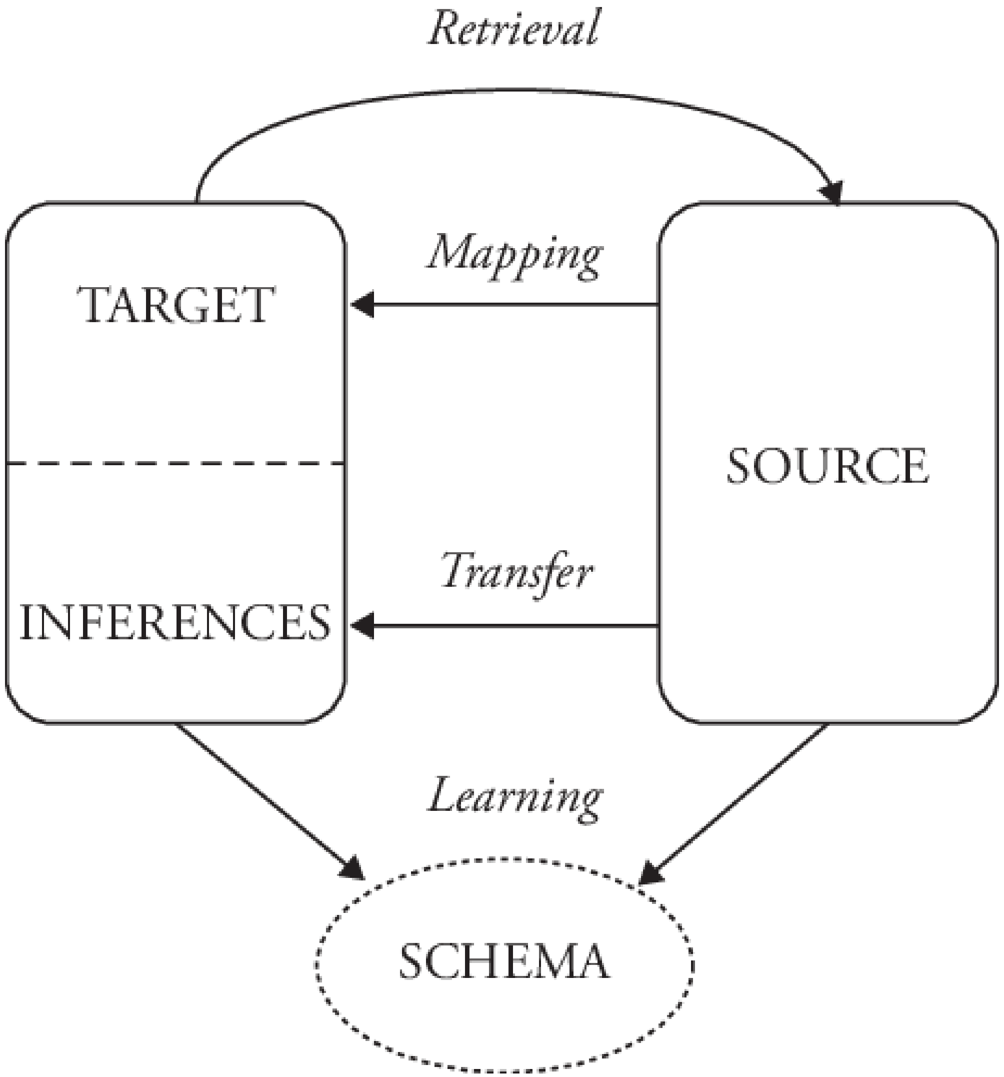
VOILA: Evaluation of MLLMs For Perceptual Understanding and Analogical Reasoning

Nilay Yilmaz, Maitreya Patel, Yiran Lawrence Luo, Tejas Gokhale, Chitta Baral, Suren Jayasuriya, Yezhou Yang

