

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

Presenter: Yuting Shi

Our work has been released online: [SSSSeki/Find-the-common \(github.com\)](https://github.com/SSSSeki/Find-the-common)

Introduction

- Background
- Problems
- Objectives
- Contribution

Background

- The recent advances in Vision and Language Models (VLMs)
Llava[10]/ MiniGPT4[11]/ InstructBLIP[12]/ GPT-4V[13]
- Visual reasoning
 - (i) Deeper and more complex inference based on visual information
 - (ii) Subtasks
Visual Question Answering[2]/ Visual Commonsense Reasoning[7]/ Compositional
Visual Reasoning[3]/ Visual Abductive Reasoning[1]
- Analyzing the reasoning capabilities of VLMs
 - Excellent performance in standard visual tasks
 - Performs strongly in zero-shot tasks

Problems

- Inductive Reasoning

generalize a group of finite observations to induce general rules in a bottom-up fashion

- Limited exploration of generation tasks for learning inductive reasoning rules from images and presenting them into human language

Objectives

- Evaluate the performance of VLMs in visual inductive reasoning
- Project Approach:
 - (i) Create a task/dataset for visual inductive reasoning
 - (ii) Explore three approaches to evaluate existing VLMs on the created task

Contribution

- **Find-the-Common** benchmark creation
- Evaluating inductive reasoning ability of state-of-the-art VLMs via several approaches and showing that IVLMs are struggle to solve the FTC task with strong visual hallucination

Benchmark Creation

- Find-The-Common Task
- Data Generation
- Dataset Quality and Statistic

Find-The-Common Task

- Input
 - (i) 3D Scenes
Four unique scenes S1, S2, S3, S4
 - (ii) Choices
Four choices C describing possible regularities in S1, S2, S3, S4.
- Output
A rule $r \in C$ that holds true for all the given scenes S1, S2, S3, S4.



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

- The green cone is on the far right. ✓**
- All objects are blue.
- The purple sphere is farthest away.
- The cube on the far left among all cubes is red.

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)

Data Generation

- Generating rule choices
- Employing Answer Set Programming(ASP) [15] for rule encoding
- Designing decoy choices.



- Q: What is the common regularity between four 3D scenes?*
- The green cone is on the far right. ✓**
 - All objects are blue.
 - The purple sphere is farthest away.
 - The cube on the far left among all cubes is red.

3 of 4 images

`r_CL_POS_SH(green,
cone, far_right)`

Not

`r_SH_POS_CL(cube, far
_left_red)`

Not `r_all_CL(blue)`

Not

`r_CL_POS_SH(purple,
sphere, blue)`

1 of 4 images

`r_CL_POS_SH(green,
cone, far_right)`

`r_SH_POS_CL(cube, far
_left_red)`

Not `r_all_CL(blue)`

Not

`r_CL_POS_SH(purple,
sphere, blue)`

Example of decoy choices setting

Dataset Quality and Statistic

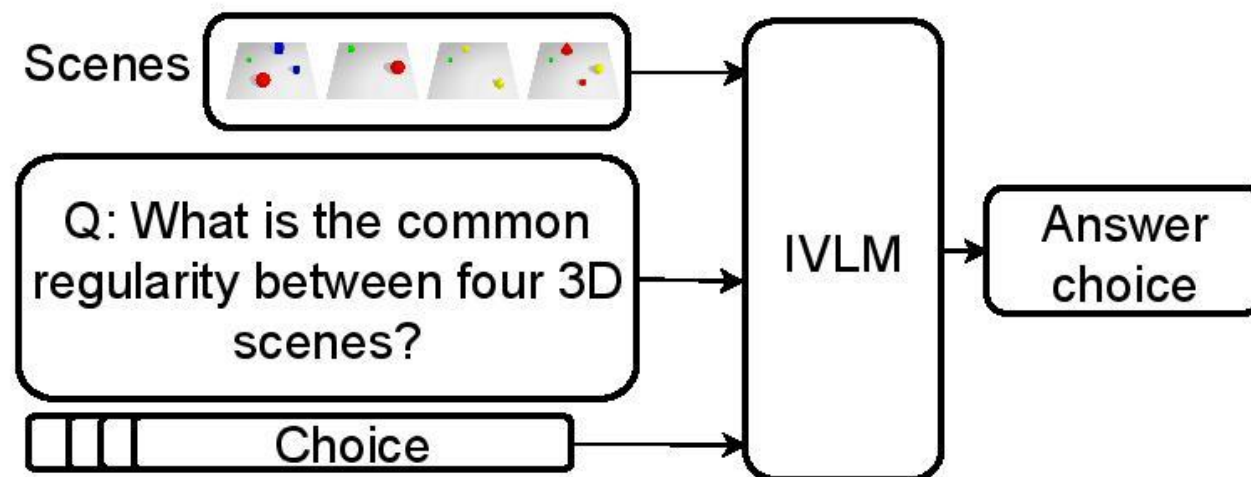
- Conduct Human baseline
 - Randomly sampled 100 instances 2 graduate students
- Results
 - Cohen's Kappa value of 0.92
- Observation
 - The evaluator misled by decoy options
- Statistics:
 - 356 instances
 - 4 scenes
 - 2-6 objects
 - 4 attributes

