Find–the–Common: Benchmarking and Assessing Inductive Reasoning Ability on Vision-Language Models

Shi Yuting[1], Wei Houjing[1], Jin Tao[1], Zhao Yufeng[1], and Naoya Inoue[1,2]
[1]JAIST [2]RIKEN
{s2210096,houjing,morgan,yfzhao,naoya-i}@jaist.ac.jp

Abstract

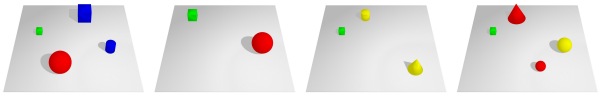
Recent advances in Instruction-fine-tuned Vision and Language Models (IVLMs) have revolutionized the landscape of integrated vision and language understanding. However, Inductive Visual Reasoning—a vital skill for text-image understanding—remains underexplored due to the absence of benchmarks. So, in this paper, we introduce Find–the–Common (FTC): a new vision and language task for Inductive Visual Reasoning. In this task, models are required to identify an answer that explains the common attributes across visual scenes. We create a new dataset for the FTC and assess the performance of several contemporary approaches including implicit reasoning, symbolic reasoning, and implicit-symbolic reasoning with various models. Extensive experiments show that even state-of-the-art models like GPT-4V can only archive with 48% accuracy on the FTC, for which, the FTC is a new challenge for the visual reasoning research community. Our dataset is available online.

1 Introduction

Instruction-tuned **V**ision **L**anguage **M**odels (IVLMs), such as MiniGPT-4 (1), InstructBLIP (2), LLaVA (3), Visual ChatGPT (4), and GPT-4V (5) have been demonstrated with excellent performance on Vision and Language tasks (6; 7; 8) and also show the strong zero-shot generalization ability to unseen tasks, such as writing HTMLs based on a hand-drawing sketch and explaining the implicit meaning of memes (1; 3).

Assessing VLMs involves various tasks testing perceptual skills like object recognition and complex analyses including counting and reasoning. Notable benchmarks like Compositional Visual Reasoning (9) evaluate compositional reasoning, Visual Spatial Reasoning (3) focuses on spatial understanding, and Visual Commonsense Reasoning (10) tests knowledge beyond visuals. Despite their comprehensiveness, these benchmarks do not fully address deductive reasoning, where models derive conclusions from given premises, often employing commonsense knowledge. Another important type of reasoning is inductive reasoning, which aims to generalize a group of finite observations to induce general rules in a bottom-up fashion (11). In the context of vision processing, we can define such competencies as **Visual Inductive Reasoning**, which requires understanding multiple visual scenes and then reasoning out common conclusions from those different scenes. We argue that visual inductive reasoning has been underexplored despite its importance, which raises the following question: *Given a set of visual scenes, can VLMs identify a common rule describing these different scenes?*

To address this issue, we propose a novel benchmark, termed **F**ind–**T**he–**C**ommon (FTC): a new task that required understanding and reasoning across 3D visual



Q: What is the common regularity between four 3D scenes?

a. The yellow object on the far right among all yellow objects is a cylinder.
b. The red sphere is in the forefront.
c. The cube on the far left among all cubes is green. ✓
d. The object farthest away is purple.

Figure 1: Example of the Find–the–Common task. Given four 3D scenes and a 4-choices question, the task is to perform inductive reasoning to identify the correct statement describing the common regularity between the 3D scenes. The choices consist of one correct choice (c) two wrong choices (a, d), and a decoy choice to fool models (b).

scenes and text reading comprehension. An example is shown in Fig. 1. Given four 3D scenes and a multiple-choice question, the task is to perform inductive reasoning to identify the correct statement describing the common regularity between the 3D scenes. Our contributions can be summarized as follows:

- We introduce a novel benchmark **Find–the–Common** for evaluating the inductive visual reasoning capabilities of IVLMs (Sec. 2.1, Sec. 2.2). Our dataset consists of 353 instances, each of which provides four 3D scenes consisting of 2-6 objects and a four-choices question, including a decoy choice that is partially true in scenes. Our dataset is available online.
- We adapt three approaches for evaluation: (i) Implicit Reasoning, (ii) Symbolic Reasoning , and (iii) Implicit-Symbolic Reasoning with the current latest state-of-the-art IVLMs. Our extensive experiments show that InstructBLIP, LLAVA, and GPT-4V reveal that even state-of-the-art IVLMs struggle to solve the FTC task (Sec. 4). Our detailed analysis also finds that GPT-4V exhibit tendencies to "hallucinate", despite their demonstrated proficiency in text-based inductive reasoning.

