DoChartDes: Fine-Grained, Automated Benchmark for Chart Description in Document Contexts

Nan Wang, Na Wang, Zeyang Yang, Yuxuan Xu, Fangzhen Peng* and Gefang Ma
CTO Organization, Lenovo Group, China
{wangnan40, wangna15, yangzy26, xuyx25, pengfz1, magf}@lenovo.com

***Abstract*—Chart understanding is crucial for evaluating multi-modal large language models' (MLLMs) capability. Mainstream benchmarks rely on closed-ended QA formats for rapid evaluation. Meanwhile, how to evaluate long-form natural language chart description is growing vital in open-ended tasks. Current open-ended evaluation suffers from vague and coarse-grained metrics, while most assess charts in isolation, ignoring interleaved document layouts. To address these, we introduce DoChartDes, *a specialized benchmark designed for open-ended chart description task*. (i) Diverse dataset: Featuring 18 diverse chart types across 6 languages and 7 document layouts, the dataset spans scenarios from simple charts to slide documents, supporting description evaluation for chart in document contexts. It includes 1.6K curated chart images, each annotated with over 25 meta-elements (42.3K total) in structured format. (ii) Fine-grained evaluation suite: This suite assesses chart description in terms of faithfulness and coverage, across 3 critical dimensions of perception and reasoning: layout understanding, factual correctness and analytical reasoning. It provides interpretable scores enabling detailed weakness analysis and hallucination detection. (iii) Intelligent judgement: The evaluation process is optimized into an intelligent way, reaching high consistency and low systemic bias with human experts while reducing time cost. With high-quality annotations and intelligent evaluation, DoChartDes provides more reliable benchmark for chart description.**

***Index Terms*—Chart Description, Multimodal Large Language Model, Open-Ended Benchmark, Document Understanding**

I. Introduction

Charts, as essential data visualization tools, offer critical insights and find extensive application in daily life [1], [2]. As volume increases and chart formats become more diverse, chart understanding is a much desired capability of multi-modal models [3]. Meanwhile, open-ended tasks like chart description are growing vital, because converting visual charts to detailed text empowers downstream tasks and enrich cross-modal resources. However, charts integrate diverse graphical elements and text components, requiring MLLMs to master two core capabilities: (i) perception of layout and factual meta-elements; (ii) reasoning of data comparisons and trends. Researchers have introduced various chart-related benchmarks to advance chart understanding, as summarized in Table I. However, problems remain in three key aspects: (i) overemphasis on closed-ended tasks, with open-ended tasks limited to short caption or summarization; (ii) limited diversity and coverage; (iii) coarse-grained evaluation metrics with insufficient human consistency checks.

*Corresponding author.

TABLE I: Comparison with existing chart-related benchmarks. "Infogra. Support": infographics (e.g., flowcharts) are evaluated; "Multi-layout": support for document layout (e.g., text+chart). "Detailed Des.": long-form chart descriptions rather than captions. "Fine-grained Eval." and "Hallu. Eval.": fine-grained and quantified hallucination measurement info.

Benchmark	Benchmark Diversity						Evaluation		
	Real Chart	Lang-uages	Chart Type	Infogra. Support	Multi-layout	Detailed Des.	Open-ended Metric	Fine-grained Eval.	Hallu. Eval.
ChartQA [4]	✓	EN	3	✗	✗	✗	✗	✗	✗
ChartBench [5]	✗	EN	9	✗	✗	✗	✗	✗	✗
CharXiv [6]	✓	EN	20+	✗	✗	✗	✗	✗	✗
Chart-to-text [7]	✓	EN	6	✗	✗	✗	NLG metrics	✗	✗
ChartLlama [8]	✗	EN	10	✗	✗	✗	GPT-score(Overall)	✗	✗
ChartX [9]	✗	EN	18	✗	✗	✗	GPT-score(Overall)	✗	✗
DoChartDes	✓	EN/DE/FR /ES/IT/PT	18	✓	✓	✓	GPT(F_1 Score, Precision, Recall)	✓	✓

Initial benchmarks such as FigureQA [10], FigCAP [11] use synthetic charts, while ChartQA [4] employs stylized charts and lacks diversity. Recently, ChartBench [5], CharXiv [6] and MMC [12] improved factualness and diversity but remain closed, restricting to multiple choice with insufficient chart content coverage. Moreover, nearly all existing benchmarks are limited in language and document layout support, focusing on English and standalone charts.

In contrast, evaluating open-ended tasks like summarization and long-form description is more challenging due to unformatted text outputs. Most prior work uses natural language generation(NLG) metrics [13], including lexical metrics [14]–[16] such as ROUGE and semantic metrics like BLEURT [17]. Inspired by text generation advancements [18]–[21], recent efforts use LLMs as evaluation tools [8]. However, these metrics suffer from vague evaluation [22]–[24] and heavily depends on the annotations' quality.

