

Find-the-Common: A Benchmark for Explaining Visual Patterns from Images

Yuting Shi¹, Houjing Wei¹, Tao Jin¹, Yufeng Zhao¹, and Naoya Inoue^{1,2}

¹JAIST ²RIKEN
{s2210096,houjing,morgan,yfzhao,naoya-i}@jaist.ac.jp

Abstract

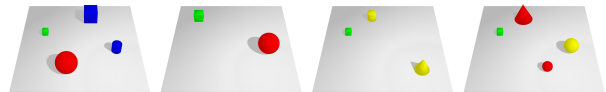
Recent advances in Instruction-fine-tuned Vision and Language Models (IVLMs), such as GPT-4V and InstructBLIP, have prompted some studies have started an in-depth analysis of the reasoning capabilities of IVLMs. However, Inductive Visual Reasoning, a vital skill for text-image understanding, remains underexplored due to the absence of benchmarks. 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 Image-Based Reasoning, Text-Based Reasoning, and Image-Text-Based 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 has been released and is available online: <https://github.com/SSSSSeki/Find-the-common>.

Keywords: Visual Inductive Reasoning, Visual and Language Task Benchmark, IVLMs (instruction-fine-tuned vision and language models)

1. Introduction

Given the recent advances in Large Language Models (LLMs), researchers have attempted to adapt them to multi-modal inputs, leading to many impressive Instruction-tuned Vision Language Models (IVLMs), such as MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), Visual ChatGPT (Wu et al., 2023), and GPT-4V (Yang et al., 2023). IVLMs demonstrate their excellent performance on traditional Vision and Language tasks (Yin et al., 2023; Xu et al., 2023b; Fu et al., 2023) 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 (Zhu et al., 2023; Liu et al., 2023).

To assess the reasoning ability of Vision Language Models (VLMs), multiple types of tasks have been well-established. These tasks are designed to test the basic perceptual abilities of VLMs such as object recognition as well as complex analysis such as object counting, comparison, and commonsense reasoning. For example, Compositional Visual Reasoning (Johnson et al., 2017) diagnoses the competency of VLMs to perform compositional reasoning on images, such as counting, comparing, and logical reasoning. Complementarily, Visual Spatial Reasoning (Liu et al., 2023) provides a focused evaluation for the understanding of spatial relationships in images. Another direction is to test the ability to use external knowledge beyond image contents, which is proposed as Visual Commonsense Reasoning (Zellers et al., 2019). It



What is the common regularity between four 3D scenes?

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

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

requires VLMs to perform a higher-order cognitive and commonsense understanding of the world and to provide a rationale to justify their answers.

However, we argue that all these benchmarks scratch one important type of reasoning skill, namely *deductive reasoning*: given premises, the model is required to derive specific conclusions (sometimes by using commonsense knowledge) in

a *top-down* manner.

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 (Xu et al., 2023a). 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 IVLMs identify a common rule describing these different scenes?*

To address this issue, we propose a novel benchmark, termed **Find-the-Common**. An example is shown in Fig. 1. Given four 3D scenes and multiple choices, 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 reasoning capabilities of IVLMs (§2.1, §2.2). Our dataset consists of 353 instances, each of which provides (i) four 3D scenes consisting of 2-6 objects and (ii) four multiple choices, including a decoy choice that is partially true in scenes (Fig. 1). Our dataset has been released and is available online: <https://github.com/SSSSSeki/Find-the-common>.
- Our extensive experiments show that state-of-the-art IVLMs, including GPT-4V, LLaVA, and InstructBLIP, struggle to solve the FTC task. 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 Definition

The benchmark focuses on the task of inductive reasoning, which involves extracting principles from specific instances. There are many possible design choices for this task. Considering the challenges brought about by the complexity and nuances in real photographs, we choose to initiate our research from simplified 3D object scenes. This allows us to concentrate more on evaluating the reasoning performance of IVLMs.

Formally, we define the FTC as follows:

- **Input:** (i) *3D Scenes*: Four unique scenes S_1, S_2, S_3, S_4 , each containing 2-6 objects. Each object has four attributes and takes one

Attribute	Possible 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)

Table 1: Attributes of objects and their possible values.

possible value. See Table 1 for further details. (ii) *Choices*: Four choices C describing possible regularities in S_1, S_2, S_3, S_4 .

- **Output:** A rule $r \in C$ that holds true for all the given scenes S_1, S_2, S_3, S_4 .

The example is given in Fig. 1. For evaluation, we use accuracy as a performance metric.

