ORIC: Benchmarking Object Recognition under Contextual Incongruity in Large Vision-Language Models

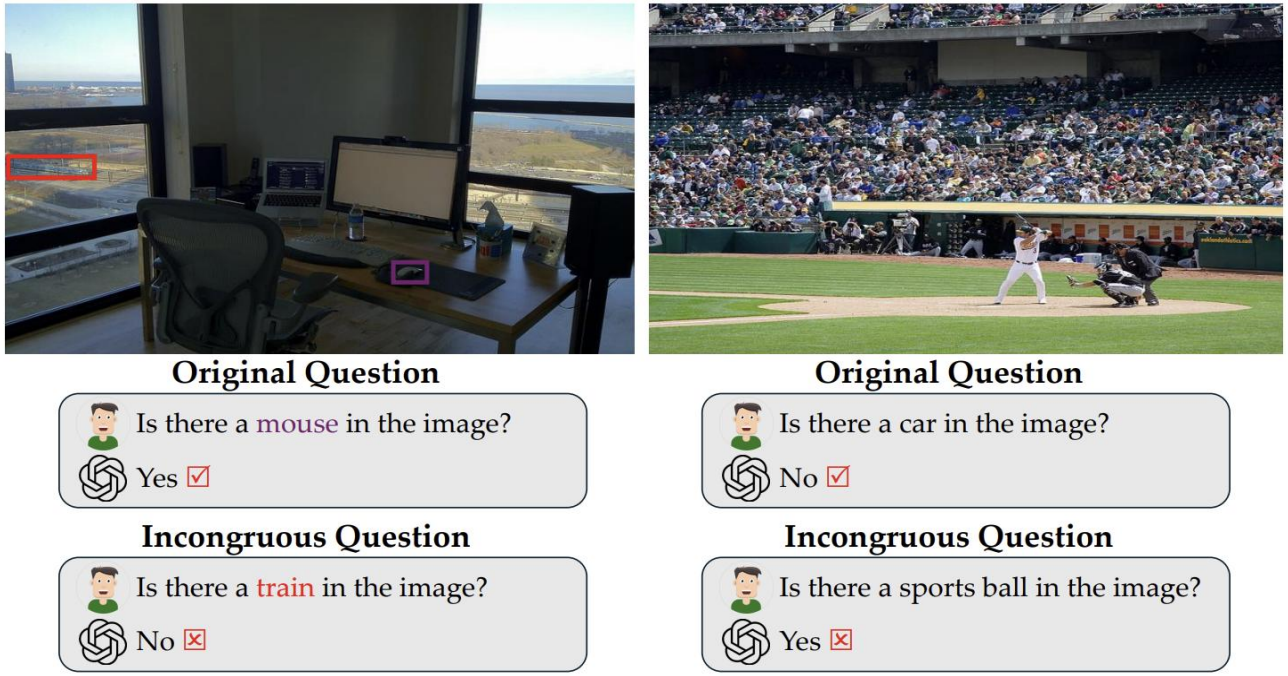
Zhaoyang Li*, Zhan Ling*, Yuchen Zhou, Litian Gong, Erdem Bıyık, Hao Su
{zhl165, z6ling, yuz256, haosu}@ucsd.edu, lgong024@ucr.edu, biyik@usc.edu

