清華大學
Tsinghua University

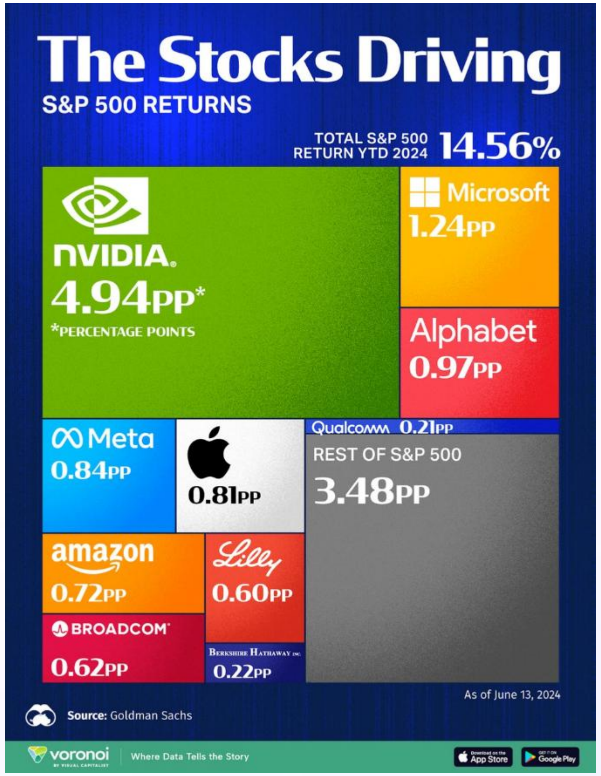
InfoChartQA: A Benchmark for Multimodal Question Answering on Infographic Charts

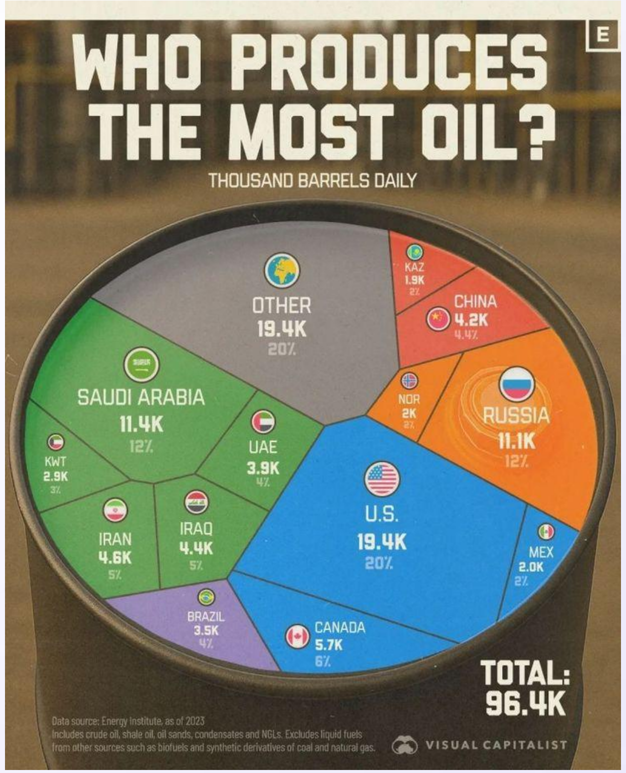
Tianchi Xie[1,*], Minzhi Lin[1,*], Mengchen Liu[2], Yilin Ye[3], Changjian Chen[4], Shixia Liu[1]

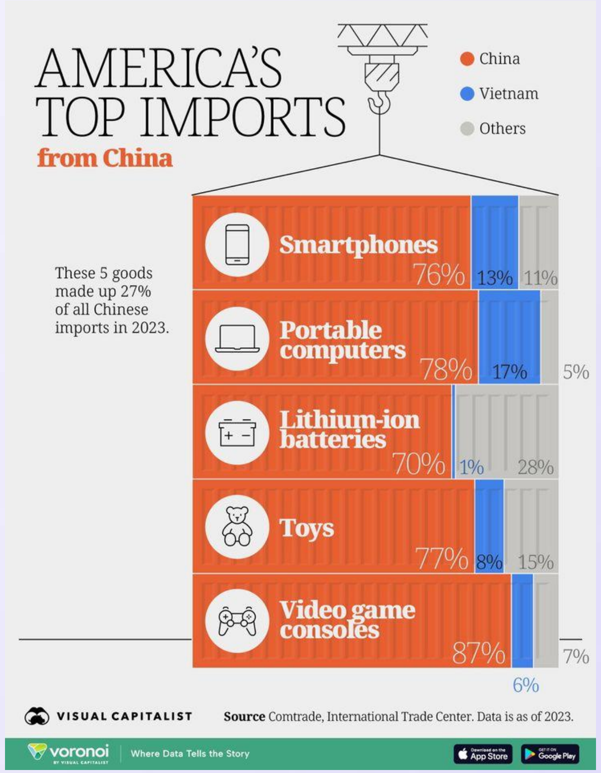
1 Tsinghua University
2 Meta
3 Hong Kong University of Science and Technology
4 Hunan University

