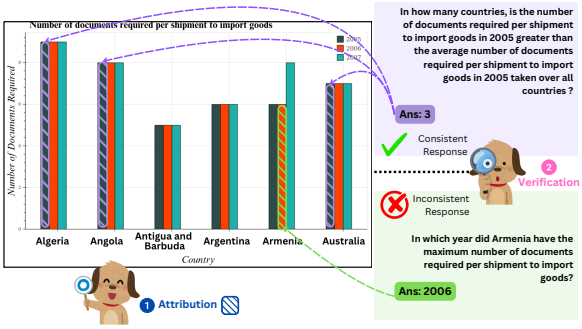


Figure 1: We introduce the task of visual attribution for charts (❶), which grounds textual responses to specific regions in the chart image. This promotes reliable understanding by enabling users to verify claims (❷), thus detect potentially hallucinated responses and identifying chart-response misalignments.

Charts, which are graphical representations of data, play a pivotal role in communicating complex insights across diverse domains, from business ana-

*Primary Research Mentor
[1]Code and data

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22447–22462

lytics to scientific research (Embarak and Embarak, 2018). As LLMs and multimodal large language models (MLLMs) increasingly handle chart-related tasks—such as chart question answering (Kafle et al., 2018; Masry et al., 2022; Kahou et al., 2017; Methani et al., 2020), captioning (Kantharaj et al., 2022; Tang et al., 2023; Hsu et al., 2021), and chart-to-table (Liu et al., 2022a) conversion—the need for robust validation mechanisms becomes paramount. Unlike textual information, charts encapsulate measurable and exact quantities and often represent intricate relationships like trends, proportions, and comparisons (Healy, 2024). Accurately interpreting these relationships requires more than superficial analysis; it demands understanding context-dependent factors like chart type, data encoding, and the layout of visual components such as axes, legends, colors, and shapes. The attribution challenge is further compounded by the need to disentangle overlapping visual elements, resolve ambiguities in labeling, and consistently map visual evidence to textual answers.

Accurate attribution in chart-related tasks is crucial for ensuring that multimodal large language models (MLLMs) generate reliable and trustworthy outputs (Bai et al., 2024). Charts often convey critical information involving exact quantities, trends, and comparisons. When an MLLM's response to a chart-related request cannot be clearly linked to specific visual elements, it becomes difficult to assess whether the answer is grounded in the chart's data or influenced by hallucinated patterns. This lack of transparency can lead to incorrect conclusions, undermining the reliability of automated systems in critical areas such as financial analysis, policy-making, and scientific research, where accurate data interpretation is essential for decision-making. Reliable attribution helps mitigate these risks by making the model's process verifiable, meaning that the model's response can be traced back to identifiable visual elements in the chart. As demonstrated in Fig 1, this allows for confirmation that the generated response is directly supported by the chart's data, reducing the potential for hallucinated or incorrect interpretations.

Main Results: We introduce the task of Post-Hoc Fine-grained Attribution for Charts. This task identifies the specific chart elements (like bars, points, or sectors) that directly support the model's answer to a given question and enable the grounding of model responses to visual elements. We focus specifically on post-hoc attribution since it provides a flexible plug-and-play mechanism agnostic of the actual multimodal chart system used underneath and decouples attribution from response generation for traceability.

We introduce ChartVA-Eval, a new benchmark designed to advance the evaluation of chart visual attribution systems. ChartVA-Eval comprises real-world chart data sourced from financial documents and policy datasets, such as SEC Filings, the World Bank Open Data, Open Government Data, and the Global Terrorism Database. The benchmark features diverse chart styles and includes retrieval, reasoning, and computation-based questions, all paired with fine-grained visual attribution annotations. We also propose ChartLens, a chart attribution methods that leverages set-of-marks prompting with multimodal LLMs to produce reliable attributions. ChartLens demonstrates significant improvements, achieving 26-66% higher accuracy compared to competitive baselines, underscoring its effectiveness in identifying the precise chart elements that support model-generated answers.

Main Contributions:

- We propose the task of Post-Hoc Fine-grained Visual Attribution for Charts, which focuses on determining the specific chart elements that support a given chart-associated textual response, improving transparency and mitigating hallucination in MLLMs.
- We present ChartVA-Eval, a comprehensive benchmark of 1200+ samples, containing real-world chart data from diverse sources. The benchmark features diverse chart styles and fine-grained attribution annotations to facilitate rigorous evaluation.
- We introduce ChartLens, a novel chart attribution algorithm based on set-of-marks prompting with multimodal LLMs. Our method achieves 26-66% improvements over existing baselines, establishing a new state-of-the-art method for chart attribution tasks.

2 Related Work

2.1 Response Attribution in LLMs

Generative LLMs now lead performance in various tasks, but their tendency to produce hallucinations remains a significant challenge (Zhang et al., 2023). To mitigate these issues, researchers have explored training LLMs to provide citations alongside their answers (Gao et al., 2023; Menick et al.,

2022; Nakano et al., 2021). Other methods augment LLMs with external tools such as retrievers (Ye et al., 2023b; Li et al., 2023b), and search engines (Nakano et al., 2021).

Three primary attribution strategies have emerged. Direct model-driven attribution allows the model to generate answers and attributions simultaneously, though this often leads to inaccuracies in both the answers and the cited sources (Peskoff and Stewart, 2023; Sun et al., 2022). Post-retrieval answering involves explicitly retrieving information first and then answering based on the retrieved data(Ye et al., 2023b; Li et al., 2023b). However, retrieval does not always equate to accurate attribution, as conflicts between the model's internal knowledge and the retrieved information can arise (Huo et al., 2023; Chen et al., 2023). In post-generation attribution, the model generates an answer first, and then a search is conducted to find supporting references, modifying the answer if necessary (Li et al., 2023a).

Additionally, recent research has focused on generating more structured attributions for data from different modalities. For example, MATSA (Mathur et al., 2024) introduces the Fine-grained Structured Table Attribution (FAST-Tab) task, where a multi-agent LLM system provides row- and column-level attributions to visually support claims derived from tables.

