ChartFI: Benchmarking Faithfulness and Insightfulness of Chart Descriptions from Multimodal Large Language Models

Fen Wang, Zekai Shao, Qiman Kang, Chunran Hu, Zhixuan Zhang, Lexu Xie, Chao Liu, and Siming Chen

Abstract— Chart descriptions are essential for accessibility, cross-modal retrieval, and assisting readers in extracting insights from complex visualizations. As multimodal large language models (MLLMs) are increasingly adopted for automated chart description generation, a critical question arises: how faithfully and insightfully do these models actually describe charts? Current benchmarks fall short on two fronts: existing datasets consist of simple, homogeneous charts paired with shallow, fact-enumerating descriptions; and prevailing metrics fail to capture the multi-faceted nature of description quality. To address these gaps, we present the Chart Faithfulness and Insightfulness Benchmark (ChartFI-Bench). We first summarize four dimensions that characterize high-quality chart descriptions: factual accuracy, salient feature emphasis, domain-informed guidance, and chart–text complementarity. Guided by these dimensions, we construct a high-quality benchmark comprising 896 chart–description pairs, which feature visually complex charts and semantically rich descriptions. Furthermore, we design four aligned evaluation metrics—Faithfulness, Coverage, Informativeness, and Acuity—to systematically assess the quality of descriptions across these dimensions. Experiments conducted on mainstream MLLMs demonstrate the effectiveness of the proposed framework and reveal common weaknesses among existing models.

Index Terms—Chart Description, Evaluation Metric, Benchmark, Multimodal large language models

1 Introduction

High-quality chart descriptions reduce cognitive load [29], facilitating rapid comprehension of the core content and assisting readers in discovering deeper insights underlying the data [42, 55, 58]. However, the automatic generation of such descriptions necessitates the integration of multiple complex tasks, including the perception of visual features, the mapping between graphical elements and numerical values, and the application of domain knowledge for reasoning. Recently, Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in visual understanding, thereby advancing research in automated generation of chart descriptions. For instance, VisText [55] and ChartCap [31] train chart-specific models to generate chart descriptions, whereas ChartInsighter [58] and Vistoryteller [52] leverage Large Language Models (LLMs) to invoke external tools for mitigating hallucinations in the generated text.

Although prior studies have explored the generation of high-quality chart descriptions from diverse methodological perspectives, a critical question remains unanswered: do current models actually produce descriptions that are both faithful and insightful? Answering this question is hindered by two compounding gaps in the existing evaluation infrastructure. On the data side, current benchmarks mostly comprise simple, stylistically homogeneous charts paired with descriptions that enumerate surface-level facts without conveying deeper analytical insights. On the metric side, traditional natural language generation metrics such as BLEU [44] and METEOR [5] measure surface-level textual similarity and are highly sensitive to stylistic variation, rendering them unreliable for assessing factual correctness or analytical depth. Critically, similarity-based metrics are semantically blind: "A outperforms B" and"B outperforms A" share most of their surface tokens yet carry opposite meanings, yet these measures assign inflated scores to semantically contradictory statements.

- *Fen Wang, Zekai Shao, Qiman Kang, Chunran Hu, Zhixuan Zhang, Siming Chen are with School of Data Science, Fudan University. Siming Chen is the corresponding author. E-mail: simingchen@fudan.edu.cn.*
- *Fen Wang and Zekai Shao contributed equally as first authors.*
- *Lexu Xie is with Zhengzhou Zhongke Institute of Integrated Circuit and System Application.*
- *Chao Liu is with School of Computer Science, Fudan University and Zhengzhou Zhongke Institute of Integrated Circuit and System Application.*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

The evaluation landscape for chart descriptions thus exhibits a structural blind spot across both dimensions. Several recent frameworks [12, 36] attempt to address semantic and factual correctness in image captioning, but they remain ill-suited for chart descriptions because they decompose descriptions into independent lexical units (words or phrases) for scoring, thereby disrupting the strong internal semantic dependencies that characterize analytical chart descriptions. ChartCap [31] proposes a chart-specific metric that evaluates descriptions by comparing a chart reconstructed from the description against the original, addressing faithfulness more directly—but this imposes stringent requirements on description granularity that most real-world descriptions cannot meet, since they summarize high-level semantic information rather than providing fine-grained reconstruction specifications. Insightfulness, meanwhile, remains entirely unaddressed by existing metrics. More recent approaches that employ LLMs as evaluators [9, 63] have not been rigorously validated for domain-specific reliability, leaving their trustworthiness uncertain.

To address these gaps, we propose the Chart Faithfulness and Insightfulness Benchmark (ChartFI-Bench), a comprehensive benchmark for evaluating the faithfulness and insightfulness of chart descriptions generated by MLLMs. To construct a high-quality benchmark, we first summarize the core characteristics of good chart descriptions to guide our data construction process. These characteristics include: factual accuracy, salient feature emphasis, domain-informed guidance and chart–text complementarity, which ensures that the text conveys high-level semantic content to complement the visual information. These dimensions serve as the foundational principles guiding both the construction of the benchmark and the design of the evaluation methodology. To build a high-quality chart description benchmark, we collect charts from arXiv and apply a systematic filtering pipeline followed by manual verification, ultimately obtaining 896 chart-description pairs with visually complex charts and semantically rich descriptions.

Finally, we construct a four-dimensional evaluation framework to systematically quantify the faithfulness and insightfulness of chart descriptions. We first decompose descriptions into atomic data facts, each represented as a structured 6-tuple, to enable fine-grained evaluation independent of phrasing differences. For assessing faithfulness, we evaluate four verification strategies through a 2×2 comparative study and identify the direct MLLM-as-a-Judge approach, which relies exclusively on the chart image and its corresponding description, as the most effective. For coverage, we match reference and generated data facts through insight-type-specific scoring rules to enable precise alignment. We further introduce an informativeness metric with context-adaptive weighting that adjusts semantic-level importance based on chart com-

Table 1: Comparison of ChartFI-Bench and Existing Datasets. Semantic Coverage refers to the four-level semantic content for chart descriptions [38]: L1 (chart construction, e.g., chart type), L2 (statistical summaries, e.g., value), L3 (perceptual and cognitive observations, e.g., trend changes), and L4 (contextual and domain-specific insights). "–" indicates that the word count is not reported in the original paper.

Dataset	Size	Word Count	Chart Types	Semantic Coverage	Multi-Chart	Evaluation Dimensions			
						Faithfulness	Coverage	Informativeness	Acuity
Chart-to-Text [23]	44K	120	6	L1–L3	✗	✗	✓	✗	✗
ChartSumm [47]	84K	45	3	L1–L3	✗	✗	✓	✗	✗
VisText [55]	12K	77	3	L1–L3	✗	✓	✓	✗	✗
ChartLlama [16]	11K	-	10	L1–L3	✗	✗	✓	✗	✗
VL2NL [27]	2K	161	10	L1–L2	✗	✗	✓	✓	✗
ChartInsighter [58]	75	157	1	L1–L3	✗	✗	✓	✓	✗
ChartCap [31]	**565K**	223	9	L1–L3	✓	✓	✓	✗	✗
ChartFI (Ours)	896	**241**	**14**	**L1–L4**	✓	✓	✓	✓	✓

plexity, and an acuity metric that evaluates domain-knowledge utilization along five progressive sub-dimensions. Subsequently, we utilize the constructed benchmark and the proposed evaluation framework to conduct a comprehensive evaluation of state-of-the-art MLLMs. The evaluation results expose shared weaknesses across current models and offer actionable directions for future improvement. We also quantify the consistency between automatic metric scores and human judgments to demonstrate the effectiveness of automatic evaluation. The benchmark is available at https://github.com/wangfen01/ChartFI. Our main contributions are as follows:

- We characterize the dual requirements of high-quality chart descriptions—faithfulness and insightfulness—and summarize four key dimensions that shape description quality, providing a foundation for both benchmark construction and evaluation design.
- We construct a high-quality benchmark of 896 chart-description pairs and propose a comprehensive four-metric evaluation framework to rigorously assess the faithfulness and insightfulness of chart descriptions.
- We conduct comprehensive evaluations of state-of-the-art MLLMs and reveal common weaknesses among existing models.

2 Related Work

This section reviews prior research on chart description, focusing on large models, benchmark datasets, and evaluation metrics.

2.1 Large Models for Chart Description

Early research on chart description primarily relied on template-based generation methods [8, 19, 35], which severely limit the diversity and flexibility of generated content. Subsequent work shifted toward data-driven generation paradigms. Chart-to-Text [42] and DataTales [53] converted the underlying structured data of charts into textual descriptions; however, this approach depends solely on data table inputs and inevitably discards critical visual information such as color and layout. While LineCap [39], Chart2Text [23] and MMCA [34] incorporated visual features, their descriptions remain limited, often failing to capture the deeper trends and insights embedded in charts. To enhance the semantic richness of generated descriptions, VisText [55] pursued this goal but is restricted to simple, single-dimensional chart types, making it difficult to handle complex charts.

