HKUST CIVAL

Chart-LLaVA: Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning

Xingchen Zeng
The Hong Kong University of Science and Technology (Guangzhou)

20/05/2025

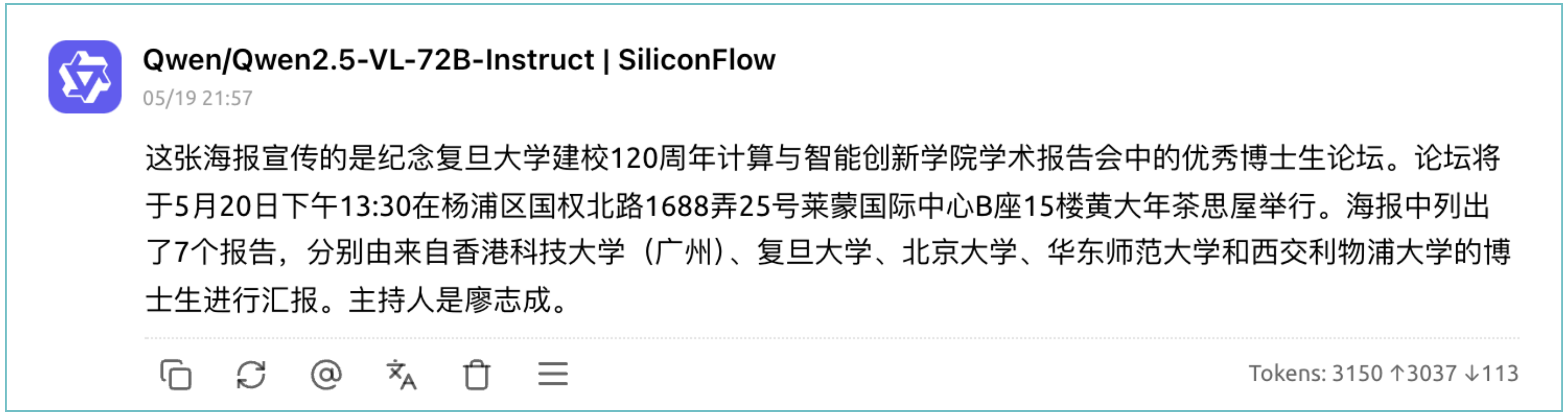
Contents

Contents

Introduction

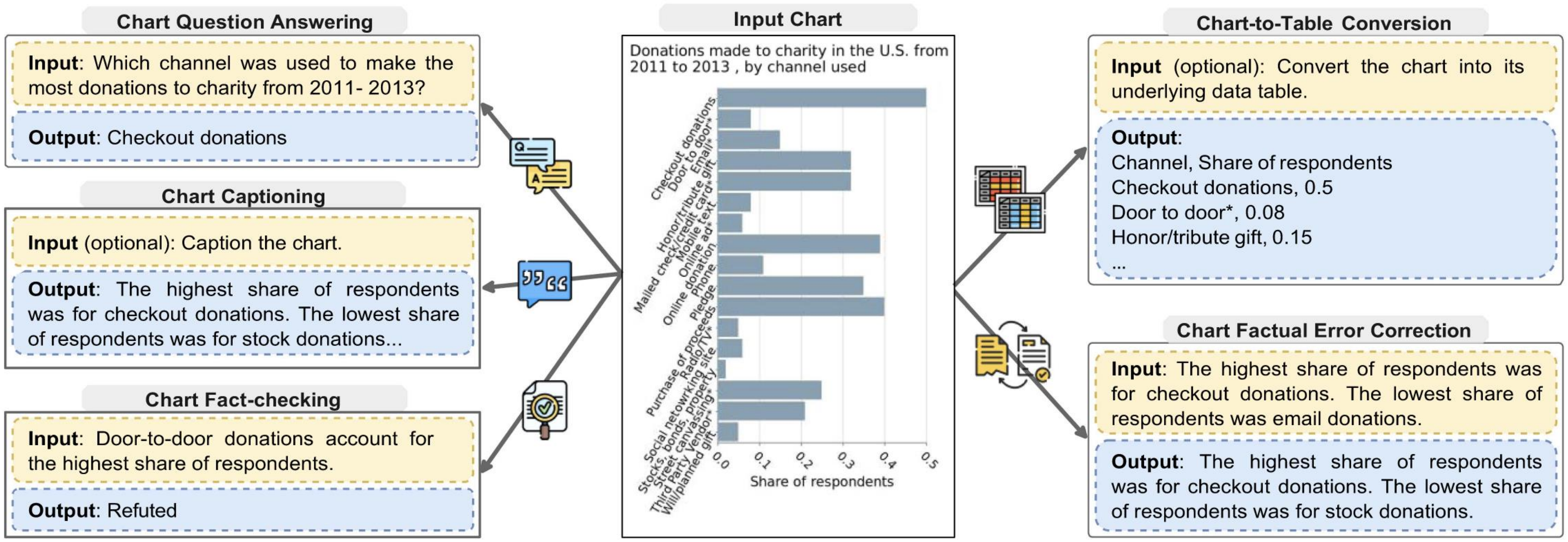
Background: Multimodal LLM

- Multimodal Large Language Models (*e.g.*, GPT4-Vision) have made remarkable strides in understanding and interpreting natural images, enabling breakthroughs in various vision-language tasks.



Background: Automatic Chart Understanding

- Automatic chart understanding is the process of using AI techniques to interpret and extract meaningful information from charts.

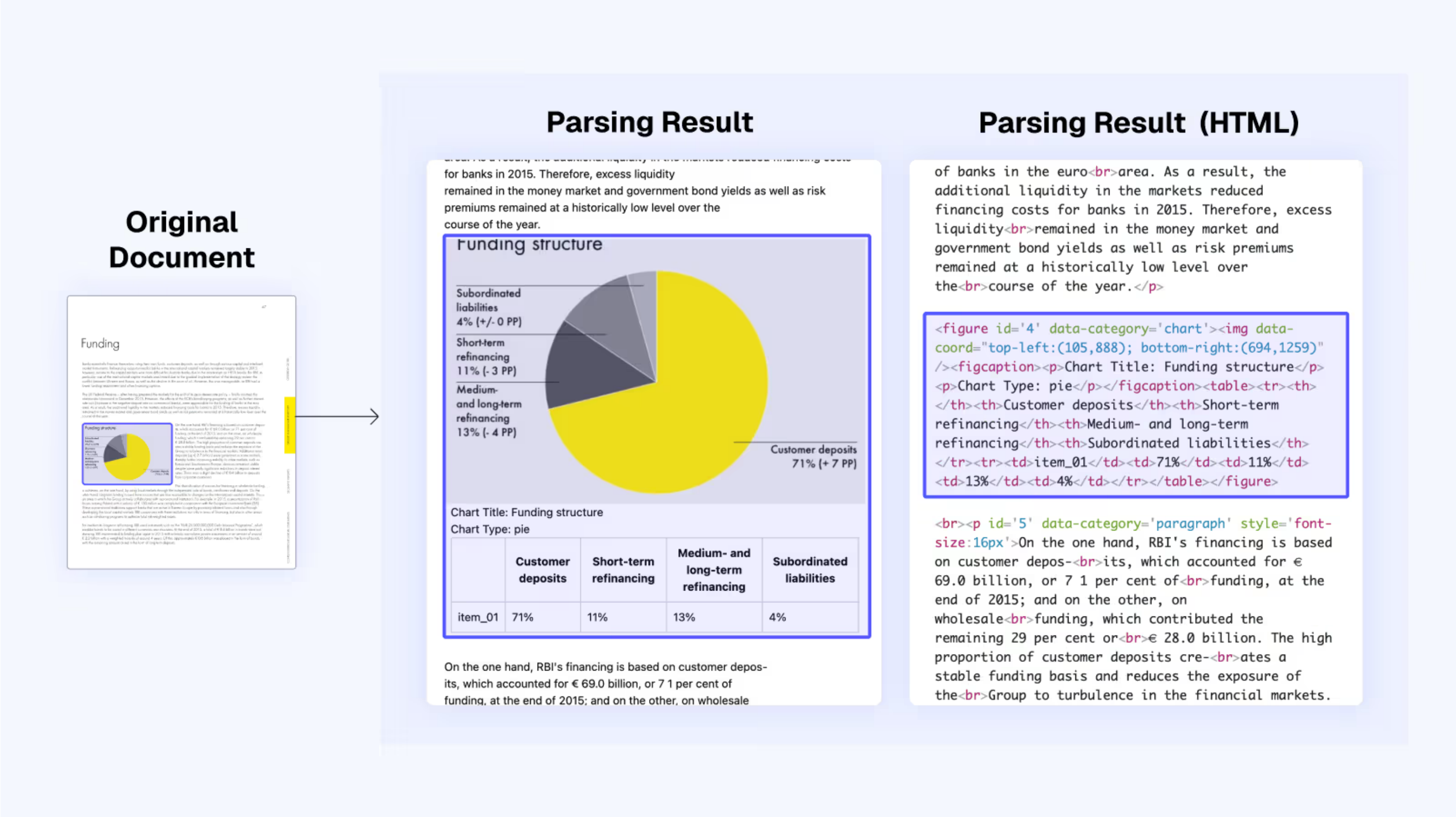


[Huang et. al, TKDE2025]

