清華大學
Tsinghua University

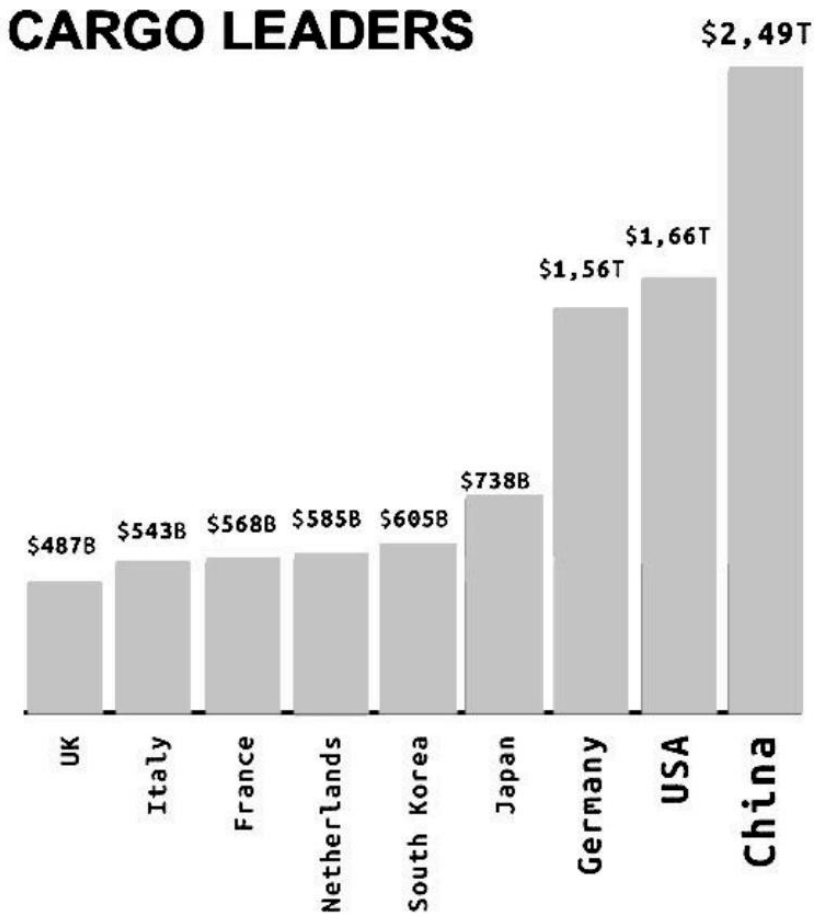
InfoDet: A Dataset for Infographic Element Detection

Jiangning Zhu[1], Yuxing Zhou[1], Zheng Wang[1], Juntao Yao[1],
Yima Gu[1], Yuhui Yuan[2], Shixia Liu[1]