To address hallucinations in chart description, recent work has introduced external reasoning mechanisms to improve factual accuracy. VL2NL [27] guided LLMs to focus on statistical features and utilized external tools for data analysis, partially alleviating numerical hallucinations. ChartInsighter [58] further invoked external modules to assist reasoning, providing deeper insights into trends. Vistoryteller [52] introduced a multi-agent collaboration framework that simulates a human-like authoring process for crafting cohesive data stories, thereby aligning role-specific reasoning with user-specified thematic intentions. In summary, current methods have evolved along three dimensions: diversity, semantic depth, and factual accuracy. However, existing benchmarks fall short in rigorously evaluating them under complex scenarios. To address this gap, we construct a comprehensive benchmark to better evaluate these methods on faithfulness and insightfulness.

2.2 Dataset for Chart Description

Early chart description datasets were constructed through rule-based or template-driven generation methods [16]. However, such synthetic data exhibits a considerable gap from real-world charts in both visual style and structural complexity. To address this limitation, several studies have collected real chart descriptions written by human experts. For instance, Chart-to-Text [23] aggregated authentic chart resources from websites such as Statista [1] and Pew [45]. VisText [55] further enriched the semantic coverage of descriptions by incorporating multi-level features that span statistical, perceptual, and cognitive dimensions. ChartInsighter [58] constructed chart descriptions for time-series data and annotated hallucination types of descriptions generated by different models at the sentence level, thereby facilitating the evaluation of hallucination in chart description generation. Nevertheless, the charts in these datasets are mostly simple and homogeneous. SciCap [19] and ArxivCap [30] constructed large-scale, domain-specific chart datasets from scientific literature, yet their textual content, which was primarily extracted from figure captions, lacks a comprehensive semantic interpretation of the charts.

With the advancement of LLMs, a growing body of research has leveraged these models to automate the generation of charts and descriptions. For instance, ChartAssistant [40] and ChartLlama [16] employed LLMs to synthesize topical datasets and generate the corresponding code for chart rendering. ChartCap [31] and ChartDiff [64] utilized carefully designed instructions to guide LLMs in generating descriptions of charts. However, existing approaches primarily enumerate surface-level statistical facts and lack deep, domain-grounded insights; furthermore, the generated text frequently exhibits hallucination. To address these limitations, we construct a high-quality benchmark for chart descriptions that integrates the knowledge of domain experts with rich semantic insights. We ensure the quality of the benchmark through manual inspection. A systematic comparison is presented in Tab. 1.

2.3 Evaluation for Chart Description

Classical metrics like BLEU [44], CIDEr [56], and BERTScore [65] rely predominantly on superficial lexical or embedding alignments. By inappropriately rewarding deceptive word overlap and penalizing valid stylistic variations, they struggle to reflect a model's true comprehension capabilities. CLIPScore [17] captures only high-level semantic alignment, rendering it ill-suited for lengthy, detail-rich chart narratives. Furthermore, metrics designed for general image captioning, such as CAPTURE [12] and CAPability [36], often evaluate texts by segmenting them into isolated phrases; this naive fragmentation disrupts the inherent semantic coherence required for chart comprehension.

Chart-specific evaluation metrics also face limitations. For instance, ChaTS-Critic [28] evaluated sentence-level faithfulness by leveraging Large Language Models (LLMs) alongside extracted underlying chart data. However, extracting underlying data directly from chart images, particularly for complex charts, inevitably leads to cascading errors and compromises the reliability of the evaluation. ChartCap [31] introduced the Visual Consistency Score (VCS), which assesses quality

via reverse chart reconstruction. However, VCS demands strict descriptive granularity. Since practical chart descriptions typically summarize macroscopic trends rather than detailing exhaustive data points, reverse reconstruction was frequently infeasible. To address these challenges, we introduce an evaluation framework specifically designed to systematically assess the faithfulness and insightfulness of chart descriptions.

3 High-quality Chart Description Characterization

Effective chart descriptions enable readers to rapidly grasp key data insights and uncover underlying trends. However, descriptions within current datasets [31,58] frequently fail to meet these expectations. They tend to present visual features as unfocused enumerations. This lack of narrative prioritization forces readers to manually distill critical information, thereby undermining the text's communicative efficacy and imposing an unnecessary cognitive burden [62]. Furthermore, these descriptions rarely integrate domain-specific context, despite substantial evidence that diverse users—including both sighted and visually impaired readers—strongly prefer descriptions enriched with deeper contextual semantics [38, 55].

To address these limitations, we characterize high-quality chart descriptions along two core dimensions: **Faithfulness** (factual alignment with the data) and **Insightfulness** (surfacing meaningful analytical takeaways). These dimensions underpin our benchmark construction and evaluation frameworks. We further operationalize Insightfulness via three progressive sub-dimensions: Salient Feature Emphasis for targeted focal selection, Domain-Informed Guidance for interpretive depth, and Chart-Text Complementarity for balancing information across visual and textual modalities.

D1: Factual Accuracy. A high-quality chart description must faithfully reflect the underlying data in the chart [33]. Because erroneous chart descriptions may mislead users, resulting in flawed decision-making [7, 20, 21, 25]. Prior work has confirmed that factual hallucinations remain a persistent challenge in automatically generated chart descriptions [42,55,58], underscoring factual accuracy as a fundamental criterion for high-quality descriptions.

D2: Salient Feature Emphasis. Rather than providing a flat, undifferentiated enumeration of all visible data, a high-quality chart description inherently prioritizes the most visually distinctive and statistically notable features [38]. Visual elements such as sharp inflection points, extreme values, or anomalous inter-group variances act as natural cognitive anchors that guide reader comprehension [24]. Accordingly, a well-crafted narrative aligns with this intrinsic visual hierarchy, ensuring that salient data signals drive the initial reading flow rather than an exhaustive list of minor details.

D3: Domain-Informed Guidance. The patterns presented in a chart often cannot be adequately interpreted from surface visual information alone; domain knowledge is required to bridge the gap between observable phenomena and their deeper insights. A chart may contain multiple discernible trends, yet not all extractable insights carry equal value. A high-quality chart description should incorporate specific context and domain expertise to prioritize the most significant phenomena, articulating why they merit attention, how they relate to the central research question, and what underlying mechanisms or theoretical implications they may reflect, rather than enumerating all possible interpretations indiscriminately [6]. Through such domain-grounded and purposeful exposition, the description guides readers from merely observing the data to understanding why the data matter.

D4: Chart–Text Complementarity. Chart-text complementarity relies on a semantic division of labor: charts excel at presenting numerical relationships and spatial distributions, whereas descriptions should provide high-level synthesis, interpretation, and contextualization. A description fails this complementary role when it mechanically reproduces low-level visual details—such as the verbatim narration of axis labels or color encodings [57]. Such overlap degrades synergy into redundancy; it forces readers into constant context-switching between modalities to verify correspondences, inducing a split-attention effect [4, 29, 54] that increases cognitive load without information gain. Referencing visual elements is encouraged, provided they serve as functional anchors for generalizations rather than redundant transcriptions.

4 ChartFI-Bench: A benchmark for chart description

Based on the characteristics of high-quality chart descriptions (Sec. 3), we construct a high-quality benchmark containing 896 pairs of charts and corresponding descriptions (Fig. 1).

4.1 Dataset Construction

Based on our preliminary study (Sec.3) and related benchmark works [9, 42, 55, 58], we identify three core requirements that the benchmark dataset should meet: 1) **Diversity in Visual Encodings:** The dataset need reflect the complexity of real-world visualizations, incorporating a wide taxonomy of chart types, diverse visual marks, and multi-panel visualizations. 2) **High-Level Analytical Insight:** The description should deliver deep, non-trivial insights rather than a simple enumeration of facts. It should reflect what a real-world reader would genuinely need to understand, grounded in domain knowledge and focused on what the data truly means, not just what it shows. 3) **Accurate ground truth:** Every claim in a description should be verifiable from the chart image or the source literature.

Data source and selection. The arXiv [10] repository hosts a massive volume of scientific preprints, offering an ideal raw data pool of authentic, complex visualizations accompanied by detailed, domain-specific explanations. Consequently, we downloaded all arXiv preprints published from January 2024 to February 2026. To ensure a high standard of data quality, we designed a rigorous filtering mechanism. We employed Semantic Scholar [26] to retrieve metadata and publication records, retaining only papers formally accepted at top-tier venues (e.g., ICLR, AAAI, TVCG). This selection criterion ensured that the charts and their corresponding descriptions had undergone a rigorous peer-review process, thereby guaranteeing a high standard of data quality. This process resulted in a collection of 54,335 papers. We subsequently employed MinerU [41] to parse the filtered PDF documents, extracting images and their surrounding textual context, which produced an initial pool of 302,977 raw figures.

Chart filtering. The raw figures were heavily saturated with non-visualizations (e.g., flowcharts) and visually homogeneous charts. To guarantee that the retained charts were both visually complex and structurally diverse, we designed a three-stage filtering pipeline. First, we employed an MLLM to filter the raw images, retaining only valid data visualizations that matched our predefined chart typologies. Second, we implemented a complexity filter to exclude simplistic charts [13]. Specifically, the LLM assessed whether each candidate contained multiple independent key insights. Charts with variable distributions or data patterns that exhibited simple, monotonic structures (e.g., a line chart with a single increasing trend) were classified as simple and removed. Third, three authors performed a final manual review to remove visually ambiguous or informationally sparse charts, such as those lacking axis labels, legends, or necessary semantic context. This step ensured that every retained chart could effectively support the generation of semantically rich descriptions, resulting in a final collection of 1,530 charts from the original 53,660 LLM-extracted candidate charts.