Introduction

Background: MLLMs-driven Chart Understanding

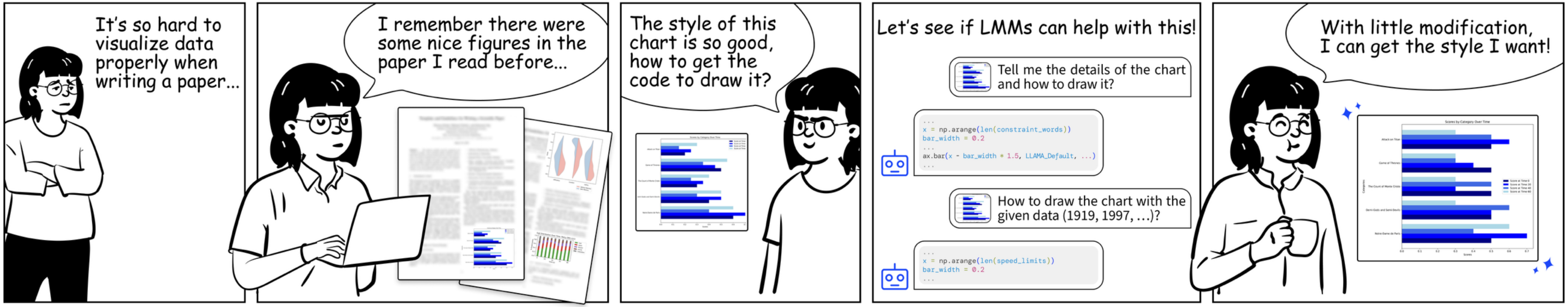
- MLLMs has revolutionized automatic chart understanding and given rise to a wide range of popular real-world applications.



[Turn Charts into LLM-Actionable Data, UpStage, 2024]

Background: MLLMs-driven Chart Understanding

- MLLMs has revolutionized automatic chart understanding and given rise to a wide range of popular real-world applications.



[ChartMimic: Chart-to-code translation, Shi et. al, ICLR2025]

Significance: MLLMs-driven Chart Understanding

- **Versatile application scenarios.**
  - Charts are ubiquitous in scientific papers, financial reports, and news articles.
  - Chart2Table, Chart2Code, Chart Captioning, Chart Question Answering....

- **Appropriate for benchmarking MLLMs' progress.**
  - MLLMs need to perform complex reasoning over numerical data, textual labels, and complex visual elements to answer difficult questions.

Benchmark	GPT-4 Evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	84.3% PaLI-17B
TextVQA VQA score (val)	78.0% 0-shot	71.8% PaLI-17B
ChartQA Relaxed accuracy (test)	78.5%[A]	58.6% Pix2Struct Large

[OpenAI, GPT4 Blog]

Contents

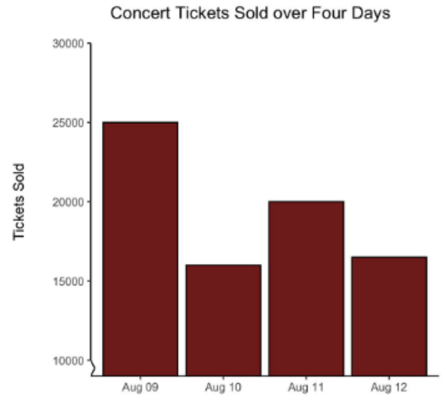
Background

Motivated Cases

inverted y-axis

- Qwen-VL-Max: 17
- GPT-4-Vision-Preview: 17
- Our model: 12

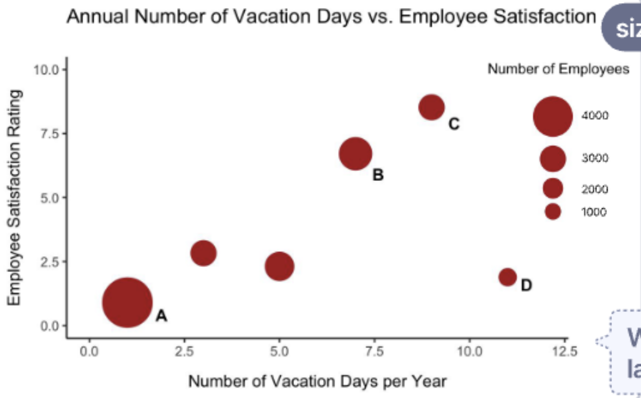
What's the FIFA Ranking of Germany in 2003?



truncated y-axis

- Qwen-VL-Max: 87.5%
- GPT-4-Vision-Preview: 60%
- Our model: 80%

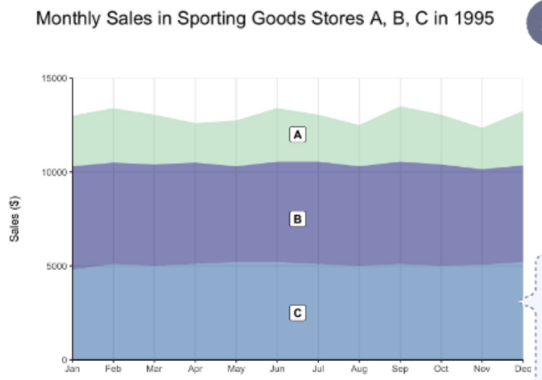
Approximately what is the number of concert tickets sold on Aug 10 as a proportion of that on Aug 11?



size of bubble

- Qwen-VL-Max: C
- GPT-4-Vision-Preview: D
- Our model: A

Which company has the largest number of employers?



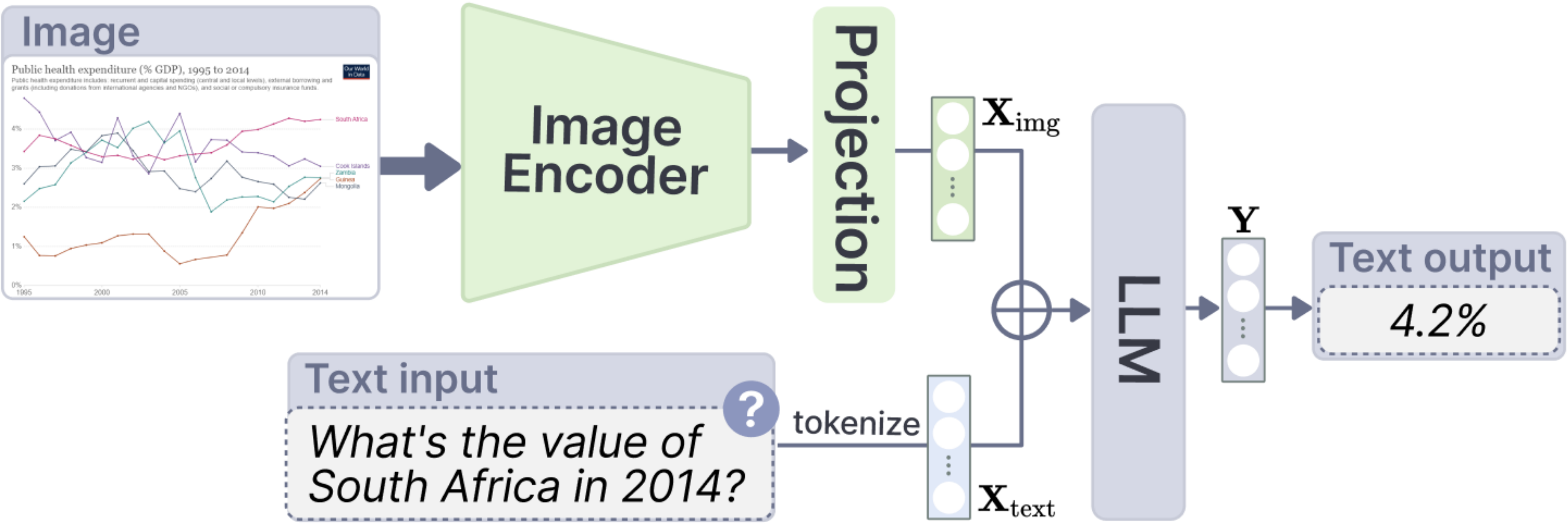
stacked area chart

- Qwen-VL-Max: 50%
- GPT-4-Vision-Preview: 33%
- Our model: 100%

Approximately what is store C's total sales as a proportion of store B's total sales in February of 1995?