Motivation

Compared with plain charts, infographic charts are better in **enhance visual engagement and communicating abstract concepts through symbolic visuals.**





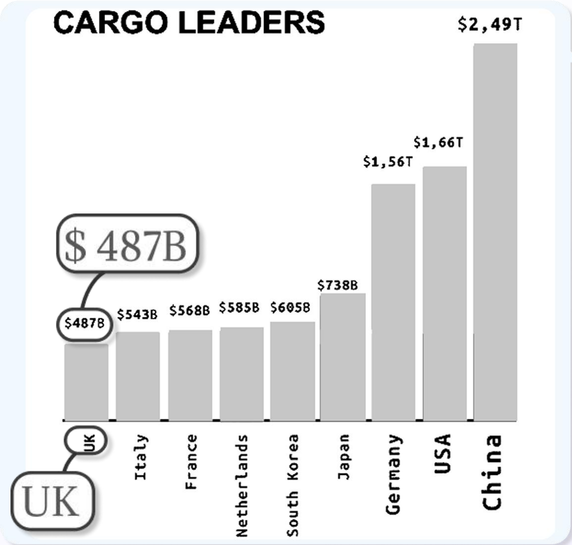


Motivation

Understanding infographic charts requires both visual recognition and reasoning, **posing challenges for MLLMs**.

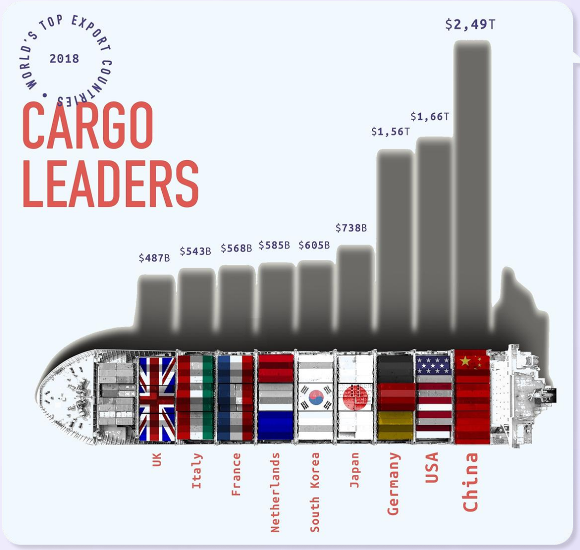
- For example, MLLMs may answer the same question on plain charts correctly, but fail on infographic charts.

What is the export volume of UK?



$ 487B ✓

What is the export volume of UK?

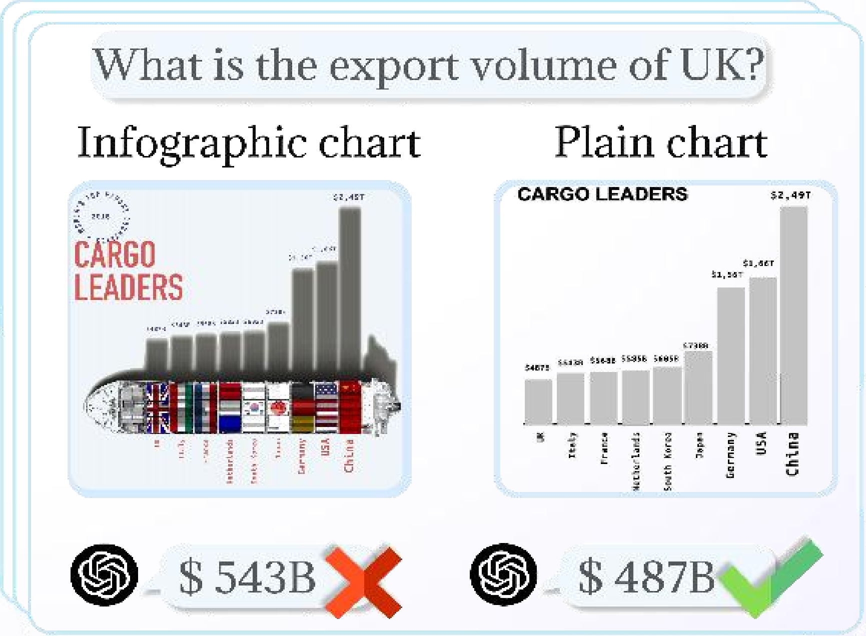


$ 543B ✗

Challenges

Existing ChartQA benchmarks fall short due to the lack of **paired plain charts** and **visual-element-based questions.**

Paired plain charts



What is the export volume of UK?

