NeurIPS 2026
Poster

ChartArena: A Unified Benchmark with Atomic-Primitive Reasoning for Chart Parsing

Shangpin Peng[1, 4, 5, *] Gengluo Li[2, *] Xingyu Wan[1] Chengquan Zhang[1, †] Hao Feng[1] Binghong Wu[1]
Huawen Shen[1] Weinong Wang[1] Ziyi Cai[4] Zhuotao Tian[4,✉] Han Hu[1] Can Ma[2] Yu Zhou[3,✉]

1 Tencent LLM Department
2 Institute of Information Engineering, CAS
3 Nankai University
4 Shenzhen Loop Area Institute
5 Hong Kong University of Science and Technology

* Equal contribution
† Project leader
✉ Corresponding author

github.com/pspdada/ChartArena

SCAN FOR CODE & DATA

01 Introduction — Motivation

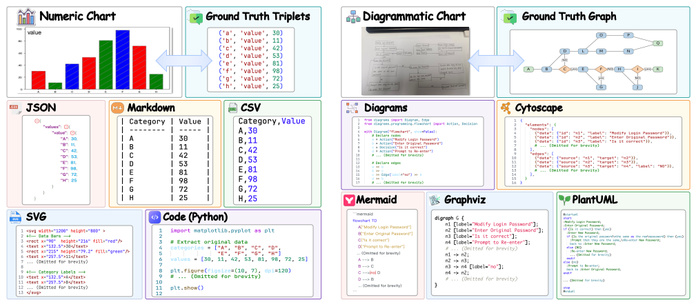
Charts carry dense quantitative and relational information, and **chart parsing** turns them into machine-readable structure for analysis and reasoning. Building a truly general parser is still hard, mainly for two reasons.

1 Fragmented evaluation

Outputs come in disjoint formats (Markdown, JSON, CSV, SVG, code, Mermaid). Benchmarks cover narrow sub-domains and pristine synthetic images, not real-world photos.

2 Limited structural reasoning

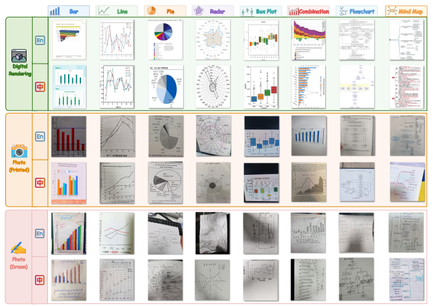
End-to-end models map pixels straight to text, skipping axis grounding and relational topology. They learn shallow mappings and break under real-world shifts.



Heterogeneous output formats for numeric charts (left) and diagrammatic charts (right)

02 The ChartArena Benchmark — Benchmark

2,400	8	3	EN + ZH
charts	chart families	visual scenarios	bilingual



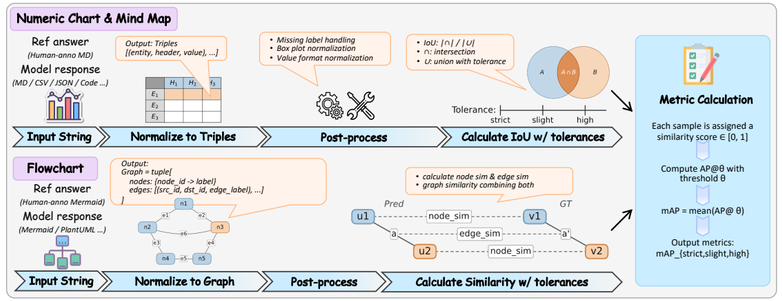
8 chart families × 3 visual scenarios × 2 languages, 50 samples per setting

Human-agent co-annotation. MLLMs draft Markdown (numeric) or Mermaid (diagrammatic) labels, then annotators refine structure and content over multiple stages; hard-to-read values are cross-checked by several annotators.

Coverage gap. Prior benchmarks cover at most 5 of the 8 families, use digital images only, and are English-only except ChartY.