ASU Arizona State University

UMBC

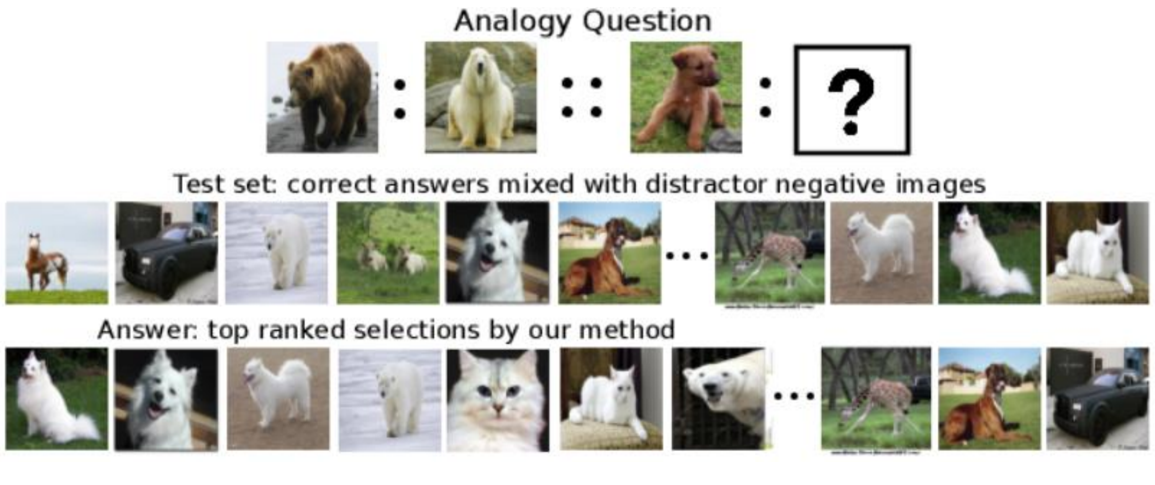
Visual Analogical Reasoning



Holyoak, Keith. (2012). Analogy and Relational Reasoning. The Oxford Handbook of Thinking and Reasoning.