UC San Diego

USC University of Southern California

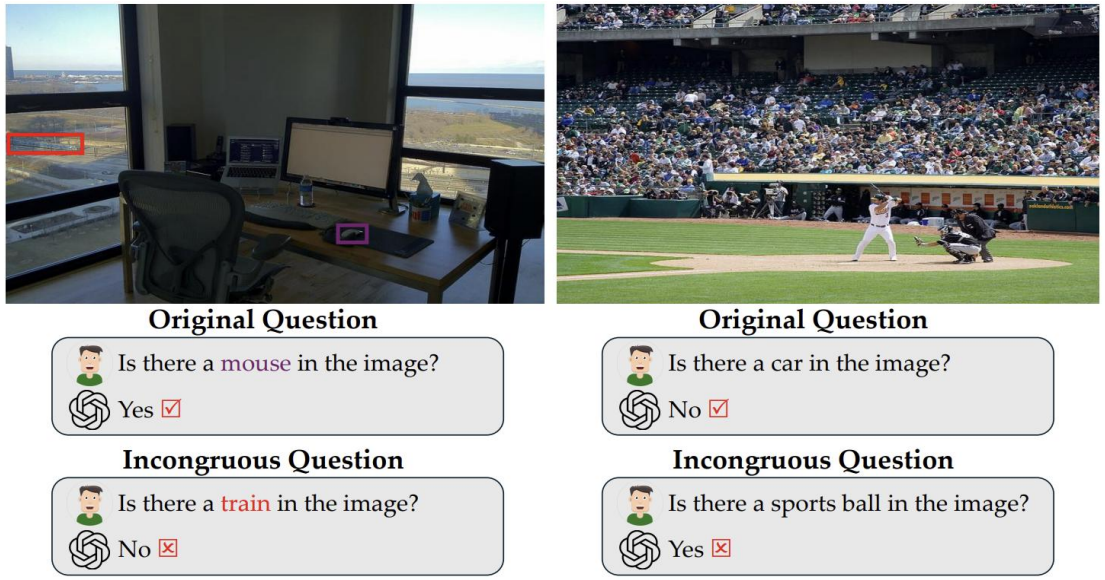
UC RIVERSIDE

Motivation: Contextual Incongruity Causes Recognition Failures



- LVLMs achieve strong VQA and captioning results, but still suffer from object misidentification and object hallucination.
- Existing benchmarks often preserve object–context compatibility. This paper studies the high-uncertainty case where ROI evidence
- This paper studies the high-uncertainty case where ROI evidence mismatches with contextual priors.

Motivation: Problem Formulation and Controlled Evidence



Binary existence query

$$P(a \mid q, I), a \in \{yes, no\}, I = (ROI, context)$$

- q: question; I: image; ROI: region of interest containing the queried object.
- In normal benchmarks, ROI evidence and context priors usually agree.
- In contextual incongruity, local evidence and scene priors were mismatched:
  - Present but unexpected object: model may answer no.
  - Absent but plausible object: model may answer yes.

Method Part 1: LLM-Guided Positive Question Construction

Label “Yes” Questions

Method Part 2: CLIP-Guided Negative Question Construction

Label “No” Questions

Method Part 3: ORIC-Driven Uncertainty Mitigation

- Fine-tune Qwen3-VL-8B-Instruct on 600 ORIC-style binary questions using verifiable rewards.
- Reward combines answer correctness and output-format compliance:

$$r_i = r_{acc,i} + r_{fmt,i}$$

- Group-normalized reward:

$$\hat{r_i} = \frac{r_i - mean(\{r_i\}_{j=1}^{G})}{std\left(\{r_i\}_{j=1}^{G}\right) + \epsilon}$$

- GRPO optimizes relative sample quality without a PPO-style critic.
- R1-style prompt forces explicit reasoning and verifiable final answer:

<REASONING> … </REASONING> <SOLUTION> yes/no </ SOLUTION >

Experimental Method: Design and Process

- Build ORIC-Bench from MSCOCO validation images.
  - Images: 1,000; Questions: 1,000 yes and 1,000 no.

- For each selected image pair:
  - generate present-object queries using LLM-guided sampling;

  - generate absent-object queries using CLIP-guided sampling.

- Evaluate 18 LVLMs and 2 open-vocabulary detectors.

- Use four prompt variants for LVLMs:
  - “Is there {object} in the image?”

  - “Does the image contain {object}?”

  - “Have you noticed {object} in the image?”

  - Can you see {object} in the image?”

- Report macro precision, recall, F1, class-wise metrics, and yes-prediction proportion.