2.2 Visual Chart Understanding

Automated chart understanding has seen significant advancements through classification-based and generation-based methods. Early classification models like IMG+QUES (Kafle et al., 2018) and Relation Networks (Santoro et al., 2017) faced out-of-vocabulary (OOV) challenges, which were mitigated by dynamic encoding techniques such as SANDY (Kafle et al., 2018) and PReFIL (Kafle et al., 2020) that incorporated OCR sub-networks . Pre-trained models like STL-CQA (Singh and Shekhar, 2020) and VisionTaPas (Masry et al., 2022) further improved performance by leveraging transformer-based architectures .

Generation-based approaches dominate tasks like chart captioning and chart-to-table conversion. Models such as Donut (Kim et al., 2022) and Pix2Struct (Lee et al., 2023a) introduced end-to-end OCR-free architectures, while UniChart (Masry et al., 2023) and MatCha (Liu et al., 2022b) incorporated chart-specific pre-training objectives.

The emergence of Multimodal Large Language Models (MLLMs) like ChartLlama (Han et al., 2023) and ChartAssistant (Meng et al., 2024) has enabled strong zero-shot performance through instruction-tuning on chart-specific datasets. Additionally, tool-augmented methods such as DePlot (Liu et al., 2022a) and StructChart (Xia et al., 2023) aid LLMs by converting charts to structured data tables. Despite these advances, challenges remain in handling domain-specific chart diversity and developing robust evaluation metrics (Wang et al., 2024).

3 Post-Hoc Visual Attribution in Charts

Problem Statement. Given a dataset $\mathcal{D}$ consisting of a set of charts $\mathcal{C}$, each chart is an image $c \in \mathcal{C}, c = \mathcal{I}^{w \times h \times 3}$, and is associated associated with a set of responses $\mathcal{R}_c$. Each response is denoted by $v, v \in \mathcal{R}_c$.

The objective is to determine the set of visual regions within the chart c that provide evidence for the response v. Specifically, for a given chart c and response v, the goal is to produce an attribution set $\mathcal{A}_{c,v}$, where:

$$\mathcal{A}_{c,v} = \{a_1, a_2, \ldots, a_n\} \tag{1}$$

and each a_i represents a distinct region corresponding to an element in the chart c (e.g., bars, lines, points, segments) that supports the response v.

The expectation is that $\mathcal{A}_{c,v}$ satisfies the following criteria:

1. **Relevance**: Each region a_i must be directly relevant to the response v.

2. **Completeness**: The set $\mathcal{A}_{c,v}$ should comprehensively cover all the visual evidence needed to justify v.

3. **Precision**: The regions a_i should be specific and exclude irrelevant parts of the chart.

The task can be summarized as finding a mapping function:

$$f : (c, v) \mapsto \mathcal{A}_{c,v} \tag{2}$$

where f identifies the precise visual elements in c that substantiate the response v.

Dataset	ChartVA- AITQA	ChartVA- PlotQA	ChartVA-ChartQA
# of Queries	301	595	348
# of Charts	301	581	266
# of Bar Charts	203	396	121
# of Pie Charts	0	0	109
# of Line Charts	98	199	118
Chart Source	Synthetic	Synthetic	Real World
Multiple Attributions	No	Yes	Yes
Avg # of Attributions	1	2.4	1.43
Max # of Attributions	1	12	8
Avg # of Data Series	1.23	2.52	2.45
Max # of Data Series	8	4	14

Table 1: Statistics for the ChartVA-Eval benchmark, reported for constituent datasets.

4 ChartVA-Eval

In this section, we introduce ChartVA-Eval, a benchmark designed to evaluate visual attribution in charts. Each data point in ChartVA-Eval consists of a chart image $c \in \mathcal{C}$, represented as $c = \mathcal{I}^{w \times h \times 3}$, a textual response $v \in \mathcal{R}c$, and a ground truth attribution set $\mathcal{A}c, v^{gt}$, which represents the ground truth attributions. Additionally, the set of all regions in the chart corresponding to different elements is denoted as $\mathcal{A}_c$, representing all potential regions within the chart.

4.1 Data Sources

The ChartVA-Eval Benchmark is constructed from a diverse set of data sources to ensure comprehensive evaluation of post hoc attribution in chart-based visual question answering (VQA). By incorporating both synthetic and real-world charts, it captures a wide range of design styles, chart types, and question-answer (QA) contexts. This diversity enables rigorous assessment of model performance across different domains and visual complexities. The benchmark draws from three key datasets: MATSA-AITQA (Mathur et al., 2024), PlotQA (Methani et al., 2020), and ChartQA (Masry et al., 2022), each offering unique characteristics and challenges.

MATSA-AITQA (Mathur et al., 2024) provides chart data derived from tabular QA over public U.S. SEC filings of major airline companies, covering the fiscal years 2017 to 2019 (Katsis et al., 2021). The tables are paired with QA pairs and annotated cells corresponding to the data points supporting the answers. From these tables, synthetic charts are generated by applying variations in themes, color palettes, fonts, and design elements like grid lines and tick styles, resulting in over $O(10^4)$ possible style combinations. Each QA pair is associated with a single visual attribution. The dataset includes chart types such as grouped bar charts, stacked bar charts, simple bar charts (both horizontal and vertical), line charts. Further details for synthetic chart generation are present in the appendix. PlotQA (Methani et al., 2020) focuses on synthetic scientific charts paired with bounding box annotations and diverse reasoning-based questions. The dataset includes line charts and bar charts (both vertical and horizontal), with one or more visual elements annotated to support each answer. The data for these charts is sourced from publicly available repositories, including the World Bank Open Data, Open Government Data, and the Global Terrorism Database. This controlled synthetic environment allows for evaluating fine-grained attribution tasks that require careful interpretation and logical reasoning. ChartQA (Masry et al., 2022) offers real-world charts accompanied by human-authored QA annotations. The charts are sourced from platforms such as Statista, Pew Research Center, Our World in Data (OWID), and the Organisation for Economic Co-operation and Development (OECD). The dataset includes a variety of chart types, particularly pie charts, line charts, and bar charts. Given the scarcity of pie charts in other datasets, they are oversampled to ensure balanced representation. ChartQA captures the complexities and variability found in real-world data visualizations, providing a realistic benchmark for evaluating attribution models.