Approaches

Three approaches examine different face of visual inductive reasoning

Image-Based

- Test the reasoning ability based on image information



Text-Based

- Test the reasoning ability based on text

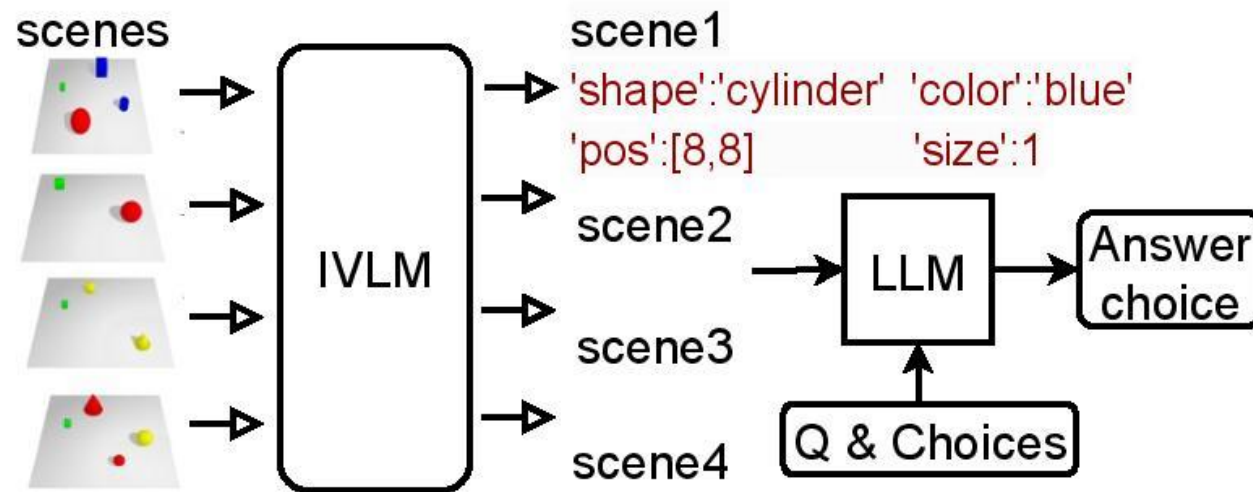
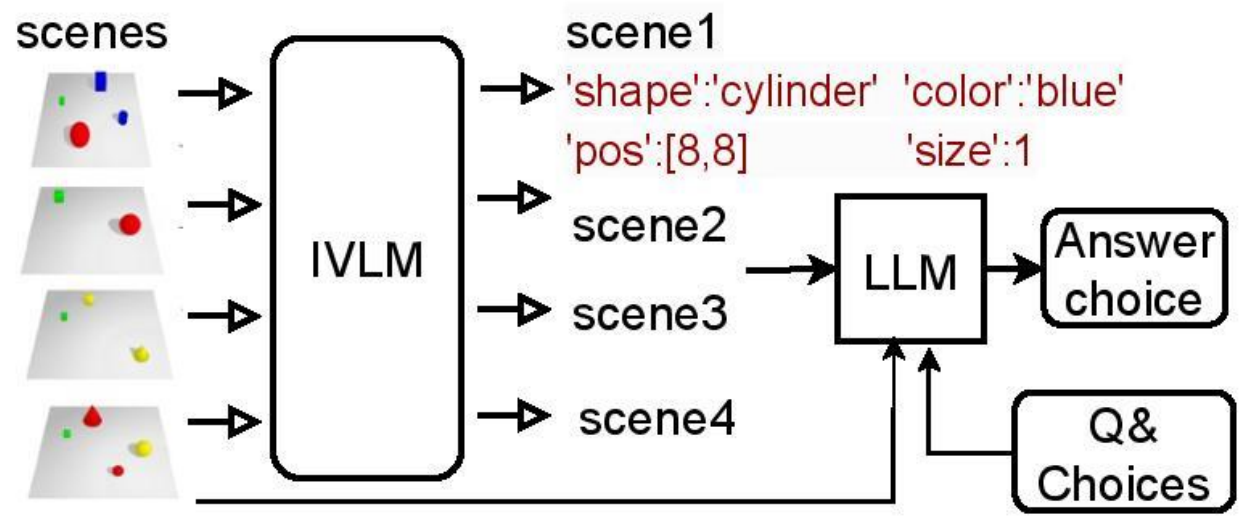