To bridge these gaps, we propose DoChartDes, a fine-grained and intelligent benchmark for evaluating the open-ended chart description task. The main contributions are:

- **Chart Diversity with Multilingual and Multi-layout Coverage**: Specifically designed for chart description task, DoChartDes spans 18 chart types across 6 languages and 7 modular document layouts, supporting evaluation on both simple charts and complex multi-element slide documents. It contains 1.6K carefully curated chart images, each annotated with an average of over 25 meta-elements which sufficient cover chart content, resulting in a total of 42.3K structured annotation. All annotations are systematically designed around distinct chart-type characteristics and formatted in JSON to enable precise parsing and automated evaluation.

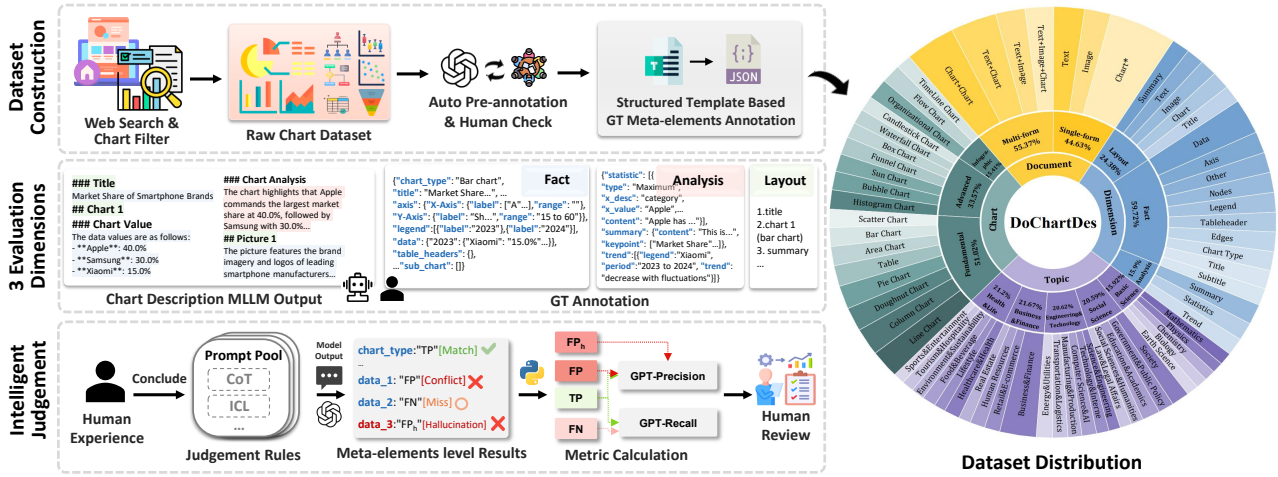


Fig. 1: Overview of DoChartDes. Dataset construction and intelligent evaluation via 3 dimensions: layout understanding, factual correctness and analytical reasoning. Dataset distribution of 5 topics, 18 chart types, 7 document layouts. *denotes chart types match single-chart ones.

- **Fine-grained Evaluation Suite**: We propose a fine-grained and interpretable evaluation suite that enables detailed analysis of chart description. Our suite evaluates 2 core criteria: faithfulness and coverage, along with 3 dimensions to evaluate models' perception (layout understanding, factual correctness) and reasoning (analytical reasoning) capabilities. Not only provides fine-grained evaluation, this suite also unifies faithfulness and coverage metrics at the same element level. It enables accurate detection of hallucinations and localizes the model's gaps in description quality.
- **Efficient Expert-aligned Intelligent Evaluation**: We optimize the evaluation process of DoChartDes to an efficient, fully automated and intelligently precise way. While maintaining operational efficiency, it achieves high consistency rate (Pearson coefficient = 0.8) and low systemic bias (MBE < 2) with human experts.

II. The DoChartDes Benchmark Construction

As shown in Fig 1, DoChartDes is a specialized benchmark designed for open-ended chart description tasks, featuring 18 diverse chart types across 6 languages and 7 document layouts, supporting evaluation from simple charts to complex multi-element slide documents. Our dataset construction workflow encompasses two steps: (i) raw dataset collection, (ii) structured template-based ground truth (GT) annotation.

A. Data Collection

The raw chart dataset is constructed from a diverse range of websites and links across the internet, including public data portals (e.g., statistical bureaus, Tableau), academic platforms (e.g., arXiv), design tools (e.g., Slideshare, Canva) and infographic websites (e.g. Creately, Venngage). Curated from these source, we employs CLIP-based filtering and manual validation to ensure multilingual representation and balanced diversity across chart types, topics and layouts (Fig. 1).

Broad Topic Domains of Real-world. Leveraging its diverse data sources, DoChartDes ensures comprehensive coverage of real-world domains through 5 high-level domains and 26 fine-grained subdomains, including: basic science, engineering & technology, social science, commerce & finance, health & lifestyle. Our dataset addresses limitations of existing chart datasets by integrating diverse sources and ensuring extensive coverage of chart types, thereby enabling broader topical scope and richer stylistic diversity.

