THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

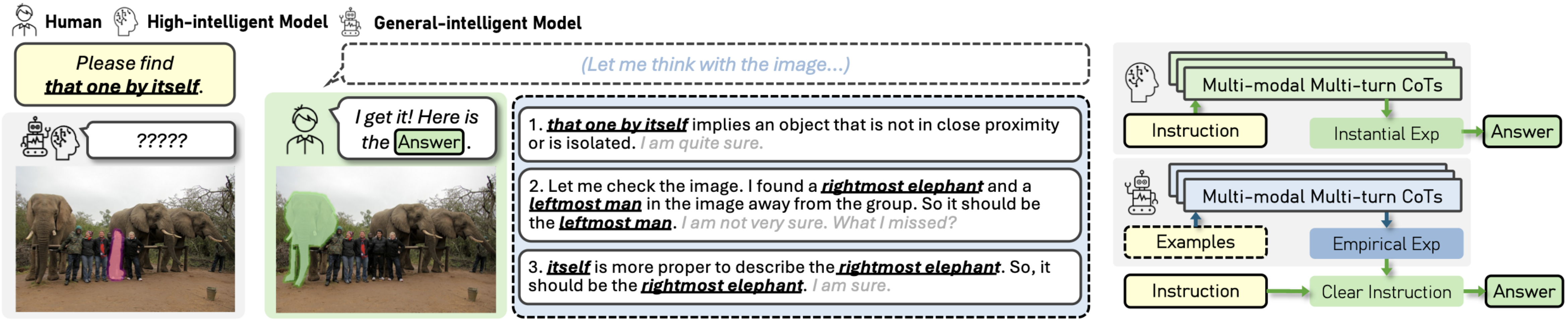
Visual-O1: Understanding Ambiguous Instructions via Multi-modal Multi-turn Chain-of-thoughts Reasoning

Minheng Ni, Yutao Fan, Lei Zhang, Wangmeng Zuo

Hong Kong Polytechnic University
Harbin Institute of Technology

1 Introduction

Overview of Understanding Ambiguous Instruction


Human
High-intelligent Model
General-intelligent Model
Please find ***that one by itself***.
?????
I get it! Here is the Answer.
(Let me think with the image...)
Multi-modal Multi-turn CoTs
Instruction
Instantial Exp
Answer
Multi-modal Multi-turn CoTs
Examples
Empirical Exp
Instruction
Clear Instruction
Answer

1. ***that one by itself*** implies an object that is not in close proximity or is isolated. *I am quite sure.*

2. Let me check the image. I found a ***rightmost elephant*** and a ***leftmost man*** in the image away from the group. So it should be the ***leftmost man***. *I am not very sure. What I missed?*

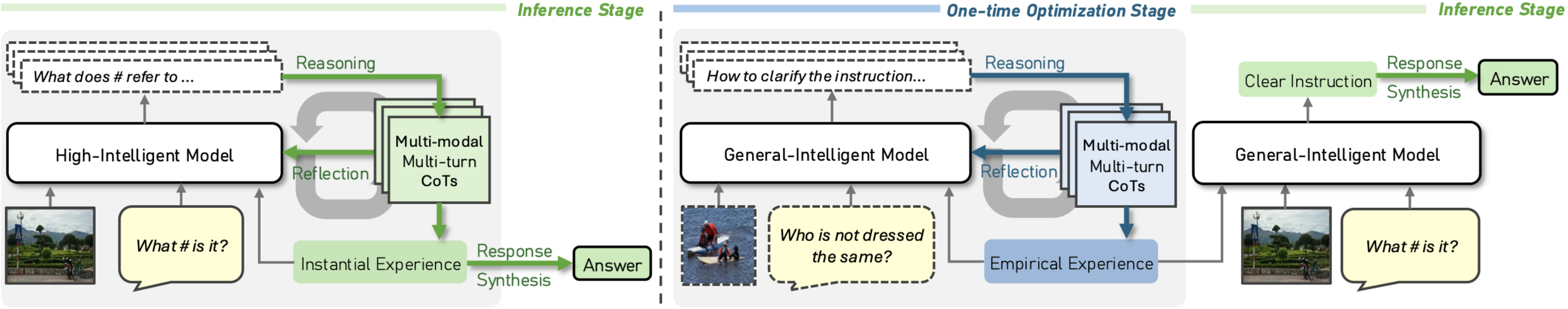
3. ***itself*** is more proper to describe the ***rightmost elephant***. So, it should be the ***rightmost elephant***. *I am sure.*

(a) Ambiguous Instruction **(b) Ambiguity Understanding by Human** **(c) Ambiguity Understanding by Visual-O1**