Infographic chart

Plain chart

$ 543B ✗

$ 487B ✓

Visual-element-based questions

Question:

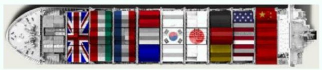
What is the export volume of the country corresponding to this container?

Answer: $ 487B

Question:

Answer: B

Which option best interprets the metaphor conveyed by the image?



A. The shadow lengths of the containers quantifies nations' cargo export volumes.

B. The shadow length of the containers quantifies each nation's export volume.

C. ...

D. ...

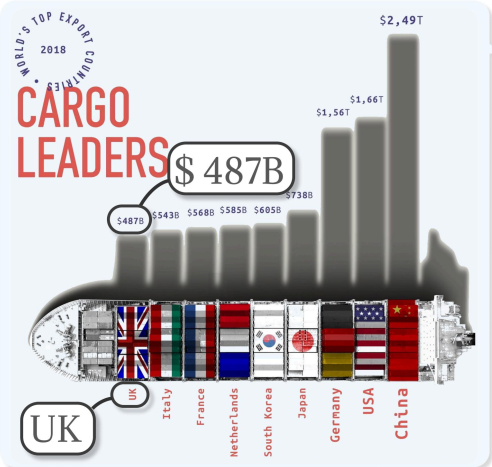
The challenge in systematically evaluating infographic charts' impact on MLLMs' performance

InfoChartQA

(a) Infographic and plain chart pairs

Enable the diagnosis of failure cases

What is the export volume of UK?



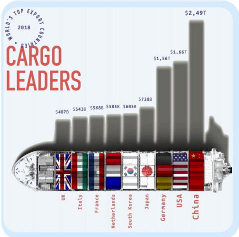
$ 543B ✗

Analysis

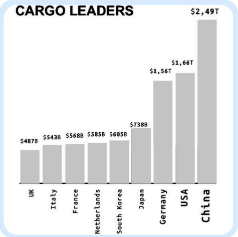
5,948 Infographic-plain chart pairs

What is the export volume of UK?

Infographic chart



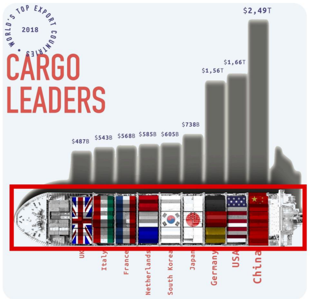
Plain chart



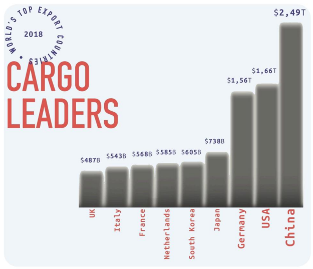
$ 543B ✗

$ 487B ✓

Hypothesis: the visual elements distract MLLMs



Remove visual elements



$ 487B ✓

Verified!

(b) Visual-element-based questions

Enable more flexible questions

7,475 basic questions

Question:

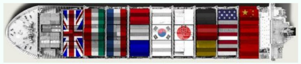
What is the export volume of the country corresponding to this container?

Answer: $ 487B

462 metaphor-related questions

Question:

Which option best interprets the metaphor conveyed by the image?



A. The shadow lengths of the containers quantifies nations' cargo export volumes.

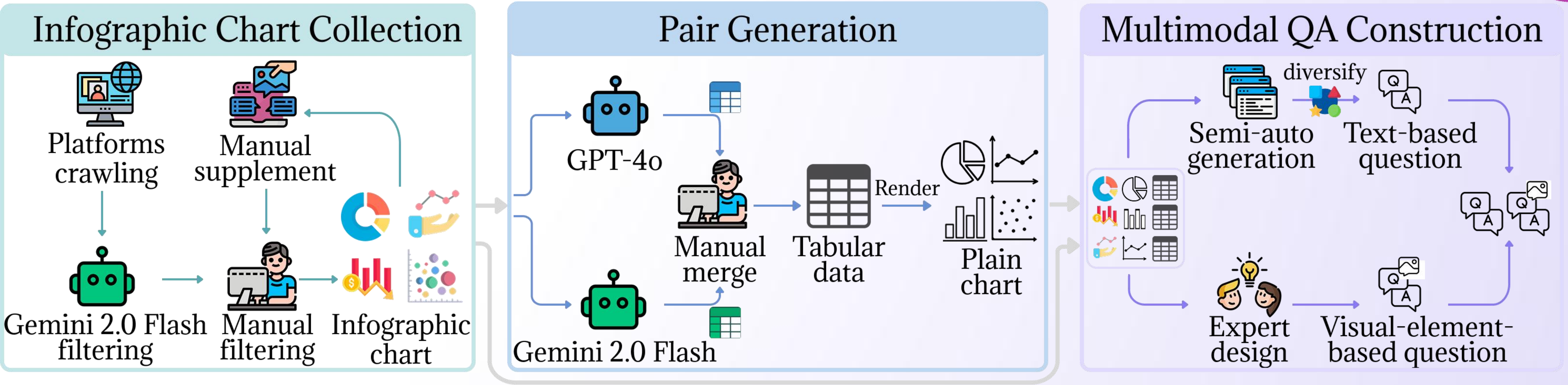
B. The shadow length of the containers quantifies each nation's export volume.