Context extraction and cleaning. Following chart selection, we extracted contextual paragraphs associated with each target chart from the source literature. Specifically, we located paragraphs within the main text that referenced the target charts and applied length-based filters: texts containing fewer than ten words were automatically removed, and paragraphs exceeding 500 words were excluded as overly verbose. Furthermore, we employed the LLM to conduct a semantic assessment of the candidate texts, retaining only those charts supported by at least three valid data insights. Finally, we extracted the core arguments most relevant to the charts to provide substantive domain knowledge for the subsequent generation of descriptions.

Description generation. Description generation followed a two-stage multi-model collaboration strategy. In the initial stage, we prompted GPT-5.2 [43] with the chart image to generate a preliminary description encompassing three semantic levels [38]: chart construction (L1), statistical summaries (L2), and perceptual and cognitive observations (L3). Although this content faithfully reflected the information directly extracted from the chart, it remained limited to visual-level observations and lacked contextual and domain-specific insights (L4).

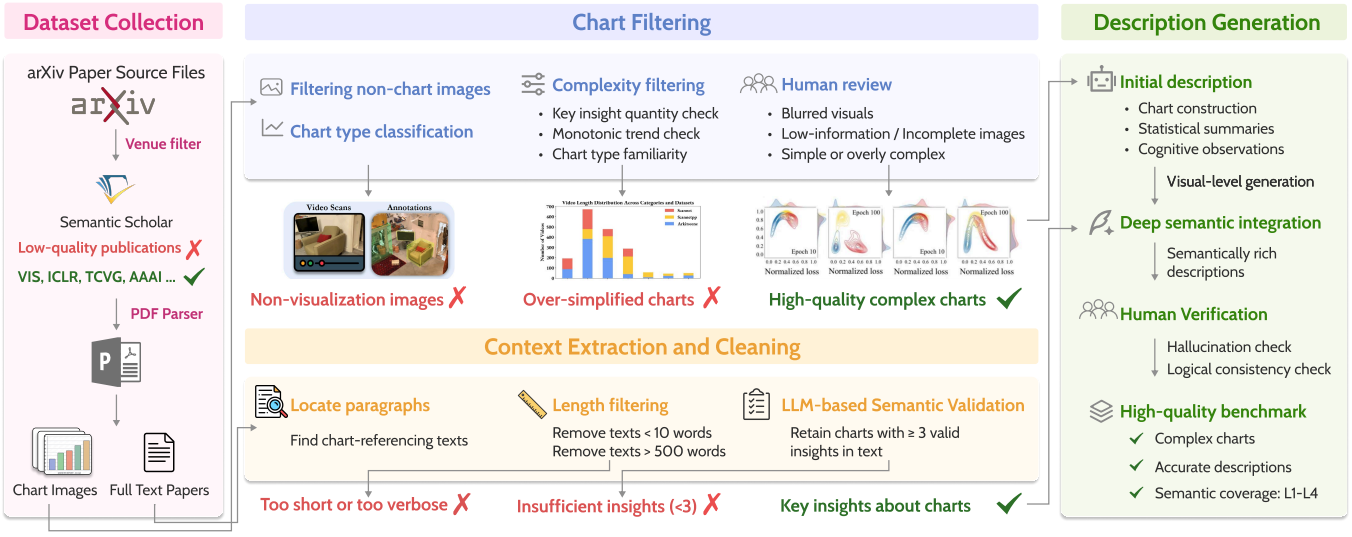

Fig. 1: Overview of the benchmark construction pipeline, consisting of three stages: dataset collection from academic papers, chart filtering with complexity and human review, and description generation with multi-level semantic coverage.

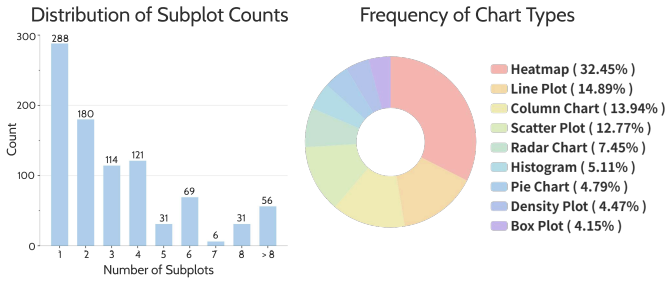

Fig. 2: Distribution of subplots per chart (left) and chart types (right).

To address this, we injected the previously distilled context into the second stage to guide Gemini-3.0-Pro [15] in deeply refining the initial description. Specifically, the model leveraged this semantic prior to identify and emphasize the most salient data insights, explain the underlying causes of the observed patterns, and suppress the unfocused enumeration of global insights. This effectively directed the narrative toward the conclusions most valuable to the reader. Ultimately, this process yielded a total of 992 chart descriptions that aligned with the high-quality characteristics defined in Sec. 3 (D2, D3, and D4).

Human verification. Given the inherent risk of hallucination in MLLMs, we implemented a rigorous manual verification process. Because the second stage of generation incorporated contextual text from the source literature, the resulting descriptions frequently contained experimental settings or background details not directly inferable from the chart images alone. Therefore, three authors conducted an item-by-item review guided by two core principles: *factual fidelity* and *narrative coherence*. To ensure factual fidelity, the reviewers cross-checked all claims against both the corresponding chart images and the source text, correcting errors such as hallucinated numerical values and misidentified extrema. Furthermore, they refined the descriptions to ensure a logical analytical narrative, discarding unverifiable claims and incoherent structures. In total, the manual stages of chart filtering and description verification amounted to approximately 182 person-hours. Through this comprehensive pipeline, we ultimately constructed a corpus of 896 high-quality chart-description pairs.

4.2 Dataset Analysis

Visual Diversity. ChartFI-Bench covers 14 chart types, providing the most extensive coverage among existing benchmarks (Tab. 1). Although line and bar charts are the most prevalent chart types in the scientific literature, they tend to be relatively simple in structure. In contrast, heatmaps, which frequently encode high-dimensional and complex data, constitute the largest share of our collection (Fig. 2, right), followed by line charts and bar charts. Beyond chart type diversity, the benchmark comprises 288 single charts and 608 multi-charts, such as scatter and pie charts in a single plot (Fig. 2, left). The source papers cover diverse domains, predominantly Computer Science (714 charts), followed by Electrical Engineering and Systems Science (85), and Statistics (46). A complete domain breakdown is in the supplementary materials.

Linguistic Diversity. ChartFI-Bench contains the longest descriptions with 241 words, longer than those in prior benchmarks. The descriptions span all four semantic levels (L1–L4) [38], distinguishing ChartFI-Bench from prior benchmarks that are limited to L1–L3. Tab. 2 presents the distribution of insight types across chart types, demonstrating the semantic richness of our descriptions; a comprehensive breakdown of insight types is provided in the supplementary materials.

Table 2: Distribution of description insights across the 12 most common chart types.

Fact Type	Column	Box Plot	Density	Heatmap	Histogram	Line Plot	Pie Chart	Radar	Scatter	Bubble	Violin	Area
Chart Construction	192	53	60	485	73	224	63	126	214	33	27	27
Value	832	144	71	1726	77	441	389	458	383	117	80	36
Distribution	208	80	290	570	210	135	71	89	367	32	64	17
Extrema	168	31	21	270	40	137	26	91	84	27	31	23
Range	122	67	22	206	25	83	15	30	126	22	43	1
Outlier	12	8	6	51	16	9	3	4	25	4	4	0
Proportion	70	6	4	33	7	10	22	1	12	9	2	9
Comparison	301	111	95	515	104	325	77	174	235	49	38	24
Trend	280	89	47	511	64	606	35	59	222	23	28	124
Correlation	56	25	16	334	15	69	11	6	126	9	4	27
Rank	129	38	16	218	27	94	206	123	131	4	11	11
Hierarchy	7	0	7	34	0	4	100	0	3	0	0	4
Domain-Specific	369	120	123	852	131	374	120	188	333	52	47	42

5 Evaluation Metrics

In this section, we introduce our evaluation framework comprising four distinct metrics: Faithfulness, Coverage, Informativeness, and Acuity.

5.1 Data Fact Formulation

Chart descriptions convey diverse insight types, such as trends, extrema, and comparisons. To enable fine-grained evaluation, we decompose each description into data facts, each capturing a single atomic insight.

Adapted from Narrative Player [50], we formalize each data fact as a 6-tuple: ⟨*type*, *parameters*, *measure*(*s*), *context*, *breakdown*(*s*), *focus*⟩, where *type* specifies the category of the insight (e.g., trend, rank or proportion) based on existing taxonomies [11,50,51,60]; *parameters* characterize the fact (e.g., an increase in a trend); *measure*(*s*) identify the dependent variables; *context* defines the data subspace; *breakdown*(*s*) specifies the grouping dimensions; and *focus* highlights the specific values emphasized within those dimensions.

