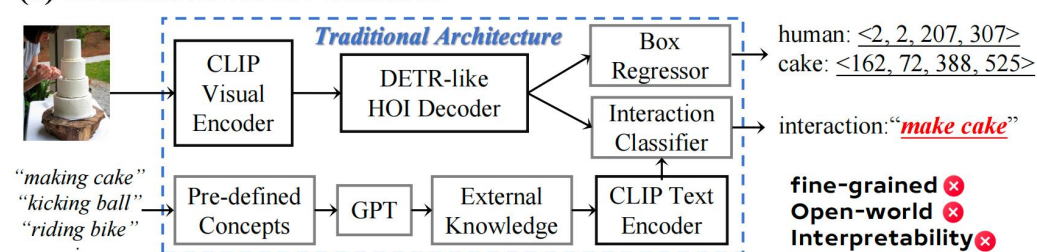


Eastman Z Y Wu<sup>1,2,3</sup>, Yali Li<sup>1,2,3</sup>, Shengjin Wang<sup>1,2,3†</sup>

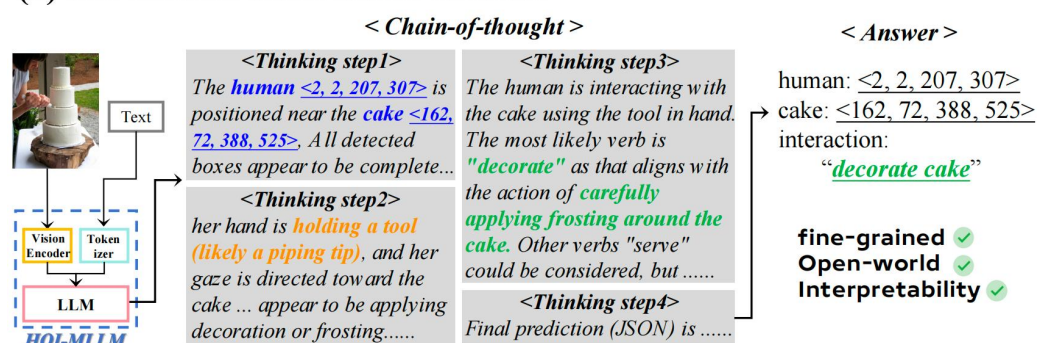
<sup>1</sup>Department of Electronic Engineering, Tsinghua University <sup>2</sup>BNRist, China <sup>3</sup>NERC-DAE Detection