C. ...

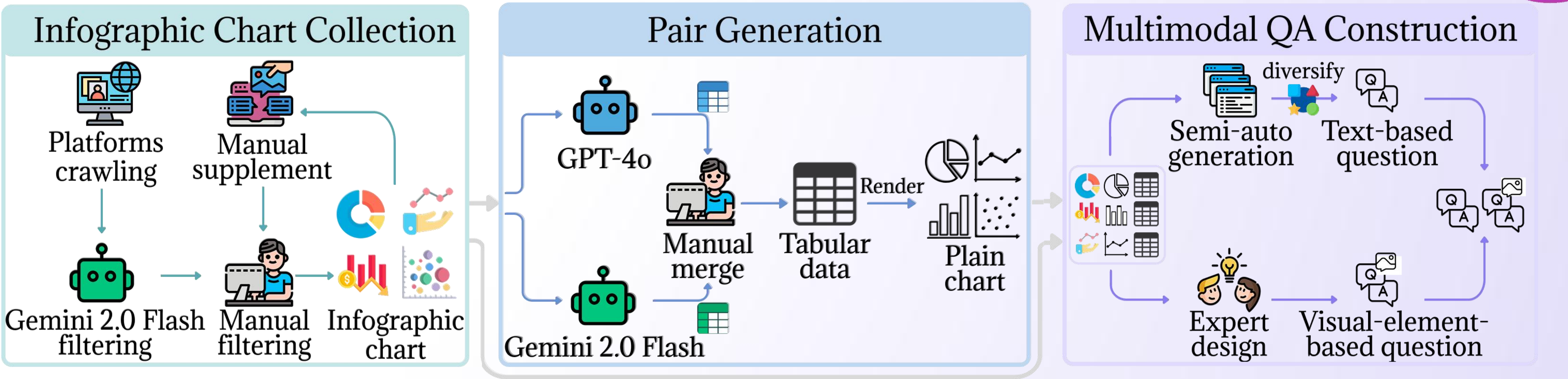
D. ...