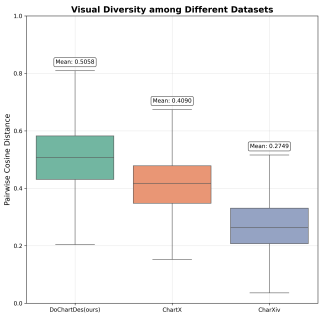
Diverse Chart Types Coverage. DoChartDes encompasses a collection of 18 distinct chart types, categorized into three groups according to their usage prevalence and application domains: **(i) 8 high-frequency fundamental types**: line, area, bar, column, pie, doughnut, scatter and table. They establish baseline requirements for understanding common data patterns. **(ii) 7 advanced types:** sun, bubble, histogram, box, funnel, waterfall, candlestick. Evaluating them addresses the lack of advanced analytical task coverage in current benchmarks. **(iii) 3 infographic types:** flow chart, organizational chart, timeline. Focused on structural storytelling and contextual integration, they capture how charts operate in real documents and fill the gap in layout-aware evaluation. To ensure alignment with real-world usage distribution, we allocate images with a 3:2:1 ratio across these 3 groups.

Multilingual and Multi-layout Inclusiveness. Unlike existing chart understanding benchmarks limited to English and isolated charts, DoChartDes systematically incorporates multilingual and multi-layout documents spanning 6 languages (English, German, French, Spanish, Italian, Portuguese) and 7 modular document layouts (from single-form only to real-world compound modular).

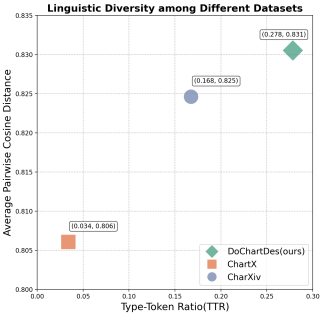
B. Data Annotation

Our GT annotation follows two key principles to ensure robust evaluation, all annotations under cross-verification by human experts.

Structured Annotation for Fine-grained Evaluation. In order to enable fine-grained evaluation across three dimensions, we design a structured JSON template that dissects charts into granular, uniform meta-elements. Each chart is



(a) Box plot of pairwise cosine distances among images.



(b) Scatter chart of TTR and cosine distance across datasets.

Fig. 2: Visual and linguistic diversity comparison across datasets. Ours DoChartDes exhibits the highest diversity.

annotated with an average of over 25 meta-elements (totaling 42.3K annotations across 1625 curated images), with pre-annotations generated by GPT-4o [25] and cross-validated by human experts against image content.

By structuring annotations in JSON, we achieve a trade-off between flexibility in annotation and accuracy in localizing chart description flaws. Our standardized template can also convert natural language chart descriptions into structured format, enabling precise and meta-element level info (false or miss) for MLLM optimization. By decomposing into three dimensions, our GT annotation supports systematic, hierarchical assessment of perception and reasoning competencies in chart description. Fig 1 provides meta-element examples across layout, fact and analysis.

Char Type Specific Customization. Unlike existing benchmarks that apply generic QA formats on diverse chart types, our benchmark customizes GT templates the unique characteristics of each chart type, enabling the key inspection points coverage that may arise across different types. For example, numerical-pattern charts (e.g., line/bar charts) emphasize data values and axis, while structural charts (e.g., flowcharts) prioritize node relationships.

This type specific customization ensures that annotations align with the unique visual and semantic properties of each chart type, avoiding generic annotations that fail to adapt to the heterogeneous nature of real-world charts.

C. Dataset Analysis

Visual Diversity. To rigorously assess and quantify the visual diversity of chart images in our proposed benchmark, we compare DoChartDes with two recently established benchmarks: ChartX [9] and CharXiv [6]. Using CLIP embeddings [26], we compute the average pairwise cosine distance between images within each dataset. Fig 2a confirms this advantage through DoChartDes (0.51) significantly outperforms CharXiv (0.27) and ChartX (0.41). These findings demonstrate DoChartDes provides substantially richer visual heterogeneity for robust chart description evaluation.

Linguistic Diversity. To further characterize linguistic richness, we employ GPT-4.1 [25] OCR to extract text and analyze lexical and semantic diversity. DoChartDes achieves 0.278 Type-Token Ratio (TTR) and 0.831 cosine distance of Sentence-BERT embeddings [27]), reflecting broader vocabulary and conceptual variability. Fig 2b shows that DoChartDes performs better on both two dimensions, demonstrating the benchmark's linguistic diversity.

III. Evaluation Strategy

Current chart understanding evaluation faces two core challenges with traditional metrics: (i) Unlike computer vision's pixel-level evaluation, description lacks unified elements, leading to inconsistent metrics calculation due to varying element granularity. (ii) The absence of hallucination scoring hinders reliability assessment.

To address these, we propose a fine-grained evaluation suite that combines coverage and faithfulness as shown in Figure 3. By unifying element granularity and hallucination detection, this suite enhances metrics consistency and interpretability. We assesses MLLMs' two core capabilities: perception (layout/fact) and reasoning (analysis), all quantitatively evaluated via coverage and faithfulness criteria.