1 Tsinghua University
2 Canva CORE

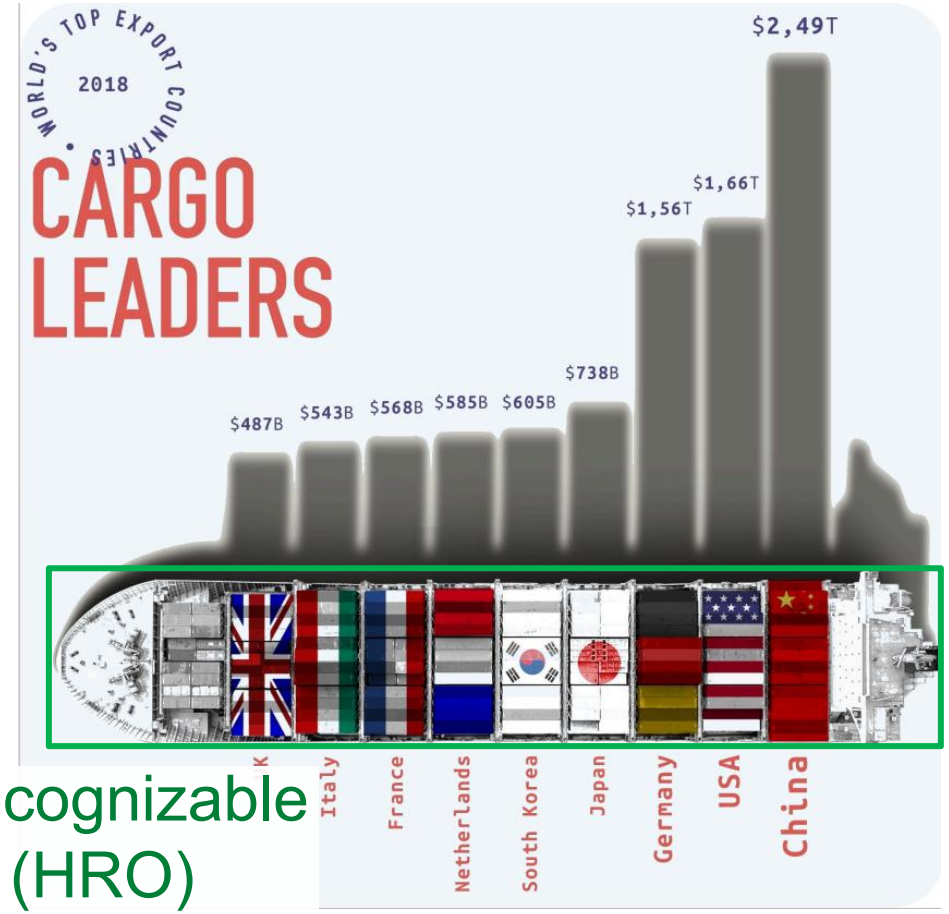
Motivation

- Charts are a fundamental medium for conveying data-driven insights
- Infographics thoughtfully arranges texts, charts, and HROs



Chart

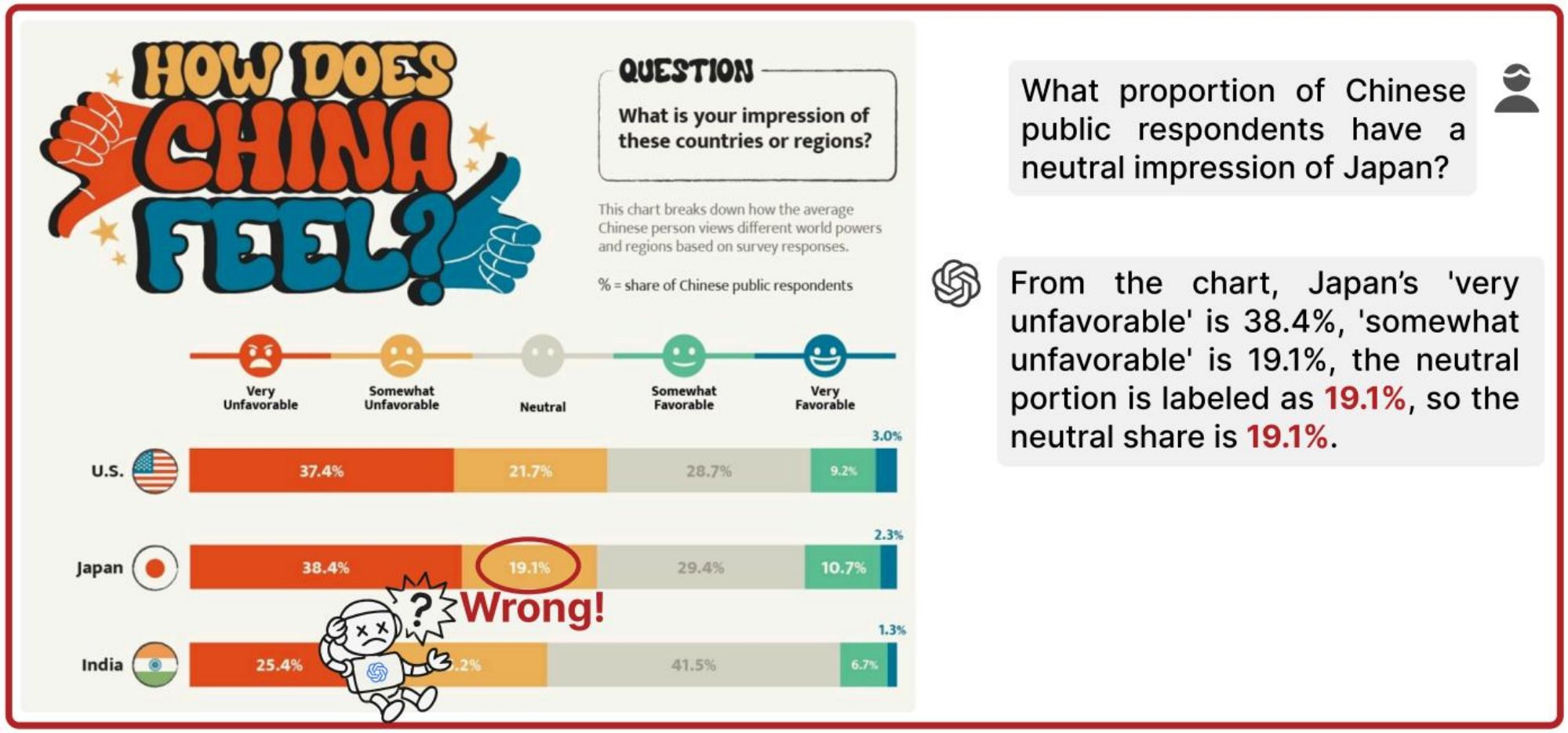
Human-Recognizable Object (HRO)