4.2 Attribution Annotation

For the ChartQA and PlotQA datasets, we employed a hybrid approach to generate attribution annotations, combining large-scale automated annotation with human verification. We utilized GPT-4o to generate initial annotations by leveraging the underlying data tables, questions, and answers. Specifically, we identified frequent question templates and designed tailored prompts for each template. For example, for cardinality related QA pairs, the model was instructed to select all data points counted in the cardinality. These automated annotations were subsequently refined through human validation. In an interactive setting, annotators reviewed the rendered bounding boxes on the charts and assessed the annotations based on two criteria: (1) Relevance — ensuring the annotated elements directly support the answer, and (2) Completeness — verifying that all necessary chart elements were included. This process ensured high-quality and precise attribution annotations for both datasets. Further details on attribution annotation are provided in the appendix.

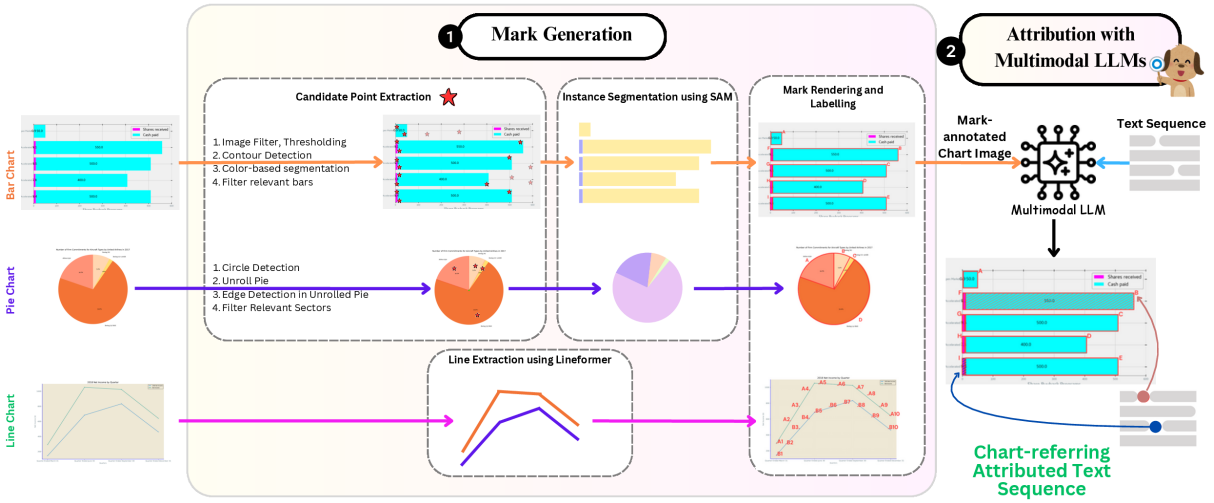


Figure 2: **ChartLens:** ❶ Chart elements, such as bars and pie sectors, are extracted through heuristic-guided methods and refined using SAM, while lines are segmented using Lineformer. ❷ The segmented elements are then marked, labeled, and used to prompt multimodal LLMs, enabling fine-grained attribution by grounding textual responses to visual regions.

5 ChartLens

ChartLens (Fig 2) facilitates fine-grained visual attribution for charts by first detecting and labeling chart elements with distinct marks. These marks, which serve as referable visual anchors, are then used to prompt MLLMs to attribute responses to specific visual elements within the chart.

5.1 Mark Generation

The goal of mark generation is to identify and tag fine-grained visual features within chart images to form a set of candidates for attribution. These marks serve as visual anchors to prompt multimodal LLMs by providing locality-based grounding. Effective mark generation requires the ability to isolate individual chart components, while ensuring robustness across various chart types and visual styles.

Heuristic-guided Instance Segmentation For segmenting bar charts, the input image is first binarized using Otsu thresholding applied to both the RGB and HSV representations. If the chart has a dark background, the binarized image is inverted to ensure foreground features, such as bars, are correctly highlighted. From the binarized image, an initial set of contours is generated. These contours are further refined by breaking them down using unique pixel values, isolating individual bars. To eliminate irrelevant or spurious contours, a filtering step based on solidity and area thresholds is applied, ensuring that only well-defined bars are retained.

For pie charts, segmentation begins by identifying the largest contour in the binarized image, which typically corresponds to the pie chart itself. We compute the minimum enclosing circle for this contour to approximate the chart's boundary. Following the approach in (Savva et al., 2011), the pie chart is unrolled along the radial axis to create a linear representation. In this unrolled form, complete edges are detected to identify sector boundaries, which are then mapped back to the original circular region. This process yields segments corresponding to individual slices of the pie chart.

While these heuristic methods leverage the structural and geometric properties of charts effectively, they suffer from several limitations. They are sensitive to noise and perform poorly on low-contrast images, often misidentifying irrelevant components such as grid lines or labels as chart elements. To address these issues, we employ the Segment Anything Model (SAM) (Kirillov et al., 2023) for instance segmentation. Specifically, n points are sampled from each detected element and used as prompts for SAM. The model generates masks that accurately enclose the objects associated with the sampled points, overcoming the shortcomings of classical methods.

SAM's architecture allows it to handle noisy and low-quality images more robustly. It produces precise masks that closely align with the boundaries of chart elements, even in complex cases. Additionally, SAM naturally suppresses background fea-

tures like grid lines by generating weaker masks (low IoU) for these elements, as they lack the spatial coherence of primary chart components. Unlike heuristic approaches, SAM generalizes well across diverse chart types and layouts without requiring extensive parameter tuning. By combining heuristic-guided preprocessing with SAM-based instance segmentation, we achieve a more flexible and accurate segmentation process that leverages the strengths of both classical computer vision and modern deep learning techniques.

Transformer-based Line Segmentation We use LineFormer (Lal et al., 2023) to extract lines from line charts. Lines present unique challenges for segmentation due to their fine structural features, such as narrow width, overlapping trajectories, and the presence of intersecting lines. These characteristics make it difficult for classical computer vision methods or point-based prompting approaches to accurately identify and segment lines, especially in dense or complex charts.

LineFormer effectively addresses these challenges. It leverages the global context provided by the transformer architecture to distinguish lines even when they are closely spaced or intersecting. After detecting candidate lines with LineFormer, we divide each line into equally spaced segments along its domain extent (horizontal range). These smaller segments serve as fine-grained marks for our attribution algorithm.