## 1. Motivation

### (a) Traditional HOID methods



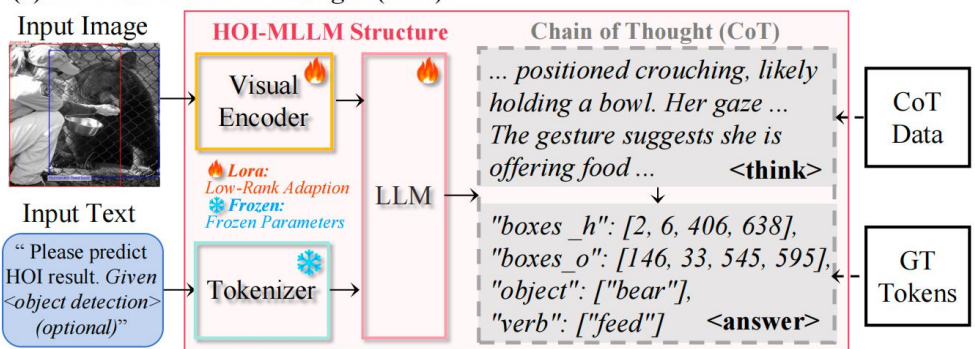
### (b) Our MLLM-based HOID methods



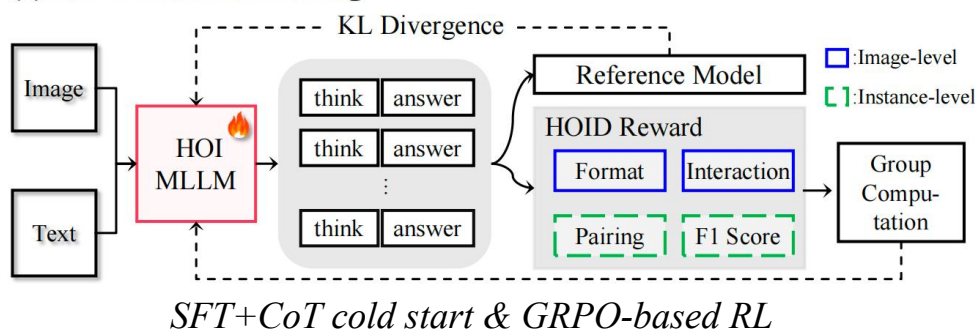
Traditional HOID methods rely on predefined concepts suffering from insufficient semantic reasoning, limited open-world knowledge and weak interpretability.

## 2. Our HOI-MLLM Framework

### (a) SFT with Chain of Thought (CoT)



### (b) GRPO-based Post-training



## 3. Quantitative Results

HICO-DET	mF1			mPre	mRec
	Full	Rare	N.Rarwane		
HOI-MLLM	24.28	25.75	23.27	26.72	26.26

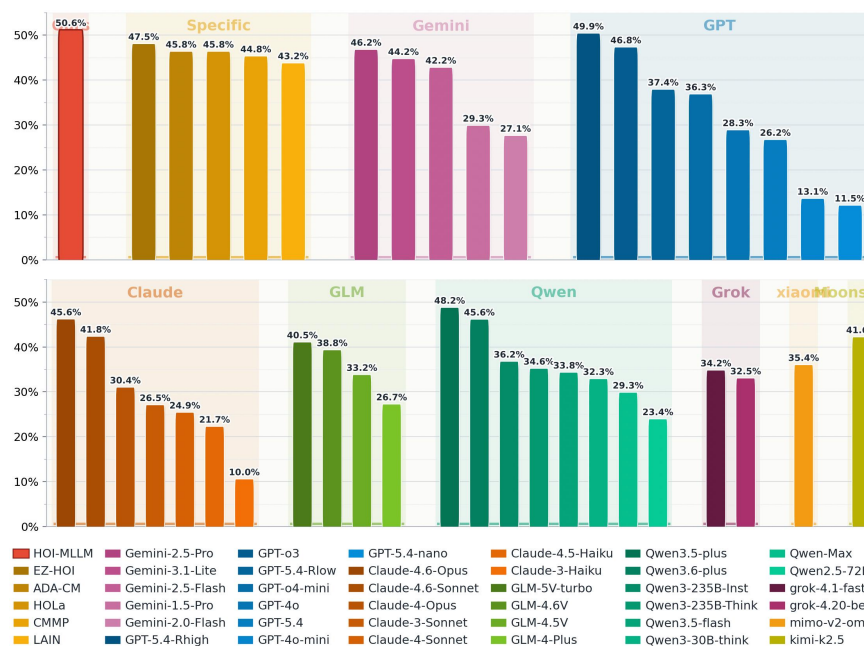
  

End-to-End OVD & HOID					
PViC [28]	28.68	23.61	30.20	25.29	44.10
CMMP [19]	26.08	23.16	26.95	24.40	36.64
CMMP+ [19]	33.24	30.24	34.14	33.24	43.52
ADA-CM [20]	28.20	24.64	29.26	27.37	35.62
ADA-CM+ [20]	<b>34.14</b>	<b>31.33</b>	<b>34.98</b>	<b>32.25</b>	<b>44.69</b>
EZ-HOI [29]	27.82	23.25	29.18	26.00	36.58
EZ-HOI+ [29]	31.44	29.43	32.04	30.57	40.20
HOI-MLLM (ours)	30.32	27.12	31.68	<b>35.71</b>	36.21