Infographic

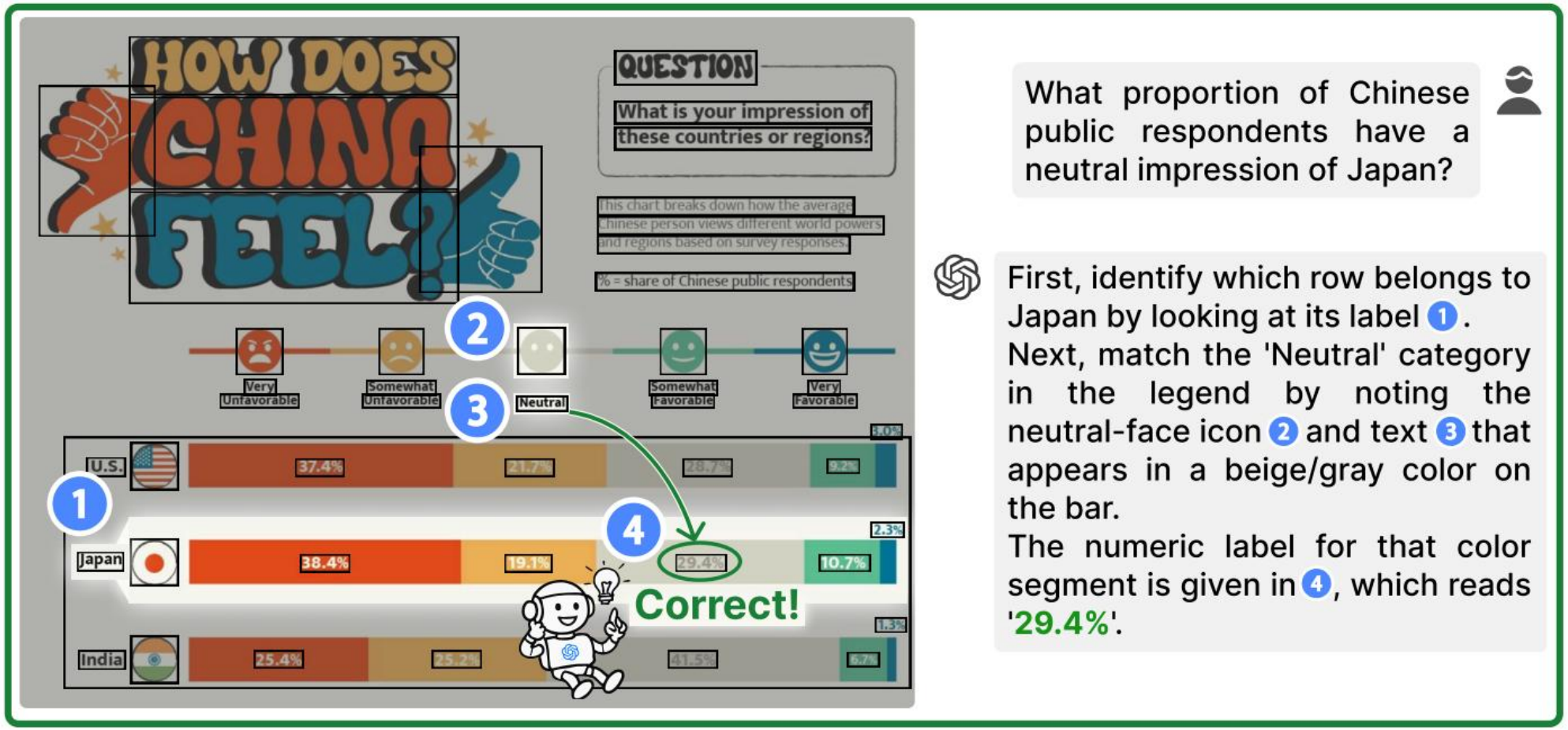
Motivation

- Limitation of VLMs: inaccurate visual grounding of infographic elements



Motivation

- If we provide grounded annotations of elements:



Need a robust object detection model for infographics!

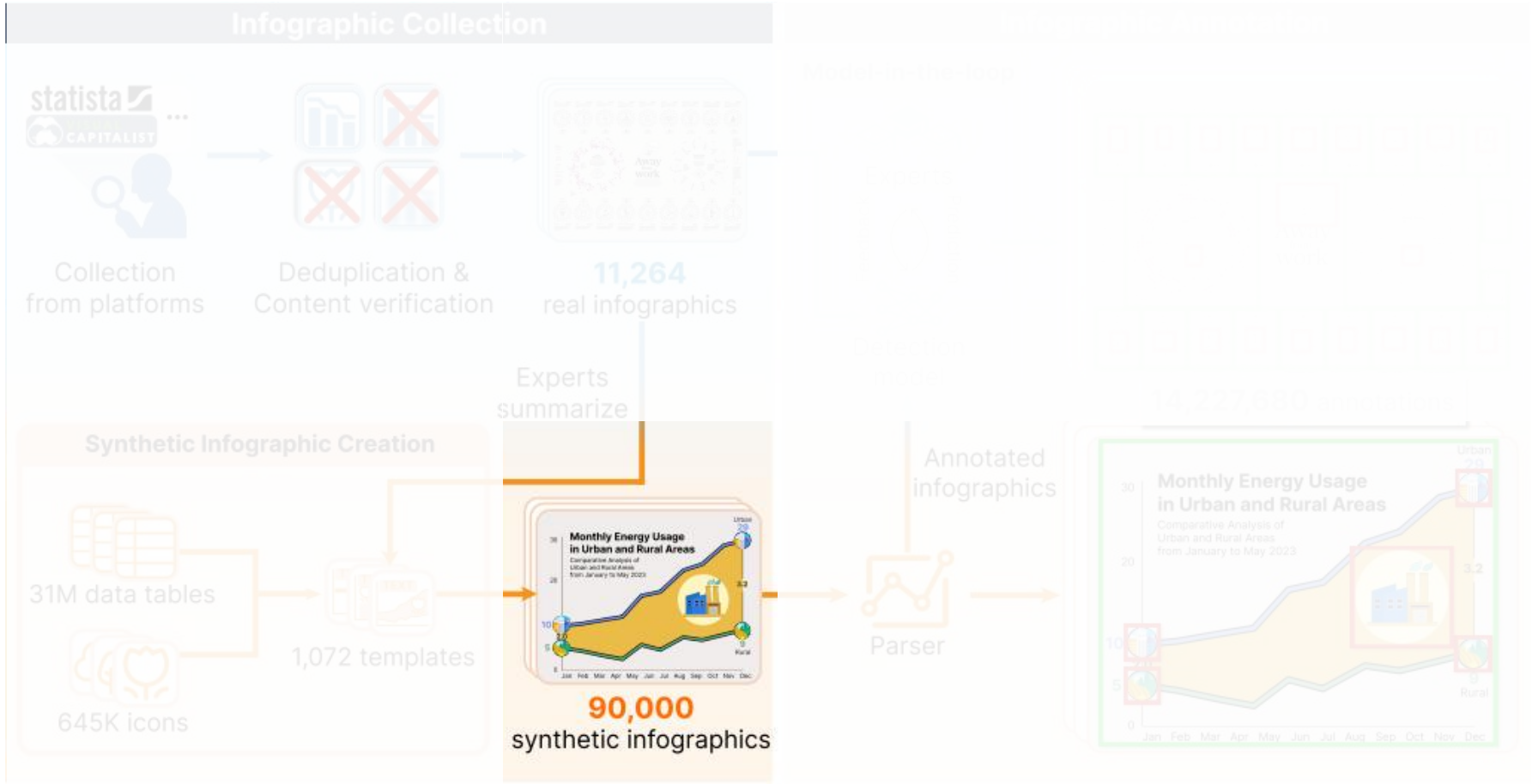
Related Work

- Developing the object detection model requires a diverse set of infographics with accurate annotations