To minimize annotation ambiguity, we define a controlled vocabulary of parameters for each *type*. Specifically, trend facts are restricted to directional states (e.g., INCREASE, DECREASE, PEAK); numeric types record exact quantities; comparison and correlation facts encode relational directions (e.g., POSITIVE, NEGATIVE) alongside the compared entities; and distribution facts capture structured degree-state pairs (e.g., HIGHLY CLUSTERED). This constrained formulation ensures that semantically equivalent observations, regardless of their natural language phrasing, thereby enabling reliable automatic matching (detailed in Sec. 5.3). Ultimately, this data-fact decomposition decouples our evaluation from surface-level linguistic variations, allowing us to strictly assess whether each atomic insight is factually supported by the chart. All subsequent metrics are grounded in these structured representations.

5.2 Faithfulness

To validate **D1** (Factual Accuracy), we evaluate the fidelity of chart descriptions to their underlying data. We design four candidate verification methods and conduct a comparative study to identify the most effective approach. These four methods arise from a 2×2 experimental design crossing two orthogonal dimensions: input representation (original descriptions vs. atomic data facts) and verification mechanism (*LLM-as-a-Judge* vs. code-based verification). Specifically, the *LLM-as-a-Judge* mechanism employs the MLLM to directly assess the chart image against either the original description (*Method 1*) or fine-grained atomic data facts extracted from that description (*Method 2*). In contrast, the code-based approach is a two-step pipeline: the MLLM first recovers structured data from the chart image, and the LLM subsequently generates executable code to verify either the full description (*Method 3*) or each atomic fact (*Method 4*) using the recovered data.

We curated a test set of 16 chart descriptions spanning 13 chart types (e.g., radar chart, heatmap and bubble chart). Sampled from the outputs of various open-source and closed-source models (e.g., GPT-5.4 [43], Qwen3.5-27B [46] and InternVL3.5-14B [59]), these descriptions yield a total of 938 distinct verification points. The ground truth labels were annotated by two authors based on the chart images. We evaluate performance using precision, recall, and F1-score. As shown in Fig. 3, *Method 1* achieves the best overall performance with an F1-score of 0.91. Regarding the verification mechanism, direct LLM judgment (*Methods 1 and 2*) consistently surpasses code-based verification (*Methods 3 and 4*). Furthermore, in terms of input representation, utilizing the original descriptions (*Methods 1 and 3*) yields higher precision than using data facts (*Methods 2 and 4*).

To obtain deeper insights into the failure modes associated with each method, we perform a case-by-case error attribution analysis on all misclassified instances, categorizing these errors into five distinct types (Fig. 3) : 1) **Verification Omission**, where the evaluated method overlooks critical error information present in the description; 2) **Code Logic Error**, occurring when the generated verification code executes successfully but relies on flawed judgment logic (e.g., verifying a cluster by checking whether at least one point falls within a region, rather than whether a sufficient number of points concentrate there); 3) **Extraction Loss**, denoting instances where the extraction process for data facts omits critical details or semantics from the original description (e.g., "significantly increase" is reduced to "increases", causing the verifier to miss degree-level errors); 4) **MLLM Misjudgment**, wherein the model's comprehension or reasoning about the chart image is fundamentally erroneous; 5) **Data Extraction Error**, wherein the model extracts an erroneous data table from the chart, consequently causing code verification to yield incorrect results.

The distribution of these failure modes varies significantly across the four evaluated methods. Notably, Code Logic Error occurs only

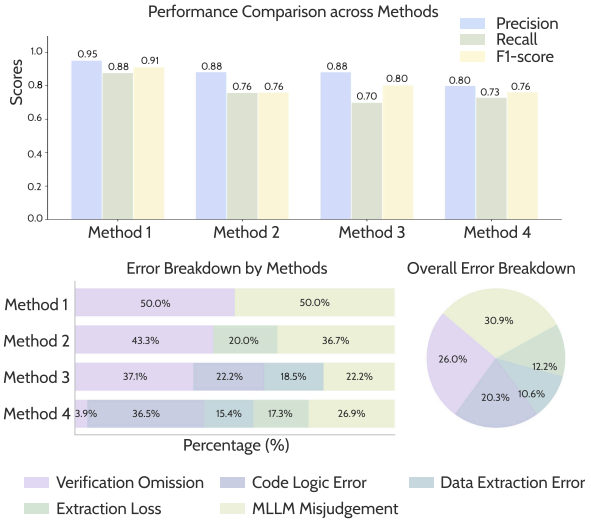

Fig. 3: Performance comparison (top) and error analysis (bottom) across methods on the Faithfulness metric.

in *Methods 3* and *4*; nevertheless, it still accounts for 20.3% of the total errors. This suggests that, although code-based verification offers deterministic and interpretable judgments in principle, the verification code generated by LLMs exhibits insufficient logical robustness, introducing errors that ultimately negate its intended benefits. Likewise, data fact extraction aims to enhance evaluation coverage by decomposing descriptions into finer-grained units; however, the information loss incurred during extraction undermines this advantage. For example, the phrase "significantly increase" may be reduced to the data fact "increases", omitting the degree modifier. When the actual error of the model lies precisely in misjudging the degree of change rather than the direction, such information loss causes the verification mechanism to overlook the error, resulting in false negatives. This explains why the recall of *Method 4* (0.73) falls below that of *Method 1* (0.88). *Method 1*, which directly compares the chart image against its corresponding description, achieves the best overall performance through the simplest processing pipeline and is therefore adopted as the primary method for completeness verification in our evaluation framework. We formally define the **Faithfulness** score as follows:

$$\text{Faithfulness} = 1 - \frac{N_{\text{error}}}{N_{\text{df}}} \tag{1}$$

where N_{df} denotes the total number of data facts in the description, and N_{error} denotes the number of erroneous data facts identified by the verification model.

5.3 Coverage

Coverage quantifies the extent to which a model-generated chart description captures the essential insights conveyed by a reference description. We decompose both the reference and generated descriptions into fine-grained atomic data facts, enabling more precise alignment at the level of individual insights. Formally, given a set of reference data facts $R = \{r_1, r_2, \ldots, r_n\}$ extracted from the reference description and a set of generated data facts $M = \{m_1, m_2, \ldots, m_k\}$, coverage measures the proportion of reference facts successfully matched by at least one fact in M. Models with low coverage tend to omit critical data patterns or trends in the chart, resulting in an incomplete summary of the chart. To systematically evaluate the completeness of generated descriptions, the **Coverage** metric is defined as follows:

$$\text{Coverage} = \frac{\left|\{\, r_i \in R \mid \exists m_j \in M,\ \phi(r_i, m_j) \geq \tau \,\}\right|}{|R|} \tag{2}$$

Central to this formulation is a multi-dimensional scoring function $\phi(s,m) \in [0,1]$, which evaluates the compatibility between the

reference fact r and the generated fact m along each data fact field independently, and subsequently aggregates the results:

$$\phi(r,m) = \sum_{d \in D} w_d \cdot \sigma_d(r,m) \tag{3}$$

where D denotes data-fact field set defined in Sec. 5.1, and w_d represents dynamically adjusted weights. We highlight three design choices.

Type-Specific Fact Matching. Each data fact type is associated with a distinct set of parameters that characterize the insight it captures; consequently, the semantics of a match differ fundamentally across types. We therefore implement type-specific scoring rules that tailor comparisons to the parameters of each category. For numeric types (e.g., value, rank, and proportion), we evaluate approximate numerical equality within a configurable tolerance. For trend facts, we extract directional keyword sequences from the relevant parameters and assess agreement through longest common subsequence (LCS) matching. For comparison and correlation facts, we verify directional consistency (e.g., POSITIVE versus NEGATIVE) while accounting for parameter-position swaps: when the specification and the model output reference the same pair of entities in reversed order, the expected direction must invert accordingly. For distribution facts, we perform structured tuple matching over (degree, state) pairs (e.g., "HIGHLY CLUSTERED"), assigning partial credit to near matches. The complete specification is provided in the supplemental material.

Schema-grounded fact normalization. Each data fact comprises multiple descriptive fields, including measure, context, breakdown, and focus, and the values of these fields are drawn from the chart description. In practice, the same entity may be represented through highly varied expressions across different descriptions (for example, "mean average precision" versus "mAP", or "accuracy" versus "acc score"), making direct field comparison unreliable. To address this, we extract a schema of canonical variable names from each chart (e.g., axis labels, legend entries, and data categories) and normalize the tokens in these fields to the canonical entries of the schema before matching. This normalization improves the accuracy of alignment by collapsing surface-level variations into unified representations. When tokens fall outside the schema, we fall back to calculating semantic similarity scores using Qwen3-Embedding-0.6B [46].

Cross-Type Semantic Matching. Similar chart insights can be expressed through distinct analytical representations. For example, a maximum value can appear as an extremum fact (A achieves the highest score) or a rank fact (A ranks first). To address this variance, we define a set of equivalence groups (e.g., extremum, rank) and implement dedicated cross-type alignment routines with calibrated conversion logic. This approach allows semantically equivalent facts expressed under different types to be recognized as valid matches.