Answer: B

Construction Pipeline

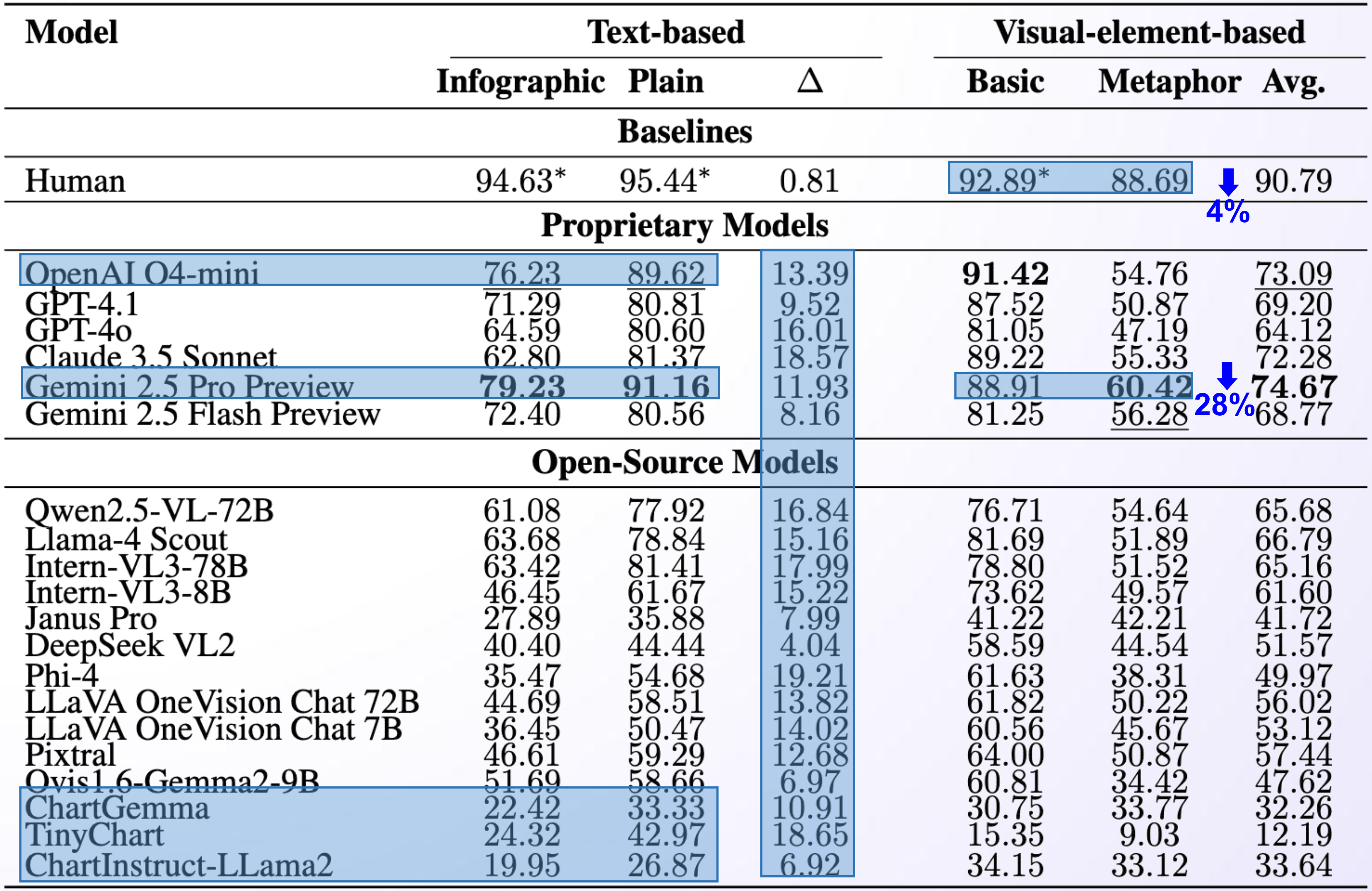


Construction Pipeline



Dataset	Chart type	Infographic charts	Text-based questions	Visual-element-based questions	HD-D	SD
ChartQA	3	×	2.5K	×	0.769	0.805
ChartBench	42	×	16.8K	×	0.630	0.743
ChartQAPro	9	✓	1.9K	×	0.828	0.864
InfographicVQA	11	✓	3.2K	1.1K	**0.837**	**0.823**
InfoChartQA	**54**	✓	**50.9K**	**7.9K**	0.812	0.800

Quantitative Results



Model	Text-based			Visual-element-based		
	Infographic	Plain	Δ	Basic	Metaphor	Avg.
Baselines						
Human	94.63*	95.44*	0.81	92.89*	88.69	90.79
Proprietary Models						
OpenAI O4-mini	76.23	89.62	13.39	**91.42**	54.76	73.09
GPT-4.1	71.29	80.81	9.52	87.52	50.87	69.20
GPT-4o	64.59	80.60	16.01	81.05	47.19	64.12
Claude 3.5 Sonnet	62.80	81.37	18.57	89.22	55.33	72.28
Gemini 2.5 Pro Preview	**79.23**	**91.16**	11.93	88.91	**60.42**	**74.67**
Gemini 2.5 Flash Preview	72.40	80.56	8.16	81.25	56.28	68.77
Open-Source Models						
Qwen2.5-VL-72B	61.08	77.92	16.84	76.71	54.64	65.68
Llama-4 Scout	63.68	78.84	15.16	81.69	51.89	66.79
Intern-VL3-78B	63.42	81.41	17.99	78.80	51.52	65.16
Intern-VL3-8B	46.45	61.67	15.22	73.62	49.57	61.60
Janus Pro	27.89	35.88	7.99	41.22	42.21	41.72
DeepSeek VL2	40.40	44.44	4.04	58.59	44.54	51.57
Phi-4	35.47	54.68	19.21	61.63	38.31	49.97
LLaVA OneVision Chat 72B	44.69	58.51	13.82	61.82	50.22	56.02
LLaVA OneVision Chat 7B	36.45	50.47	14.02	60.56	45.67	53.12
Pixtral	46.61	59.29	12.68	64.00	50.87	57.44
Ovis1.6-Gemma2-9B	51.69	58.66	6.97	60.81	34.42	47.62
ChartGemma	22.42	33.33	10.91	30.75	33.77	32.26
TinyChart	24.32	42.97	18.65	15.35	9.03	12.19
ChartInstruct-LLama2	19.95	26.87	6.92	34.15	33.12	33.64