Dataset	# Real	# Synthetic	Infographic?
Borkin et al. (2016)	393	0	✓
FigureQA (Kahou et al., 2018)	0	100,000	-
PlotQA (Methani et al., 2020)	0	224,377	-
Beagle (Battle et al., 2018)	41,000	0	-
VisImages (Deng et al., 2023)	12,267	0	-
VG-DCU (Dou et al., 2024)	4,515	10,682	-
InfoDet (ours)	11,264	90,000	✓

Dataset Construction

- **Problem**: Existing infographic datasets rely on manual annotation, which limits their scale
- **Solution**: Integrate real and synthetic infographics



Annotation Quality

- The annotation precision and recall of the real infographics are comparable to those of commonly used object detection datasets

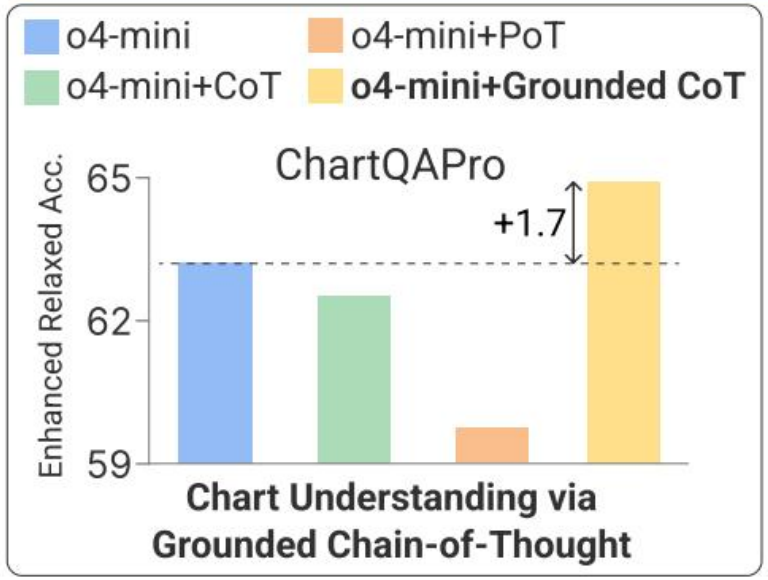
	Precision (%)	Recall (%)
COCO	71.9	83.0
Objects365	91.7	92.0
InfoDet (ours)	93.9	96.7