Beyond these scoring functions, two additional mechanisms ensure robust evaluation. First, through dynamic weight reallocation, when both the reference fact and the model fact share an N/A value in a specific field, the weight of the corresponding dimension is redistributed proportionally among the remaining active dimensions. This adjustment prevents inapplicable fields from distorting the overall score. Second, through global optimal assignment, rather than matching facts sequentially, we compute the matching score $\phi(r_i, m_j)$ for all pairs, sort the scores in descending order, and greedily assign matches subject to a one-to-one constraint and a minimum threshold τ. This strategy resolves the issue wherein an early suboptimal pairing precludes a subsequent superior match.

Based on the resulting alignment, we derive three metrics. Let k denote the number of matched pairs:

$$\text{Precision} = \frac{k}{|M|}, \quad \text{Recall} = \frac{k}{|F|}, \quad \text{F}_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Recall directly quantifies coverage, representing the fraction of reference insights that the model successfully reproduces. Precision complements this by measuring the proportion of generated facts present in the reference set, reflecting the strict adherence of the model to the reference standard. Furthermore, we report the F1 score, computed as the harmonic mean of Precision and Recall, to provide a single, balanced summary metric. Through extensive empirical calibration, we establish the matching threshold at $\tau = 0.7$ and allocate the weights of the dimensions as $\{0.3, 0.3, 0.2, 0.1, 0.1\}$ for parameters, measure(s), context, breakdown(s), and focus, respectively.

5.4 Informativeness

The complementarity principle (D4) established in Sec. 3 states that high-quality chart descriptions should not reproduce low-level visual details that are readily apparent from the chart, but should instead focus on higher-level synthesis content. To operationalize this qualitative principle into a quantifiable evaluation criterion, we propose the Informativeness Metric. The core approach decomposes a generated chart description into constituent units across different semantic levels [38], categorized from L1 to L4. Specifically, the units at levels L2 and L3 are further divided into the fine-grained data facts defined in Sec. 5.1. We assign a base weight b_l to each level. Through extensive experiments across diverse types of charts and styles of description, we empirically determine these weights to be $b_l \in \{1, 6, 7, 7\}$.

However, fixed base weights alone are insufficient to accommodate the structural differences across descriptions of charts with varying levels of complexity. Descriptions of complex charts typically allocate a larger proportion of content to the visual encoding and layout details of L1, thereby reducing the relative proportion of higher-level insights. Conversely, descriptions of structurally simple charts require less space for the exposition of visual elements, which allows the analytical content of L3 and L4 to occupy a greater proportion. Without accounting for such variation in distribution, a single set of fixed weights introduces systematic bias across charts of diverse complexity. To address this limitation, we introduce a mechanism for context-adaptive weight modulation driven by the distribution of levels in the reference description. This mechanism enables the metric to automatically calibrate the relative importance of each level according to the analytical emphasis of the specific chart. Let p_l denote the proportion of units at level l in the reference description. The raw context-aware weight is defined as:

$$\tilde{w}_l = b_l \exp(p_l) \tag{5}$$

When a specific level is frequently represented in the reference description (p_l is large), the exponential term amplifies the weight of this level in a super-linear manner. This amplification automatically steers the metric toward the types of insights that are most relevant to the given chart. Conversely, when a level is absent from the reference ($p_l = 0$), the exponential term reduces to one, and the weight reverts to the base weight b_l. This soft prior mechanism addresses a specific challenge: the reference description might not exhaustively cover all valid high-level insights due to length constraints or the focus of the annotators. The proposed formulation avoids penalizing these uncovered insights with a zero reward; instead, it preserves a base-level credit that is commensurate with the cognitive depth of the insights. Subsequently, the raw weights are normalized:

$$w_l = \frac{b_l \exp(p_l)}{\sum_{k=1}^{L} b_k \exp(p_k)} \tag{6}$$

Given a sequence of facts $Y = \{y_1, y_2, \ldots, y_{|Y|}\}$ extracted from the candidate description, let $l(y_i)$ denote the semantic level [38] of the i-th fact. The Informativeness Score is defined as the mean weight of the semantic levels across the generated sequence:

$$\text{Informativeness} = \frac{1}{|Y|} \sum_{i=1}^{|Y|} w_l(y_i) \tag{7}$$

This score quantifies the expected cognitive value of each generated fact. By utilizing the total count of facts, denoted as $|Y|$, as the denominator, the formulation penalizes low-quality verbosity. Specifically, when the output of a model contains numerous low-level facts concerning visual appearances, these low-weight items dilute the overall mean. Consequently, high scores are exclusively achieved by descriptions that deliver high-value insights with expressive economy.

Table 3: Performance comparison across models on reference-based metrics and ChartFI.

Model	Reference-based Metrics				ChartFI			
	BLEU	METEOR	ROUGE	BLEURT	Faithfulness	Coverage	Informativeness	Acuity
GPT-5.4	**0.0731**	0.2894	**0.2206**	-0.4832	**0.9501**	**0.3358**	**0.3188**	**3.1233**
Gemini-3-Flash	0.0723	**0.3076**	0.2163	**-0.4342**	0.9471	0.3145	0.2762	3.0719
Claude-Sonnet-4.6	0.0651	0.2852	0.2139	-0.4870	0.8712	0.3038	0.3033	2.8919
Qwen3.5-plus	0.0615	0.2821	0.1939	-0.4803	0.8638	0.3059	0.2854	2.9930
Qwen3.5-27B	0.0574	0.2687	0.1938	-0.4849	0.8591	0.2901	0.2747	2.8249
Llama-4-Scout-17B-16E	0.0450	0.2541	0.1915	-0.4745	0.7348	0.2539	0.3058	2.1833
InternVL3.5-14B	0.0538	0.2578	0.2001	-0.4764	0.8134	0.2461	0.2667	1.7589

5.5 Acuity

We propose Acuity to measure whether a chart description can bridge the gap between *what the chart shows* and *what the chart means*. We operationalize this metric through five sub-dimensions, each evaluated by Gemini-3.1-Pro [15] on a 5-point Likert scale. These sub-dimensions are organized along a progressive cognitive chain from knowledge mobilization to cross-variable synthesis:

Domain Knowledge Accuracy & Relevance. This sub-dimension assesses whether the description incorporates domain concepts that are technically accurate and directly pertinent to the provided chart. It evaluates the effective mapping of visual elements to established phenomena, problems, or theoretical constructs within the field, rather than relying on generic or loosely associated terminology [37].

Knowledge Integration Efficiency. Possessing relevant knowledge alone is insufficient; it must be utilized efficiently. This sub-dimension evaluates whether the introduced domain context demonstrably enhances the understanding of the reader regarding specific chart phenomena. This avoids both under-explanation (where necessary knowledge is absent) and over-elaboration (where accurate but dispensable background dilutes the analytical signal).

Insight Prioritization. A chart typically presents multiple valid observations, yet these observations vary significantly in value for research or decision-making [24]. Identifying the findings that deserve prominence requires domain expertise, such as recognizing that an unexpected drop in performance at a specific scale is more consequential than a general upward trend. This sub-dimension assesses whether the description leverages domain knowledge to emphasize the most analytically significant findings and articulate why they merit attention, rather than merely reporting visually prominent features without selection informed by domain knowledge.

Etiological Explanation. Beyond identifying *what* patterns exist, understanding *why* they emerge represents the most critical application of domain knowledge [18]. This sub-dimension evaluates whether the description provides plausible, domain-grounded explanations for the observed phenomena, and whether it delineates a clear boundary between interpretations supported by evidence and speculative assertions.

Multivariate Synthesis. When a chart displays multiple variables or contains several subplots, the interpretation of each panel in isolation obscures insights that emerge exclusively from cross-panel comparisons [61]. Crucially, determining the variables or conditions that warrant comparison, and understanding the significance of their joint patterns, depends fundamentally on domain expertise. This sub-dimension evaluates whether the description synthesizes information across variables or subplots to generate higher-order insights informed by domain knowledge, rather than treating them as isolated observations.

The overall Acuity score aggregates the evaluations across these sub-dimensions. Specifically, Acuity is defined as the mean of the valid sub-dimension scores:

$$\text{Acuity} = \frac{1}{N}\sum_{i=1}^{5} A_i \tag{8}$$

where N denotes the number of applicable sub-dimensions out of the total five. If the chart does not involve multiple variable dimensions, the Multivariate Synthesis sub-dimension is excluded from the aggregation, reducing the denominator to $N = 4$. Otherwise, $N = 5$.

6 Evaluation

We evaluate seven MLLMs, validate our metrics, and compare metric changes using author-written chart descriptions as references.

6.1 Setup

Models. To comprehensively assess the capabilities of chart summarization across diverse model architectures, we evaluate both proprietary and open-source MLLMs. Regarding proprietary models, we benchmark GPT-5.4 [43], Gemini-3-Flash [15], Claude-Sonnet-4.6 [3] and Qwen3.5-plus [46]. As for open-source models, we select Qwen3.5-27B [46], Llama-4-Scout-17B-16E [2] and InternVL3.5-14B [59]. Following established practices in prior work [9,61], we set the temperature to 0 and the top-p value to 1 for all models to ensure deterministic and reproducible outputs.