A. Template-based Meta-elements Matching

We establish chart-type-specific structured meta-element annotation templates, defining a meta-element $e = (c, d, a, v)$ and a template $\mathcal{T} = \{e_i\}_{i=1}^{T}$. $c \in \{perception, reasoning\}$ denotes the core capability required to process the meta-element: $perception$ corresponds to recognizing observable elements in the chart, whereas $reasoning$ corresponds to deriving implicit insights from the chart content. $d \in \{layout, fact, analysis\}$ denotes the evaluation dimension, corresponding to layout organization, factual element extraction, and analytical interpretation, respectively. a is the attribute identifier referring to a specific chart meta-element (e.g., `"x-axis label"`). v is the corresponding attribute value. T denotes the total number of potential elements in the template for a given chart type.

The GT annotation set $\mathcal{G}$ derived from applying $\mathcal{T}$ to a real chart image, generating its meta-elements set:

$$\mathcal{G} = \{g_k = (c_k, d_k, a_k, v_k^{\text{GT}})\}_{k=1}^{K} : \tag{1}$$

where K is actual meta-elements quantity in the chart image.

The matching mechanism has two phases: extraction and matching. Firstly, based on the structured meta-elements annotation template $\mathcal{T}$, we utilize MLLM as extractor $\theta_{\text{MLLM}}^{\text{Ext}}$ to standardize model output $\mathcal{O}$ to a unified level output meta-elements set $\tilde{\mathcal{O}}$:

$$\tilde{\mathcal{O}} = \theta_{\text{MLLM}}^{\text{Ext}}(\mathcal{O}, \mathcal{T}) = \{o_i = (c_i, d_i, a_i, v_i)\}_{i=1}^{N} \tag{2}$$

where N is the total meta-elements number in $\tilde{\mathcal{O}}$.

We acknowledge the critical role of the NL-to-meta-element extraction step. Early experiments revealed that free-form parsing can be inconsistent across models and languages, motivating our structured, chart-type-specific templates (Section III) to standardize meta-element extraction and ensure cross-model and cross-lingual consistency. The strong alignment between automated and human scores (Table III) indicates that extraction errors are neither significant nor systematic; moreover,

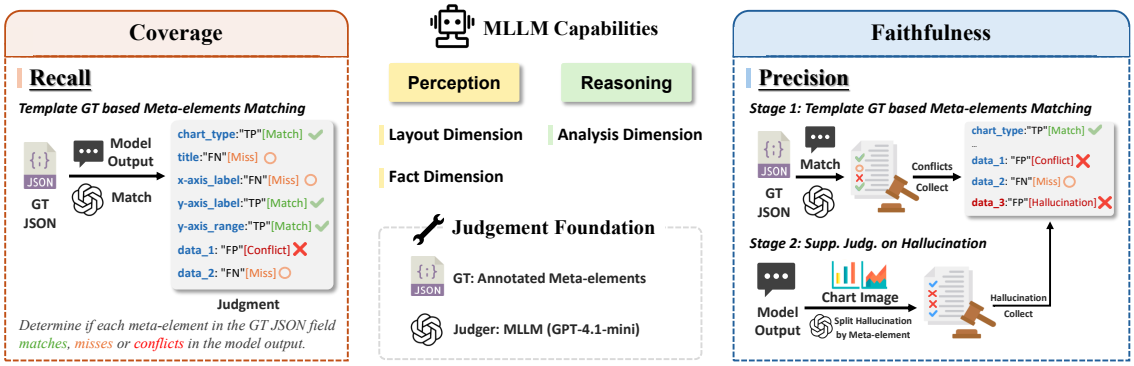


Fig. 3: DoChartDes's Chart Description Evaluation Suite via 2 Core Criteria: Faithfulness and Coverage

applying the same extraction protocol to all models preserves ranking fairness.

Secondly, we equip the MLLM judger $\theta^{\text{Jud}}_{\text{MLLM}}$ with a comprehensive prompt pool as the rule set $\mathcal{R}$, which formalizes detailed human expert evaluation experience. Guided by rule set, MLLM execute judgment at meta-element level, evaluating each meta-element g_k in GT set by comparing it with the corresponding element in the standardized output set $\tilde{\mathcal{O}}$:

$$\mathcal{J}(g_k) = \theta^{\text{Jud}}_{\text{MLLM}} \begin{pmatrix} g_k \\ \tilde{\mathcal{O}} \\ \mathcal{R} \end{pmatrix} \rightarrow \begin{cases} \text{TP}: & \text{Correct match} \\ \text{FP}: & \text{Value conflict} \\ \text{FN}: & \text{Missed} \end{cases} \quad (3)$$

B. Supplemental Judgment of Hallucination

Template-based evaluation identifies matches, misses, and conflicts to compute metrics, but overlooks hallucinations generated by MLLMs that neither exist nor can be inferred from the image. Approaches rely on GPT to segment output for precision computation, reflecting the faithfulness of MLLM. It has two limitations: granularity misalignment between precision and recall (due to uncertainty of decomposition) and inability to quantify hallucinations.