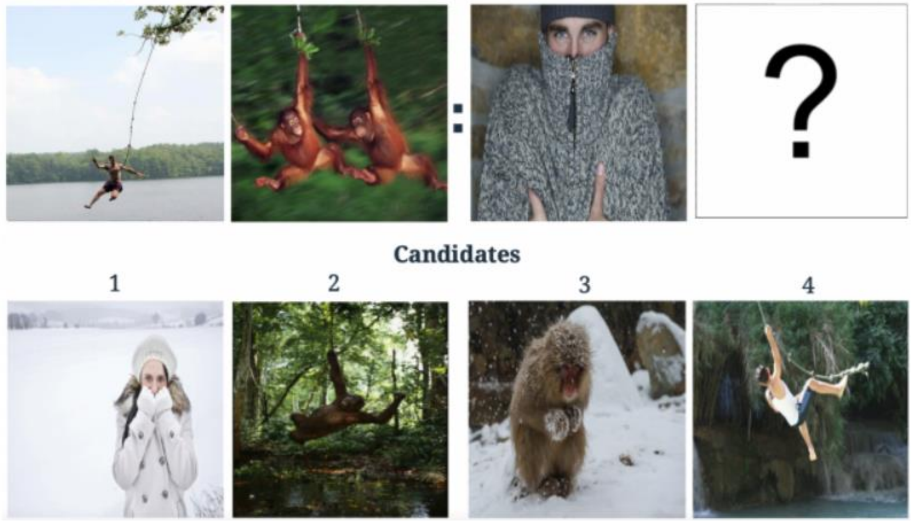
Related Work

VISALOGY



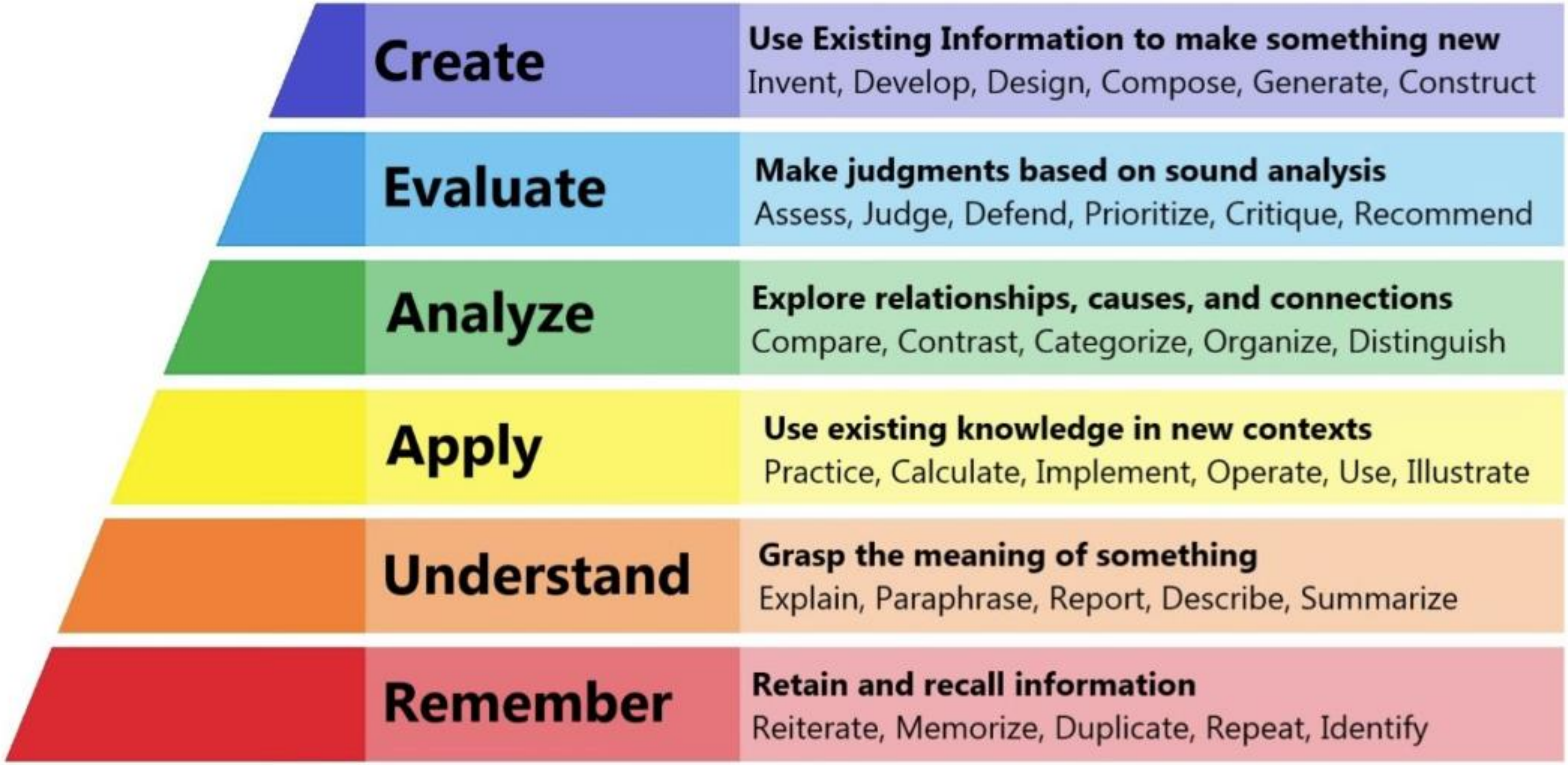
Analogy Question
: :: :
?
Test set: correct answers mixed with distractor negative images
Answer: top ranked selections by our method

VASR



Sadeghi, Fereshteh, et al. (2015). VISALOGY: Answering Visual Analogy Questions. NIPS 2015
Bitton, Yonatan, et al. (2022). VASR: Visual Analogies of Situation Recognition. AAAI 2023

Bloom’s Learning Objective Taxonomy



Create

Use Existing Information to make something new
Invent, Develop, Design, Compose, Generate, Construct

Evaluate

Make judgments based on sound analysis
Assess, Judge, Defend, Prioritize, Critique, Recommend