03 Unified Evaluation Protocol — Evaluation

Predictions and references are normalized to a **triple view** (numeric charts, mind maps) or a **graph view** (flowcharts), then scored with **tolerance-aware IoU** (strict, slight, high; EM and mAP) or node and edge similarity, so that results stay comparable across output formats.



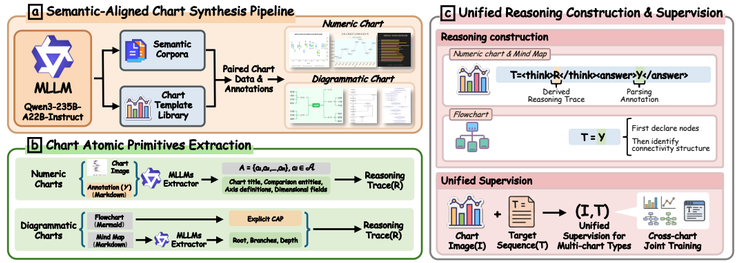
Inputs are normalized to triples or graphs, then scored under three tolerances

04 Chart Atomic Primitives — Method

CAP is a compact semantic scaffold (title, entities, axes, dimensions, key statistics) without low-level visual detail. The model infers it first, then writes the final parse, giving it explicit structure.

$$A = \{a_1, \ldots, a_n\},\ a_i \in \mathcal{A} \qquad R = \mathcal{G}(A)$$

$$T = \texttt{<think>}\ R\ \texttt{</think>}\ \texttt{<answer>}\ Y\ \texttt{</answer>}$$



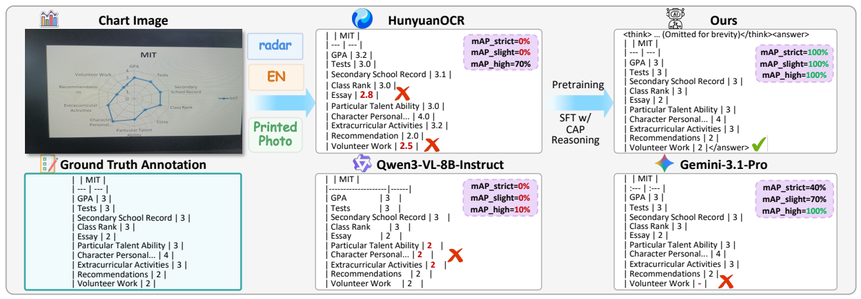
05 Experiments and Results — Experiments

Main results on ChartArena (mAP_{high})