5.2 Attribution with MLLMs

To facilitate accurate attribution in chart-based tasks, we employ *Set-of-Marks (SoM) prompting*, a visual prompting technique designed to leverage the visual grounding capabilities of multimodal LLms. Inspired by (Yang et al., 2023), SoM prompting partitions an image into regions of varying granularity using interactive segmentation models like SEEM or SAM. These segmented regions are then overlaid with visual marks, such as alphanumeric labels, masks, or bounding boxes. This marked image is presented as input to the multimodal LLM. SoM prompting is effective because it enables explicit localization within the image, helping the model isolate distinct regions and understand their spatial relationships. Additionally, by labeling these elements, the technique simplifies reasoning for the model, making it easier to reference specific components during visual grounding tasks. The combination of these factors enhances the model's ability to interpret and connect visual information with textual queries.

In our approach, we prompt multimodal LLMs with chart images overlaid with marks. The prompts are structured to achieve two primary goals: validation and attribution. The prompt first explains the concept of chart attribution, providing a few-shot set of textual examples of question-answer (QA) pairs along with their corresponding attribution. Next, the model is instructed to follow a chain-of-thought (CoT) reasoning process to perform step-wise validation and attribution.

Validation involves verifying whether the QA pair is consistent with the information in the chart image. The model evaluates if the answer aligns with the visual elements and data presented in the chart.

Attribution requires the model to identify and mention the specific labeled elements within the chart that support the given answer. By explicitly referencing these elements, the model's response becomes more transparent and easier to verify.

6 Experiments

6.1 Baselines

Zero-shot GPT-4o Bounding Box Prompting: As a baseline, we prompt GPT-4o (OpenAI, 2024) to predict normalized bounding box coordinates for chart components (e.g., lines, bars, pie sectors) based on input text and the visual chart. This approach aligns with prior work for zero-shot localization tasks.

Kosmos-2: Kosmos-2 (Peng et al., 2023) is a multimodal large language model (MLLM) trained on grounded image-text data (GrIT) that integrates text-to-visual grounding capabilities. By representing object locations as Markdown links, it enables tasks such as referring expression comprehension, phrase grounding, and multimodal reasoning, and generates bounding boxes for visual grounding tasks.

LISA: LISA (Large Language Instructed Segmentation Assistant) (Li et al., 2023b) is a reasoning-based segmentation model that generates masks from implicit and complex textual queries. By introducing a <SEG> token and leveraging the embedding-as-mask paradigm, LISA extends MLLM capabilities to reasoning segmentation with robust zero-shot performance and further improves with minimal task-specific fine-tuning.

Baseline	ChartVA - AITQA			ChartVA - PlotQA			ChartVA - ChartQA		
	P	R	F1	P	R	F1	P	R	F1
Zero-shot ChatGPT4o	22.33	23.23	22.77	3.40	3.21	3.30	8.13	7.41	7.75
KOSMOS2	0.51	0.51	0.51	1.60	0.74	1.01	3.31	2.96	3.13
LISA	0.83	29.29	1.62	0.18	6.39	0.34	0.52	30.37	1.01
ChartLens	**79.86**	**61.17**	**69.28**	**35.38**	**33.94**	**34.65**	**74.51**	**56.30**	**64.14**

Table 2: Comparison of ChartLens with baselines on the ChartVA-Eval benchmark for bar charts.

Baseline	ChartVA - AITQA		ChartVA - PlotQA		ChartVA - ChartQA	
	Detection % (↑)	Chart Ar % (↓)	Detection % (↑)	Chart Ar % (↓)	Detection % (↑)	Chart Ar % (↓)
Zero-shot ChatGPT4o	18.28	1.94	6.79	**8.63**	3.39	**1.15**
KOSMOS2	74.19	46.03	38.83	27.06	87.29	41.49
LISA	**94.62**	63.18	50.21	40.92	50.21	40.92
ChartLens	59.14	**1.25**	**51.84**	9.98	**77.8**	5.34

Table 3: Comparison of ChartLens with baselines on the ChartVA-Eval benchmark for line charts.

Baseline	ChartVA - ChartQA		
	P	R	F1
Zero-shot ChatGPT4o	8.94	5.99	7.17
KOSMOS2	20.18	8.24	11.70
LISA	1.32	13.86	2.41
ChartLens	**53.33**	**44.57**	**48.56**

Table 4: Comparison of ChartLens with baselines on the ChartVA-Eval benchmark for pie charts.

6.2 Evaluation

Bar Charts and Pie Charts: Detected regions are first matched to ground truth regions (e.g., bars or sectors in the chart) based on a threshold Intersection over Union (IoU) value of IoU ≥ 0.9. The matched regions are treated as discrete items, where each detected region is assigned a unique label corresponding to its ground truth region. The performance is evaluated using Precision, Recall, and F1-score, computed over the set of filtered detected regions and ground truth regions. Let D denote the set of detected regions after filtering, and G denote the set of ground truth regions. Precision (P), Recall (R), and F1-score ($F1$) are defined as $P = \frac{|D \cap G|}{|D|}, \quad R = \frac{|D \cap G|}{|G|}, \quad F1 = \frac{2 \cdot P \cdot R}{P+R}$.

Line Charts: Unlike bar charts and pie charts, where detected regions can be matched to discrete ground truth regions, the task for line charts involves referring to singular points. Since grounding models generate bounding boxes or regions, it is challenging to precisely match these regions to individual ground truth points without ambiguity. To address this, two metrics are defined for evaluation:

1. **Detection Rate:** Measures the proportion of ground truth points covered within the detected region(s), analogous to recall.

2. **Average Area Detected:** Large detected areas indicate low precision, even if the recall is high. This is quantified as the average percentage of the input chart image covered. Larger detected areas result in higher recall but lower precision, making this metric critical for evaluating the trade-off between precision and recall in line chart attribution tasks.

For ChartLens, Large Multimodal Models (LMMs) are prompted to detect pairs of marked points on the line between which the attribution lies. These point pairs are considered corners of a bounding box, and the same metrics are used.

6.3 Implementation Details

The base multimodal language model (MLLM) for ChartLens is `ChatGPT-4o`, which is used for zero-shot bounding box detection and attribution tasks. For LISA, we use the `xinlai/LISA-13B-llama2-v1` checkpoint with 8-bit quantization and mixed-precision (`fp16`) to optimize memory efficiency during inference. Kosmos-2 is implemented using the `microsoft/kosmos-2-patch14-22` checkpoint. We use the `facebook/sam-vit-large` checkpoint for SAM. All experiments are conducted using a single NVIDIA A5000, over 6 hours. Default hyperparameters from each model's implementation are used unless stated otherwise.