2 Benchmark: Find–the–Common

2.1 Task Formulation

The benchmark focuses on the task of visual inductive reasoning, which involves extracting principles from specific instances. Considering the challenges brought about by the complexity and nuances in real photographs, we initiate our research from simplified 3D object scenes. This approach allows us to construct inductive reasoning tasks with clarity. By omitting extraneous details, we can concentrate more on evaluating the reasoning performance of IVLMs. More specifically, we formulate the FTC task as follows: Given an input instance $Z = (I, Q, C)$ consisting of an image I, a question Q with respect to a four-choice answer C.

Image. Each image consists of four 3D scenes $I = \{S_1, S_2, S_3, S_4\}$, each scene containing 2-6 objects. Each object has four attributes including Color, Shape, Size and Position, and each attribute corresponds with one value. We defer to Tab. 1 for further details.

Table 1: Attributes of objects and their values.

Attributes	Values
Color	Red, Green, Blue, Yellow, Purple
Shape	Cube, Sphere, Cylinder, Cone
Size	Small, Medium, Large
Position	(1,1), (1,2), ..., (7,7), (8,8)

Question and Answer. An answer is a set of choices $C = \{c, w_1, w_2, d\}$ represents the regularities within $\{S_1, S_2, S_3, S_4\}$. Four choices $\{c, w_1, w_2, d\}$, comprising one *correct* choice c, two *wrong* choices w_1, w_2, and one *decoy* choice d. The decoy choice is designed to be partially true for given scenes, ensuring models examine all scenes to determine the correct choice.

Given a question Q, models aim to identify an answer $c \in C$ that holds true across all the provided scenes S_1, S_2, S_3, S_4. An example is given in Fig. 1.

2.2 Dataset Creation

To generate 3D scenes and multiple-choices answers, we adopt a two-step approach: (1) **Answer Generation**: generate choices based on set pre-defined linguistic rule, and (2) **Scene Generation**: generate scenes that satisfy the correct choice and do not satisfy the wrong choices.

Answer Generation. We construct 13 linguistic templates with attribute placeholders at three levels:

- Four one-attribute template (e.g., "All objects are [Shape].").
- Six two-attributes template (e.g., "All objects are [Color] [Shape].").
- Three three-attributes template (e.g., "The [position] [color] is [shape].").

We then randomly select a template and fill in the placeholders with different values. For [position], we use relative choices like "on the far left", "on the far right", "forefront", and "farthest away" for diversity.

Scene Generation. We generate two types of 3D scenes: (i) one scene satisfying c but not d, w_1, w_2, and (ii) three scenes satisfying c, d but not w_1, w_2. We formulate this as a constraint satisfaction problem using Answer Set Programming (ASP)(12), a logic-based framework for such problems. We create an ASP program where each answer set corresponds to one scene configuration. We define predicates like shape(X, S) i.e., object X has shape S and encodes rules as ASP rules. An example of ASP is shown

in table4 in Appendix. To generate scenes that satisfy (or do not satisfy) these rules, we use integrity constraints, such as:

- To satisfy the rule "All objects must be red":
  `:- not r_all_SH(red).`
- To not satisfy the rule "All red objects must be cube":
  `:- r_all_CL_are_SH(red, cube).`

We then randomly sample one answer set using `clingo`[1] and convert it back to a scene configuration. Scenes are rendered using `pyrender`[2]. The final dataset comprises 353 instances.

2.3 Dataset Quality

To ensure that our visual inductive reasoning problems are consistently solvable by humans, we conduct a human evaluation study. We randomly sample 100 instances and ask two graduate school students to solve them. The inter-annotator agreement between these evaluators indicates Cohen's Kappa (13) of 0.92, indicating almost agreement. The accuracy scores of the two evaluators are 0.98 and 0.88, respectively. An evaluator with an accuracy of 0.88 is frequently fooled with decoy choices, which results in lower accuracy.

3 Approach

We employ three approaches to assess the zero-shot generalization capability of IVLMs on visual inductive reasoning:

Implicit Reasoning. Tests the models' ability to identify common rules among scenes via utilizing visual perception (Fig. 2(a)). The task involves selecting the correct answer from the given four options by reasoning from the distinct four 3D scenes.

Symbolic Reasoning. Evaluates the models' capacity to convert visual data into textual descriptions for logical reasoning (Fig. 2(b)). Initially, four scenes are fed together into IVLMs to generate their scene descriptions. Then these descriptions, along with four options and a question, are processed by a Large Language Model (LLM) to produce the correct choice.