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

2 Methodology

Overview of Visual-O1



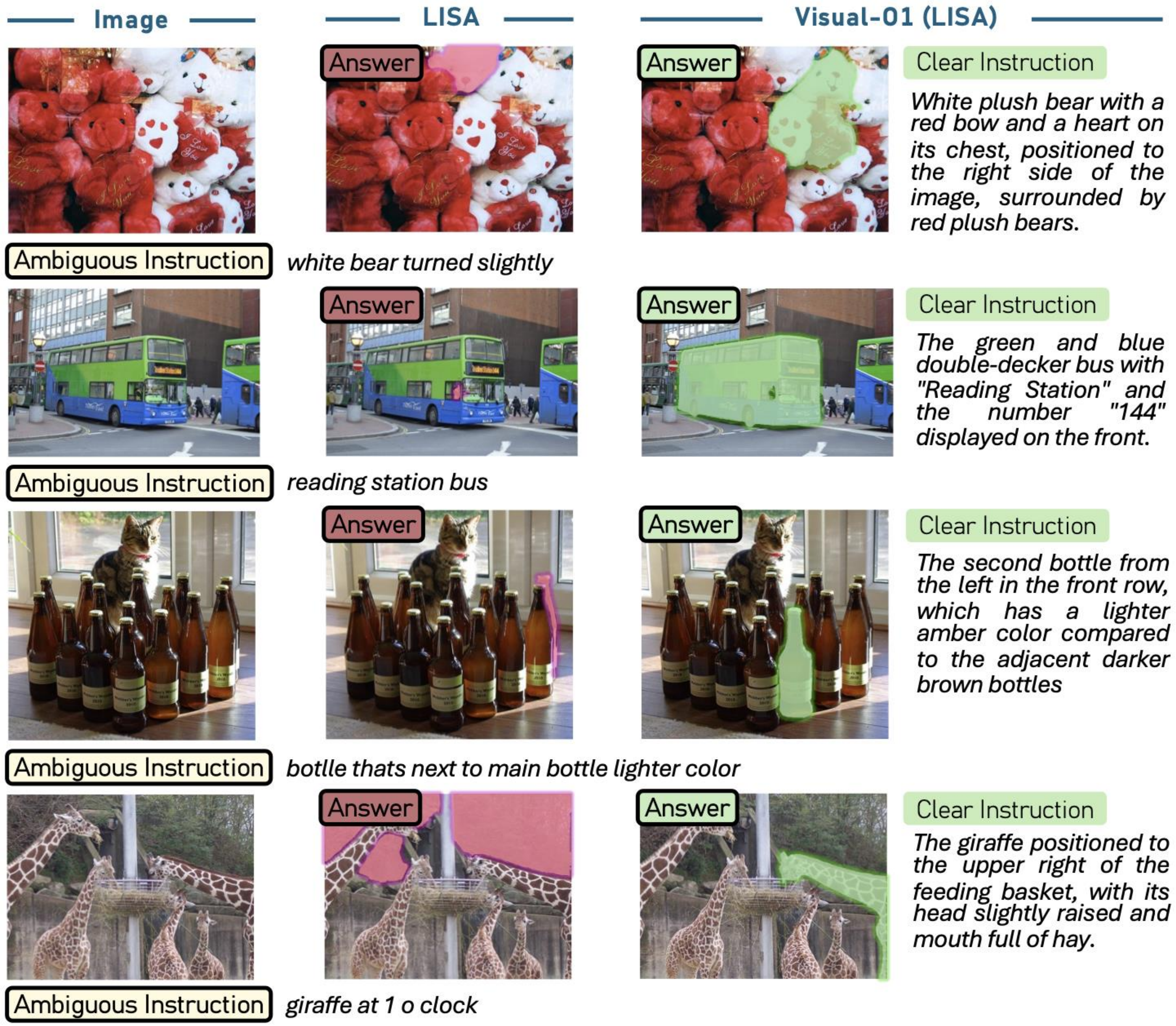
(a) Visual-O1 for High-Intelligent Model with Instantial Experience

(b) Visual-O1 for General-Intelligent Model with Empirical Experience

3

Results on RIS

Visualized Results on RIS



Ambiguous Instruction: *white bear turned slightly*

Clear Instruction: *White plush bear with a red bow and a heart on its chest, positioned to the right side of the image, surrounded by red plush bears.*

Ambiguous Instruction: *reading station bus*

Clear Instruction: *The green and blue double-decker bus with "Reading Station" and the number "144" displayed on the front.*