2.2. Dataset Creation

To generate 3D scenes and multiple choices of rules, we take a two-step approach: (i) generating choices, and then (ii) generating scenes satisfying the correct choice and not satisfying the wrong choices.

Firstly, we generate four choices of rules $\{c, w_1, w_2, d\}$, including 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 IVLMs to examine all the scenes to arrive at the correct choice.

To generate choices, we first prepare 13 linguistic templates with attribute placeholders at three levels: (i) one-attribute rule (e.g., “All objects are [shape].”) (4/13), (ii) two-attributes rule (e.g., “All objects are [color] [shape].”) (6/13), and three-attributes rule (e.g., “The [position] [color] is [shape].”) (3/13). We then randomly select one template and randomly determine the value of each placeholder. For [position], because it is not straightforward to represent the absolute position as a sentence, we resort to relative choices: “on the far left”, “on the far right”, “forefront”, and “farthest away”. For the diversity of choices, we ensure all the placeholders will be filled in with different values.

Secondly, we generate two types of 3D scenes: (i) one scene satisfying c but not d, w_1, w_2 , and (ii) three scene satisfying c, d but not w_1, w_2 . This problem can be naturally formulated as a conventional constraint satisfaction problem: finding the best configuration (i.e. attributes) of objects that satisfy (or unsatisfy) given rules. We thus use Answer Set Programming (ASP), a logic-based modeling framework for constraint satisfaction problems.

We create an ASP program so that each answer set corresponds to one scene configuration. To represent one object, we define a predicate, such

as $\text{shape}(X, S)$ (i.e., object X has shape S) and then encode rules as ASP rules. Examples of ASP rules include:

- *All objects are [shape]:*
 $\text{r_all_SH}(S) \text{ :- shape}(X, S) \text{ ; obj}(X)$
- *A [color] object exists, and all [color] objects are [shape]:*
 $\text{r_all_CL_are_SH}(C, S) \text{ :- r_CL_exists}(C), \text{shape}(X, S) \text{ ; color}(X, C).$

To generate a scene satisfying (or not satisfying) these rules, we use integrity constraint as follows:

- The rule “*All objects must be red*” must be satisfied:
 $\text{:- not r_all_SH}(\text{red}).$
- The rule “*All red objects must be cube*” must not be satisfied:
 $\text{:- r_all_CL_are_SH}(\text{red}, \text{cube}).$

We then randomly sample one answer set using `clingo`¹ and convert it back to a scene configuration. We used `pyrender`² to render the sampled 3D scenes. We generate 353 instances as a final dataset.

2.3. Dataset Quality

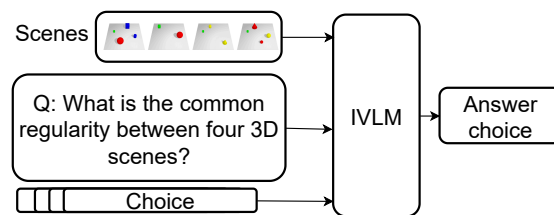
To ensure that our visual inductive reasoning problems are consistently solvable by humans, we conduct a human evaluation study. We randomly sample 100 instances and ask two graduate school students to solve them.

The inter-annotator agreement between these evaluators indicates Cohen’s Kappa of 0.92, indicating “almost perfect 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.

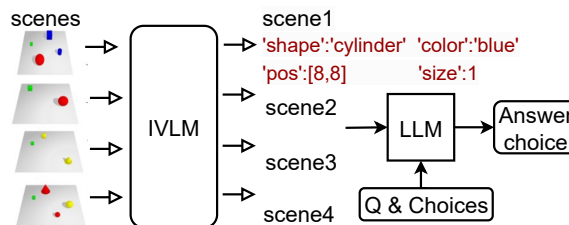
3. Approaches

In this study, we employ three approaches to assess the zero-shot generalization capability of IVLMs on visual inductive reasoning:

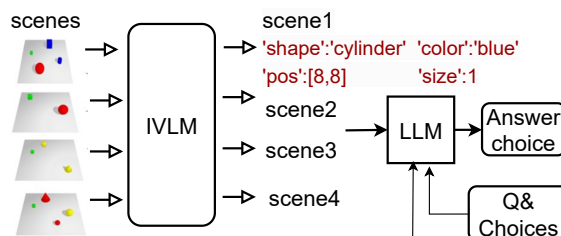
Image-Based Reasoning Tests the models’ ability to identify common rules among scenes via utilizing visual perception (Fig. 2a).