Analyze

Explore relationships, causes, and connections
Compare, Contrast, Categorize, Organize, Distinguish

Apply

Use existing knowledge in new contexts
Practice, Calculate, Implement, Operate, Use, Illustrate

Understand

Grasp the meaning of something
Explain, Paraphrase, Report, Describe, Summarize

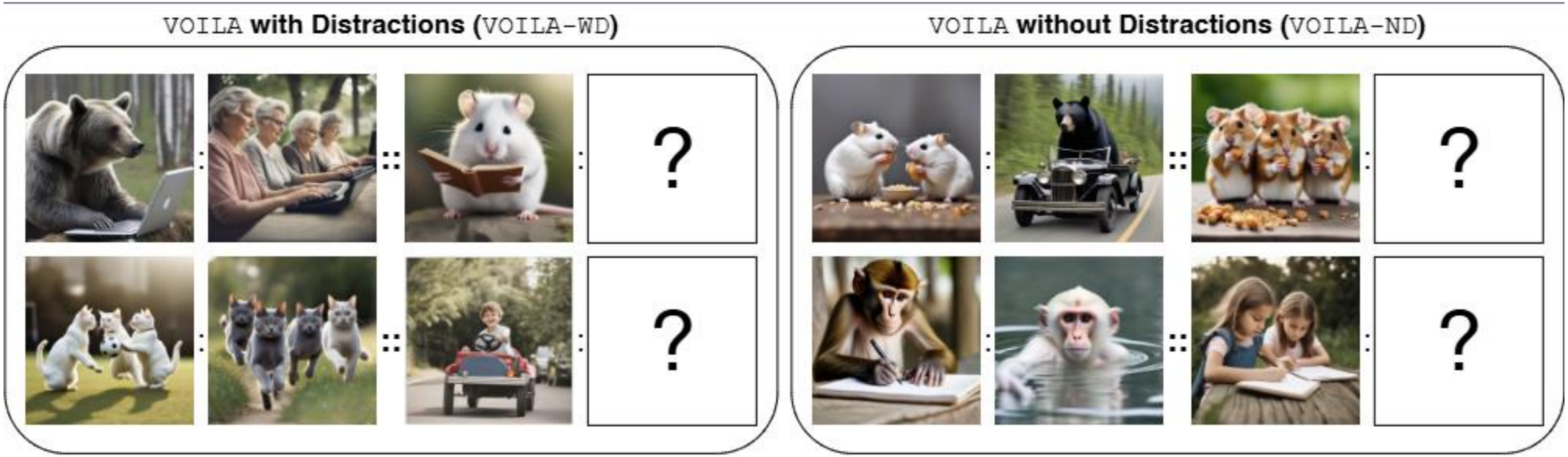
Remember

Retain and recall information
Reiterate, Memorize, Duplicate, Repeat, Identify

B. S. Bloom et al. Taxonomy of educational objectives. The classification of educational goals.
Handbook 1: Cognitive domain. Longmans Green, New York, 1956.

VOILA: a large-scale, open-ended, dynamic benchmark, designed to evaluate MLLMs’ perceptual understanding and abstract relational reasoning with application of analogical mapping strategy to the visual domain.

VOILA Benchmark


VOILA with Distractions (VOILA-WD)
VOILA without Distractions (VOILA-ND)
?
?
?
?