Spearman correlation coefficient 0.893

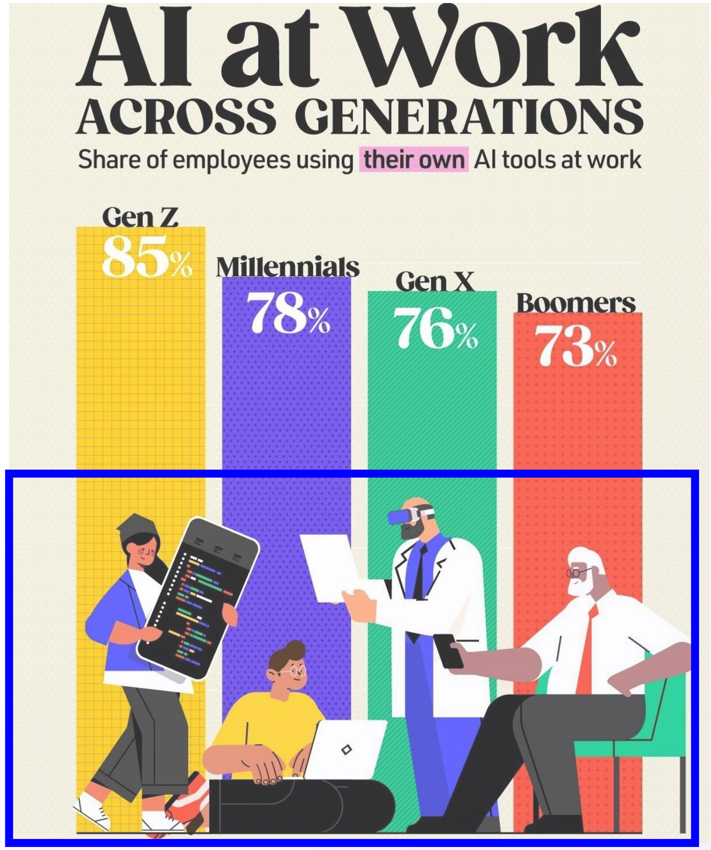
Why?

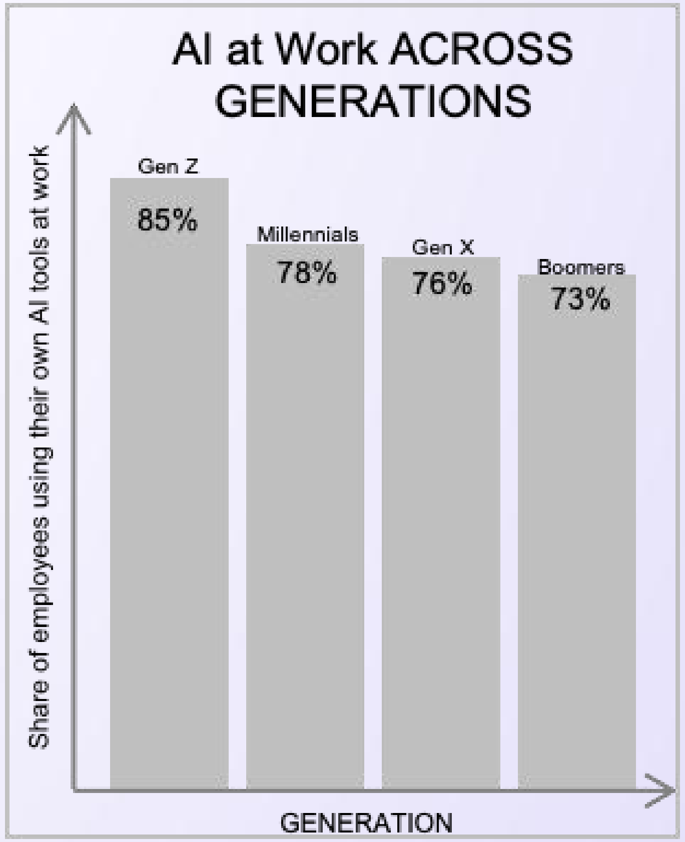
- The performance of MLLMs **degraded** on infographic charts
- Strong performance on **text-based questions** is foundational to strong performance on **visual-element-based questions**.
- **Metaphor-related questions** are challenging for MLLMs

Ablation Study 1

Question: **Why** MLLMs perform worse on infographic charts than on plain charts?

Hypothesize: **Pictorial visual elements** primarily contribute to the degradation.





