ICLR
International Conference On
Learning Representations

AQuA: Toward Strategic Response Generation for Ambiguous Visual Questions

Jihyoung Jang, Hyounghun Kim

Motivation

- **Question**

What color is this bat?

Motivation

- **Question**

What color is this bat?

- **Human**

There are a lot of bats, so what kind are they?

Motivation

- **ChatGPT**

The bat in the foreground (leaning against the wall, farthest to the right) is black with a red handle

- **Gemini**

Based on the image, the bats in the foreground are a variety of colors, including natural wood, ……

- **Qwen**

The bat in the image is red and black.

Introduction

- **Question**

What color is this bat?

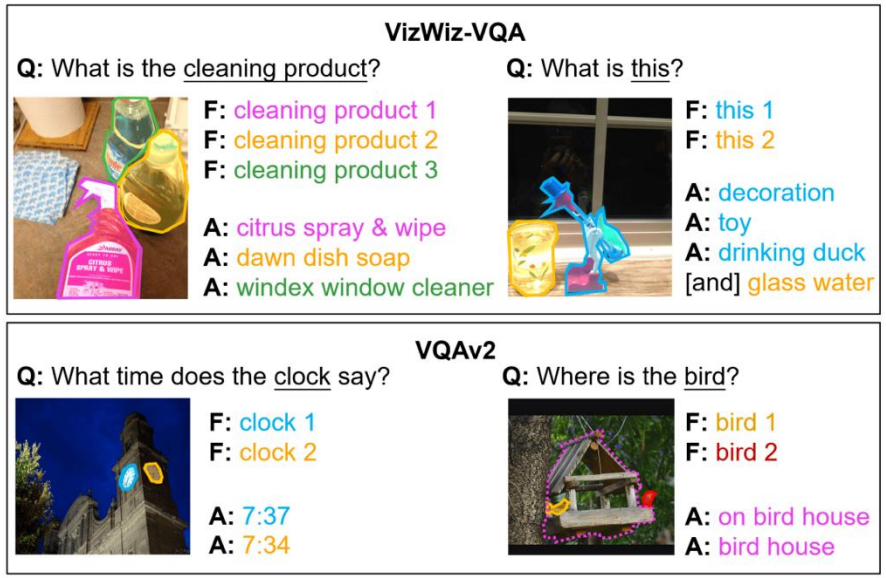
- **Human**

There are a lot of bats, so what kind are they?

In the given image context, ‘this bat’ is ambiguous.

Introduction

- Ambiguous visual question answering is challenging for VLMs.
- Pervious work focuses on asking clarification questions when the model is uncertain.



FocusAmbiguity (Chen et al., 2025)

ClearVQA (Jian et al., 2025)

AQuA

- Unlike prior datasets, we propose a dataset that enables strategic responses based on ambiguity levels.
- Ambiguity doesn’t always require clarification!

AQuA

Q: What food is on the baking tray?
A: A pizza is on the baking tray.

Level 0 (Unambiguous Questions)

No ambiguity. Clear VQA.

AQuA

Q: What toppings are on this?
A: The term “this” points to the main item, the hot dog in the bun, and it’s topped with mustard and ketchup.

Level 1 (Low-Level Referential Ambiguity)

Ambiguous terms, but resolvable from context.
Inference is sufficient, no clarification required.

AQuA

Q: What is this player doing right now?
A: The reference might be to the Mariners fielder in white at first, ……,
or to the runner in gray with a blue helmet and number 49, ……

Level 2 (Multiple Valid Interpretations)

2-3 of plausible targets. Enumerating all candidates is more efficient than asking for clarification.