Architecture of MLLMs

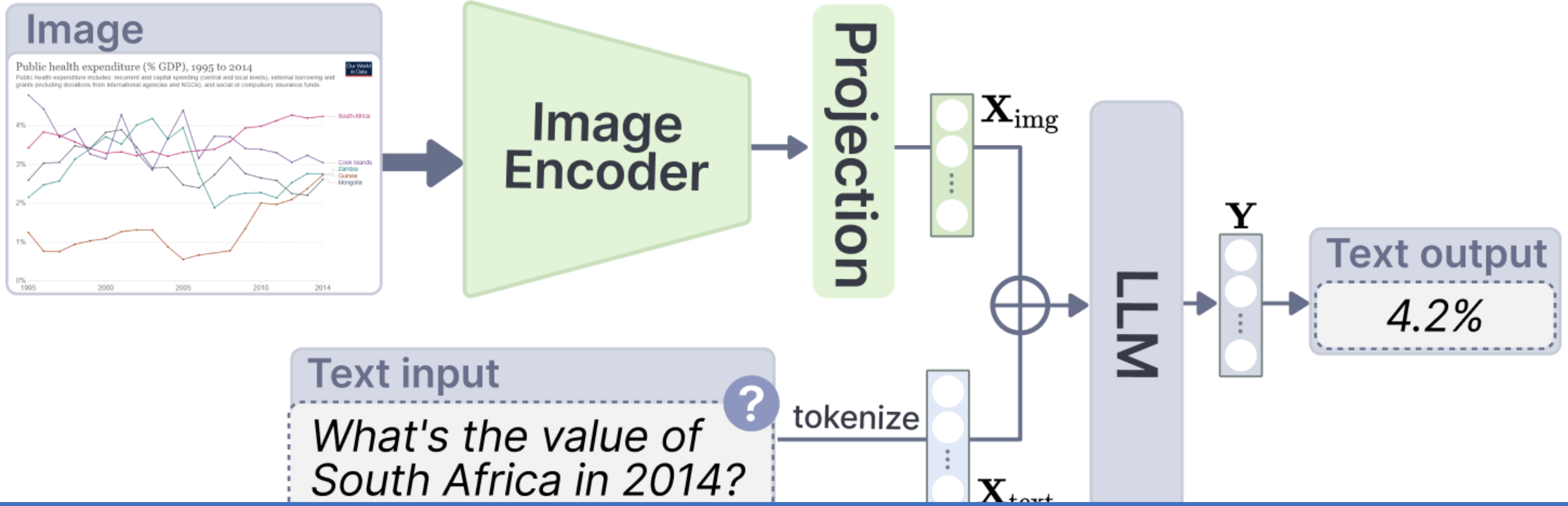
- MLLM architectures consist of three core components: ***Vision Encoder***, ***Vision-to-Language Projector***, and ***Large Language Model*** (LLM).



Training Data of MLLMs

- **Instruction data** is the basis of MLLM training.
  - Generally: <*target image*, *text task description*, *text output*>

  Chart QA data is naturally in the instruction format.
  - Chart QA: < chart , question , answer >



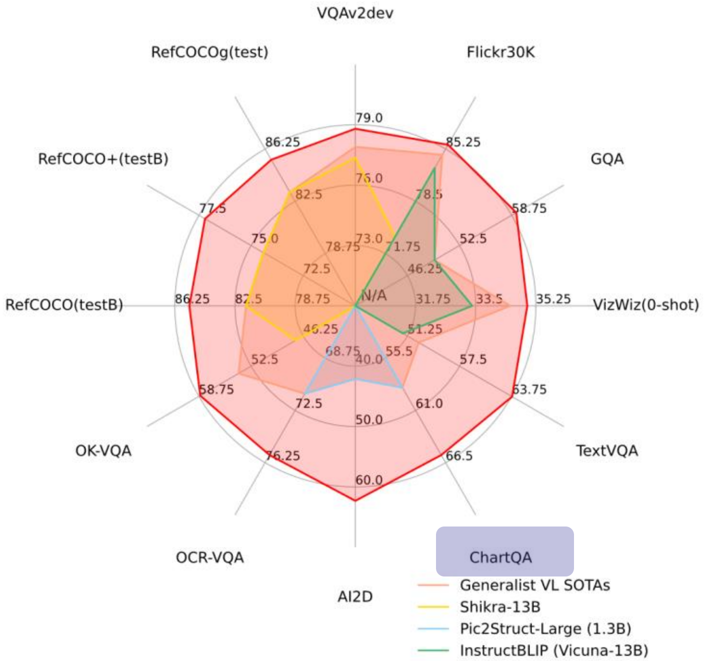
RQ: What makes effective visual instructions for CQA?

Introduction

Existing Chart Data

- The representative CQA dataset ChartQA, is used by leading open-sourced and commercial MLLMs as their training data or benchmark for chart understanding.

Qwen-VL,
InternVL,
Seed-VL
......



Benchmark	GPT-4 Evaluated few-shot	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	84.3% PaLI-17B
TextVQA VQA score (val)	78.0% 0-shot	71.8% PaLI-17B
ChartQA Relaxed accuracy (test)	78.5%[A]	58.6% Pix2Struct Large

RQ: Are existing Chart QA datasets good enough?

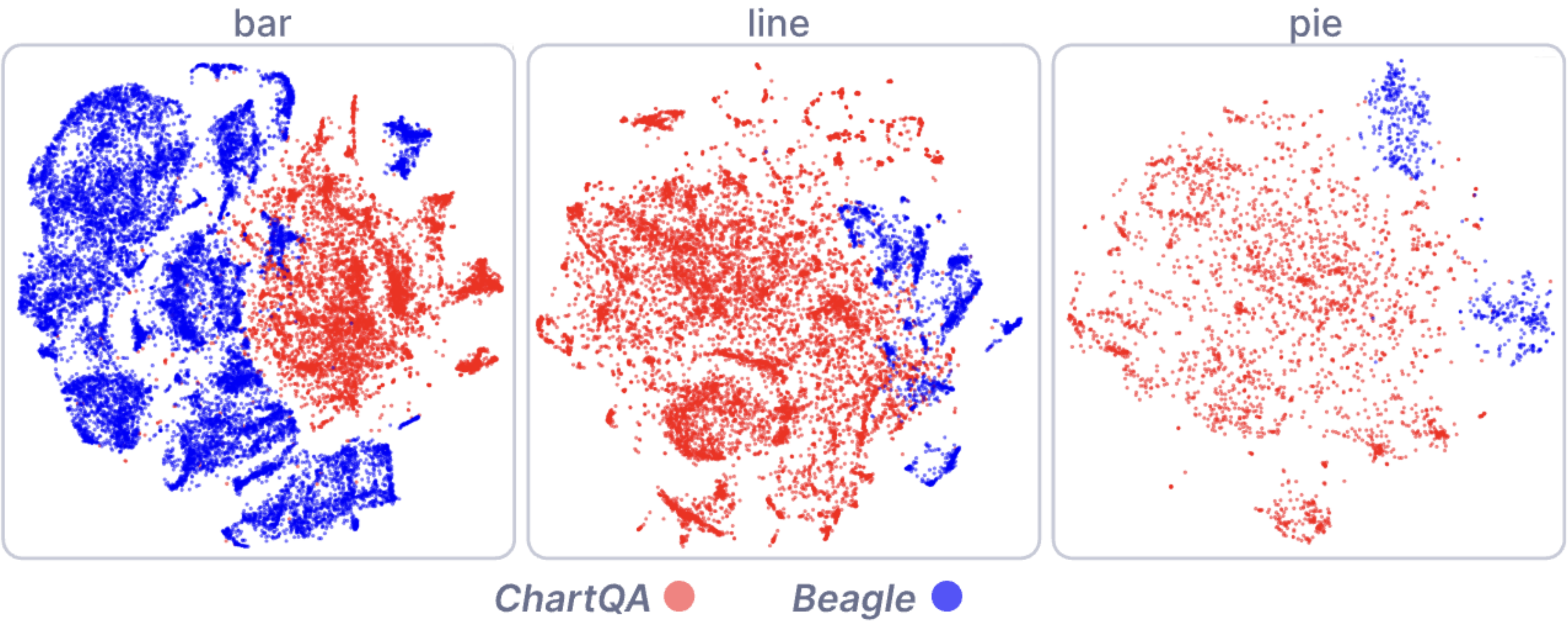
Research Questions

- RQ1: Are existing Chart QA datasets good enough?
  - Computational Analysis of Existing Dataset
- RQ2: What makes effective visual instructions for CQA?
  - Instruction Tuning Ablations