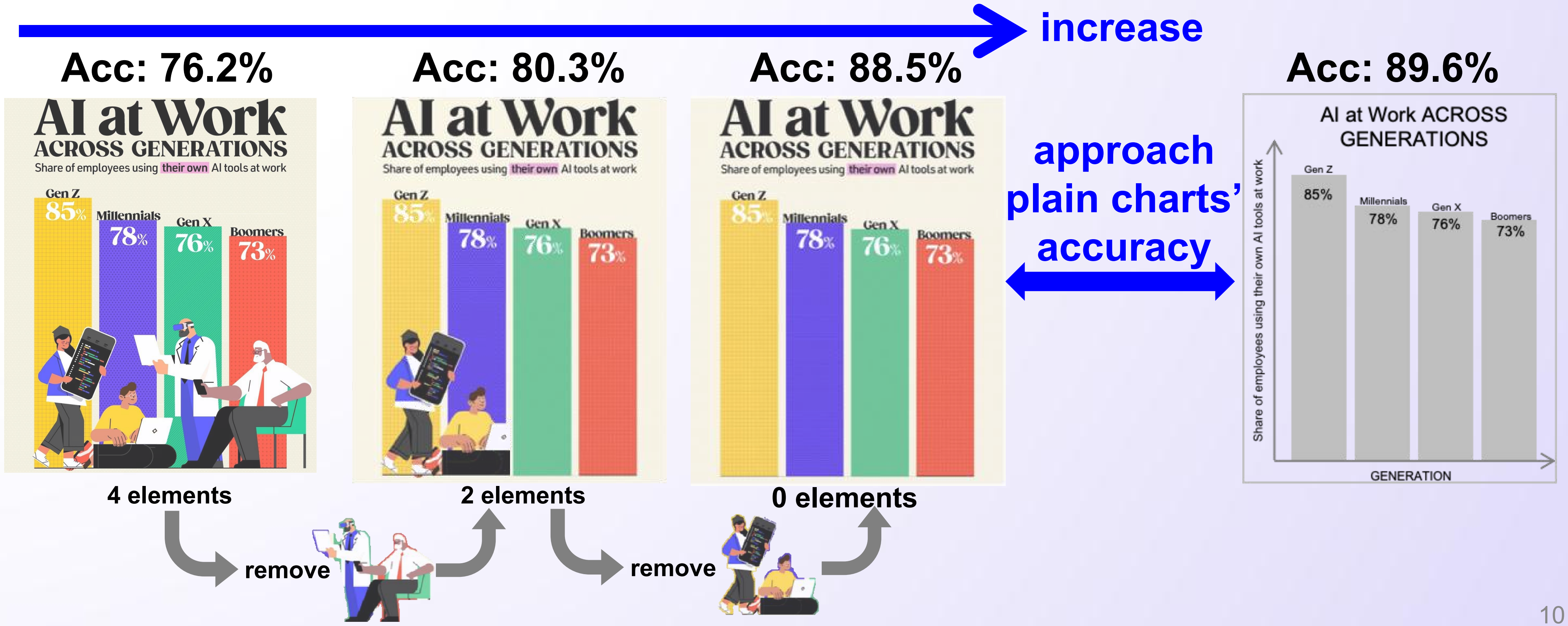
Ablation Study 1

Question: **Why** MLLMs perform worse on infographic charts than on plain charts?

Hypothesize: **Pictorial visual elements** primarily contribute to the degradation.

verify!

- gradually removed elements as bellow:



increase

Acc: 76.2%

Acc: 80.3%

Acc: 88.5%

Acc: 89.6%

approach plain charts' accuracy

4 elements

2 elements

0 elements

remove

remove

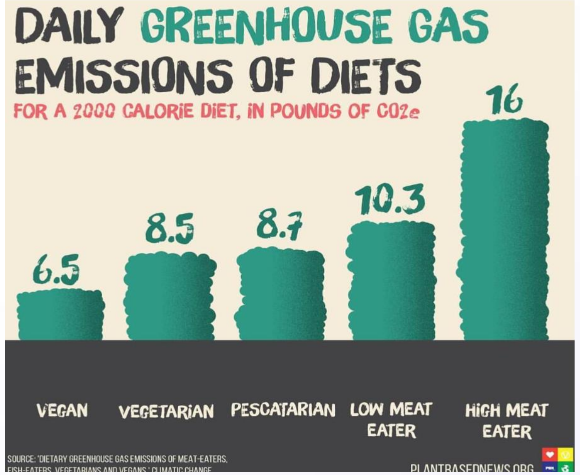
Ablation Study 2

Question: **How** pictorial visual elements affect MLLMs?

Hypothesize: **Clearer connections between text and visual elements improve understanding**, and elements may affect MLLMs **by disturbing the connections.**

- applied modifications to **introduce varying connection perturbations**

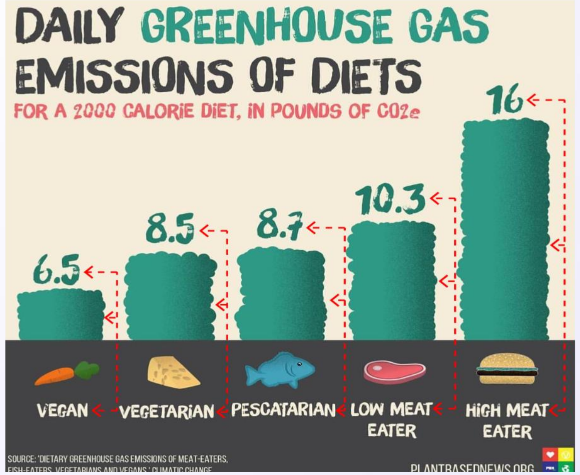
verify!