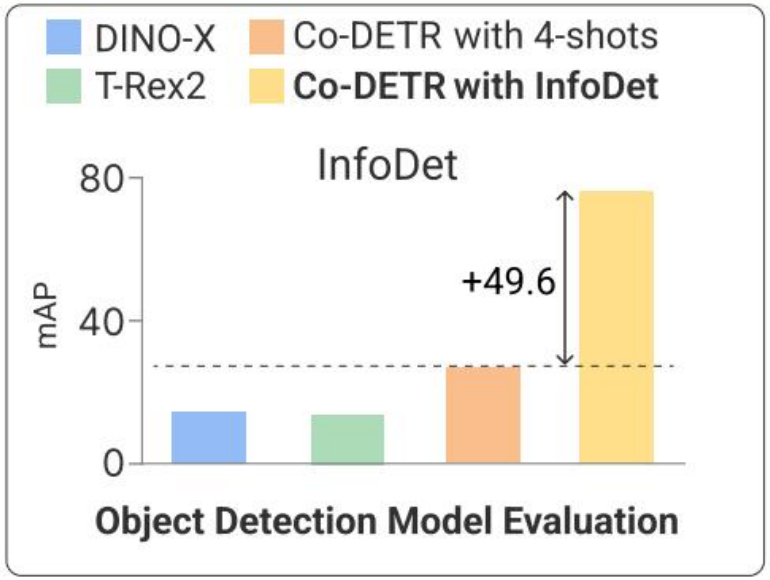
Applications: Overview

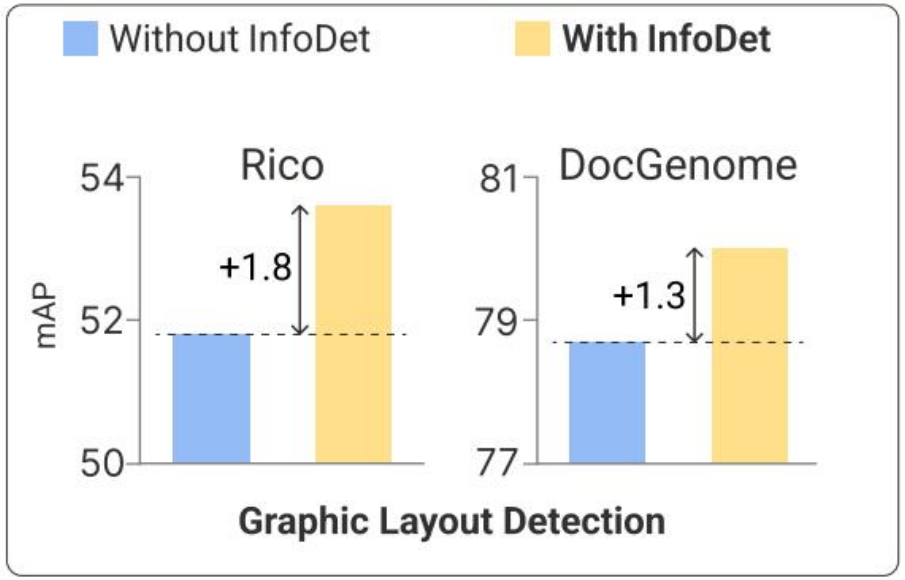
InfoDet

11,264 real infographics
90,000 synthetic infographics

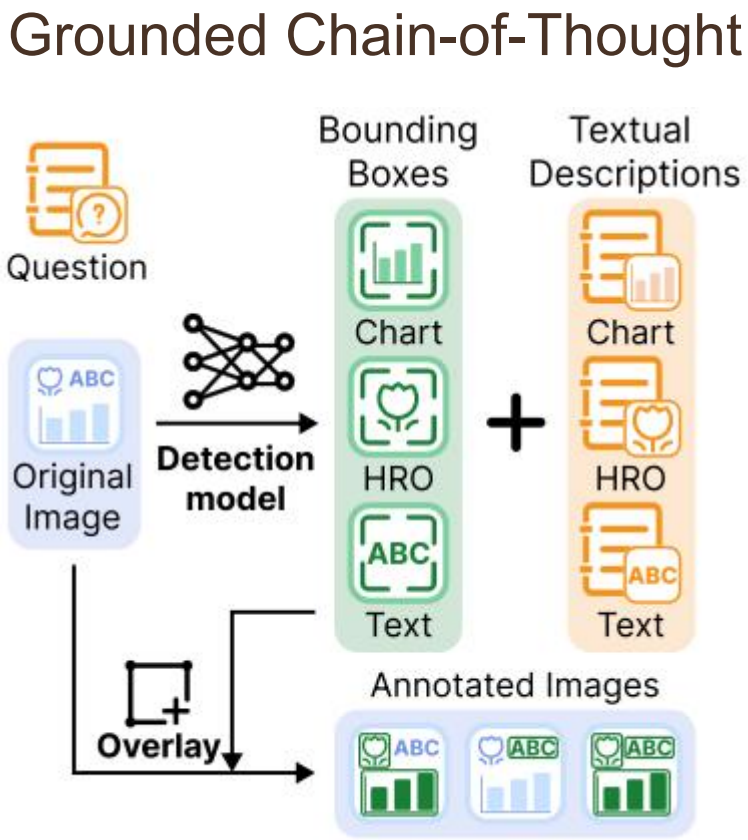
14,227,680 annotations







Application 1: Thinking-with-Boxes



- Dataset: ChartQAPro
- Models: o1, o3, o4-mini
- Baselines:
  - Direct prompting
  - Chain-of-Thought
  - Program-of-Thought