Empirical Study

Computational Analysis of ChartQA Dataset

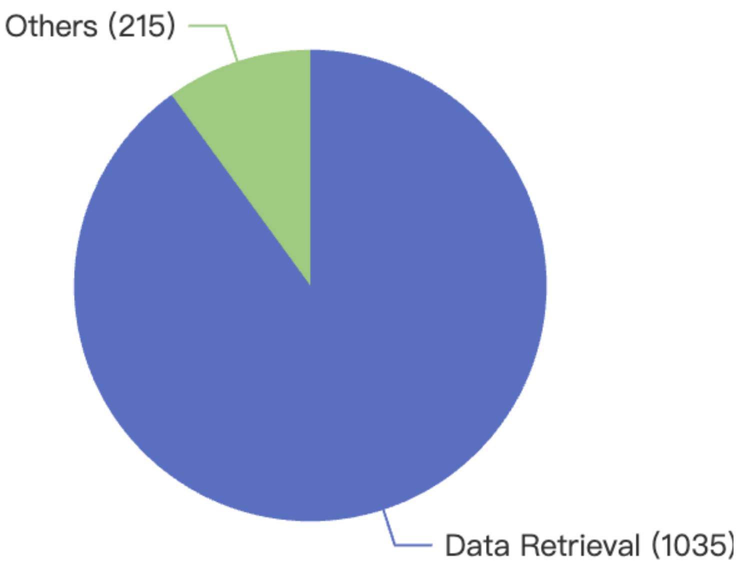
- We identified notable biases in both ChartQA's **chart** and QA pairs distribution.
  - Feature comparison between ChartQA and a real-world chart dataset (*i.e.*, Beagle).
  - ChartQA only include bar, pie, and line charts.



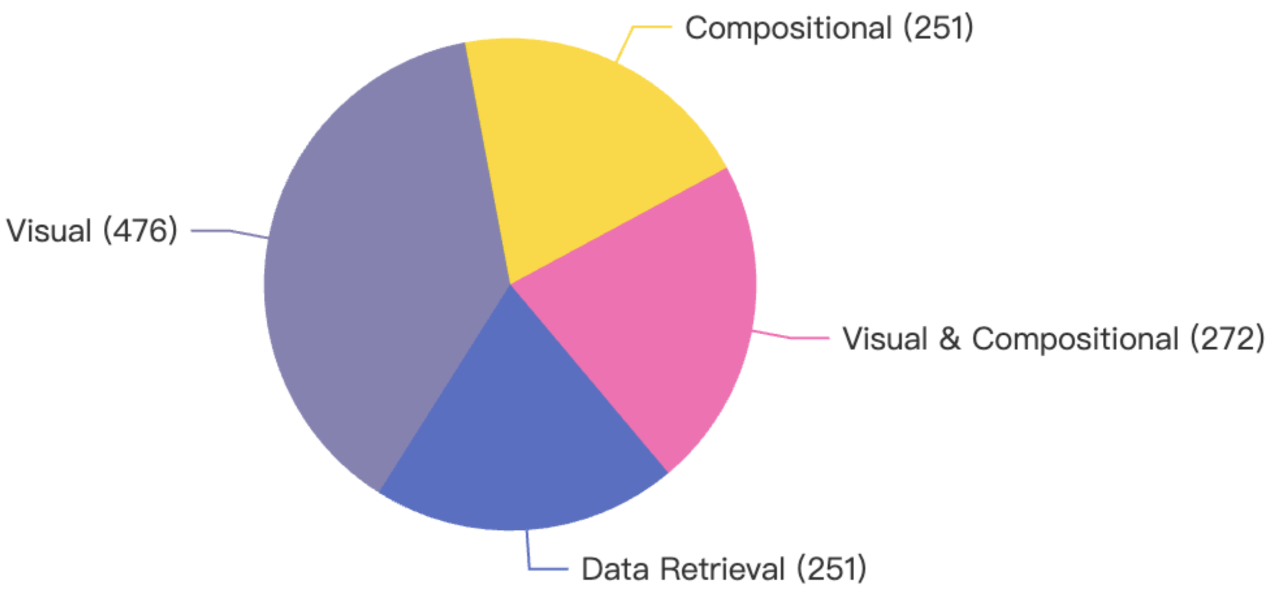
Computational Analysis of ChartQA Dataset

- We identified notable biases in both ChartQA's **chart** and QA pairs distribution.
  - ChartQA consists of two sets: Machine-generated and human-authored.
  - The distributions of question types are biased across the two sets.

Machine-generated



Human-authored



Empirical Study

Computational Analysis of ChartQA Dataset

- On the evaluation side, we further break down a series of MLLMs' scores on ChartQA according to the question type we annotated.

Model	ChartQA-M	ChartQA-H				Literacy
		Data Retrieval	Compositional	Visual	Visual-Compositional	
LLaVA-1.6-13b	72.16%	70.28%	30.24%	66.10%	10.81%	22.13%
LLaVA-1.6-34b	77.52%	71.49%	44.35%	71.19%	35.14%	35.11%
Qwen-VL-Chat	85.36%	66.67%	23.79%	62.29%	18.92%	25.19%
Qwen-VL-Plus	70.32%	51.00%	24.60%	60.81%	24.32%	24.42%
GPT4-Vision	87.25%	68.43%	25.96%	68.84%	20.85%	41.98%

- High scores on ChartQA-M may be "misleading".
- MLLMs underperform in:
  - scenarios require **numerical computation** or **visual comparison**.
  - visual literacy tasks which contain out-of-distribution chart types and question tasks.

Empirical Study

Instruction Tuning Ablation on ChartQA

Fine-tuning Settings:

- LLaVA-1.5 is chosen as the base model it was not trained on any chart data.
- Data combinations of ChartQA-M, ChartQA-H, ChartQA-Table.

Table 2: Results on *ChartQA-H* test set with models trained on individual and different combinations of training datasets in ChartQA.

Model	Data Retrieval	Compositional	Visual	Visual-Compositional
Baseline LLaVA-1.5	24.50%	9.27%	28.60%	13.51%
LLaVA-1.5 + *ChartQA-H*	32.93%	15.73%	47.25%	8.11%
LLaVA-1.5 + *ChartQA-M*	31.33%	10.08%	38.77%	8.11%
LLaVA-1.5 + *Chart2Table*	36.55%	9.68%	47.46%	13.51%
LLaVA-1.5 + *ChartQA-H & ChartQA-M*	43.37%	15.73%	51.91%	5.41%
LLaVA-1.5 + *ChartQA-H & Chart2Table*	42.17%	16.94%	51.91%	13.51%
LLaVA-1.5 + *ChartQA-H & ChartQA-M & Chart2Table*	48.59%	18.55%	54.66%	13.51%

Findings:

- **The effect of ChartQA-H is a “superset” of ChartQA-M.**
- Moreover, Chart2Table serves as an accompanying effective instruction task if the data tables are available.

Overall Motivation

- Informed by the empirical studies:
  1. Current chart data are biased in visual feature and chart type.
  2. Complex QA are more effective.

- We propose:
  1. ***Data filtering*** for efficiently utilizing the existing data;
  2. ***Data generation*** for optimizing the data distribution.

↓

A dataset of appropriate size while encompassing the real-world chart features and QA tasks.

Data Engine

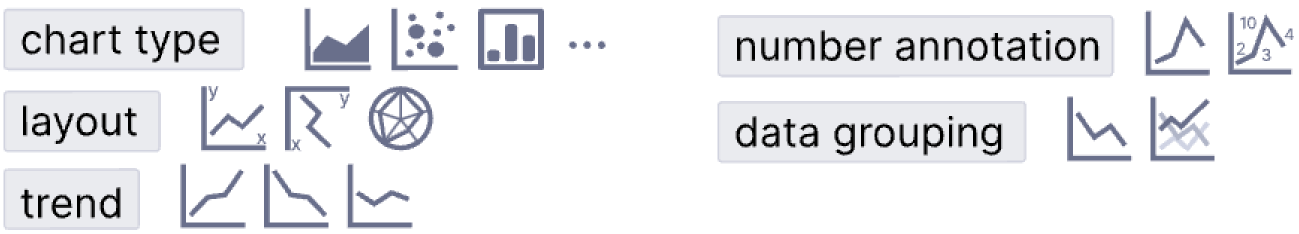
Part1: Data Filtering

Motivation:

- LLaVA as a leading generic MLLM only requires ***1223K*** instruction data. Chart-specialized Models, UniChart and ChartAssistant, use about *6900K* and *39400K* chart-related data.
- This disparity highlights that it’s impractical to **incorporate all available chart data into generic MLLMs’ training**.