Implicit-Symbolic Reasoning. Can be seen as a combination pipeline of the above two approaches, where the four 3D scenes are provided to models twice (Fig. 2(c)). Specifically, in addition to generated scene descriptions, 3D scenes are also provided to VLMs for predicting the final answer. We showcase how it works on GPT-4V in this work. The goal is to assess the models' proficiency in integrating visual and textual information to deduce the correct option.

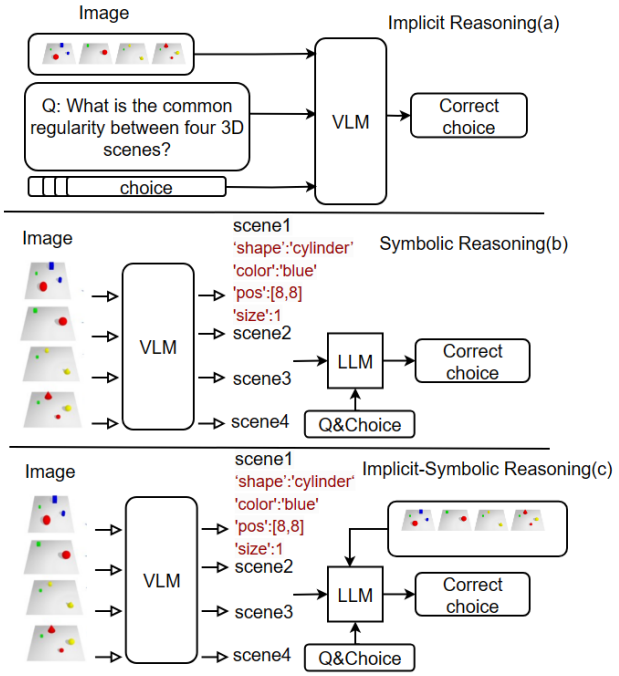


Figure 2: Baseline Reasoning

4 Experiment

In this section, we present our experimental setup and results, focusing on evaluating three state-of-the-art Integrated IVLMs using the baseline approaches outlined in Sec. 3, comparison with human evaluation is also included. Also, we build ground truth experiments to create a controlled environment that isolates the models' logical reasoning faculties.

4.1 Setting

To evaluate models on our Find–the–Common benchmark, we randomly sampled 50 instances and tested them by using three baseline approaches discussed in Section 3, using one or more of three types of Visual Language Models (VLMs): (i) InstructBLIP (2), a smaller IVLM; (ii) LLAVA, which combines a vision encoder with Large Language Models; and (iii) GPT-4V(ision)[3], an extremely large VLM. We

1) https://potassco.org/clingo/
2) https://github.com/mmatl/pyrender
3) https://openai.com/research/gpt-4v-system-card

use Accuracy for evaluation. Further details of prompts and model parameters, can be found in the Appendix. Additionally, we use Chain-of-Thought (CoT) prompting (14) without any few-shot demonstrations (henceforth, Zero-shot CoT) in implicit reasoning.

4.2 Results and Analysis

Table 2: Accuracy of different models and approaches

Approach	Model	Accuracy
Symbolic	GPT-4V	48.0%
	InstructBLIP	-
	LLaVA	12.0%
Implicit	GPT-4V	44.0%
	GPT-4V (CoT)	42.0%
	InstructBLIP	24.0%
	InstructBLIP (CoT)	-
	LLaVA	30.0%
	LLaVA (CoT)	-
	Human	91.5%
Implicit-Symbolic	GPT-4V	46.0%

The performance of models for three baselines is shown in Table 2. InstructBLIP and LLAVA, as smaller IVLMs, face challenges, achieving only 21% and 30% accuracy respectively in the Implicit baseline, indicating limited accuracy compared to the human baseline. GPT-4V achieves 44% accuracy, showing improvement but still lagging behind human performance. These results suggest limited zero-shot generalization capability in current visual instruction fine-tuning methods for visual inductive reasoning.

Table 3: Ground Truth Approach

Model	Accuracy
Implicit-Symbolic	46.0%
Symbolic-Ground truth	74.0%
Implicit-Symbolic-Ground Truth	92.5%

To analyze the reasoning capabilities of VLMs further, we conducted two experiments, replacing VLM-generated scene descriptions with ASP-generated scene parameters. Using GPT-4V as the VLM and GPT-4 as the LLM, we observed:

1) Including image information significantly enhances model performance: In Table 4, the Implicit+Symbolic + Ground Truth approach shows 92.5% accuracy, a 16.5% increase over the 76% in the Symbolic baseline ground truth, indicating images provide essential background information or visual cues.