To address these challenges, hallucination detection employs a multi-stage pipeline and is not decided solely by matching output meta-elements to the GT template. We define hallucination as an independent error category (FP_h), including attribute ($a_i \notin \mathcal{T}$) and value hallucinations ($v_i \notin \mathcal{I}$).

$$\mathcal{J}(o_i) = \theta^{\text{Jud}}_{\text{MLLM}}(o_i, \mathcal{I}, \mathcal{R}) \rightarrow \text{FP}_h : a_i \notin \mathcal{T} \vee v_i \notin \mathcal{I} \quad (4)$$

First, the judger takes the whole image $\mathcal{I}$ as input; the template $\mathcal{T}$ enforces objective, aligned granularity; and the extracted meta-elements are compared with the ground truth. We explicitly distinguish visually grounded inferences (e.g., inferring maximum from the tallest bar) from hallucinations (e.g., fabricating absent categories or values), ensuring mutual exclusivity. Finally, we remove duplicate detections to prevent overcounting, enabling accurate metric computation.

C. Metrics of Coverage and Faithfulness

To quantify chart description quality under our meta-element matching protocol, we report two complementary metrics: GPT-Recall and GPT-Precision. GPT-Recall characterizes coverage by measuring the proportion of ground-truth meta-elements correctly recovered, and it decreases with both omissions (FN) and value conflicts (FP). GPT-Precision characterizes faithfulness by measuring the proportion of predicted meta-elements supported by the chart, while penalizing both conflicts (FP) and hallucinations (FP_h). Accordingly, we compute the metrics from TP, FP, FN, and FP_h as follows.

$$\text{GPT-Recall} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \quad (5)$$

$$\text{GPT-Precision} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FP}_h} \quad (6)$$

IV. Experiment

A. Experimental Setup

Models. We benchmark a diverse closed-source and open-source models in chart description. Closed-source models include: state-of-the-art performers and their efficiency-optimized variants: (i) GPT-4.1 and GPT-4.1-mini [25], (ii) GPT-4o and GPT-4o-mini, (iii) Gemini 2.0 Flash [28]; Open-source models are categorized by parameter size. Models with parameters greater than 50B: (i) Qwen2.5-VL-72B [29], (ii) InternVL3-78B [30]. In the 7-15B parameter range, we evaluate:(i) Qwen2.5-VL-7B [29], (ii) InternVL3-8B [30], (iii) Ovis2-8B [31], (iv) Gemma3-12B [32], (v) Phi-4-multimodal-instruct [33]. In addition, we also evaluate chart-specific MLLMs: (i) ChartGemma [34], (ii) TinyChart [35].

Implementation Details. We employ GPT-4.1-mini for intelligent evaluation, chosen for its cost–accuracy balance. Our method exhibits no GPT bias and supports easy substitution of evaluators. Hyperparameters mainly follow VLMEvalKit [36] settings. Using 20 parallel threads, this benchmark can process approximately 100 images in 5 minutes. To trade off between faithfulness and coverage, we use the F_1 score as the harmonic metric of GPT-Precision and GPT-Recall.

B. Results Analysis

We conducted extensive experiments with our proposed evaluation suite. The main results are presented in Table II.

Meta-element Level Analysis To demonstrate the distinct strengths and weaknesses of MLLM, we conducted a detailed meta-element level analysis of the top-performing models from

TABLE II: Evaluation results F_1 score (%) on DoChartDes across languages (main headers) and evaluation subsets (sub-headers). For each language, results are reported across 3 fine-grained dimensions: layout understanding, factual correctness and analytical reasoning. Models are grouped into 4 categories for comparison. The highest score within each model category is highlighted in bold.