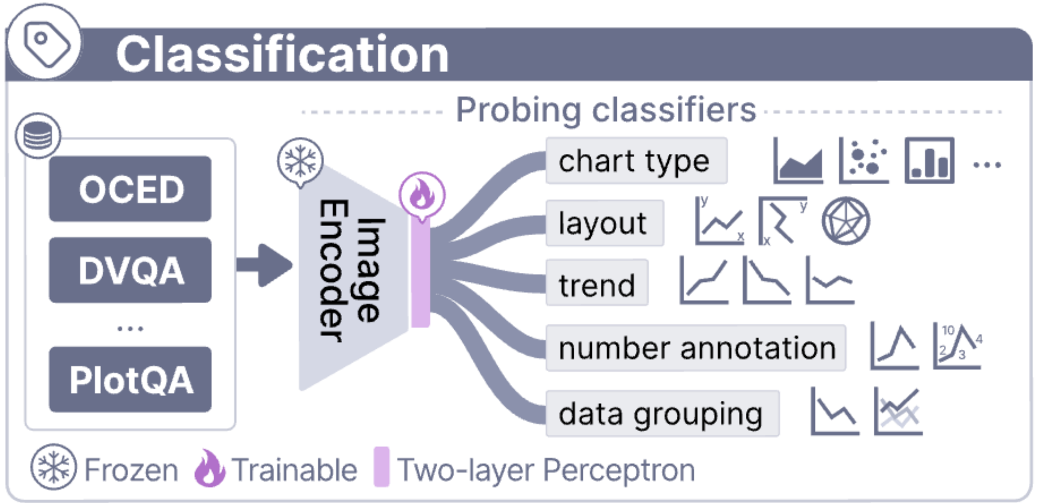
Challenge:

- Directly sample image data in the feature space can lead to **unbalanced chart distribution regarding**

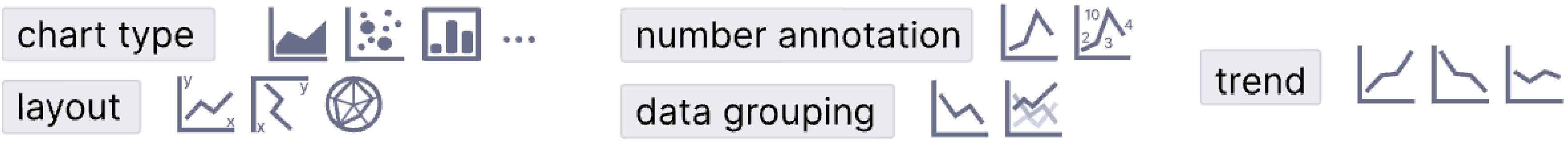


Data Engine

Part1: Data Filtering

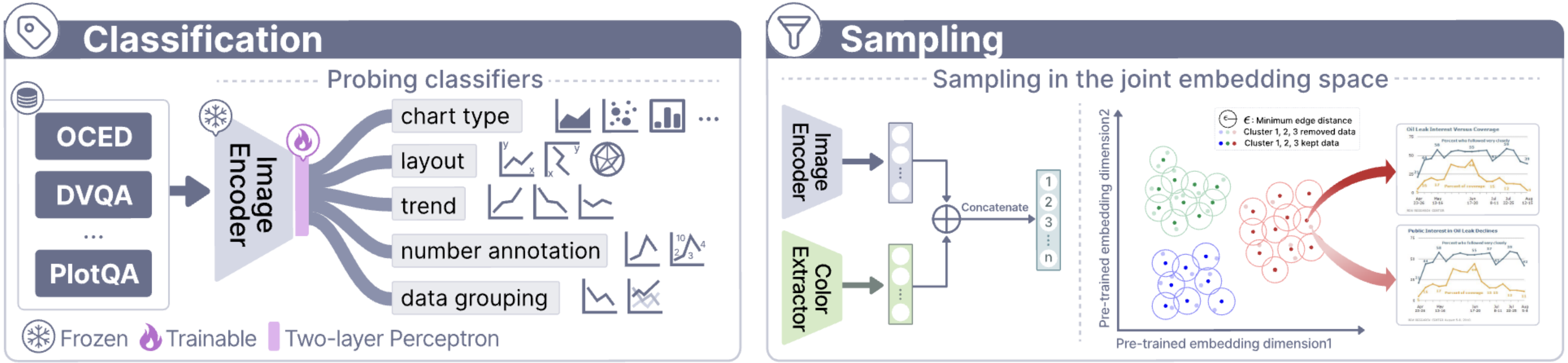


- To achieve balanced sampling, we first define key attributes that are important for chart understanding and build classifiers for them.

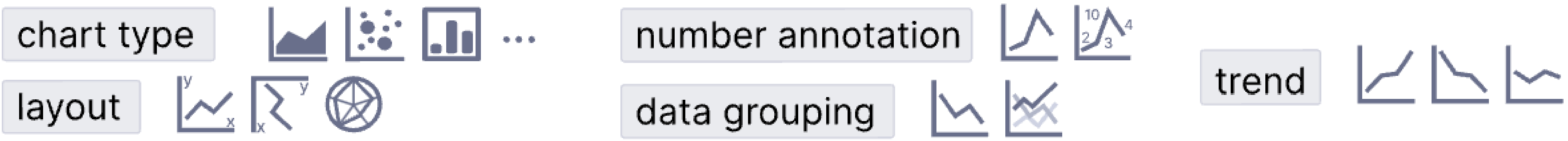


Data Engine

Part1: Data Filtering



- To achieve balanced sampling, we first define key attributes that are important for chart understanding and build classifiers for them.



- Then, we conduct stratified sampling based on the predicted attributes.

Data Engine

Part1: Data Filtering

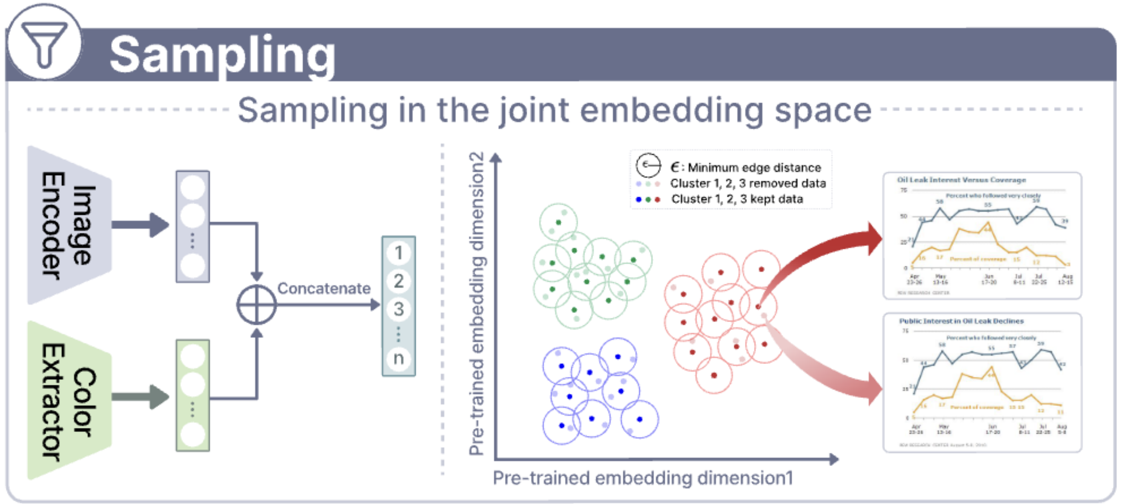


Table 3: Statistics of existing datasets, only considering the training set if dataset splits (*i.e.*, train-test) exist. Data counts consider the data tables and QA pairs associated with images. For example, a chart may be attached with its data table and two QA pairs, and then it is counted three times in total.

Dataset	***Chart tables***	***Chart QA pairs***
Statista, OECD, OWID	144,147	679,420
PlotQA	155,082	2,414,359
Unichart	189,792	2,218,468
Beagle	3,972	51
ChartInfo	1,796	21,949
VisText	9,969	0
ExcelChart	106,897	0
Total existing	611,655	5,334,247
Filtered dataset	69,418	68,223

Data Engine

Part2: Data Generation

Seed Examples → High-quality Expansion → Covering the Chart-task space

Table 4: The chart-task space of our dataset, which is summarized by the visual literacy research *VLAT* [33].