AQuA

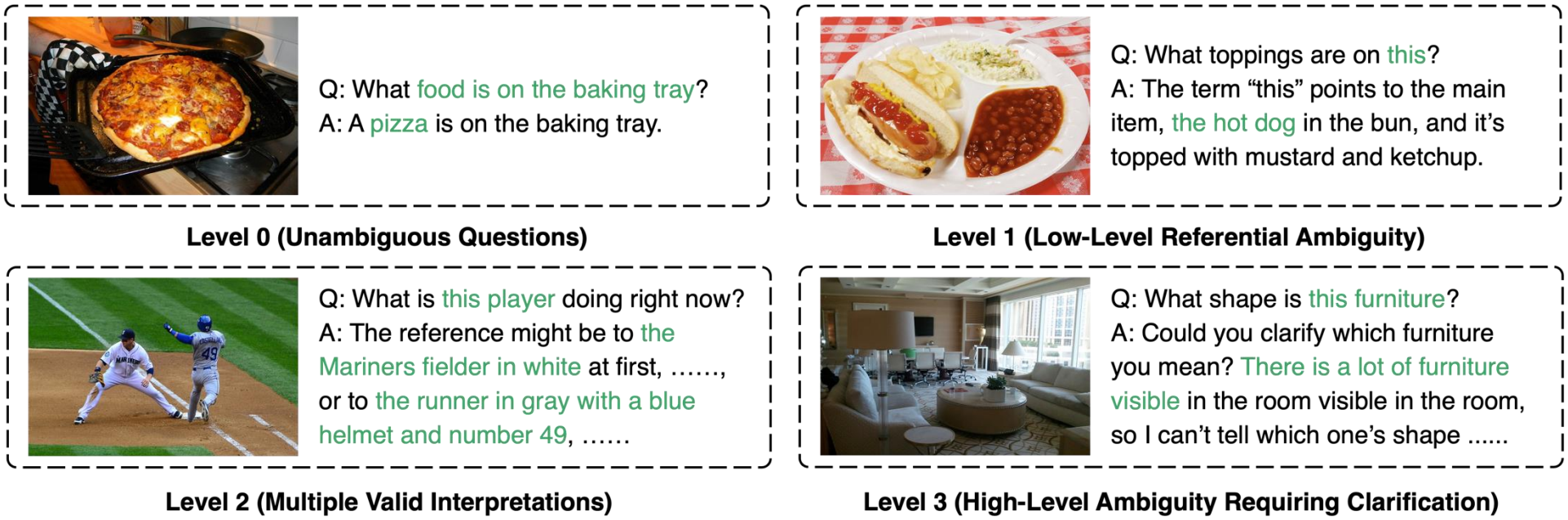
Q: What shape is this furniture?
A: Could you clarify which furniture you mean? There is a lot of furniture visible in the room visible in the room, so I can't tell which one's shape .....

Level 3 (High-Level Ambiguity Requiring Clarification)

Too many plausible targets.
Clarification is necessary to resolve ambiguity.