7 Results

In this section, we present a detailed comparison of the proposed ChartLens model against several baselines, including Zero-shot ChatGPT4o, KOSMOS2, and LISA, across three different chart types: bar charts, line charts, and pie charts. The results demonstrate that ChartLens consistently outperforms the baselines across all chart types, highlighting its robustness and effectiveness in visual

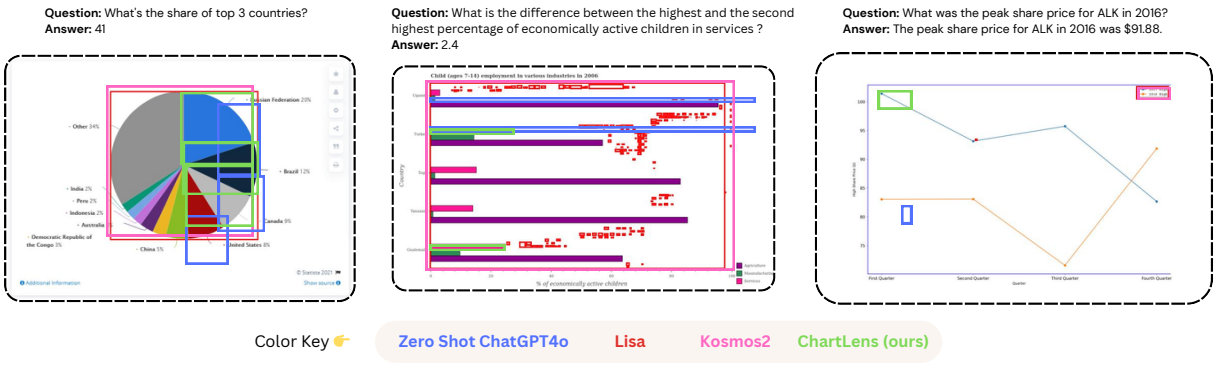


Figure 3: Qualitative comparison of our ChartLens with the baselines. ChartLens is able to effectively localize relevant, complete and precise attributions in the chart images.

chart understanding.

Bar Charts ChartLens demonstrates significant performance improvements over all baselines in bar charts, achieving F1 scores of 69.28 on ChartVA-AITQA, 34.65 on ChartVA-PlotQA, and 64.14 on ChartVA-ChartQA (Table 2). In contrast, Zero-shot ChatGPT4o achieves much lower F1 scores of 22.77, 3.30, and 7.75, reflecting its limitations in numerical reasoning and visual attribution. KOSMOS2 and LISA perform poorly, with F1 scores below 5 across benchmarks, highlighting their inability to handle bar charts due to insufficient grounding of visual and numerical reasoning.

Line Charts For line charts (Table 3), ChartLens achieves strong detection accuracy of 59.14%, 51.84%, and 77.8% on ChartVA-AITQA, PlotQA, and ChartQA, respectively, with low chart area errors of 1.25%, 9.98%, and 5.34%. While LISA and KOSMOS2 achieve high detection rate, this can largely be explained by the high Chart% area covered by their attributions; covering large areas of the chart makes capturing specific points non-trivial but reduces the specificity of attributions, making them less effective at fine-grained localization. In contrast, ChartLens reduces Chart% area by $\approx$ 3-50 times.

Pie Charts ChartLens outperforms baselines in pie charts, achieving an F1 score of 48.56, significantly higher than Zero-shot ChatGPT4o (7.17), KOSMOS2 (11.70), and LISA (2.41) (Table 4). Its precision (53.33) and recall (44.57) confirm its ability to attribute pie chart segments accurately. In contrast, Zero-shot ChatGPT4o and KOSMOS2 struggle with interpreting proportions, while LISA's extremely low performance highlights its difficulty in handling pie chart geometry and segmentation tasks.

Figure 3 shows a qualitative comparison of ChartLens with the baselines across bar charts, line charts, and pie charts. ChartLens consistently identifies and attributes relevant chart elements more accurately than the baselines, demonstrating a clear understanding of numerical and visual relationships. Zero-shot ChatGPT4o attempts to make fine-grained specific selections, however fails to exhibit robust localization since it expresses attributions using text based coordinates. LISA and KOSMOS2 consistently refer to typical chart components, like the pie as a whole, or the entire area but do not exhibit sensitivity to given queries.

8 Conclusion and Future Work

In this work, we introduced the task of Post-Hoc Fine-grained Visual Attribution for Charts, addressing the challenge of grounding chart-related responses to specific visual elements. To facilitate this, we proposed ChartLens, a novel attribution algorithm leveraging segmentation techniques and set-of-marks prompting with multimodal LLMs. Additionally, we presented ChartVA-Eval, a comprehensive benchmark featuring real-world and synthetic charts across diverse domains, enabling rigorous evaluation of visual attribution methods. Our experiments demonstrated that ChartLens significantly outperforms competitive baselines, achieving improvements of 26-66%. By enhancing transparency and mitigating hallucinations in MLLMs, our work lays a foundation for reliable chart interpretation in critical applications such as financial analysis, policy-making, and scientific research. Future work will explore extending these methods to other forms of visual data and improving robustness across chart styles and complexities.

9 Ethics Statement

This research utilizes publicly available datasets, ensuring compliance with their respective licenses. The identities of human evaluators remain confidential, and no personally identifiable information (PII) is used at any stage of our experiments. Our work is designed specifically for fine-grained visual attribution applications and does not extend to other use cases. We acknowledge the broader challenges associated with large language models (LLMs), including potential risks related to misuse and safety, and encourage readers to consult relevant literature for a detailed exploration of these issues (Kumar et al., 2024; Cui et al., 2024; Luu et al., 2024).

10 Limitations

While our work makes significant strides in fine-grained visual attribution for charts, it has certain limitations.

First, the system relies on segmentation as a core component, and any inaccuracies in the segmentation process may result in imperfect or incomplete attributions. However, as segmentation is modular, it can be improved or replaced with more advanced methods in future iterations.