Visualization	Visualization Task									Note of X†
	Data Retrieval	Find Extremum	Determine Range	Characterize Distribution	Find Anomalies	Find Clusters	Find Correlations/ Trends	Make Comparisons	ETC	
Line Chart	X	X	X				X	X		
Bar Chart	X†	X	X					X		
Stacked Bar Chart	X†	X	X					X†		† Both Absolute Value and Relative Value
100% Stacked Bar Chart	X†	X†						X†		† Only Relative Value
Pie Chart	X†	X†						X†		† Only Relative Value
Histogram	X†	X†		X				X†	Identify the Characteristic of Bins	† Only Derived Value
Scatterplot	X	X	X	X	X	X	X	X		
Area Chart	X	X	X				X	X		
Stacked Area Chart	X†	X	X				X	X†		† Both Absolute Value and Relative Value
Bubble Chart	X	X	X	X	X	X	X	X		
Treemap	X†	X†						X†	Identify the Hierarchical Structure of Dataset	† Only Relative Value

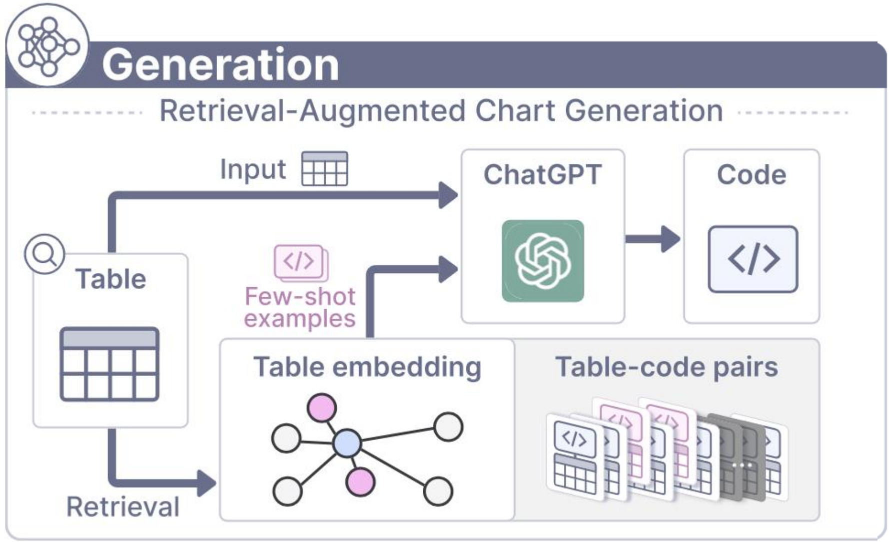
Part2: Data Generation

Table 4: The chart-task space of our dataset, which is summarized by the visual literacy research *VLAT* [33].

Visualization	Visualization Task									Note of $X^{\dagger}$
	Data Retrieval	Find Extremum	Determine Range	Characterize Distribution	Find Anomalies	Find Clusters	Find Correlations/ Trends	Make Comparisons	ETC	
Line Chart	X	X	X				X	X		
Bar Chart	$X^{\dagger}$	X	X					X		
Stacked Bar Chart	$X^{\dagger}$	X	X					$X^{\dagger}$		† Both Absolute Value and Relative Value
100% Stacked Bar Chart	$X^{\dagger}$	$X^{\dagger}$						$X^{\dagger}$		† Only Relative Value
Pie Chart	$X^{\dagger}$	$X^{\dagger}$						$X^{\dagger}$		† Only Relative Value
Histogram	$X^{\dagger}$	$X^{\dagger}$		X				$X^{\dagger}$	Identify the Characteristic of Bins	† Only Derived Value
Scatterplot	X	X	X	X	X	X	X	X		
Area Chart	X	X	X				X	X		
Stacked Area Chart	$X^{\dagger}$	X	X				X	$X^{\dagger}$		† Both Absolute Value and Relative Value
Bubble Chart	X	X	X	X	X	X	X	X		
Treemap	$X^{\dagger}$	$X^{\dagger}$						$X^{\dagger}$	Identify the Hierarchical Structure of Dataset	† Only Relative Value

Data Engine

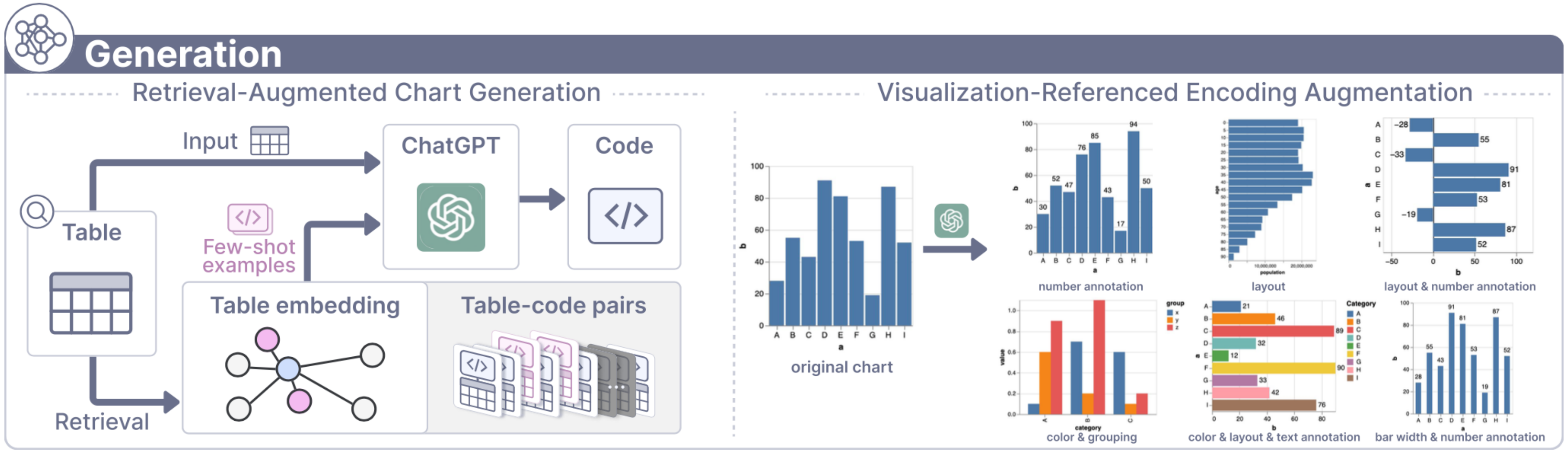
Part2: Data Generation



- **Retrieval-Augmented Chart Generation**
  - Collecting high-quality seed examples from previous studies and authoritative chart libraries (Vega-Lite, Matplotlib, Seaborn, and ECharts).
  - Retrieving seed examples according to table similarity to serve as in-context learning examples to improve generation accuracy.