(a) Image-Based Reasoning



(b) Text-Based Reasoning



(c) Image-Text-Based Reasoning

Figure 2: An overview of approaches.

Text-Based Reasoning Evaluates the models’ capacity to convert visual data into textual descriptions for logical reasoning (Fig. 2b). 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 an LLM to produce the correct choice.

Image-Text-Based Reasoning Can be seen as a combination pipeline of the above two approaches, where the four 3D scenes are provided to models twice (Fig. 2c). Specifically, in addition to generated scene descriptions, 3D scenes are also provided to IVLMs for predicting the final answer. The goal is to assess the models’ proficiency in integrating visual and textual information to estimate the correct option.

4. Experiments

4.1. Setup

To evaluate IVLMs on our Find-the-Common benchmark, we randomly sample 50 instances and test them by using three baseline approaches discussed in §3. We use the following IVLMs: (i)

¹<https://potassco.org/clingo/>

²<https://github.com/mmatl/pyrender>

Image-Based Reasoning 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.
-

Text-Based Reasoning Approach

Step 1:

Please analyze the provided image with 4 scenes of objects on a flat surface. Divide each scene's edge into 10 equal parts to create a grid and describe each object's position, color, shape, and relative size (from 1 to 4) using JSON format. The bottom left corner should be considered the origin (0,0), and the z-axis value is always 0. Format the description as a JSON object with keys for 'color', 'pos' (as an array), 'shape', and 'size'.

Step 2:

"There are four scenes describing the same common rule. " + {question}&{choices} + " The scenes information is below:" + {json-formatted scene descriptions}

Image-Text-Based Reasoning Approach

From the given image and Json-formatted scene description, 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.
-

Table 2: Prompt examples for baseline approaches.

InstructBLIP (Dai et al., 2023), a smaller IVLM; (ii) LLaVA, which combines a vision encoder with an LLM (Liu et al., 2023); and (iii) GPT-4V(ision)³, an extremely large IVLM. See Table 2 and Table 7 for further details of prompts and model parameters. Additionally, we use Chain-of-Thought (CoT) prompting without any few-shot demonstrations (Kojima et al., 2022) (henceforth, Zero-shot CoT) in Image-Based Reasoning. We use Accuracy for evaluation. To ensure the reliability of our results, we conduct each experiment five times and calculate the average accuracy and the standard deviation.

4.2. Results

The results are shown in Table 3. Overall, it shows that InstructBLIP, LLaVA and GPT-4V all struggle

³[gpt-4-1106-vision-preview.
https://openai.com/research/
gpt-4v-system-card](https://openai.com/research/gpt-4v-system-card)

Approach	Model	Acc. _{std}
Image-Based	GPT-4V	45.6 _{3.9}
Image-Based	LLaVA	30.8 _{3.4}
Image-Based	InstructBLIP	22.4 _{2.9}
Text-Based	GPT-4V	46.5 _{4.4}
Text-Based	LLaVA	12.6 _{1.7}
Text-Based	InstructBLIP	-
Image-Text-Based	GPT-4V	46.8 _{4.2}
Human	-	91.5

Table 3: Accuracy of the FTC.

Approach	Model	Acc. _{std}
Image-Based	GPT-4V	44.4 _{5.3}
Image-Based	LLaVA	-
Image-Based	InstructBLIP	-

Table 4: Impact of Chain of Thought (CoT).

with our benchmark in comparison to human. InstructBLIP, as a relatively smaller IVLMs, has a near-random chance accuracy, and GPT-4V has some more improvement, yet there is a huge gap to the human baseline. This suggests that the current way of visual instruction fine-tuning does not provide a strong zero-shot generalization ability for visual inductive reasoning.

Table 4 shows that CoT does not enhance the accuracy for all models, despite the effectiveness in the NLP domain (Wei et al., 2022; Kojima et al., 2022). It also shows that InstructBLIP and LLaVA are unable to comprehend CoT instructions, failing to generate reasoning processes.

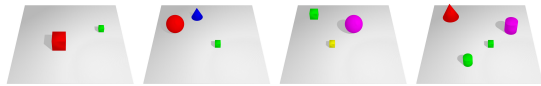
4.3. Analysis

To obtain further insights, we manually examine the reasoning process of GPT-4V CoT for all 50 instances. We observe object hallucination in 82.0% of them (41/50). The object hallucination error can be categorized into three types: (a) recognizing objects that do not exist in a scene, (b) missing objects that should have been recognized, and (c) incorrectly recognizing the attributes of objects, such as position and color.