Second, our approach primarily focuses on visual chart elements, such as bars, points, or sectors, and does not account for textual components like captions, labels, or titles. Addressing this limitation and integrating text-based reasoning alongside visual attribution remains a promising direction for future research.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023. Complex claim verification with evidence retrieved in the wild. *arXiv preprint arXiv:2305.11859*.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. 2024. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *Preprint*, arXiv:2401.05778.

Dr Ossama Embarak and Ossama Embarak. 2018. The importance of data visualization in business intelligence. *Data analysis and visualization using python: analyze data to create visualizations for BI systems*, pages 85–124.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

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

Kieran Healy. 2024. *Data visualization: a practical introduction*. Princeton University Press.

Ting-Yao Hsu, C Lee Giles, and Ting-Hao'Kenneth' Huang. 2021. Scicap: Generating captions for scientific figures. *arXiv preprint arXiv:2110.11624*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*.

Siqing Huo, Negar Arabzadeh, and Charles Clarke. 2023. Retrieving supporting evidence for generative question answering. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 11–20.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656.

Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. 2020. Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 1498–1507.

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

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

Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2021. Ait-qa: Question answering dataset over complex tables in the airline industry. *Preprint*, arXiv:2106.12944.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The ethics of interaction: Mitigating security threats in llms. *Preprint*, arXiv:2401.12273.

Jay Lal, Aditya Mitkari, Mahesh Bhosale, and David Doermann. 2023. Lineformer: Line chart data extraction using instance segmentation. In *International Conference on Document Analysis and Recognition*, pages 387–400. Springer.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023a. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023b. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. *arXiv preprint arXiv:2311.03731*.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023b. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838*.

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

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

Quan Khanh Luu, Xiyu Deng, Anh Van Ho, and Yorie Nakahira. 2024. Context-aware llm-based safe control against latent risks. *Preprint*, arXiv:2403.11863.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

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

Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. 2024. Matsa: Multi-agent table structure attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–258.

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

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

OpenAI. 2024. Hello, gpt-4o! https://openai.com/index/hello-gpt-4o/.

Ankur P Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *arXiv preprint arXiv:2004.14373*.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Denis Peskoff and Brandon M Stewart. 2023. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Manolis Savva, Nicholas Kong, Arti Chhajta, Li Fei-Fei, Maneesh Agrawala, and Jeffrey Heer. 2011. Revision: Automated classification, analysis and redesign of chart images. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 393–402.

Hrituraj Singh and Sumit Shekhar. 2020. Stl-cqa: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*.

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

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023a. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.

Xi Ye, Ruoxi Sun, Sercan Ö Arik, and Tomas Pfister. 2023b. Effective large language model adaptation for improved grounding. *arXiv preprint arXiv:2311.09533*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

A Details on Benchmark Construction

A.1 Datasets

A.1.1 MATSA

The TabCite dataset from MATSA (Mathur et al., 2024),consists of tables derived from various sources such as Wikipedia pages and SEC filings. The TabCite benchmark is built by reformulating existing datasets like TOTTO (Parikh et al., 2020), FetaQA (Nan et al., 2022), and AITQA (Katsis et al., 2021) to create QA pairs with human-generated questions, free-form answers, and ground truth row/column attributions. The dataset focuses on fine-grained table structure attribution, particularly identifying rows and columns for accurate reasoning and table-based question answering. MATSA, in comparison to other models, performs well across multiple settings, achieving the best F1 scores for fine-grained attribution, indicating its effectiveness for reasoning over tables with complex structures.

For ChartVA-Eval, we selected the AITQA subdataset due to its simpler table structure, which allowed for easier and more traceable conversion

from tables to charts without losing attribution. Additionally, AITQA is the only subdataset that contains numerical values in every cell, making it ideal for generating charts that can be directly derived from the data. The numerical consistency ensures that the table-to-chart conversion maintains the integrity of the data, allowing for accurate visualization and reasoning over the tables.

A.1.2 PlotQA

PlotQA (Methani et al., 2020) is a dataset designed for the task of reasoning over real-world plots. It includes data sourced from various online platforms like World Bank Open Data and the Global Terrorism Database, covering a wide range of indicator variables such as fertility rates, rainfall, and coal production across different years, countries, and regions. The dataset comprises 841 unique variables and 160 entities, with data spanning from 1960 to 2016. These statistics are represented in three main plot types: bar plots, line plots, and scatter plots. The plots vary in their visual elements, including legend positions, fonts, grid lines, and color schemes, allowing for rich and diverse plot representations. In total, 224,377 unique plots were generated, ensuring a comprehensive coverage of data.

To facilitate the creation of complex reasoning tasks, the PlotQA dataset also features a collection of 7,000 crowd-sourced questions, which were generated by workers on Amazon Mechanical Turk. These questions were categorized into three types: structural understanding, data retrieval, and reasoning. By analyzing the crowd-sourced questions, the authors extracted 74 question templates that were manually paraphrased to ensure natural phrasing. This process aimed to ensure that the dataset more accurately reflects real-world challenges in plot interpretation, providing a rich resource for training and evaluating machine learning models for visual reasoning tasks. The resulting dataset is notable for its realistic question vocabulary, longer questions, and diverse set of answers, making it a significant step forward in the field of plot-based question answering.

A.1.3 ChartQA

ChartQA (Masry et al., 2022) is a benchmark dataset designed to evaluate question answering (QA) models over chart images. It consists of a diverse collection of charts crawled from four sources: Statista, Pew Research, Our World In Data (OWID), and the OECD. These charts cover various topics such as economics, politics, and global issues, and include bar, line, and pie charts. To enhance the dataset's coverage, two main methods of annotation were employed: human-authored QA pairs collected via Amazon Mechanical Turk (AMT) and machine-generated questions derived from human-written chart summaries. The dataset focuses on two types of questions: compositional (involving logical or mathematical operations) and visual (related to chart attributes like color or height), which are designed to test complex reasoning abilities.

The dataset also employs data augmentation through the fine-tuning of a T5 model on SQuAD to generate diverse, human-like questions from chart summaries. This process helps to introduce linguistic variations and enriches the dataset with syntactic complexity. A significant feature of ChartQA is its coverage of both simple and complex charts, with the latter including multi-column charts like stacked bars and multi-line graphs. With 6,150 unique tokens in questions and 4,319 in answers, ChartQA presents a challenging task for QA models, reflecting real-world scenarios where questions require intricate reasoning over chart data. The dataset's broad topic coverage and diverse question types make it an essential resource for advancing research in visual question answering and complex reasoning over data visualizations.