Model	English					German					French					Spanish					Italian					Portuguese				
	Chart		Document			Chart		Document			Chart		Document			Chart		Document			Chart		Document			Chart		Document		
	Fac.	**Ana.**	**Lay.**	**Fac.**	**Ana.**	**Fac.**	**Ana.**	**Lay.**	**Fac.**	**Ana.**	**Fac.**	**Ana.**	**Lay.**	**Fac.**	**Ana.**	**Fac.**	**Ana.**	**Lay.**	**Fac.**	**Ana.**	**Fac.**	**Ana.**	**Lay.**	**Fac.**	**Ana.**	**Fac.**	**Ana.**	**Lay.**	**Fac.**	**Ana.**
Closed-Source Models																													
Gemini-2.0-Flash	87.6	75.3	86.0	86.7	72.9	85.3	75.8	88.2	81.5	74.6	87.7	81.0	86.2	88.4	87.5	90.5	79.1	87.9	86.8	80.8	87.3	77.1	88.4	87.9	81.3	86.7	76.6	87.1	89.9	80.9
GPT-4o	87.1	77.0	87.5	84.9	72.2	85.4	73.4	90.8	84.1	77.3	83.7	75.0	86.6	87.0	87.7	87.1	71.4	88.9	87.8	78.2	84.9	72.6	**91.6**	89.4	78.2	85.2	70.8	88.4	87.8	81.0
GPT-4o-mini	**79.8**	66.8	84.5	**78.4**	**68.6**	77.0	63.6	88.2	74.2	68.9	76.7	70.6	84.3	81.1	82.0	79.0	66.7	**90.3**	83.1	70.4	74.2	67.7	89.2	81.7	76.1	77.0	63.8	87.4	82.0	74.8
GPT-4.1	**89.0**	77.1	**88.2**	87.7	**75.3**	**90.5**	**77.8**	**91.4**	86.9	77.9	88.2	78.7	**87.0**	**89.0**	86.7	**91.2**	79.3	89.0	87.2	80.1	88.2	76.0	90.7	88.9	**81.9**	88.3	74.5	**88.8**	**91.1**	82.3
GPT-4.1-mini	**88.4**	**78.2**	**87.2**	**87.8**	**74.6**	90.0	76.9	91.0	**88.5**	**78.7**	**89.0**	**82.0**	86.9	88.2	**88.0**	90.2	**81.2**	87.3	**90.6**	**84.8**	**89.0**	**77.8**	90.6	**89.3**	81.1	**89.7**	**77.5**	88.3	89.6	**82.4**
Open-Source Models																													
Qwen2.5-VL-72B	83.1	74.0	**85.9**	81.6	71.9	81.5	68.8	77.6	80.9	70.0	85.0	76.2	82.4	**85.0**	83.6	**88.3**	76.1	83.0	**85.5**	**82.1**	**82.4**	72.6	**86.0**	**84.6**	78.2	**86.5**	73.3	87.1	**86.4**	**80.8**
InternVL3-78B	**87.1**	**75.5**	85.8	**84.4**	**75.0**	**83.2**	**75.3**	**89.5**	**81.3**	**76.8**	**85.7**	**80.2**	**83.5**	84.8	**87.0**	87.3	**78.1**	**86.2**	83.2	80.1	82.3	**75.2**	85.9	82.1	**78.6**	83.6	**75.5**	**87.2**	86.3	**80.8**
Qwen2.5-VL-7B	82.2	**71.3**	81.9	74.9	**71.6**	74.9	**68.3**	82.4	**75.8**	67.3	76.0	**74.6**	84.5	77.0	**78.9**	81.3	**72.5**	80.7	79.9	**74.7**	77.0	69.9	83.4	**79.9**	76.5	**81.6**	68.1	84.8	79.0	77.0
InternVL3-8B	**84.6**	69.5	83.7	**79.8**	71.4	**78.1**	68.1	84.4	72.6	**67.7**	79.0	72.4	81.8	**78.9**	81.6	81.9	71.0	82.0	78.7	73.7	78.2	**70.9**	83.1	75.1	**77.6**	81.2	69.5	84.9	**82.4**	**79.8**
Ovis2-8B	80.9	68.3	82.8	74.2	68.4	70.7	67.3	85.4	67.5	65.3	73.0	73.8	83.0	71.2	80.9	78.9	69.7	82.5	74.3	73.5	73.3	67.8	85.3	70.9	70.7	74.9	**70.3**	86.4	73.6	75.7
Gemma3-12B	82.1	68.6	**84.9**	74.6	66.5	76.1	63.3	**87.3**	67.7	65.0	**81.8**	69.9	**85.6**	77.9	80.3	**83.7**	70.7	**87.7**	**80.5**	73.2	**79.7**	68.7	**88.1**	78.8	72.9	78.0	69.5	**87.1**	78.9	74.6
Phi-4-multimodal	74.2	68.2	84.1	69.5	67.1	56.1	59.2	79.2	44.5	61.2	57.2	68.1	76.7	59.3	73.0	60.2	62.9	78.5	58.9	63.1	40.3	53.6	68.1	36.1	55.6	52.3	53.1	80.9	51.9	65.5
Chart-Specific Models																													
ChartGemma	**32.1**	**32.9**	**67.6**	**25.7**	**35.8**	**16.0**	**28.4**	**69.6**	**20.2**	**34.6**	**23.1**	**29.2**	**70.9**	**24.4**	**40.4**	**23.3**	**32.0**	**69.3**	**18.4**	**26.9**	**17.8**	**21.7**	**67.5**	**14.5**	**32.6**	**16.9**	**24.2**	**66.3**	**17.2**	**29.6**
TinyChart	22.9	25.3	56.1	17.0	22.0	13.0	11.6	50.4	8.4	9.9	17.4	12.4	52.6	14.5	6.3	17.6	13.9	50.5	11.9	10.0	13.5	12.8	49.6	10.0	11.1	15.9	11.3	52.5	9.4	7.9

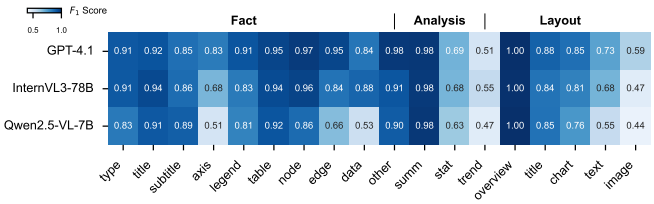


Fig. 4: Heatmap of meta-element level F_1 score comparisons among the SOTA models across categories. summ: summary, stat: statistic.