Experimental Results: Main ORIC-Bench Performance

Model	POPE				ORIC-Bench			
	Precision	Recall	F1 Score	YP (%)	Precision	Recall	F1 Score	YP (%)
Closed-source								
GPT-5-2025-08-07	89.06	88.60	**88.56**	44.62	79.50	78.75	**78.61**	42.12
Encoder-based								
InternVL3-9B	88.8	88.69	**88.68**	47.96	77.33	76.95	76.87	44.60
Qwen3-VL-8B-Instruct	88.13	88.04	88.03	47.66	79.93	79.61	**79.55**	44.94
Encoder-free								
Emu3-Chat	87.43	86.72	**86.66**	43.25	67.74	65.79	**64.78**	33.41

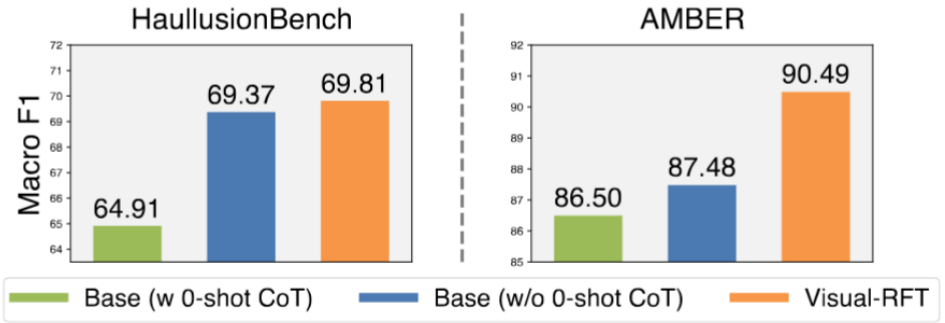
Model	POPE			ORIC-Bench		
	Small	Medium	Large	Small	Medium	Large
Emu3-Chat	68.22	80.97	94.19	38.73	56.61	71.99
GPT-5-2025-08-07	78.24	88.48	94.30	67.85	71.69	84.34
InternVL3-9B	82.29	90.43	96.34	63.63	77.61	86.45
Qwen3-VL-8B-Instruct	79.96	89.71	96.40	69.96	77.67	85.24

- Qwen3-VL-8B-Instruct achieves the best overall F1: 79.55.
- Even GPT-5 remains below 80 F1, showing persistent difficulty.
- Detectors are competitive but weaker at holistic absence reasoning.
- Performance drops across object sizes and model families, indicating that contextual incongruity, not only scale, drives uncertainty

Experimental Results: ORIC-Driven Uncertainty Mitigation

- Approach: Visual-RFT (GRPO + verifiable rewards)
  - Fine-tune Qwen3-VL-8B on 600 ORIC-style training samples
- Visual-RFT mitigates uncertainty-driven errors
- More aligned with human reasoning
- Strong generalization

Method	Overall			
	Precision	Recall	F1	YP (%)
(a) Standard ORIC-Bench Evaluation				
w 0-shot CoT	78.69	78.50	78.46	46.23
w/o 0-shot CoT	79.93	79.61	79.55	44.94
Visual-RFT	**83.55**	**82.88**	**82.79**	43.05
(b) Human-Labeled Ground Truth on ORIC-Bench				
w/o 0-shot CoT	78.70	78.63	78.63	47.14
Visual-RFT	**84.03**	**83.64**	**83.62**	44.72



Deficiencies: Limitations of the Current Method

- **Dataset scope:** ORIC-Bench is constructed mainly from MSCOCO; broader domains remain underexplored.
- **Binary setting:** current questions focus on yes/no object existence, not open-ended recognition or localization.
- **Mitigation scale:** Visual-RFT uses 600 samples; stronger training regimes and larger incongruity datasets may further improve robustness.

Future Research: Improvement Directions

- Extend ORIC to more diverse datasets, including indoor scenes, robotics, medical images, and long-tail categories.
- Improve training with larger ORIC-style data and human preference signals.
- Apply contextual-incongruity evaluation to safety-critical embodied AI and robotics.

CVPR JUNE 3-7, 2026 DENVER COLORADO

Thank You

Key Takeway

ORIC shows that contextual incongruity is a major source of uncertainty in LVLM object recognition, and ORIC–style Visual–RFT can improve robustness

https://github.com/ZhaoyangLi-1/ORIC