A.2 Human Annotation

We employed three graduate student annotators, aged 23-26, with prior experience working with charts across various domains. The annotators were fairly compensated at the standard Graduate Assistant hourly rate, following their respective graduate school policies.

The purpose of the annotation process was to ensure high-quality and precise visual attributions by refining automated annotations and verifying their correctness. Specifically, the annotators were tasked with reviewing bounding box annotations to assess Relevance, ensuring that the annotated chart elements directly supported the provided answers, and Completeness, verifying that all necessary chart elements were included. The annotation process was conducted in an interactive setting where annotators could inspect rendered visualizations and iteratively refine the annotations as required. An overview of the annotation instructions can be seen in Fig 4.

To evaluate the consistency and reliability of the annotations, we calculated inter-annotator agreement metrics using Cohen's Kappa (κ). For Relevance, the overall agreement was near perfect with a Kappa score of 0.89. Similarly, for Completeness, the agreement remained strong, achieving a Kappa score of 0.84. Additionally, pairwise Kappa scores between the three annotators were computed to further validate consistency: Annotator 1 and Annotator 2 achieved a κ of 0.87, Annotator 1 and Annotator 3 reported a κ of 0.85, while Annotator 2 and Annotator 3 achieved a κ of 0.83.

A.3 Synthetic Chart Construction

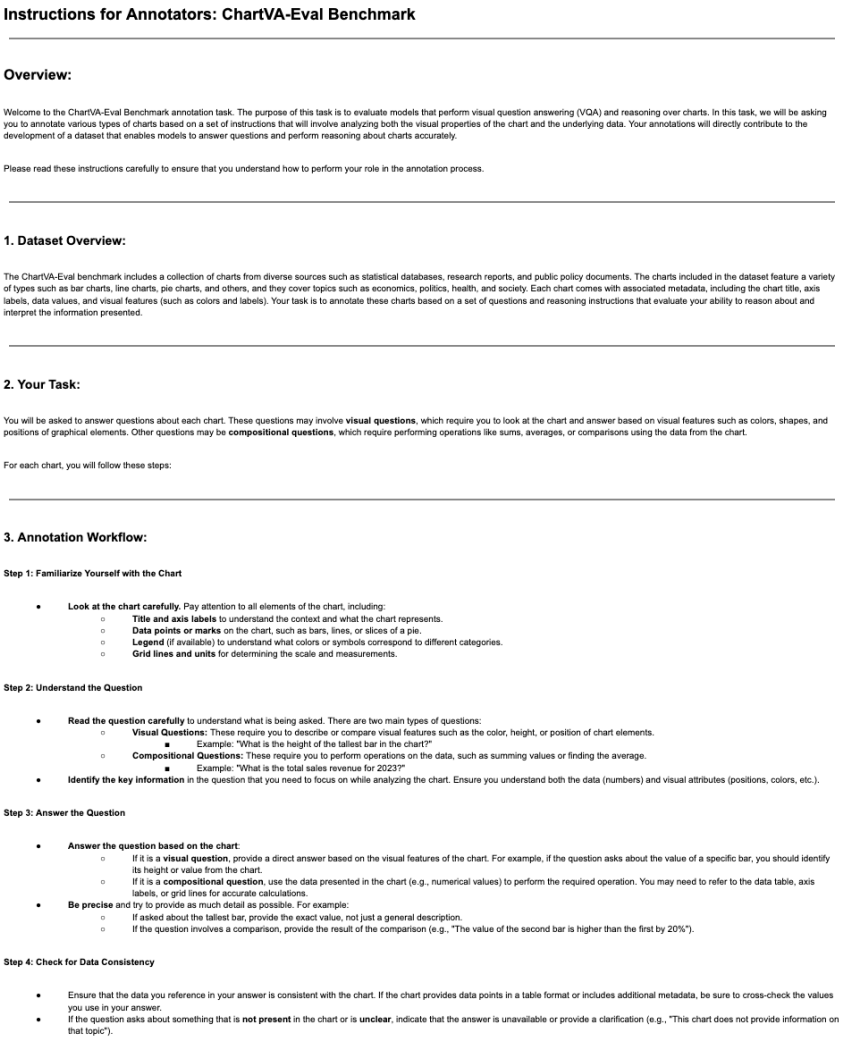
Fig 5 represents the design decision space for MATSA synthetic charts. It illustrates the wide range of charts generated, showcasing the visual diversity, variations in layouts, and the impact of different design choices on the chart's structure and appearance.

B Algorithmic Heuristics for Point Extraction

Algorithms 1 and 2 depict the algorithmic workflow for mark identification and rendering, utilizing heuristics and SAM-based segmentation techniques. These algorithms effectively segment and label chart elements, enabling downstream fine-grained attribution.

Algorithm 1 Detect Bounding Boxes in Bar Charts

procedure DETECTBARBOUNDINGBOXES(image_path, predictor)
    Input: Image path $image\_path$, predictor model $predictor$
    Output: Processed image, list of bounding boxes
    Step 1: Preprocess Image
    Load image and check validity
    Convert to grayscale and apply thresholding
    Perform morphological operations to clean noise
    Step 2: Detect Initial Contours
    Find contours in the thresholded image
    Filter contours by area to identify bar-like shapes
    Step 3: Process Bar Contours
    for each bar contour **do**
        Expand bounding box for analysis
        Mask region and extract unique colors
        for each unique color **do**
            Create mask and detect sub-contours
            if contour is rectangular **then**
                Store bounding box
            end if
        end for
    end for
    Step 4: Refine Bounding Boxes
    Remove overlapping boxes and sort by position
    for each box **do**
        Use $SAMpredictor$ to refine boxs
        if valid contour found in mask **then**
            Add final bounding box
        end if
    end for
    Step 5: Finalize Output
    Label bounding boxes and draw on image
    Return: Processed image, final list of bounding boxes
end procedure



Instructions for Annotators: ChartVA-Eval Benchmark

Overview:

Welcome to the ChartVA-Eval Benchmark annotation task. The purpose of this task is to evaluate models that perform visual question answering (VQA) and reasoning over charts. In this task, we will be asking you to annotate various types of charts based on a set of instructions that will involve analyzing both the visual properties of the chart and the underlying data. Your annotations will directly contribute to the development of a dataset that enables models to answer questions and perform reasoning about charts accurately.