Data Engine

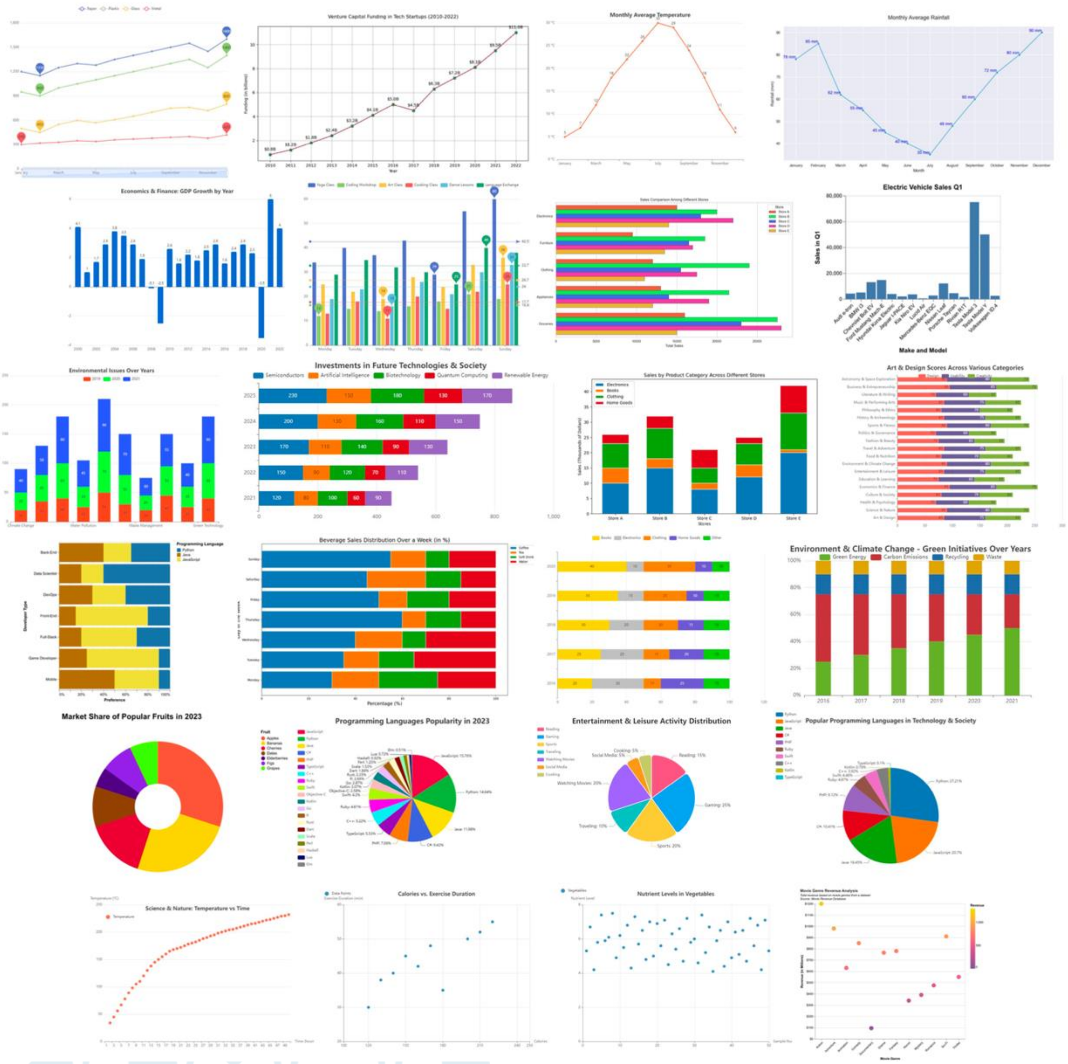
Part2: Data Generation



- **Visualization-referenced Encoding Augmentation**
  - Prompting the LLM all the reasonable modifications it can apply to a specific chart type.
  - Covering the visual encoding space and optimizing the chart distribution.

Data Engine

Examples of the Generated Data



Data Engine

Examples of the Generated Data

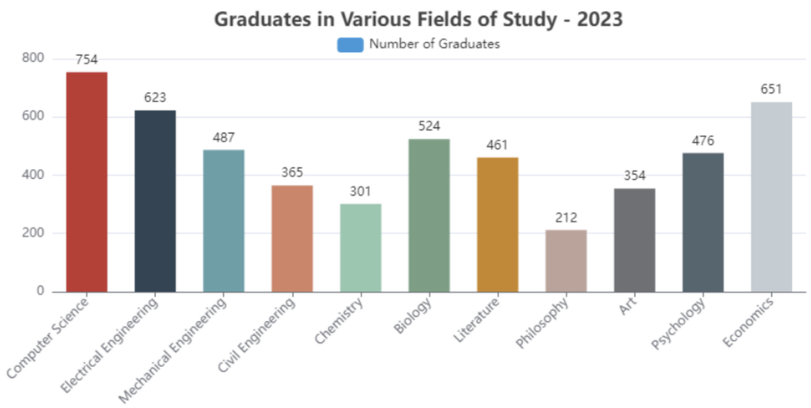


Chart-to-table translation

```
"from": "human",
"value": "Please extract the underlying table from the given chart"
```

```
"from": "gpt",
"value": "| Field of Study | Number of Graduates |\n|---|---|\n|
Computer Science | 754.0 |\n| Electrical Engineering | 623.0 |\n|
Mechanical Engineering | 487.0 |\n| Civil Engineering | 365.0 |\n|
Chemistry | 301.0 |\n| Biology | 524.0 |\n| Literature | 461.0 |\n|
Philosophy | 212.0 |\n| Art | 354.0 |\n| Psychology | 476.0 |\n|
Economics | 651.0 |\n|"
```

Generated questions

```
"from": "human",
"type_of_question": "Data Retrieval",
"value": "What is the number of graduates in Civil Engineering?"
```

```
"from": "gpt",
"value": "Based on the chart, the bar corresponding to Civil Engineering has a
         height that indicates the number of graduates is 365.",
"label": "365"
```

```
"from": "human",
"type_of_question": "Find Extremum",
"value": "Which field of study has the highest number of graduates?"
```

```
"from": "gpt",
"value": "By comparing the heights of the bars, the field with the tallest bar
         is Computer Science. The number of graduates in Computer Science is
         754, which is the highest among the listed fields of study."
"label": "Computer Science"
```

```
"from": "human",
"type_of_question": "Determine Range",
"value": "What is the range of the number of graduates across all fields of
         study?"
```

```
"from": "gpt",
"value": "The smallest number of graduates is in Philosophy with 212 gradu
          ates, and the largest is in Computer Science with 754 graduates.
          Therefore, the range is computed as 754 - 212, which is 542.",
"label": "542"
```

Customizing Training Receipts for Charts

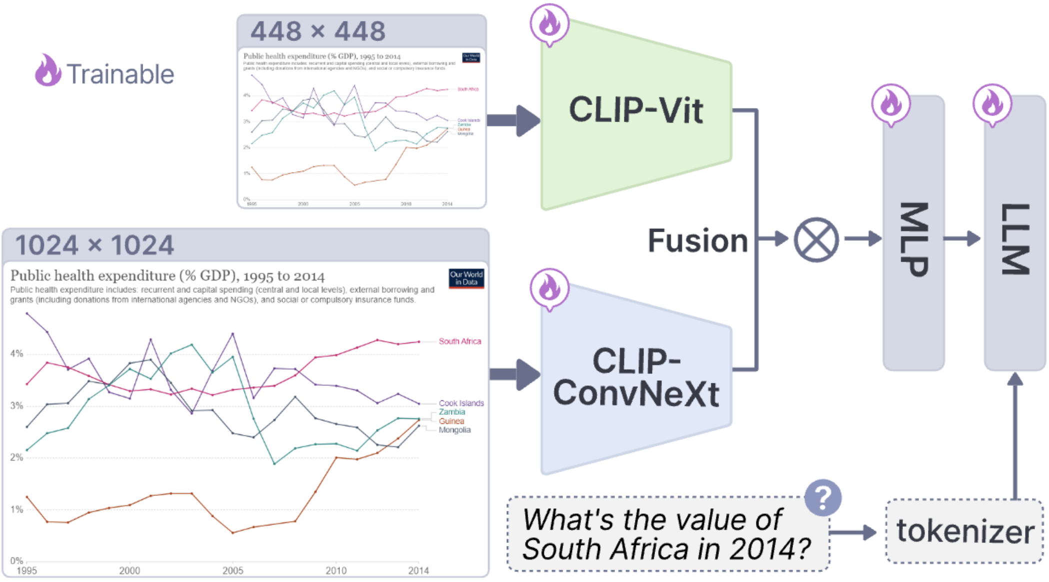


Fig. 7: Architecture of the MLLM adopted in our work. High-resolution and normal-resolution features of the input image are fused to facilitate the efficient recognition of fine-grained features. During the training phase, vision encoders are unfreezed to enable the adaptation to chart characteristics.

1. Data Composition
   - LLaVA665K+**80KChart2Table+119K Chart QA pairs**

2. The pre-trained CLIP underperforms in visualization scenarios.
   - **Unfreezing the visual encoders.**

2. Charts are text-rich and require fine-grained recognition.
   - **Integrating a mixture-of-resolution adaptation strategy for enhanced recognition.**

Evaluation

Benchmarks: ChartQA and Chart-to-Table

- We surpass the current leading models, but with much less data, showcasing our data filtering and generation effectiveness.
- The ablation studies demonstrate the effectiveness of our two model design choices.

Table 7: Results on traditional benchmarks (*i.e.*, ChartQA and Chart-to-table). We compare our work with the previous open-source models and present results of ablations on data, training, and model design.

Model	ChartQA Aug.	ChartQA Human	ChartQA Average	Chart-to-table
Chart-T5	74.4	31.8	52.95	37.5
Donut	78.1	29.8	53.95	38.2
Matcha	88.9	38.8	63.85	39.4
Unichart	87.8	43.9	65.85	91.1
ChartLLaMa	90.4	48.9	69.7	90.0
ChartAst-D (39.4M CQA data)	91.3	45.3	68.3	**92.0**
ChartAst-S (39.4M CQA data)	92.0	58.2	75.1	91.6
No Unfreezing vision encoder	77.4	47.1	62.3	44.6
No High Resolution	88.6	55.8	72.2	87.9
No Filtered Data	91.0	61.4	76.2	90.3
No Generation Data	92.7	63.7	78.2	91.2
Our model (199K CQA data)	**93.5**	**64.9**	**79.2**	91.8

Evaluation

Our Benchmark

Table 6: Results on our benchmarks.