Metrics. We employ two categories of evaluation metrics: traditional natural language generation metrics and four metrics we proposed (Sec. 5). Regarding the traditional metrics, we adopt BLEU [44] for n-gram precision, METEOR [5] for alignment quality with the awareness of synonyms and stemming, ROUGE [32] for recall-oriented overlap at multiple granularities, and BLEURT [49] (specifically BLEURT-base-128) for the learning-based assessment of fluency and semantic adequacy. For the proposed metrics, we evaluate Faithfulness and Acuity using Gemini-3.1-Pro [15] as the adjudicator model.

Prompt. We utilize the prompt "Write a description of the chart(s) in a paragraph of at least 150 words" for all evaluated models. Extensive experiments indicate that without such a length constraint, the models tend to generate overly brief descriptions that fail to capture the complete informational content of the charts. By specifying a requirement for length, we encourage the models to produce more comprehensive descriptions that better reflect the true capabilities of the models in chart understanding. This strategy is consistent with the prompting protocol adopted by CAPTURE [12].

Qualitative analysis. To better understand the model limitations reflected by Acuity, we examined the generated analytical descriptions of representative proprietary and open-source models, including GPT-5.4, Qwen3.5-plus, and Qwen3.5-27B, under extended thinking mode. We manually analyzed the models' generated descriptions for 30 charts whose average Acuity scores across the three models were below 3. To support a fine-grained analysis of these limitations, we sampled charts across different average Acuity-score levels within the below-3 range, while ensuring diversity in domains and visualization types. Detailed case analyses are provided in the supplementary material.

6.2 Main Results

MLLMs often produce hallucinations when generating chart descriptions. As shown in Tab. 3, the top-performing model achieves a Faithfulness score of 0.9501. In contrast, open-source models exhibit a noticeable performance drop, with Llama-4-Scout-17B-16E scoring 0.7348, highlighting a substantial gap between the leading proprietary and open-source models. These results suggest that hallucinations remain a persistent issue. For example, when a chart encodes multiple variables, MLLMs often confuse the mappings between colors and categories. Moreover, MLLMs often misread specific numerical values from chart elements, such as the heights of bars or the positions of lines, especially in the absence of explicit data labels. These low-level reading errors can lead to higher-level analytical failures,

Table 4: Breakdown of the Acuity score across five domain-specific dimensions. The abbreviations denote Domain Knowledge Accuracy & Relevance (Acc.), Knowledge Integration Efficiency (Integ.), Insight Prioritization (Insight), Etiological Explanation (Etiol.), and Multivariate Synthesis (Multi.).

Model	Acc.	Integ.	Insight	Etiol.	Multi.
GPT-5.4	3.10	**3.31**	**3.32**	2.44	**3.45**
Gemini-3-Flash	**3.18**	3.17	3.20	**2.54**	3.27
Claude-Sonnet-4.6	2.99	2.96	2.98	2.47	3.06
Qwen3.5-plus	3.08	3.18	3.03	2.38	3.31
Qwen3.5-27B	2.86	2.99	2.76	2.38	3.13
Llama-4-Scout	2.18	2.23	2.16	2.01	2.34
InternVL3.5-14B	1.85	1.91	1.73	1.32	1.99

affecting the correctness of subsequent trend characterization, comparison reasoning, and extrema identification. Beyond such visual reading errors, hallucinations also emerge in domain-grounded interpretation. As shown in Tab. 4, MLLMs achieve only moderate scores on the Domain Knowledge Accuracy & Relevance dimension; specifically, Gemini-3-Flash scores 3.18, while GPT-5.4 scores 3.10. This suggests that MLLMs may also introduce inaccurate or contextually irrelevant domain explanations when interpreting chart patterns.

MLLMs struggle to prioritize the most analytically meaningful insights. Across all evaluated MLLMs, Coverage and Informativeness scores remain relatively low, indicating that MLLMs do not merely omit key information from charts but also struggle to determine which chart details are analytically worth emphasizing. In practice, MLLMs show a strong preference for generating L1 content. Although such descriptions may be factually accurate, they remain superficial and contribute little to explaining the main findings conveyed by the chart. This tendency is particularly pronounced in open-source models. Additionally, rather than prioritizing analytically meaningful insights, they often attend to visually salient features, such as the largest radar-chart regions or the darkest heatmap cells, and treat them as important without verifying whether they correspond to the central analytical findings. However, visual salience does not necessarily imply analytical importance; determining whether a visually prominent feature is analytically meaningful often requires an understanding of the variables, the chart context, and relevant domain knowledge. Current MLLMs, however, often fail to incorporate such knowledge into their reasoning process. This limitation is reflected in the relatively modest scores on Acuity's Insight Prioritization dimension, ranging from 1.73 to 3.32 (Tab. 4).

More importantly, MLLMs often fail to synthesize multiple dimensions into a coherent analytical account. They can often identify that trends differ across subplots, but they rarely explain these differences by considering all relevant dimensions jointly. Instead, they focus on one or two dimensions while overlooking additional ones that jointly determine the observed insights. As a result, the generated descriptions capture only partial insights rather than complete explanations of the underlying multivariate relationships. This limitation remains evident even for the strongest proprietary models, with GPT-5.4 achieving a Multivariate Synthesis score of 3.45 (Tab. 4).

MLLMs struggle to incorporate domain knowledge into chart interpretation. Although MLLMs can often identify salient visual patterns, they frequently struggle to connect these patterns with relevant domain knowledge. On the Etiological Explanation dimension (Tab. 4), even the best-performing model, Gemini-3-Flash, achieves only 2.54, suggesting that MLLMs rarely explain what the observed observations imply or why they matter in the corresponding domain. More importantly, even when MLLMs introduce relevant domain context, they often fail to translate it into concrete explanations that clarify specific chart findings and help readers better understand the chart. The Knowledge Integration Efficiency scores in Tab. 4, ranging from only 1.91 to 3.31, further underscore this limitation. For example, MLLMs may state that "larger models benefit from stronger reasoning ability," but they often do not connect this claim to the specific performance in the chart. As a result, their descriptions may remain factually plausible, but they are often analytically hollow, offering generic commentary rather than domain-informed interpretive insights.

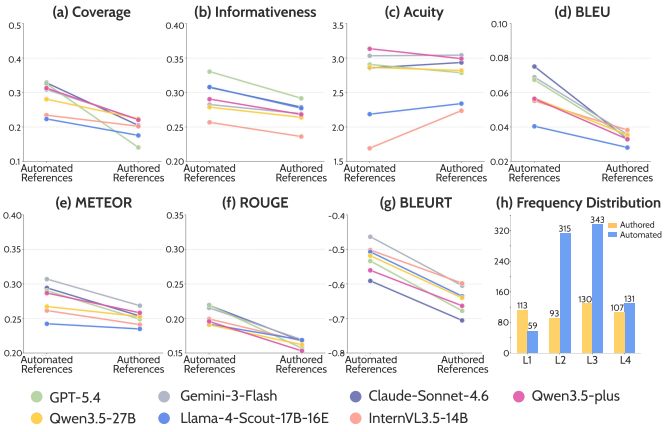

Fig. 4: Comparison of MLLM scores between our pipeline-generated references (Automated) and author-written references (Authored) across evaluation metrics (a)-(g); frequency distribution of the four semantic levels across Authored and Automated references (h).

6.3 Metric Analysis

Consistency with human evaluation. To validate the effectiveness of the proposed automatic evaluation metrics, we conducted a human evaluation study. We randomly sampled 80 chart descriptions from our benchmark, generated by seven models (described in Sec. 6.1). The domain distribution of the sampled descriptions was guided by the full benchmark, resulting in a CS to non-CS ratio of 54:26. Six domain experts (aged 20–30), each with demonstrated experience in reading and producing chart descriptions, participated in the study. Each description was assigned to an annotator with matching domain expertise. Before the annotation process, we thoroughly explained the scoring criteria of the four evaluation dimensions to the annotators. The annotators first reviewed representative examples to resolve ambiguities and ensure a consistent understanding of the rating standards. Then, each annotator independently scored the descriptions on a 5-point Likert scale across all four dimensions of the evaluation metrics. To quantify the consistency between the proposed automatic metrics and the human judgments, we adopted the Spearman rank correlation coefficient (SRCC), a widely acknowledged metric utilized in prior studies [9, 14, 48]. SRCC measures the degree to which the predicted ranking of the scores aligns with the human-assigned ranking, with values falling within the range of [-1, 1]: a value of 1 indicates perfect rank agreement, whereas -1 indicates a completely inverted ordering.