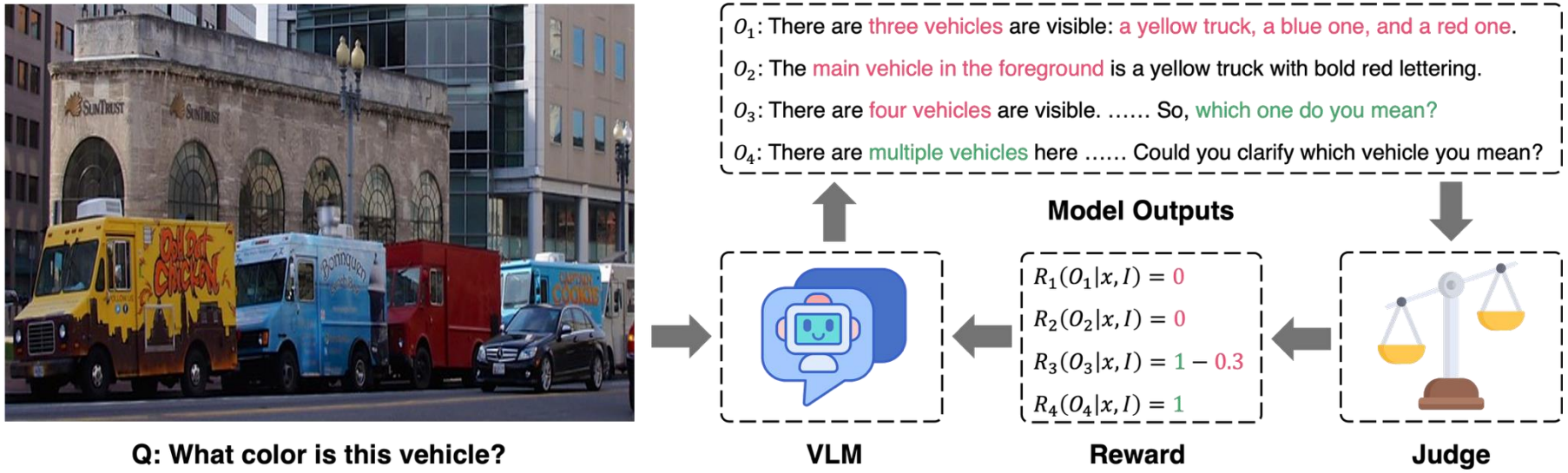
AQuA

- We collect 7.2K samples in total: 3.6K for training and 3.6K for evaluation.
- Each split is evenly balanced across the four ambiguity levels, with 0.9K instances per level.



Model

- We train VLMs using a two-stage pipeline consisting of SFT followed by GRPO.
- Why? SFT alone does not reliably select the right strategy under different levels of ambiguity.
- GRPO provides explicit rewards for strategy-aware outputs, improving contextually appropriate decision-making.



Experiments

- **Factual consistency**
  - Indicates that the model's response is faithful to the content of the given image, even if not all details are included, and is judged in a binary manner (Grounded or Ungrounded).

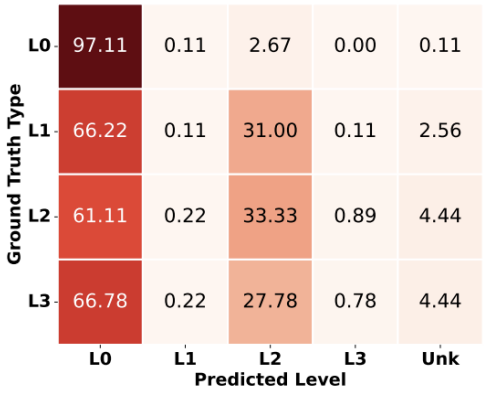
- **Strategic accuracy**
  - Measures whether the response strategy matches the ground-truth ambiguity level. If a response cannot be reliably mapped to any of the four defined levels, it is assigned an Unknown label.

Experiments

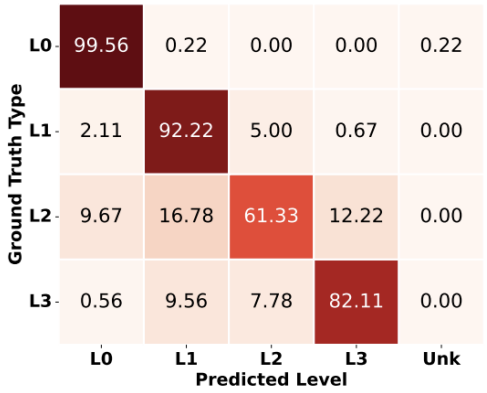
Model	Factual Acc.		Strategic Acc.					
	Grounded	Ungrounded	Level 0	Level 1	Level 2	Level 3	Overall	Unk
Zero-shot								
Qwen2.5-VL-3B-Instruct	79.86	20.14	97.11	0.11	33.33	0.78	32.83	104
Qwen2.5-VL-72B-Instruct	89.33	10.67	99.56	0.56	2.11	0.89	25.78	12
InternVL3-2B-Instruct	76.63	23.37	96.0	2.33	3.56	1.89	25.95	138
InternVL3-78B-Instruct	80.5	19.5	96.0	2.11	3.0	5.67	26.7	133
GPT-5	98.4	1.6	89.67	0.67	0.33	0.78	22.86	178
Gemini 2.5 Flash	91.89	8.11	99.00	5.22	4.44	0.89	27.39	9
Chain-of-Thought (CoT)								
Qwen2.5-VL-3B-Instruct	78.22	21.78	95.89	8.33	5.67	3.78	28.42	60
Qwen2.5-VL-72B-Instruct	86.97	13.03	93.0	13.78	2.78	1.33	27.72	10
InternVL3-2B-Instruct	76.08	23.92	97.67	2.44	1.33	1.11	25.64	54
InternVL3-78B-Instruct	79.75	20.25	96.78	5.22	3.67	12.33	29.5	74
GPT-5	98.83	1.17	97.33	3.78	0.67	1.11	25.72	14
Gemini 2.5 Flash	91.64	8.36	98.0	7.89	3.56	0.22	27.42	22
Strategy Prompting								
Qwen2.5-VL-3B-Instruct	88.08	11.92	99.78	0.22	0.22	1.44	25.42	8
Qwen2.5-VL-72B-Instruct	91.5	8.5	99.78	5.89	17.11	46.11	42.22	12
InternVL3-2B-Instruct	68.42	31.58	93.33	1.22	4.0	10.11	27.17	152
InternVL3-78B-Instruct	86.44	13.56	96.89	5.56	5.89	14.11	30.61	64
GPT-5	99.17	0.83	94.56	59.0	10.67	4.78	42.25	19
Gemini 2.5 Flash	94.08	5.92	99.11	8.0	10.68	30.11	36.98	35
AQuA Tuned Models								
Qwen2.5-VL-3B-Tuned	81.06	18.94	99.56	77.0	82.22	86.33	86.28	1
InternVL3-2B-Tuned	80.44	19.56	98.78	80.0	59.67	78.0	79.11	12