Table 5 shows examples of these cases, e.g., (a) shows that GPT-4V says "All four scenes have at least one red object", but the third scene does not have a red object. Regardless of hallucination, GPT-4V's inductive reasoning over each scene's description is accurate. This leads to the following question: is the main source of error the lack of the basic perception skills needed for visual inductive reasoning?

To further investigate the reasoning capabilities of IVLMs, we conduct a series of extra experiments

Example (a): Recognizing objects that do not exist in a scene



Let's analyze the four 3D scenes step by step based on the given choices:

(a) The cube farthest away among all cubes is green.
 - In the first scene, the cube farthest away is red.
 - In the other three scenes, the cube that's farthest away is green.

- This means the statement is not consistently true across all scenes.

(b) A red object exists.

- All four scenes have at least one red object. So, this statement is true for all scenes.

(c) The cylinder on the far left among all cylinders is purple.

- In the third scene, the cylinder on the far left is purple.

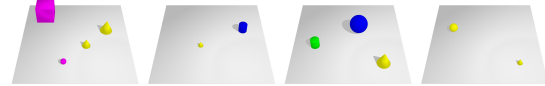
- In the fourth scene, the cylinder on the far left is red.

(...omitted due to space limitation...)

From the above analysis:

The only statement that holds true across all four scenes is (b) "A red object exists."

Example (b): Missing objects that should have been recognized



Alright, let's break this down step by step.

(a) A sphere exists.

In the first image, there's a purple sphere. In the second image, there's no sphere. In the third image, there's a blue sphere. In the fourth image, there's no sphere. This option is not consistent across all four images.

(b) A cone exists, and all cones are yellow.

In the first image, there are two yellow cones. In the second image, there's one yellow cone. In the third image, there's one yellow cone. In the fourth image, there are two yellow cones. This option is consistent across all four images. Every scene has a cone, and all cones are yellow.

(...omitted due to space limitation...)

There are no red cubes in any of the images. Given the breakdown, the only statement that remains consistently true across all four 3D scenes is: (b) A cone exists, and all cones are yellow.

Table 5: Examples of object hallucination made by GPT-4V.

Approach	Acc.
Image-Text-Based	92.5
Text-Based	74.0

Table 6: GPT4-V performance with ground truth scene description.

on GPT-4V wherein scene descriptions generated by IVLMs are substituted with scene descriptions derived from ASP. As shown in Table 6, adding ground-truth scene description to the Image-Text reasoning approach greatly improves the model's accuracy over human baseline, up to 92.5%. This is a 16.5% increase compared to using only Text-Based reasoning approach.

This result highlights the importance of image information for describing scenes. Even when text descriptions are accurate, image information still greatly helps the model's reasoning. However, without ground-truth scene description, the performance of the Image-Text-Based approach (46.8%) is not much better than the Text-Based approach (46.5%). This suggests that if the text descriptions don't match the images or are misleading, the VLMs might not be able to use the additional information from the images effectively.

5. Conclusions

In this paper, we introduce a unique benchmark **Find-the-Common** well-tailored to assess the visual inductive reasoning of IVLMs. Our findings show that IVLMs, even the top-performing GPT-4V, encounter difficulties, and failures of GPT-4V can be largely attributed to object hallucination.

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 inspire 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. And then, considering the object hallucination tendencies observed, a potential direction for improvement in the future may lie in training regimes that emphasize precise visual reasoning over textual inference.

6. Limitations

While our study offers insights into GPT4-V's behavior on visual tasks, some limitations exist: (a)

Dataset Nature: The way our dataset was curated may not fully represent real-world complexities. Future studies should ensure a broader spectrum of image types and scenarios, such as expanding the **Find-the-Common** task with more realistic images collected from real-world scenarios or generated by diffusion models. (b) Inductive pattern complexity: In the future, we should explore more complex patterns to deeply assess the inductive reasoning abilities of IVLMs. (c) Bias and Fairness of dataset: Our study relied on two master’s students to test data quality, using their average scores as a human baseline. This approach may not sufficiently represent human performance on our tasks and may not robustly evaluate the dataset’s quality.

7. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 19K20332. The authors would like to thank the anonymous reviewers for their insightful feedback.

8. Bibliographical References

- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#).
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023a. [Are large language models really good logical reasoners? a comprehensive evaluation and beyond](#).
- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. Lvlm-hub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. [The dawn of Imms: Preliminary explorations with gpt-4v\(ision\)](#).
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

9. Appendices

9.1. Hyperparameters

The hyperparameters used in our experiments are shown in Table 7.

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

Table 7: Hyperparameter settings.