As shown in Tabs. 5 and 6, the SRCC values across all four dimensions exceed 0.69, demonstrating a strong positive correlation between the automatic metrics and human assessments. Furthermore, we assessed inter-expert consistency using four experts, including two from computer science and two from electrical energy storage systems. For each domain, the two corresponding experts independently scored the same six samples, and their pairwise SRCC values were averaged across domains, yielding average inter-expert correlations above 0.8. Specifically, a detailed breakdown of the Acuity metric (Tab. 6) reveals that all of its five sub-dimensions also achieve high consistency with human scores. These results confirm that the proposed evaluation framework reliably captures the quality aspects intended by each dimension, indicating that the automatic metrics can serve as trustworthy proxies for human evaluation in the assessment of chart descriptions. Additionally, to mitigate single-judge bias and verify metric robustness, we further evaluated the same set of 80 chart descriptions using GPT-5.4 [43]. As shown in Tab. 6, the pairwise SRCC among GPT-5.4, Gemini-3.1-Pro [15], and humans across all five Acuity sub-dimensions exceeds 0.69, confirming the metric's robustness to the choice of judge model.

Consistency across domains. Given that the benchmark is predominantly composed of arXiv-sourced CS charts, we partition our evaluation results into CS and non-CS subsets to examine potential domain bias and its impact on model evaluation (Fig. 5). The evaluated metrics exhibit broadly consistent trends across the two subsets, with no obvious domain bias.

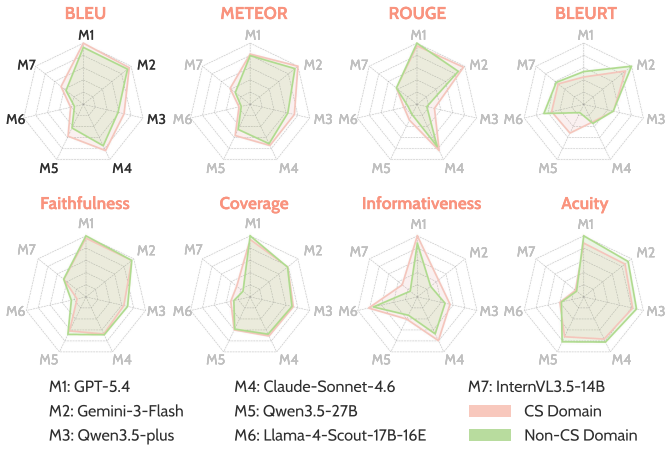

Fig. 5: Evaluation results of MLLMs across CS ($n = 714$) and non-CS ($n = 182$) domains.

Table 5: SRCC across four evaluation dimensions, including method–human agreement and inter-expert consistency.

Evaluators	Faithfulness	Coverage	Informativeness	Acuity
Our method vs. Human	0.809	0.785	0.786	0.858
Expert vs. Expert	0.857	0.800	0.800	0.829

Table 6: SRCC between human and model evaluations on Acuity and its five sub-dimensions.

Evaluators	Acc.	Integ.	Insight	Etiol.	Multi.	Acuity
Gemini vs. Human	0.790	0.842	0.826	0.790	0.767	0.858
GPT vs. Human	0.734	0.709	0.764	0.704	0.693	0.784
GPT vs. Gemini	0.898	0.864	0.857	0.792	0.819	0.908

6.4 Evaluation with Author-written References

Our benchmark construction pipeline uses chart-related text written by the original authors to construct chart descriptions. However, the resulting descriptions may still not fully recover the authors' communicative intent. This limitation may affect the benchmark's ability to evaluate how well MLLMs generate chart descriptions aligned with the authors' intended insights. To examine this potential limitation, we contacted the authors of 441 charts from our benchmark, successfully collecting 49 author-written descriptions. Upon manual inspection, we excluded 6 responses that were not valid chart descriptions (e.g., those including contextual details uninferable from the chart). The remaining 43 descriptions were utilized as ground-truth references, alongside the synthesized descriptions generated by our pipeline, to re-evaluate the MLLMs. Since the Faithfulness metric is evaluated reference-free, our subsequent analysis focuses on the remaining seven reference-based metrics (detailed in Sec. 6.2).

As shown in Fig. 4-h, the semantic level distribution (L1–L4) of author-written references differs from pipeline-generated references. Authors typically include a higher proportion of L1 content, and emphasize the advantages of their proposed methods while devoting less attention to other data dimensions. This compositional shift alters the dynamic weight allocation across L1–L4 within Informativeness metric, leading to corresponding score variations (Fig. 4-a). However, the relative performance rankings of the evaluated MLLMs remain stable. Notably, overall Coverage scores decline when evaluated against author-written references(Fig. 4-b). This reveals a critical evaluation bias: our pipeline-generated references are denser in L2 and L3 content, which models like GPT-5.4 and Claude-Sonnet-4.6 excel at capturing, yielding higher Coverage scores on the pipeline-generated references. Conversely, models like Qwen3.5-27B and Qwen3.5-plus, despite their lower Informativeness scores, achieve high Coverage against author references by successfully matching a substantial volume of L1 content. Regarding the Acuity metric(Fig. 4-c), despite minor fluctuations across different models, the relative rankings of the MLLMs remain stable despite minor cross-model fluctuations. Furthermore, under traditional reference-based metrics, the scores of all evaluated models drop significantly. This is because chart authors employ highly contextualized and linguistically diverse phrasing that exposes the limitations of standard n-gram matching metrics, rendering them inadequate for assessing the capabilities in generating chart descriptions.

7 Discussion

In this section, we discuss the limitations and broader implications of our work, along with potential directions for future research.

Tailoring Descriptions to Author Intent and Reader Backgrounds. Our description generation framework builds on the four-level semantic content hierarchy [38] and uses chart-related text written by the authors as reference. However, as discussed in Sec. 6.4, automatically generated chart descriptions may still exhibit subtle deviations from expert-authored descriptions in their semantic-level composition and analytical emphasis. This suggests that a single reference description may not fully capture the authors' intent or the expectations of readers with different levels of background knowledge. Future work could explore intent-aware and audience-aware chart description generation, where the depth and emphasis of each semantic level are adjusted according to the intended communicative goal reader profile.

Advancing chart description generation with ChartFI. Our evaluation framework can provide long-term support for assessing semantic quality in chart descriptions, particularly hallucination and insightfulness. Our current quantitative results and failure analysis reveal key limitations of current models, offering insights for designing targeted improvement strategies such as chain-of-thought prompting, multi-agent workflows, and evaluation-guided optimization. In addition, our dataset can serve as a seed set for automatically deriving synthetic training data to fine-tune future models.

Limitations of Evaluation Metrics. Our framework provides a systematic evaluation of chart descriptions across multiple quality dimensions. However, as discussed in Sec. 5.2, detecting fine-grained hallucinations remains challenging. ChartVE [22] assesses factual consistency only at a holistic level, without localizing hallucinations in specific text segments. Future work could fine-tune verification models with explicit error taxonomies to improve fine-grained hallucination detection. Moreover, Coverage metric relies on LLMs to extract data facts from references and the generated descriptions. However, semantically similar insights may be expressed differently and assigned to different types, leading to potential mismatches during the alignment phase. Future research could investigate more robust semantic matching strategies to improve the reliability of the coverage evaluation.

Dataset Scale and Diversity. Our dataset is primarily composed of scientific charts sourced from arXiv. Given that our filtering criteria prioritize chart complexity, and computer science dominates the literature on arXiv while charts from other disciplines are often comparatively simple, the dataset predominantly consists of visually complex charts from the computer science domain. Future work could apply our construction pipeline to broader sources and domains, while involving domain experts to contribute and annotate high-quality descriptions, thereby expanding both the scale and diversity of the dataset.

8 Conclusion

We introduce ChartFI-Bench, a comprehensive benchmark designed for the evaluation of chart descriptions along two dimensions: Faithfulness and Insightfulness. We first systematically summarize the characteristics of high-quality chart descriptions. Based on these principles, we construct a rigorously curated dataset of 896 chart–description pairs sourced from arXiv. We further propose a four-dimensional evaluation framework to assess the Faithfulness and Insightfulness of the descriptions. Experiments on mainstream MLLMs demonstrate the effectiveness of the proposed framework and expose shared weaknesses across current models. Our framework and evaluation results can advance research in high-quality chart description generation.

ACKNOWLEDGMENTS

The authors want to thank the reviewers for their suggestions. This work was supported by the AI for Science Program of the Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02028), the Ji Hua Laboratory S&T Program (No. X250881UG250), the National Natural Science Foundation of China (No. 62472099), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM904).