Subset of the main table; Doc. = document parsing MLLMs

	Model	bar EN	bar ZH	line EN	line ZH	pie EN	pie ZH	radar EN	radar ZH	box plot EN	box plot ZH	comb. EN	comb. ZH	flowchart EN	flowchart ZH	mind map EN	mind map ZH	Average EN	Average ZH
General	GPT-5	35.1	52.3	48.1	65.1	81.1	78.9	**32.0**	41.5	19.8	12.8	14.2	46.5	58.1	35.3	76.6	33.5	45.6	45.8
General	Qwen3-VL-8B-Ins.	27.5	58.6	35.5	61.1	77.3	84.7	16.8	42.6	11.6	12.1	13.2	47.9	50.0	41.5	66.4	54.6	37.3	50.4
General	Qwen3-VL-235B-A22B	38.4	67.9	52.3	73.8	82.6	85.5	23.2	52.4	14.1	14.1	29.1	58.2	57.9	49.8	70.8	65.2	46.0	58.4
General	Qwen3.5-35B-A3B (think)	46.2	65.3	60.3	77.6	89.7	88.4	25.2	57.8	42.2	50.6	31.5	56.9	62.5	56.5	75.1	70.9	54.1	65.5
General	Seed2.0 Pro	40.3	73.3	56.5	80.7	91.5	90.5	21.3	54.7	**44.5**	**55.2**	32.4	62.2	62.6	61.3	83.1	**85.8**	54.0	70.5
General	Kimi K2.5	45.2	70.3	60.9	79.8	87.2	86.7	30.2	59.7	40.6	47.6	33.6	63.6	59.9	57.9	80.8	79.4	54.8	68.1
General	Gemini 3.1 Pro	**57.9**	**78.7**	**67.0**	**85.3**	**92.5**	**95.1**	31.8	**62.7**	32.5	45.2	**39.7**	**70.3**	**65.6**	**63.1**	**86.8**	85.2	**59.2**	**73.2**
Doc.	dots.mocr (3B)	28.3	40.9	41.8	60.1	68.8	**78.3**	**20.3**	**43.1**	24.1	16.0	**26.9**	47.1	26.2	20.6	28.7	19.6	33.1	40.7
Doc.	PaddleOCR-VL (1B)	31.8	49.3	43.0	51.6	57.5	75.2	14.4	29.0	11.7	20.7	21.3	**54.0**	--	--	--	--	22.5	35.0
Doc.	HunyuanOCR (1B)	**32.4**	**58.9**	**48.6**	**68.0**	**70.8**	72.5	19.0	40.9	**43.9**	**43.9**	20.1	48.3	**38.3**	**35.5**	**55.0**	**46.6**	**41.0**	**51.8**
Expert	RRVF (7B)	35.8	**66.5**	41.5	54.3	51.6	75.3	16.6	40.3	14.7	14.1	23.5	**61.2**	36.4	32.4	**68.4**	**63.8**	36.0	51.0
Expert	MSRL (7B)	32.7	45.2	35.2	34.3	41.2	67.9	25.9	48.0	11.2	13.0	16.7	35.2	23.2	12.4	31.0	18.8	27.1	34.3
Expert	**Ours (1B)**	43.2	63.8	**57.8**	**73.5**	**74.3**	**82.0**	**28.1**	**54.7**	**54.3**	**59.1**	**26.3**	55.1	**50.8**	**51.3**	61.4	50.9	**49.5**	**61.3**

Qualitative result: printed-photo radar chart (EN)



After pretraining and CAP-based SFT, our model recovers every value; Gemini 3.1 Pro replaces uncertain entries with "-".

Ablation on training strategy

Setting	Numeric charts: Digital	Numeric charts: Printed	Numeric charts: Drawn	Diagrammatic charts: Digital	Diagrammatic charts: Printed	Diagrammatic charts: Drawn
MLLM backbone (HunyuanOCR)	56.3	49.7	37.4	55.0	50.4	26.0
(a) + Pretrain	62.5	50.2	38.1	58.9	51.8	27.7
(b) + SFT, direct (no CAP)	63.1	57.8	44.9	59.6	55.8	37.9
(c) + SFT, CAP on numeric only	64.2	60.8	47.8	60.1	56.3	38.5
(d) + SFT, CAP on diagrams only	62.9	58.2	44.3	62.3	57.9	41.2
(e) + SFT, CAP on both	**64.1**	**60.4**	**48.6**	**62.5**	**58.1**	**40.9**

Key findings

- **Best expert parser.** With only 1B parameters, our model reaches 49.5 EN / 61.3 ZH on average, the highest among all expert chart parsers.
- **Beats 8B generalists.** It surpasses InternVL3.5-8B (29.3 / 40.2) and Qwen3-VL-8B (37.3 / 50.4) on the overall average, and is competitive with much larger open models.
- **CAP adds robustness.** CAP reasoning lifts hand-drawn numeric charts from 44.9 to 48.6 and hand-drawn diagrams from 37.9 to 40.9.
- **A gap remains.** Frontier models such as Gemini 3.1 Pro still lead (59.2 / 73.2), mostly on mind maps and knowledge-heavy slices.

Conclusion

1. **ChartArena** unifies eight chart families, three visual scenarios, and two languages.
2. The **format-agnostic protocol** makes outputs in any format directly comparable.
3. **CAP reasoning** lets a 1B parser infer structure first and parse more robustly.

Limitations. Single-page images only (no multi-page charts); scatter plots are not covered yet.