3 groups. As shown in Fig 4, GPT-4.1 maintains balanced performance across all dimensions. while InternVL3-78B excels in fact dimension (e.g., "data" and "node"), it underperforms in "axis" and "image" parsing compared to GPT-4.1. Meanwhile, Qwen2.5-VL-7B shows balanced performance but identifies areas for enhancement. These findings offer valuable insights for optimization in specific meta-elements.

Results Observations. Closed-source models lead overall, while open-source models catching up. Leading closed-source models, like GPT-4.1, achieve top-tier performance across nearly all tasks. Concurrently, premier open-source models like InternVL and Qwen follow closely. Chart-specific models significantly underperform general-purpose MLLMs.

All models excel at factual descriptions but struggle with analytical reasoning. Both closed-source and open-source models score higher on perception (Lay. and Fac.)than reasoning (Ana.). Meanwhile, open-source models are competitive but lag in reasoning, pointing to a future development direction and underscoring DoChartDes as a challenging benchmark.

Evaluation Consistency Analysis. We sampled 200 images, ensuring balanced coverage across chart types, language and difficulty. Six language-proficient evaluators cross validated description generated by Qwen2.5VL-72B, GPT-4o-mini and Gemini 2.0 Flash. We used 7 metrics to assess ranking consistency, individual numerical error and overall bias between intelligent and human evaluation: Pearson correlation coefficient (PCC) r, Spearman's rank correlation coefficient ρ, Kendall's τ (Kd τ), Sample τ (Sp τ), root mean square error (RMSE), mean absolute error (MAE) and mean bias error (MBE). Sp τ computes Kd τ for each sample's metrics independently, and use the average value as final result. In addition to our intelligent evaluation suite, we also apply GPT-score for comparison, a metric previously used in chart description tasks [8], [9].

TABLE III: Comparison of intelligent evaluation consistency with human experts. Our method demonstrates significantly higher correlation and lower error than GPT-score.

Metric	PCC r	Sp ρ	Kd τ	Sp τ	RMSE	MAE	MBE
GPT-score [9]	0.09	0.17	0.14	0.05	25.61	19.80	13.84
Ours	0.80	0.65	0.56	0.42	10.10	5.62	1.73

To address parsing inconsistencies observed in preliminary free-form extraction experiments, we introduce chart-type-specific structured templates that ensure robust, consistent meta-element extraction (Section III-A). As shown in Table III, our method outperforms GPT-score across all dimensions: strong ranking consistency with human evaluation (PCC $r = 0.80$, Sp $\rho = 0.65$, Kendall $\tau = 0.56$) and substantial error reduction (60.6% in RMSE, 71.6% in MAE and 87.5% in MBE). Strong human alignment validates both the extraction pipeline's reliability and framework's fidelity.

V. Conclusion

Chart description in document contexts is pivotal for multimodal data-to-text, current benchmarks remain constrained by closed-response paradigms that fail to capture real-world complexity. We bridge this critical gap with DoChartDes, the benchmark designed for fine-grained chart description evaluation, featuring 3 contributions: (i) Meaningful chart diversity with multilingual and multi-layout; (ii) Fine-grained evaluation suite: metrics assess faithfulness and coverage through structured annotations; (iii) Efficient expert-aligned intelligent evaluation process: high consistency with human experts while reducing time. DoChartDes sets a direction for MLLM evaluation, enabling precise weaknesses identification.

REFERENCES

[1] K. Davila, S. Setlur, D. Doermann, B. U. Kota, and V. Govindaraju, "Chart mining: A survey of methods for automated chart analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 3799–3819, 2021.

[2] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang, "Towards natural language interfaces for data visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, p. 3121–3144, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2022.3148007

[3] K.-H. Huang, H. P. Chan, Y. R. Fung, H. Qiu, M. Zhou, S. Joty, S.-F. Chang, and H. Ji, "From pixels to insights: A survey on automatic chart understanding in the era of large foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2403.12027

[4] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning," Mar. 2022, arXiv:2203.10244 [cs]. [Online]. Available: http://arxiv.org/abs/2203.10244

[5] Z. Xu, S. Du, Y. Qi, C. Xu, C. Yuan, and J. Guo, "ChartBench: A Benchmark for Complex Visual Reasoning in Charts," Jan. 2024, arXiv:2312.15915 [cs] version: 2. [Online]. Available: http://arxiv.org/abs/2312.15915