Ambiguous Instruction: *botlle thats next to main bottle lighter color*

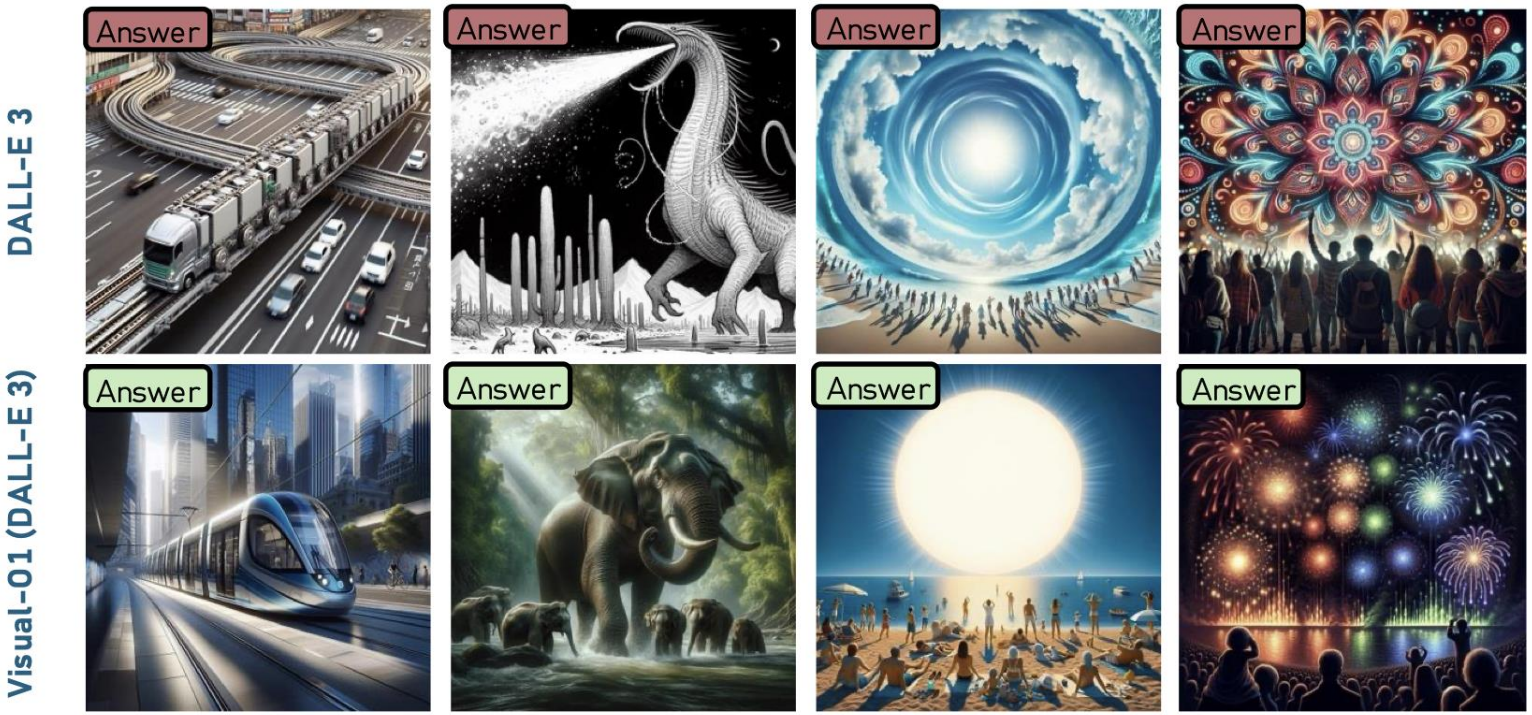
Clear Instruction: *The second bottle from the left in the front row, which has a lighter amber color compared to the adjacent darker brown bottles*

Ambiguous Instruction: *giraffe at 1 o clock*

Clear Instruction: *The giraffe positioned to the upper right of the feeding basket, with its head slightly raised and mouth full of hay.*

4 Results on Visual Synthesis

Visualized Results on Visual Synthesis



Ambiguous Instruction

A long vehicle with multiple connected sections moves along tracks in an urban setting.

Ambiguous Instruction

Large animal with long nose sprays water, small ones nearby.

Ambiguous Instruction

Bright circle high in sky, people on beach look up.

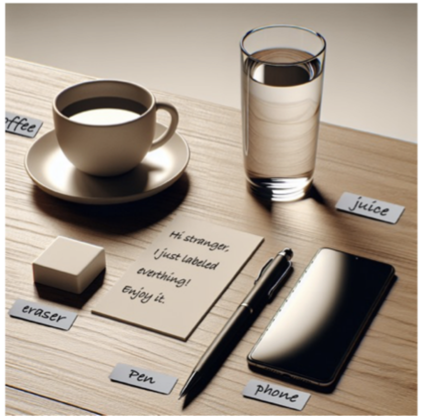
Ambiguous Instruction

Colored lights in sky make patterns, people watch at night.

5 Results on Challenging Case

Challenging Case

Image



Ambiguous Instruction

Lisa poured a glass of water, a porcelain cup of coffee, and prepared a pen, a mobile phone, and an eraser for her friend's visit. Then, Lisa labeled all the items and left an amusing note. But here's something strange for her friend. What does Lisa think it substantially be?

A. water B. coffee C. juice D. eraser E. phone

GPT-4o

Answer

Response: C. juice

GPT-4o + CoT

Answer

Response: C. juice

Visual-O1 (GPT-4o)

Answer

Response: A. water

Conclusion

- We reveal the capabilities of multi-modal models in analyzing and executing ambiguous instructions by setting up a novel benchmark for understanding ambiguous instructions in various multi-modal tasks.
- We propose Visual-O1, a multi-modal multi-turn chain-of-thought reasoning method, to build instantial or empirical experience for high-intelligent or general-intelligent models, enabling them to correctly understand ambiguous instructions.
- Experimental results show that our method improves the performance of models with varying intelligence levels on ambiguous instruction datasets and enhances their performance on general datasets.

Thank you

Code & dataset is available at https://github.com/kodenii/Visual-O1.
Please do not hesitate to contact me via minheng.ni@connect.polyu.hk if you have any questions.