Please read these instructions carefully to ensure that you understand how to perform your role in the annotation process.

---

1. Dataset Overview:

The ChartVA-Eval benchmark includes a collection of charts from diverse sources such as statistical databases, research reports, and public policy documents. The charts included in the dataset feature a variety of types such as bar charts, line charts, pie charts, and others, and they cover topics such as economics, politics, health, and society. Each chart comes with associated metadata, including the chart title, axis labels, data values, and visual features (such as colors and labels). Your task is to annotate these charts based on a set of questions and reasoning instructions that evaluate your ability to reason about and interpret the information presented.

---

2. Your Task:

You will be asked to answer questions about each chart. These questions may involve **visual questions**, which require you to look at the chart and answer based on visual features such as colors, shapes, and positions of graphical elements. Other questions may be **compositional questions**, which require performing operations like sums, averages, or comparisons using the data from the chart.

For each chart, you will follow these steps:

---

3. Annotation Workflow:

Step 1: Familiarize Yourself with the Chart

- **Look at the chart carefully.** Pay attention to all elements of the chart, including:
  - **Title and axis labels** to understand the context and what the chart represents.
  - **Data points or marks** on the chart, such as bars, lines, or slices of a pie.
  - **Legend** (if available) to understand what colors or symbols correspond to different categories.
  - **Grid lines and units** for determining the scale and measurements.

Step 2: Understand the Question

- **Read the question carefully** to understand what is being asked. There are two main types of questions:
  - **Visual Questions:** These require you to describe or compare visual features such as the color, height, or position of chart elements.
    - Example: "What is the height of the tallest bar in the chart?"
  - **Compositional Questions:** These require you to perform operations on the data, such as summing values or finding the average.
    - Example: "What is the total sales revenue for 2023?"
- **Identify the key information** in the question that you need to focus on while analyzing the chart. Ensure you understand both the data (numbers) and visual attributes (positions, colors, etc.).

Step 3: Answer the Question

- **Answer the question based on the chart**:
  - If it is a **visual question**, provide a direct answer based on the visual features of the chart. For example, if the question asks about the value of a specific bar, you should identify its height or value from the chart.
  - If it is a **compositional question**, use the data presented in the chart (e.g., numerical values) to perform the required operation. You may need to refer to the data table, axis labels, or grid lines for accurate calculations.
- **Be precise** and try to provide as much detail as possible. For example:
  - If asked about the tallest bar, provide the exact value, not just a general description.
  - If the question involves a comparison, provide the result of the comparison (e.g., "The value of the second bar is higher than the first by 20%").

Step 4: Check for Data Consistency

- Ensure that the data you reference in your answer is consistent with the chart. If the chart provides data points in a table format or includes additional metadata, be sure to cross-check the values you use in your answer.
- If the question asks about something that is **not present** in the chart or is **unclear**, indicate that the answer is unavailable or provide a clarification (e.g., "This chart does not provide information on that topic").

Figure 4: Overview of annotation guidelines provided to annotators for ensuring accurate and consistent visual attributions.

Color Schemes

Font Family Selection

- Comic Neue
- Courier Prime
- Fira Sans
- Merriweather
- Poppins
- PT Sans
- Ubuntu
- Caladea
- Cantarell
- Carlito
- Droid Sans
- Liberation Mono
- Liberation Serif
- Nimbus Mono PS
- Nimbus Roman
- Nimbus Sans Narrow
- URW Bookman

Matplotlib Style Sheets

- ggplot
- seaborn
- seaborn-dark
- seaborn-whitegrid
- seaborn-poster
- classic
- bmh
- dark_background
- fivethirtyeight
- grayscale
- tableau-colorblind10
- fast
- Solarize_Light2
- seaborn-darkgrid
- seaborn-ticks
- seaborn-bright
- default

Font Sizes

- Title size: Random integer between 16 and 22
- Label size: Random integer between 12 and 16
- Tick label size: Random integer between 10 and 14

Caption and Legend

- Caption position: top, bottom
- Legend position: best
- Legend frame visibility: True or False

Line Styles and Marker Options

- Marker styles: o, s, d, ^, v, x, *

Spine Settings

- Top spine visibility: True or False
- Right spine visibility: True or False
- Bottom spine visibility: True
- Left spine visibility: True
- Spine color: black, gray, blue
- Spine linewidth: Random float between 0.8 and 1.8

Value Display

- Show values: True or False

Tick Settings

- Tick direction: in, out, inout
- Tick length: Random float between 4 and 8
- Tick width: Random float between 0.8 and 1.5

Figure 5: The design decision option space for MATSA synthetic charts, illustrating the various configurable elements and parameters available for customizing chart generation. This visual representation highlights the flexibility in chart design, encompassing aspects such as chart type, data presentation styles, and visual encoding options.

Algorithm 2 Detect and Label Pie Chart Sectors

procedure DETECTPIECHARTSECTORS(image_path, predictor)
    Input: Image path $image\_path$
    Output: Processed image with labeled pie chart sectors
    Step 1: Preprocess Image
    Load the image and convert it to grayscale
    Apply binary thresholding with Otsu's method
    Detect external contours and find the largest one
    Step 2: Identify Center and Radius
    Compute the minimum enclosing circle of the largest contour
    Extract the center (x, y) and radius r
    Step 3: Extract Pie Chart Region
    Create a circular mask based on the detected center and radius and isolate the pie chart region
    Step 4: Unroll the Pie Chart
    Define sampling angles θ in $[0, 2\pi]$
    Sample pixel intensities along concentric ellipses at varying radii
    Store the unrolled intensities as a 2D array
    Step 5: Detect Sector Boundaries
    Apply the Sobel operator to detect horizontal edges in the unrolled image
    Compute a binary edge map by thresholding strong edges
    Identify complete edges that span most of the unrolled height
    Map these edges back to angles in $[0, 2\pi]$
    Step 6: Process and Refine Sectors
    for each candidate sector **do**
        Use $SAMpredictor$ to refine sector
        if valid sector found in mask **then**
            Add final sector
        end if
    end for
    Compute midpoint angles for labeling sectors
    Step 7: Finalize Output
    for each sector **do**
        Label sectors on chart image
    end for
    Return: Processed image with labeled sectors, final list of bounding boxes
end procedure