Dataset Creation Pipeline

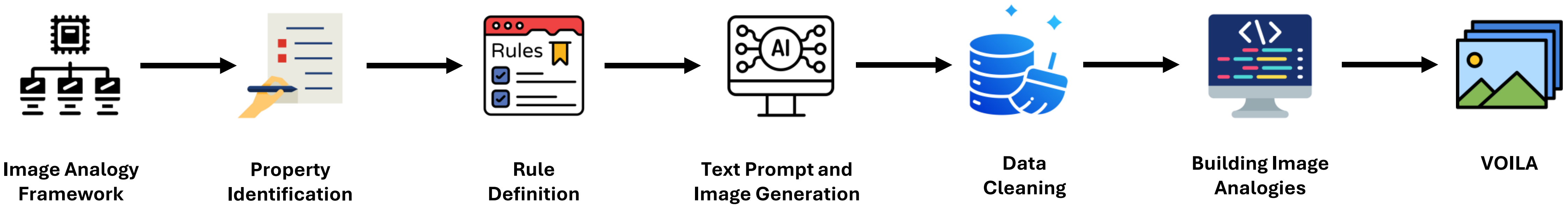
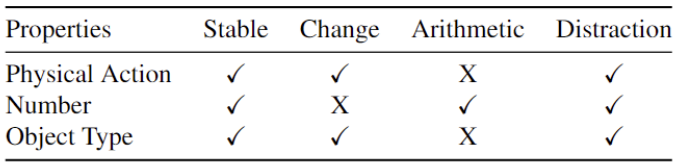
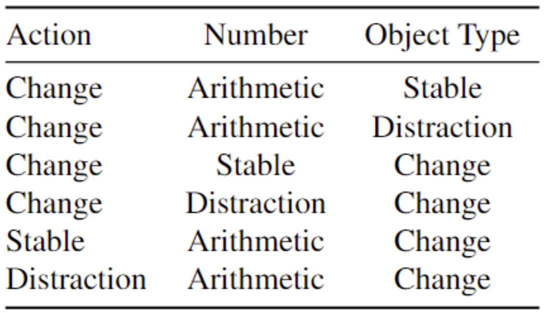


Image Analogy Framework → Property Identification → Rule Definition → Text Prompt and Image Generation → Data Cleaning → Building Image Analogies → VOILA



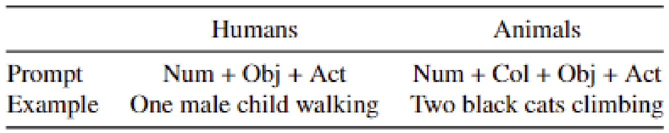
Properties	Stable	Change	Arithmetic	Distraction
Physical Action	✓	✓	X	✓
Number	✓	X	✓	✓
Object Type	✓	✓	X	✓