Application 1: Thinking-with-Boxes

Chart Group	o1				o3				o4-mini			
	Direct	CoT	PoT	Grounded CoT (ours)	Direct	CoT	PoT	Grounded CoT (ours)	Direct	CoT	PoT	Grounded CoT (ours)
Plain, Single	57.8	57.8	56.1	**60.1**	56.8	**57.7**	57.5	57.2	58.1	57.9	55.3	**60.6**
Plain, Multiple	63.7	65.1	62.2	**65.4**	62.8	61.0	58.8	**63.4**	66.7	66.1	62.3	**66.9**
Infographic, Single	66.4	64.3	60.9	**67.8**	64.9	59.5	64.2	**67.7**	67.4	64.4	67.5	**68.4**
Infographic, Multiple	66.0	67.6	66.8	**71.9**	66.0	64.9	64.2	**68.8**	70.6	69.2	64.7	**72.5**
Overall	61.4	61.9	60.0	**64.1**	60.6	60.0	59.5	**61.6**	63.2	62.5	59.7	**64.9**

- Prompting the latest VLMs to think step-by-step or write Python code does not significantly improve their performance
- Our method enhances chart understanding performance by providing grounded infographic elements

Application 2: Evaluating Object Detection Models

- Models
  - Seven foundation models: RegionCLIP, Detic, Grounding DINO, GLIP, MQ-GLIP, T-Rex2, and DINO-X
  - Four traditional models: Faster R-CNN, YOLOv3, RTMDet, Co-DETR
- Evaluation protocol
  - Zero-shot prompting
  - Few-shot prompting
  - Standard fine-tuning

Application 2: Evaluating Object Detection Models

(a) Zero-shot prompting

Model	Average Precision (AP) Chart	Average Precision (AP) HRO	Average Recall (AR) Chart	Average Recall (AR) HRO
RegionCLIP	0.8	3.6	13.9	24.9
Detic	1.8	4.4	23.7	11.3
Grounding DINO	12.6	12.2	**63.2**	**46.0**
GLIP	13.5	11.2	44.9	33.2
MQ-GLIP	13.5	11.2	44.9	33.2
DINO-X	**14.0**	**15.0**	29.4	29.1

(b) Few-shot prompting, 4-shots

Model	Average Precsion (AP) Chart	Average Precsion (AP) HRO	Average Recall (AR) Chart	Average Recall (AR) HRO
MQ-GLIP	**16.2**	15.5	**43.5**	**40.7**
T-Rex2	12.2	**16.2**	21.8	24.7

(c) Standard fine-tuning, 4-shots

Model	Average Precsion (AP) Chart	Average Precsion (AP) HRO	Average Recall (AR) Chart	Average Recall (AR) HRO
RegionCLIP	6.8	14.7	15.5	22.9
Detic	19.6	14.2	37.0	22.8
Faster R-CNN	3.4	1.0	10.8	1.5
YOLOv3	5.5	4.0	16.2	13.1
RTMDet	12.8	18.9	44.2	49.1
Co-DETR	**27.6**	**25.5**	**53.4**	**49.7**

(d) Standard fine-tuning, InfoDet

Model	Average Precsion (AP) Chart	Average Precsion (AP) HRO	Average Recall (AR) Chart	Average Recall (AR) HRO
RegionCLIP	10.1	23.3	17.5	28.6
Detic	39.6	34.3	57.4	47.7
Faster R-CNN	78.9	49.0	80.8	52.7
YOLOv3	14.7	25.5	43.2	35.7
RTMDet	83.7	53.6	86.4	59.9
Co-DETR	**88.2**	**64.0**	**89.8**	**69.5**

- Zero-shot and few-shot prompting exhibit limited performance
- Standard fine-tuning improves performance

Application 3: Graphic Layout Detection

- Datasets: Rico, DocGenome, PosterLayout

Pre-Training	Rico	DocGenome	PosterLayout
-	42.1	69.0	62.5
InfoDet	50.6	74.4	73.6
ImageNet-22K, Objects365, COCO	51.8	78.7	74.9
ImageNet-22K, Objects365, COCO, InfoDet	53.6	80.0	76.0
ImageNet-22K, Objects365, COCO, InfoDet w. HGM	**54.4**	**80.8**	**76.4**

- Pre-training on InfoDet improves model performance when fine-tuned on Rico, DocGenome, and PosterLayout

Thank you!

InfoDet: A Dataset for Infographic Element Detection

Jiangning Zhu[1], Yuxing Zhou[1], Zheng Wang[1], Juntao Yao[1],
Yima Gu[1], Yuhui Yuan[2], Shixia Liu[1]

1 Tsinghua University
2 Canva CORE