Experiments

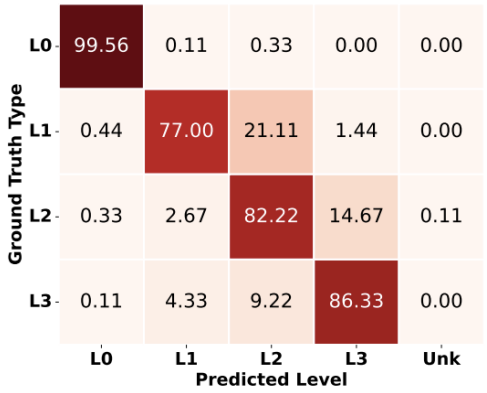
Model	Factual Acc.		Strategic Acc.					
	G	U	Level 0	Level 1	Level 2	Level 3	Overall	Unk
Qwen2.5-VL-3B-Tuned (SFT)	82.78	17.22	99.56	92.22	61.33	82.11	83.81	2
Qwen2.5-VL-3B-Tuned (SFT+GRPO)	81.06	18.94	99.56	77.0	82.22	86.33	86.28	1
InternVL3-2B-Tuned (SFT)	66.08	33.92	99.22	82.67	37.67	74.11	73.42	2
InternVL3-2B-Tuned (SFT+GRPO)	80.44	19.56	98.78	80.0	59.67	78.0	79.11	12



(a) Zero-shot



(b) SFT



(c) SFT+GRPO

After SFT+GRPO,
VLMs learn to select appropriate strategies.

Examples

Q: What color is this round cake on the cooling rack?
Zero-shot: The round cake on the cooling rack is pink.
Ours: There are multiple round cakes on the cooling rack, ... Could you specify which cake you mean .....

Q: What color is this helmet?
Zero-shot: I am not sure what color the helmet is. It can be seen yellow, green, or black.
Ours: There are many helmets visible on different people ..... Could you specify which helmet you mean perhaps by the person's position .....

Q: How many of these are on the plate?
Zero-shot: There are three items on the plate: two doughnuts and three strawberries.
Ours: It could mean the two ring-shaped pastries with red fruit filling ..... or the three bright red strawberries .....

Highlights

- Current VLMs struggle to handle ambiguity strategically.
- To address this, we propose AQuA.
- AQuA systematically organizes response strategies by levels of ambiguity.
- Models trained on AQuA learn to choose appropriate strategies based on the degree of ambiguity.

Thank you

Paper

Project Page

Jihyoung Jang, Hyounghun Kim

{jihyoung, h.kim}@postech.ac.kr

ICLR
International Conference On
Learning Representations

The PATHFINDER 1986 – 2026
POSTECH
포항공과대학교