[6] Z. Wang, M. Xia, L. He, H. Chen, Y. Liu, R. Zhu, K. Liang, X. Wu, H. Liu, S. Malladi, A. Chevalier, S. Arora, and D. Chen, "CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs," Jun. 2024, arXiv:2406.18521. [Online]. Available: http://arxiv.org/abs/2406.18521

[7] S. Kantharaj, R. T. K. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty, "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization," Apr. 2022, arXiv:2203.06486 [cs]. [Online]. Available: http://arxiv.org/abs/2203.06486

[8] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang, "ChartLlama: A Multimodal LLM for Chart Understanding and Generation," Nov. 2023, arXiv:2311.16483. [Online]. Available: http://arxiv.org/abs/2311.16483

[9] R. Xia, B. Zhang, H. Ye, X. Yan, Q. Liu, H. Zhou, Z. Chen, M. Dou, B. Shi, J. Yan, and Y. Qiao, "ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning," Feb. 2024, arXiv:2402.12185 [cs]. [Online]. Available: http://arxiv.org/abs/2402.12185

[10] S. E. Kahou, V. Michalski, A. Atkinson, A. Kadar, A. Trischler, and Y. Bengio, "FigureQA: An Annotated Figure Dataset for Visual Reasoning," Feb. 2018, arXiv:1710.07300 [cs]. [Online]. Available: http://arxiv.org/abs/1710.07300

[11] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, T. Yu, R. Rossi, and R. Bunescu, "Figure captioning with reasoning and sequence-level training," 2019. [Online]. Available: https://arxiv.org/abs/1906.02850

[12] F. Liu, X. Wang, W. Yao, J. Chen, K. Song, S. Cho, Y. Yacoob, and D. Yu, "MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning," Apr. 2024, arXiv:2311.10774 [cs]. [Online]. Available: http://arxiv.org/abs/2311.10774

[13] Z. Xiao, S. Zhang, V. Lai, and Q. V. Liao, "Evaluating evaluation metrics: A framework for analyzing nlg evaluation metrics using measurement theory," 2023. [Online]. Available: https://arxiv.org/abs/2305.14889

[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013/

[15] 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*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[16] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," 2015. [Online]. Available: https://arxiv.org/abs/1411.5726

[17] T. Sellam, D. Das, and A. P. Parikh, "Bleurt: Learning robust metrics for text generation," 2020. [Online]. Available: https://arxiv.org/abs/2004.04696

[18] Z. Luo, Q. Xie, and S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for text summarization," 2023. [Online]. Available: https://arxiv.org/abs/2303.15621

[19] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," 2023. [Online]. Available: https://arxiv.org/abs/2303.16634

[20] H. Song, H. Su, I. Shalyminov, J. Cai, and S. Mansour, "Finesure: Fine-grained summarization evaluation using llms," 2024. [Online]. Available: https://arxiv.org/abs/2407.00908

[21] K.-H. Huang, P. Laban, A. R. Fabbri, P. K. Choubey, S. Joty, C. Xiong, and C.-S. Wu, "Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles," 2024. [Online]. Available: https://arxiv.org/abs/2309.09369

[22] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," 2020. [Online]. Available: https://arxiv.org/abs/2004.04228

[23] K.-H. Huang, S. Singh, X. Ma, W. Xiao, F. Nan, N. Dingwall, W. Y. Wang, and K. McKeown, "Swing: Balancing coverage and faithfulness for dialogue summarization," 2023. [Online]. Available: https://arxiv.org/abs/2301.10483

[24] H. Qiu, K.-H. Huang, J. Qu, and N. Peng, "AMRFact: Enhancing summarization factuality evaluation with AMR-driven negative samples generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 594–608. [Online]. Available: https://aclanthology.org/2024.naacl-long.33/

[25] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin *et al.*, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," Feb. 2021, arXiv:2103.00020 [cs]. [Online]. Available: http://arxiv.org/abs/2103.00020

[27] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[28] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk *et al.*, "Gemini: A family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805

[29] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang *et al.*, "Qwen2.5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[30] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian *et al.*, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2504.10479

[31] S. Lu, Y. Li, Q.-G. Chen, Z. Xu, W. Luo, K. Zhang, and H.-J. Ye, "Ovis: Structural embedding alignment for multimodal large language model," 2024. [Online]. Available: https://arxiv.org/abs/2405.20797

[32] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[33] Microsoft, A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," 2025. [Online]. Available: https://arxiv.org/abs/2503.01743

[34] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty, "Chartgemma: Visual instruction-tuning for chart reasoning in the wild," 2024. [Online]. Available: https://arxiv.org/abs/2407.04172

[35] L. Zhang, A. Hu, H. Xu, M. Yan, Y. Xu, Q. Jin, J. Zhang, and F. Huang, "Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning," 2024. [Online]. Available: https://arxiv.org/abs/2404.16635

[36] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang *et al.*, "Vlmevalkit: An open-source toolkit for evaluating large multi-modality models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 198–11 201.