Models	Data Retrieval	Find Extremum	Determine Range	Characterize Distribution	Find Anomalies	Find Clusters	Find Correlations/ Trends	Make Comparisons
LLaVA1.6-34b	37.69	35.83	3.85	20.00	21.43	27.27	51.95	48.84
GPT-4-vision-preview	**56.92**	**60.96**	30.77	**36.67**	**42.86**	**36.36**	**68.83**	56.40
Qwen-VL-Plus	43.08	21.39	11.54	10.00	7.14	13.64	41.56	34.30
Our model	46.15	53.48	**35.57**	30.00	**42.86**	**36.36**	64.94	**58.14**

- Table 6 showcases comparative results on our benchmark, illustrating that our model outperforms commercial models in most tasks and achieve comparable performances with GPT4-Vision.

Takeaways

- High-quality and appropriate-sized instruction data for chart understanding is critical.
  - ***A data engine that supports managing existing data and generating new data to cover the real-world chart-task space.***

- CLIP-Vit underperforms in visualization scenarios.
  - ***Unfreezing visual encoders.***

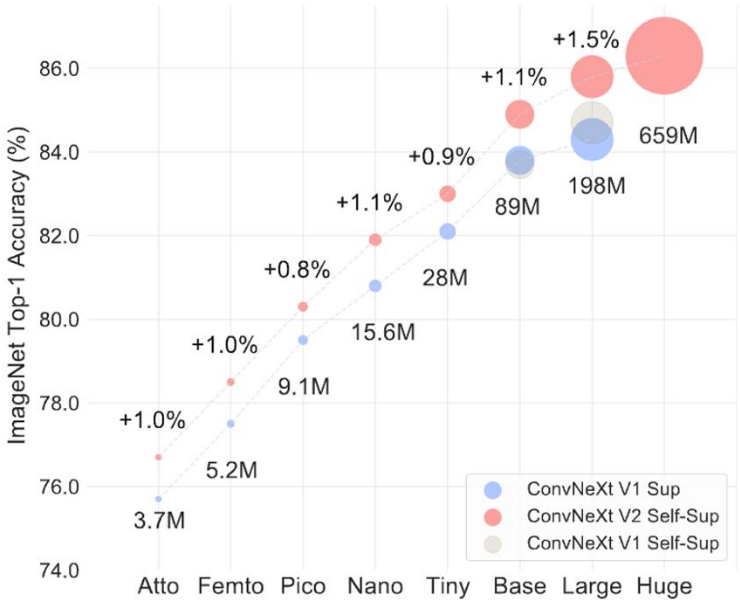
- Understanding charts require fine-grained recognition.
  - ***High-resolution.***

Contents

Future Work

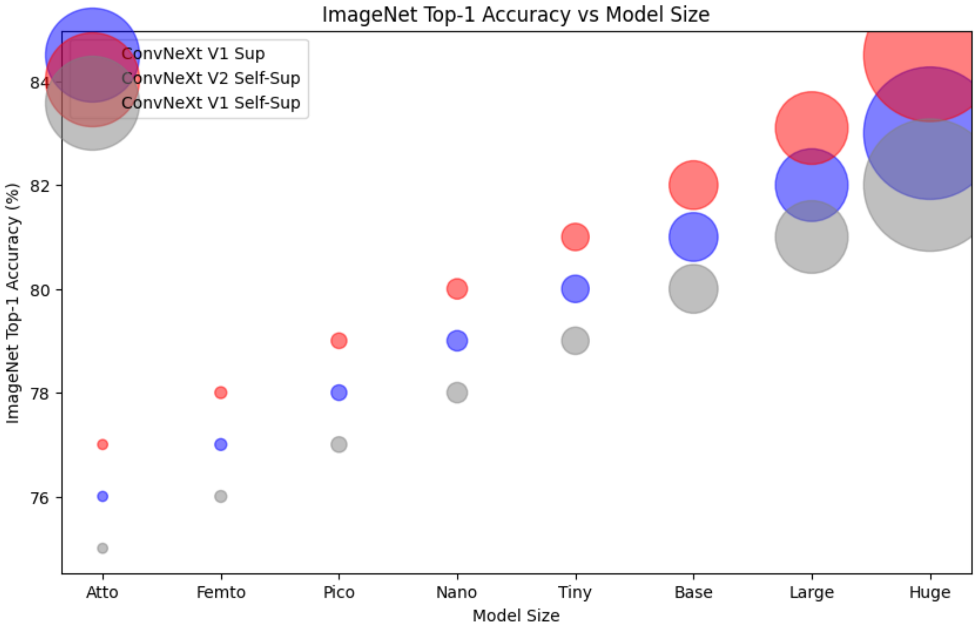
Chart-to-Code Generation

- Large Complex Code-format Visualization Corpus



User input

vs.

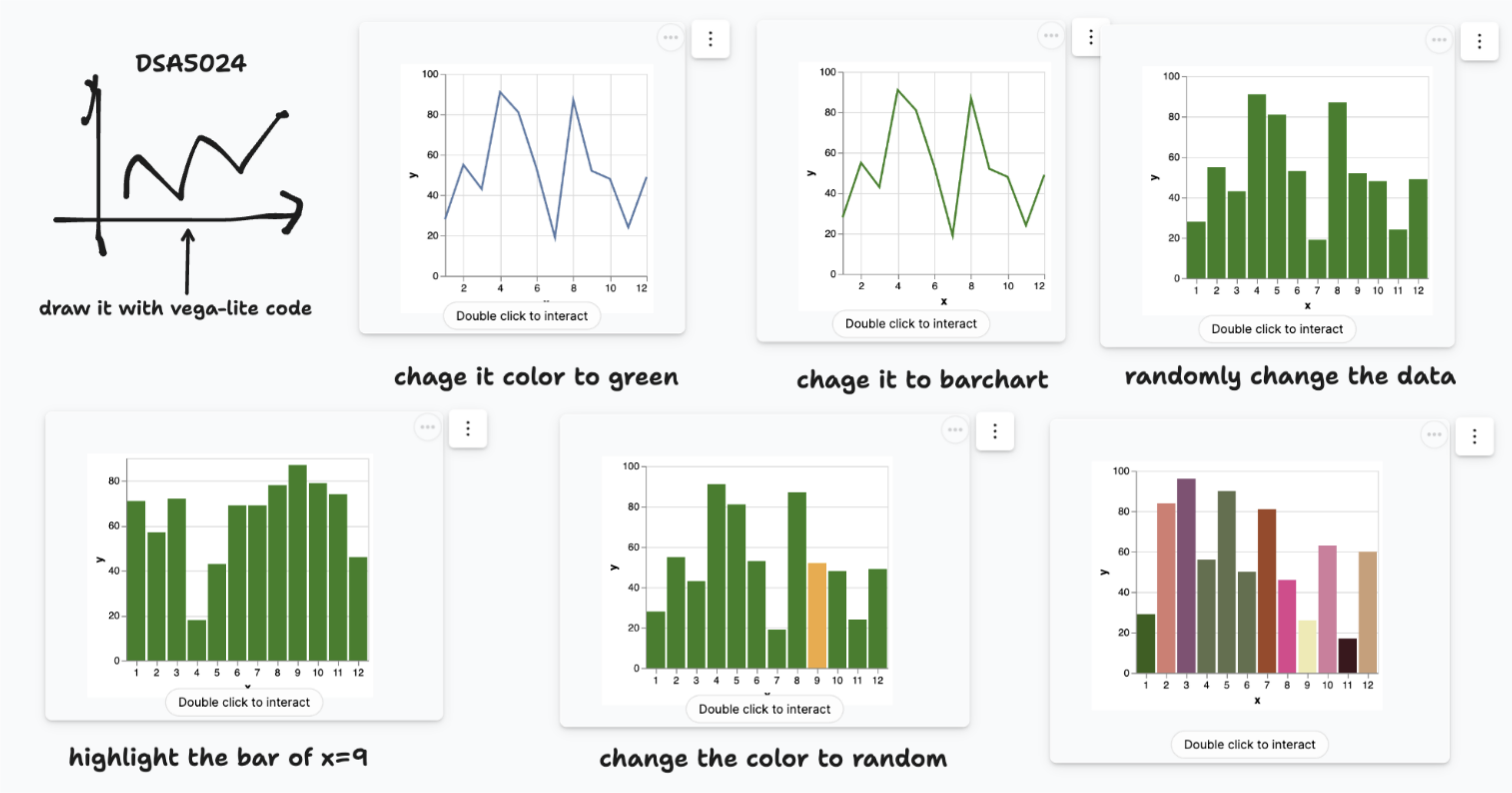


Model output

Future Work

Multi-modal Interaction Over Visualization

- Visual prompt-based Interaction (*e.g.*, Sketch)



Thank you for listening!

Codes and datasets are open-sourced.

Discussions and future collaborations are welcome.`

zengxingchen.github.io