Image-Text Based

- Test the reasoning ability based on integrated visual and textual information.



Experiment

- Experiment Setting
- Result
- Analysis – Manual Analysis
- Analysis - Ground Truth Extra Experiment

Experiment Setting

- Models
 - (i) InstructBLIP [3], a smaller IVLM
 - (ii) LLAVA which combines a vision encoder with Large Language Models [11]
 - (iii) GPT4V(vision), an extremely large VLM.
- Accuracy for evaluation
- Chain-of-Thought[14] outputs reasoning steps



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

- The green cone is on the far right. ✓
- All objects are blue.
- The purple sphere is farthest away.
- The cube on the far left among all cubes is red.

Prompt of Image-Based Approach

Step1:

(Image)

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'.

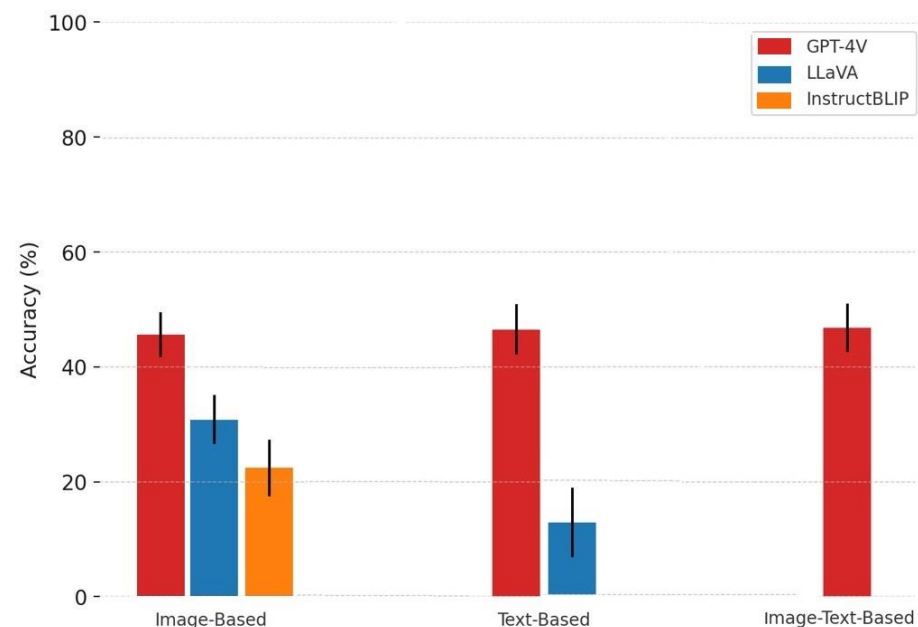
Step2:

"There are four scenes describing the same common rule." + question & options + " The scenes information is below:\n" + Json(from Step1)

Prompt of Text-Based Approach

Result

- All models perform far from human
- GPT-4V shows greater capabilities on all approaches
- The current way of visual instruction fine-tuning does not provide a strong zero-shot generalization ability
- CoT does not enhance the accuracy



Accuracy of the FTC

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

Impact of Chain of Thought (CoT)

Analysis – Manual Analysis

- Visual hallucination
- Regardless of hallucination, GPT-4V's inductive reasoning over each scene's description is accurate

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.**

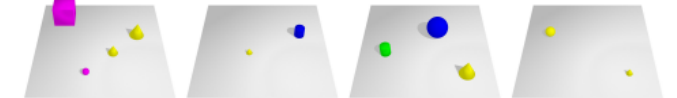
- **In the fourth scene, the cylinder on the far left is red.**

(...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.