(a) Obstructions Removal

GPT-4.1: $\uparrow 0.79$

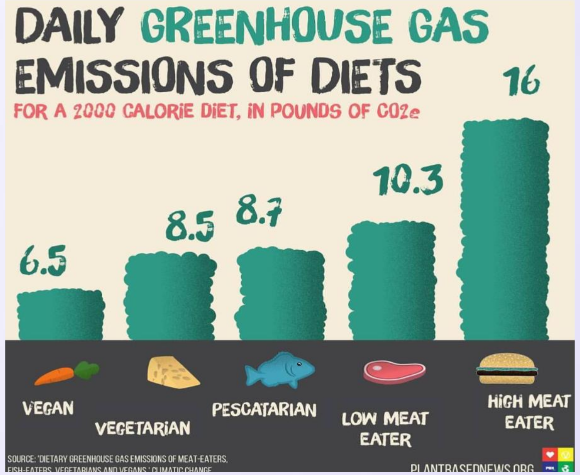
Tinychart: $\uparrow 3.23$



(b) Auxiliary Line

GPT-4.1: $\uparrow 0.85$

Tinychart: $\uparrow 3.08$



(c) Position Perturbation

GPT-4.1: $\downarrow 2.95$

Tinychart: $\downarrow 2.78$

improved clarity enhanced performance **introducing distortions reduced it**

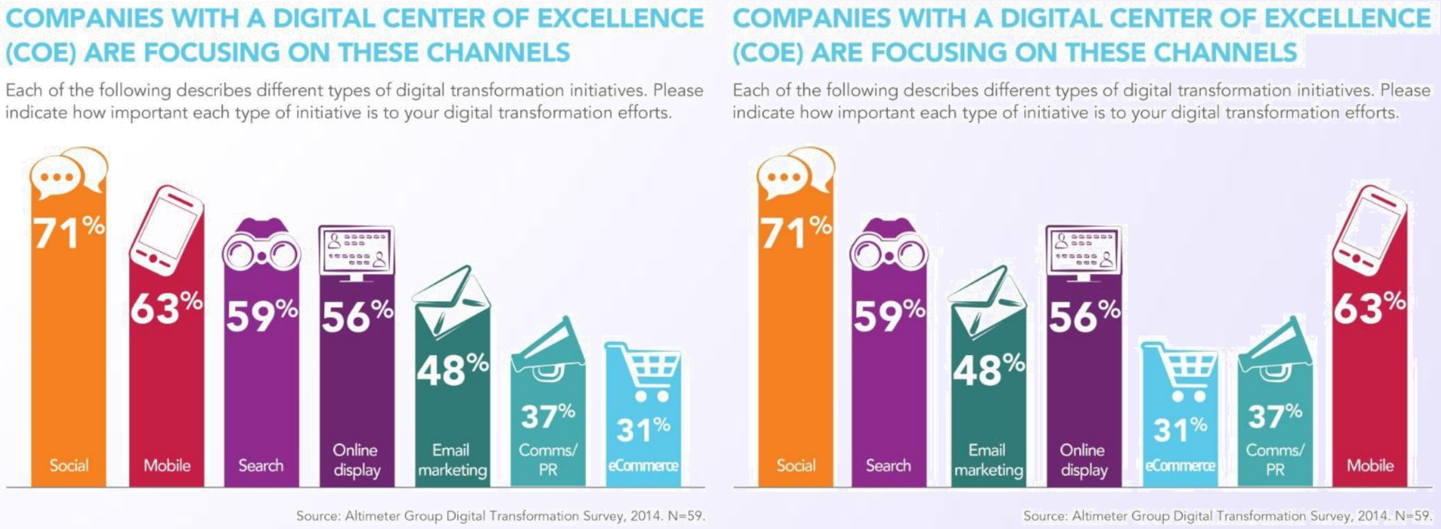
Ablation Study 3

Question: Why are **rank** questions exhibited accuracies the worst?

Hypothesize: MLLMs are sensitive to **the orders of text labels**

- Randomly selected 200 charts where GPT-4.1 ranked correctly and **applied random spatial permutations to the labels**

verify!



(a) Before shuffle (Acc: 100%)

(b) After shuffle (Acc: 76.3%)

Performance drops when the text labels are shuffled

Conclusion and Future Work

- **Systematically study how infographics affect model performance**
  - **Paired infographics and plain charts** reveal the effects of visual elements
  - **Visual-element-based questions** specifically designed for infographics
  - Ablation studies to **identify and analyze performance degradation** of MLLMs on infographic charts

- **Future work**
  - Expand metaphor questions
  - Enhance textual diversity
  - Broaden user studies

清華大學
Tsinghua University

Thank you!

InfoChartQA: A Benchmark for Multimodal Question Answering on Infographic Charts

Tianchi Xie[1,*], Minzhi Lin[1,*], Mengchen Liu[2], Yilin Ye[3], Changjian Chen[4], Shixia Liu[1]

1 Tsinghua University
2 Meta
3 Hong Kong University of Science and Technology
4 Hunan University