Oracle Setting					
PViC+ [28]	44.41	46.49	43.79	39.73	61.63
CMMP [19]	36.55	37.27	36.34	30.74	59.22
CMMP+ [19]	44.76	46.48	44.25	40.77	60.32
ADA-CM [20]	39.78	36.24	40.84	36.68	53.68
ADA-CM+ [20]	45.81	42.57	46.78	41.97	61.39
EZ-HOI [29]	42.72	33.42	45.50	37.82	59.48
EZ-HOI+ [29]	47.52	43.81	48.62	43.32	<b>63.24</b>
HOI-MLLM (ours)	<b>50.61</b>	<b>51.16</b>	<b>49.90</b>	<b>51.37</b>	56.47

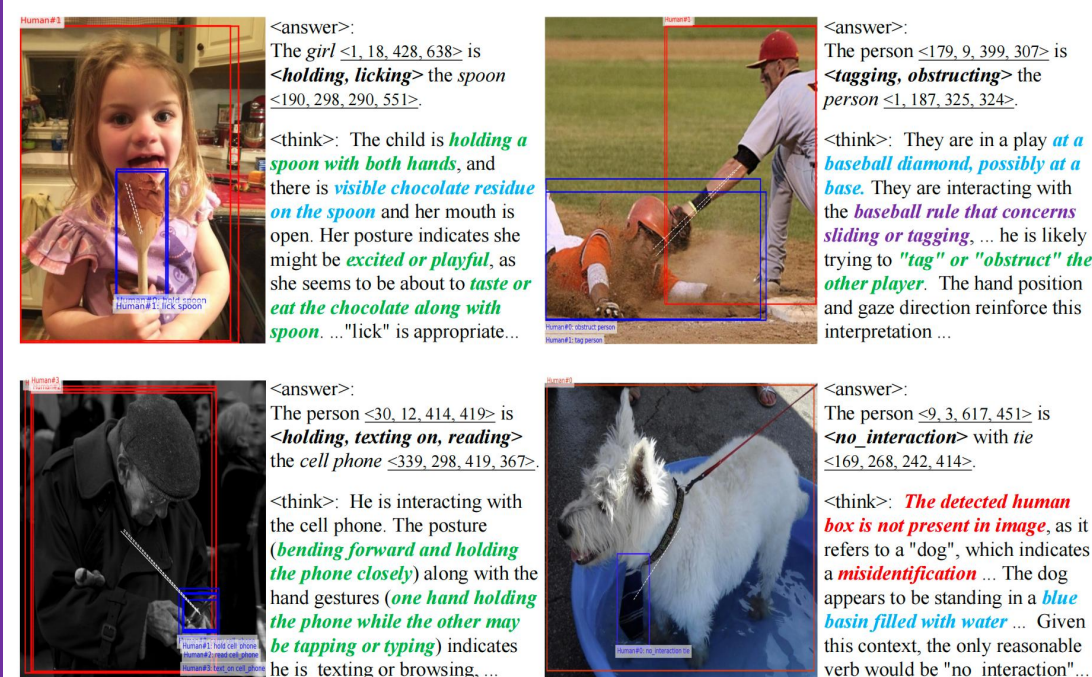
## 4. Comparison with MLLMs



## 5. Ablation Results

#	SFT	Mixed	CoT	GRPO	mF1	mPre	mRec
A1	×	×	×	×	31.31	41.92	27.46
A2	✓	×	×	×	64.07	71.00	59.46
A3	✓	✓	×	×	66.97	68.66	66.07
A4	✓	✓	✓	×	68.10	73.85	63.71
A5	✓	✓	✓	✓	69.28	74.44	65.56

## 6. Chain-of-thought Visualization



Our HOI-MLLM showcases advantages beyond traditional methods: blue indicates open-vocabulary detections, purple reflects open-world knowledge, red demonstrates reflective reasoning, and green emphasizes fine-grained visual cues such as posture, gesture.

## 7. Zero-shot Results