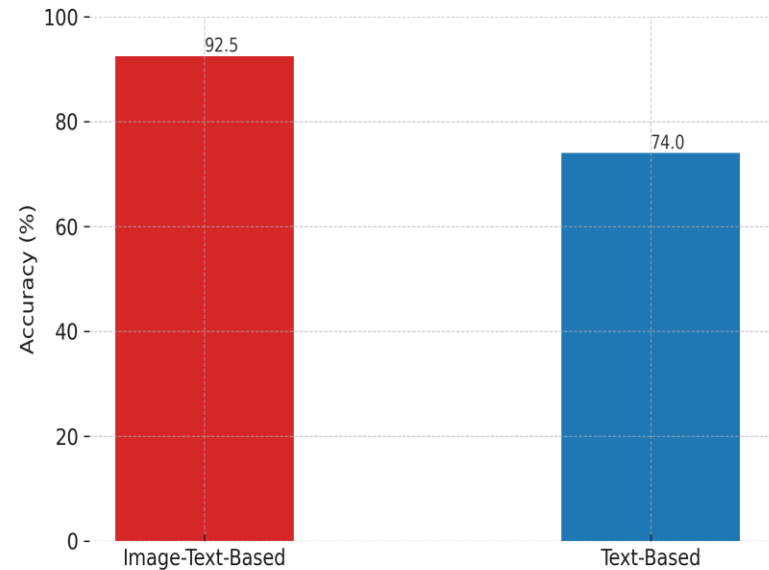
(...omitted due to space limitation...)

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.

Examples of object hallucination made by GPT-4V

Analysis - Ground Truth Experiment

- Ground truth experiment design
Scene descriptions generated by IVLMs are substituted with scene descriptions derived from ASP
- Analysis
Adding ground-truth scene description to the Image-Text reasoning approach greatly improves the model's accuracy



GPT4-V performance with ground truth scene description

Conclusion & Future Work

Conclusion

- Even advanced models like GPT-4V that have made progress in visual tasks still face significant challenges in object detection and scene interpretation compared to human benchmarks.
- Smaller models like InstructBLIP and LLaVA show deficiencies in handling complex reasoning tasks

Future Work

- Dataset Improvement

Refining the dataset with a well-designed hierarchy complexity

- Better Training Regimes

Training regimes that emphasize precise visual reasoning over textual inference

Reference

1. Liang, C., et al. (2022). Visual abductive reasoning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
2. Johnson, J., et al. (2017). CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
3. Małkiński, M., & Mańdziuk, J. (2022). Deep Learning Methods for Abstract Visual Reasoning: A Survey on Raven's Progressive Matrices. arXiv preprint arXiv:2201.12382.
4. Zerroug, A., et al. (2022). A benchmark for compositional visual reasoning. arXiv preprint arXiv:2206.05379.
5. Kumar, A. (2022). The Illustrated Image Captioning using transformers. Retrieved from <https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/>
6. Papineni, K., et al. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. ACL.
7. Wang, T., Huang, J., Zhang, H., & Sun, Q. (2020). Visual Commonsense R-CNN. arXiv preprint arXiv:2002.12204. Retrieved from <https://arxiv.org/abs/2002.12204>
8. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N. (2023). Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. arXiv preprint arXiv:2303.04671. Retrieved from <https://arxiv.org/abs/2303.04671>

Reference

9. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., & Wang, L. (2023). MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action. arXiv preprint arXiv:2303.11381. Retrieved from <https://arxiv.org/abs/2303.11381>
10. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual Instruction Tuning. arXiv preprint arXiv:2304.08485. Retrieved from <https://arxiv.org/abs/2304.08485>
11. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv preprint arXiv:2304.10592. Retrieved from <https://arxiv.org/abs/2304.10592>
12. Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv preprint arXiv:2305.06500. Retrieved from <https://arxiv.org/abs/2305.06500>
13. OpenAI. (n.d.). GPTV System Card. Retrieved from https://cdn.openai.com/papers/GPTV_System_Card.pdf
14. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv preprint arXiv:2201.11903. Retrieved from <https://arxiv.org/abs/2201.11903>
15. Gelfond, M., & Lifschitz, V. (1988). The Stable Model Semantics for Logic Programming. In R. Kowalski & K. Bowen (Eds.), Proceedings of the Fifth International Conference and Symposium on Logic Programming (pp. 1070-1080). MIT Press.

Thank you for listening