Table 2. Rules applying to the properties



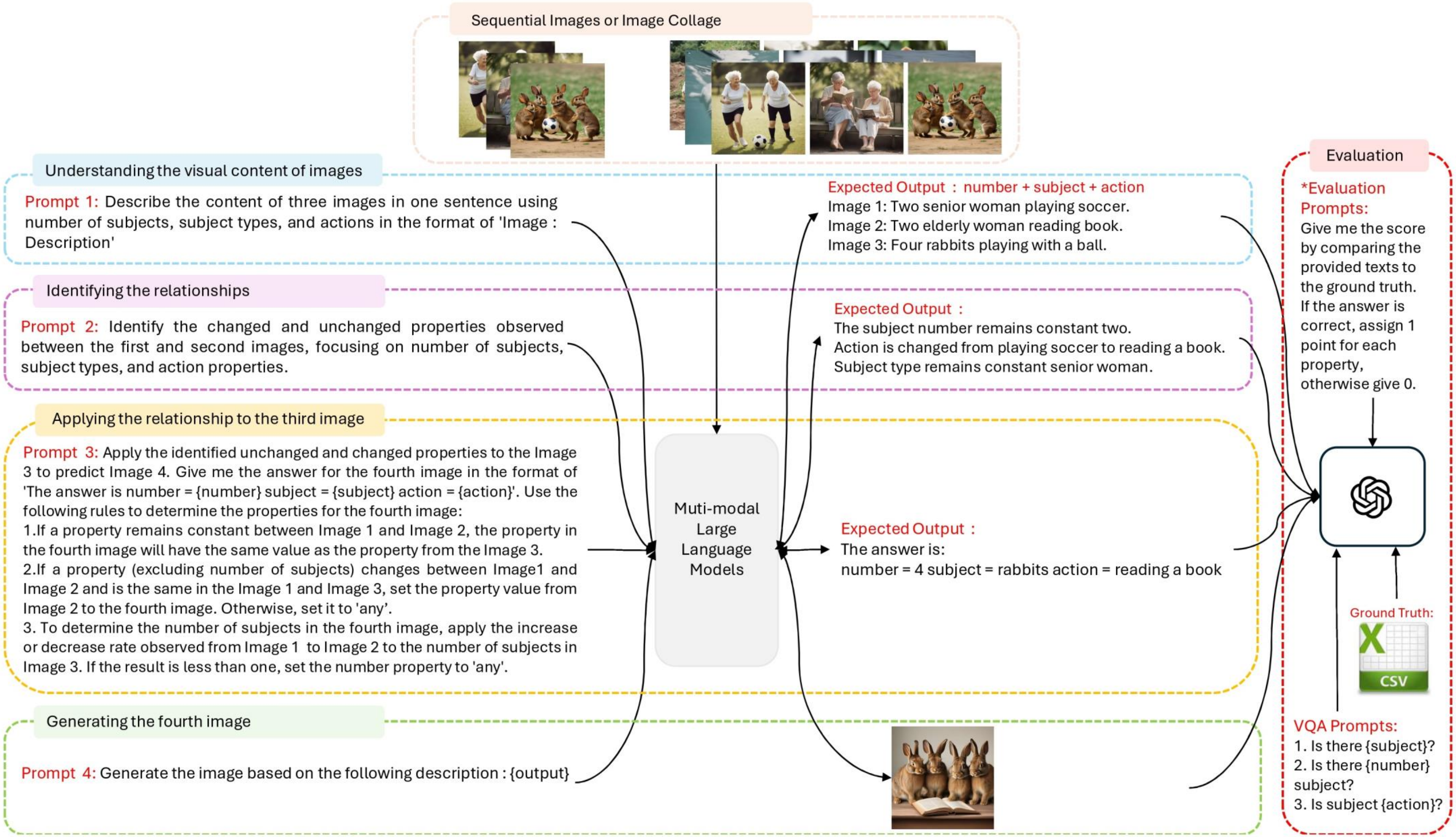
Action	Number	Object Type
Change	Arithmetic	Stable
Change	Arithmetic	Distraction
Change	Stable	Change
Change	Distraction	Change
Stable	Arithmetic	Change
Distraction	Arithmetic	Change

Table 4. In total 6 cases are required to generate visual analogy questions that change the 2 property values at the same time.



	Humans	Animals
Prompt	Num + Obj + Act	Num + Col + Obj + Act
Example	One male child walking	Two black cats climbing

Multi-step Reasoning and Evaluation Pipeline



Sequential Images or Image Collage

Understanding the visual content of images

Prompt 1: Describe the content of three images in one sentence using number of subjects, subject types, and actions in the format of 'Image : Description'

Expected Output : number + subject + action

Image 1: Two senior woman playing soccer.

Image 2: Two elderly woman reading book.

Image 3: Four rabbits playing with a ball.

Identifying the relationships

Prompt 2: Identify the changed and unchanged properties observed between the first and second images, focusing on number of subjects, subject types, and action properties.

Expected Output :

The subject number remains constant two.

Action is changed from playing soccer to reading a book.

Subject type remains constant senior woman.

Applying the relationship to the third image

Prompt 3: Apply the identified unchanged and changed properties to the Image 3 to predict Image 4. Give me the answer for the fourth image in the format of 'The answer is number = {number} subject = {subject} action = {action}'. Use the following rules to determine the properties for the fourth image:

1.If a property remains constant between Image 1 and Image 2, the property in the fourth image will have the same value as the property from the Image 3.

2.If a property (excluding number of subjects) changes between Image1 and Image 2 and is the same in the Image 1 and Image 3, set the property value from Image 2 to the fourth image. Otherwise, set it to 'any'.

3. To determine the number of subjects in the fourth image, apply the increase or decrease rate observed from Image 1 to Image 2 to the number of subjects in Image 3. If the result is less than one, set the number property to 'any'.