2) A gap exists in reasoning accuracy using scene descriptions: Automatically generated descriptions by VLMs compared to actual scenes show a noticeable accuracy gap, around 40% regardless of the method used. A manual review of GPT-4V CoT-generated descriptions showed 82% (41/50) cases of object hallucination, suggesting potential misinterpretations in complex visual information processing.

3) Language module ability in VLMs affects reasoning: InstructBLIP and LLAVA struggle with COT understanding, with LLAVA capable of generating, albeit inaccurately, JSON scene descriptions. In contrast, GPT-4V demonstrates stronger COT understanding and JSON file generation. GPT-4's 74% accuracy in the Symbolic-Ground truth (Table 4) indicates its proficiency in processing JSON-formatted textual information for linguistic inductive reasoning tasks.

5 Conclusion

In our **Find–the–Common** benchmark test, the evaluation of IVLMs' visual inductive reasoning capabilities revealed key findings. Even advanced models like GPT-4V, while making progress in visual tasks, still face significant challenges compared to the human baseline, particularly in object detection and scene interpretation. This highlights the need for improvement in the field of visual inductive reasoning.

Our study also underscores the importance of accurate scene information in enhancing model performance and the crucial role of effective interaction between images and text prompts in increasing accuracy. Additionally, smaller models like InstructBLIP and LLAVA show deficiencies in handling complex reasoning tasks, indicating a need for further optimization in model design.

All of these findings inspired the following for our future work. First, given the challenges with multiple visual scenes and hard-to-perceive objects, refining the dataset with a well-designed hierarchy complexity will better evaluate VLMs' adaptability and comprehension. Then, considering the object hallucination tendencies observed, a potential direction for improvement in the future lies in training regimes that emphasize precise visual reasoning over textual inference.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19K20332. Additionally, we thank Surawat Pothong and Tien Dang Huu, members of RebelsNLU lab for their support.

References

[1]Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. **arXiv preprint arXiv:2304.10592**, 2023.

[2]Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[3]Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[4]Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. **arXiv preprint arXiv:2303.04671**, 2023.

[5]Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

[6]Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. **arXiv preprint arXiv:2306.13549**, 2023.

[7]Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. **arXiv preprint arXiv:2306.09265**, 2023.

[8]Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. **arXiv preprint arXiv:2306.13394**, 2023.

[9]Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016.

[10]Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 6720–6731, 2019.

[11]Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond, 2023.

[12]Ilkka Niemelä. Answer set programming: A declarative approach to solving challenging search problems. In **2011 41st IEEE International Symposium on Multiple-Valued Logic**, pp. 139–141, 2011.

[13]M.L. McHugh. Interrater reliability: the kappa statistic. **Biochem Med (Zagreb)**, Vol. 22, No. 3, pp. 276–282, 2012.

[14]Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.

Table 4: ASP program example

Template: All objects are [SHAPE]: Rule: `r_all_SH(S) :- shape(X, S) : obj(X)`
Template: A [COLOR] object exists, and all [COLOR] objects are [SHAPE]: Rule: `r_all_CL_are_SH(C, S) :- r_CL_exists(C), shape(X, S) : color(X, C).`

Table 5: Prompt examples for various baseline approaches

Implicit Baseline Approach What is the common regularity between four 3D scenes? Choose one correct answer from the following choices: (a) The green cube is on the far left. (b) A blue object exists. (c) The purple cylinder is farthest away. (d) The object in the forefront is cone.
Implicit Baseline Approach (COT) What is the common regularity between four 3D scenes? Choose one correct answer from the following choices: (a) The green cube is on the far left. (b) A blue object exists. (c) The purple cylinder is farthest away. (d) The object in the forefront is cone. Let's think step by step.
Symbolic Baseline Approach Step 1: Please analyze the provided image with 4 scenes of objects on a flat surface...

Table 6: Hyperparameter Settings for Various Models

Model	**Hyperparameters**
InstructBLIP	num_beams = 5 max_new_tokens = 500, min_length = 10, top_p = 0.9, repetition_penalty = 1.5, length_penalty = 1.0, temperature = 1
GPT-4V	temperature = 0.7, max_tokens = 100, top_p = 1.0, frequency_penalty = 0.0, presence_penalty = 0.0
LLAVA	do_sample = True, temperature = 0.2, max_new_tokens = 1024, use_cache = True, stopping_criteria = [stopping_criteria]
GPT-4	temperature = 0.7, max_tokens = 1000, top_p = 1.0, frequency_penalty = 0.0, presence_penalty = 0.0