REFERENCES

[1] Statista. https://www.statista.com/, 2026. Accessed: 2026-3-31. 2

[2] A. Adcock, A. Srivastava, A. Dubey, A. Jauhri, A. Pande, A. Pandey et al. The llama 4 herd: Architecture, training, evaluation, and deployment notes. *arXiv preprint arXiv:2601.11659*, 2026. 7

[3] Anthropic. Claude-sonnet-4.6. https://claude.ai/, 2026. Accessed: 2026-3-31. 7

[4] P. Ayres and J. Sweller. The split-attention principle in multimedia learning. *The Cambridge handbook of multimedia learning*, 2:135–146, 2005. doi: 10.1017/CBO9781139547369.011 3

[5] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005. 1, 7

[6] L. Battle and A. Ottley. What do we mean when we say "insight"? a formal synthesis of existing theory. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6075–6088, 2023. doi: 10.1109/TVCG.2023.3326698 3

[7] H. P. Chan, Q. Zeng, and H. Ji. Interpretable automatic fine-grained inconsistency detection in text summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6433–6444. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.402 3

[8] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, T. Yu et al. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*, 2019. 2

[9] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311, 2024. doi: 10.1109/TVCG.2024.3456320 1, 3, 7, 8

[10] C. B. Clement, M. Bierbaum, K. P. O'Keeffe, and A. A. Alemi. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*, 2019. 3

[11] A. K. Das, M. Tarun, and K. Mueller. Charts-of-thought: Enhancing llm visualization literacy through structured data extraction. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2025.3634813 5

[12] H. Dong, J. Li, B. Wu, J. Wang, Y. Zhang, and H. Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024. 1, 2, 7

[13] J. Ellemose and N. Elmqvist. Eye of the beholder: Towards measuring visualization complexity. *IEEE Transactions on Visualization and Computer Graphics*, 2025. doi: 10.1109/TVCG.2025.3634789 3

[14] X. Fu, Y. Wang, H. Dong, W. Cui, and H. Zhang. Visualization assessment: A machine learning approach. In *2019 IEEE Visualization Conference (VIS)*, pp. 126–130. IEEE, 2019. doi: 10.1109/VISUAL.2019.8933570 8

[15] Google. Gemini 3 pro. https://gemini.google.com/, 2026. Accessed: 2026-3-31. 4, 7, 8

[16] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu et al. Chartllama: A multimodal llm for chart understanding and generation. 2023. doi: 10.1109/IJCNN64981.2025.11227777 2

[17] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.595 2

[18] E. Hoque and M. S. Islam. Natural language generation for visualizations: State of the art, challenges and future directions. In *Computer Graphics Forum*, vol. 44, p. e15266. Wiley Online Library, 2025. doi: 10.1111/cgf.15266 7

[19] T.-Y. Hsu, C. L. Giles, and T.-H. Huang. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3258–3264, 2021. doi: 10.18653/v1/2021.findings-emnlp.277 2

[20] K.-H. Huang, H. P. Chan, M. Fung, H. Qiu, M. Zhou, S. Joty et al. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2550–2568, 2024. doi: 10.1109/TKDE.2024.3513320 3

[21] K.-H. Huang, H. P. Chan, and H. Ji. Zero-shot faithful factual error correction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5660–5676. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.311 3

[22] K.-H. Huang, M. Zhou, H. P. Chan, Y. Fung, Z. Wang, L. Zhang et al. Do lvlms understand charts? analyzing and correcting factual errors in chart captioning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 730–749, 2024. doi: 10.18653/v1/2024.findings-acl.41 9

[23] S. Kantharaj, R. T. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque et al. Chart-to-text: A large-scale benchmark for chart summarization. pp. 4005–4023. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.277 2

[24] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2021. doi: 10.1145/3411764.3445443 3, 7

[25] K. Kim, S. Lee, K.-H. Huang, H. P. Chan, M. Li, and H. Ji. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*, 2024. 3

[26] R. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski et al. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*, 2023. 3

[27] H.-K. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim et al. Natural language dataset generation framework for visualizations powered by large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–22. Association for Computing Machinery, 2024. doi: 10.1145/3613904.3642943 2

[28] S. Krichene, F. Piccinno, F. Liu, and J. Eisenschlos. Faithful chart summarization with chats-pi. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8705–8723, 2024. doi: 10.18653/v1/2024.acl-long.472 2

[29] S. Latif, Z. Zhou, Y. Kim, F. Beck, and N. W. Kim. Kori: Interactive synthesis of text and charts in data documents. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):184–194, 2021. doi: 10.1109/TVCG.2021.3114802 1, 3

[30] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong et al. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.775 2

[31] J. Lim, J. Ahn, and G. Kim. Chartcap: Mitigating hallucination of dense chart captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13171–13182, 2025. 1, 2, 3

[32] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004. 7

[33] C. Liu, Y. Guo, and X. Yuan. Autotitle: An interactive title generator for visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5276–5288, 2023. doi: 10.1109/TVCG.2023.3290241 3

[34] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho et al. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1287–1310. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.70 2

[35] M. Liu, D. Chen, Y. Li, G. Fang, and Y. Shen. Chartthinker: A contextual chain-of-thought approach to optimized chart summarization. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, pp. 3057–3074. European Language Resources Association (ELRA) and ICCL, 2024. doi: 10.63317/5mr38o4exfgq 2

[36] Z. Liu, C.-W. Xie, B. Wen, F. Yu, P. Li, B. Zhang et al. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *Advances in Neural Information Processing Systems*, 38, 2026. 1, 2

[37] Y. Lu, L. Zhong, J. Yang, W. Li, P. Wei, Y. Wang et al. Domaincqa: Crafting knowledge-intensive qa from domain-specific charts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, pp. 32347–32355, 2026. 7

[38] A. Lundgard and A. Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics*, 28(1):1073–1083, 2021. doi: 10.1109/TVCG.2021.3114770 2, 3, 4, 6, 9

[39] A. Mahinpei, Z. Kostic, and C. Tanner. Linecap: Line charts for data visualization captioning models. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 35–39. IEEE, 2022. doi: 10.1109/VIS54862.2022.00016 2

[40] F. Meng, W. Shao, Q. Lu, P. Gao, K. Zhang, Y. Qiao et al. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7775–7803. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.463 2

[41] J. Niu, Z. Liu, Z. Gu, B. Wang, L. Ouyang, Z. Zhao et al. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 13–42, 2026. 3

[42] J. Obeid and E. Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 138–147. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.inlg-1.20 1, 2, 3

[43] OpenAI. Gpt-5.4. `https://chatgpt.com/`, 2026. Accessed: 2026-3-31. 3, 5, 7, 8

[44] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135 1, 2, 7

[45] Pew Research Center. Pew research center. `https://www.pewresearch.org/`, 2026. Accessed: 2026-3-31. 2

[46] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. 5, 6, 7

[47] R. Rahman, R. Hasan, A. Al Farhad, M. T. R. Laskar, M. H. Ashmafee, and A. R. M. Kamal. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. In *Canadian AI*, 2023. 2

[48] E. R. RECALL. Beyond memorability: Visualization recognition and recall. *IEEE Transactions on Visualization and Computer Graphics*, 22(1), 2016. doi: 10.1109/TVCG.2015.2467732 8

[49] T. Sellam, D. Das, and A. Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7881–7892, 2020. doi: 10.18653/v1/2020.acl-main.704 7

[50] Z. Shao, L. Shen, H. Li, Y. Shan, H. Qu, Y. Wang et al. Narrative player: Reviving data narratives with visuals. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):6781–6795, 2025. doi: 10.1109/TVCG.2025.3530512 5

[51] L. Shen, E. Shen, Z. Tai, Y. Xu, J. Dong, and J. Wang. Visual data analysis with task-based recommendations. *Data Science and Engineering*, 7(4):354–369, 2022. doi: 10.1007/s41019-022-00195-3 5

[52] Y. Shi, C. Zheng, Z. Yang, K. Xu, and N. Cao. Vistoryteller: Designing data stories with llm agent-based generation and interactive user control. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pp. 1141–1156, 2026. doi: 10.1145/3742413.3789086 1, 2

[53] N. Sultanum and A. Srinivasan. Datatales: Investigating the use of large language models for authoring data-driven articles. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 231–235. IEEE, 2023. doi: 10.1109/VIS54172.2023.00055 2

[54] J. Sweller. Implications of cognitive load theory for multimedia learning. *The Cambridge handbook of multimedia learning*, 3(2):19–30, 2005. 3

[55] B. Tang, A. Boggust, and A. Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7268–7298, 2023. doi: 10.18653/v1/2023.acl-long.401 1, 2, 3

[56] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015. doi: 10.1109/CVPR.2015.7299087 2

[57] A. Z. Wang, G. J. Quadri, M. Zhu, C. Tseng, and D. A. Szafir. Characterizing visualization perception with psychological phenomena: Uncovering the role of subitizing in data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2026. doi: 10.1109/TVCG.2025.3634807 3

[58] F. Wang, B. Wang, X. Shu, Z. Liu, Z. Shao, C. Liu et al. Chartinsighter: An approach for mitigating hallucination in time-series chart summary generation with a benchmark dataset. *IEEE Transactions on Visualization and Computer Graphics*, 31(6):3733–3745, 2025. doi: 10.1109/TVCG.2025.3567122 1, 2, 3

[59] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 5, 7

[60] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma et al. Datashot: Automatic generation of fact sheets from tabular data. *IEEE transactions on visualization and computer graphics*, 26(1):895–905, 2019. doi: 10.1109/TVCG.2019.2934398 5

[61] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024. 7

[62] S. Wiseman, S. M. Shieber, and A. M. Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2253–2263, 2017. doi: 10.18653/v1/D17-1239 3

[63] R. Xia, H. Ye, X. Yan, Q. Liu, H. Zhou, Z. Chen et al. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 2025. doi: 10.1109/TIP.2025.3607618 1

[64] R. Ye. Chartdiff: A large-scale benchmark for comprehending pairs of charts. *arXiv preprint arXiv:2603.28902*, 2026. 2

[65] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 2