Muti-modal Large Language Models

Expected Output :

The answer is:

number = 4 subject = rabbits action = reading a book

Generating the fourth image

Prompt 4: Generate the image based on the following description : {output}

Evaluation

*Evaluation Prompts:

Give me the score by comparing the provided texts to the ground truth. If the answer is correct, assign 1 point for each property, otherwise give 0.

Ground Truth:

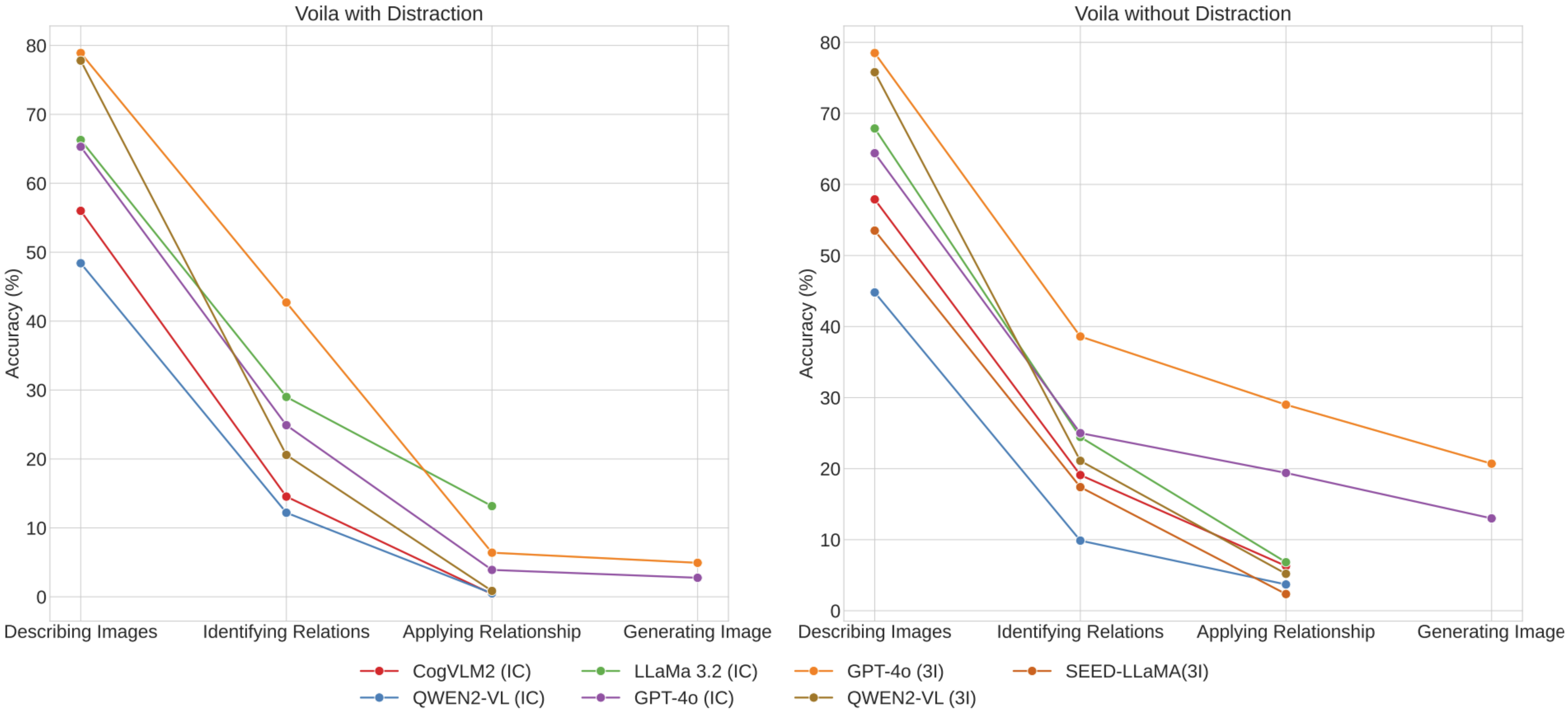
VQA Prompts:

1. Is there {subject}?

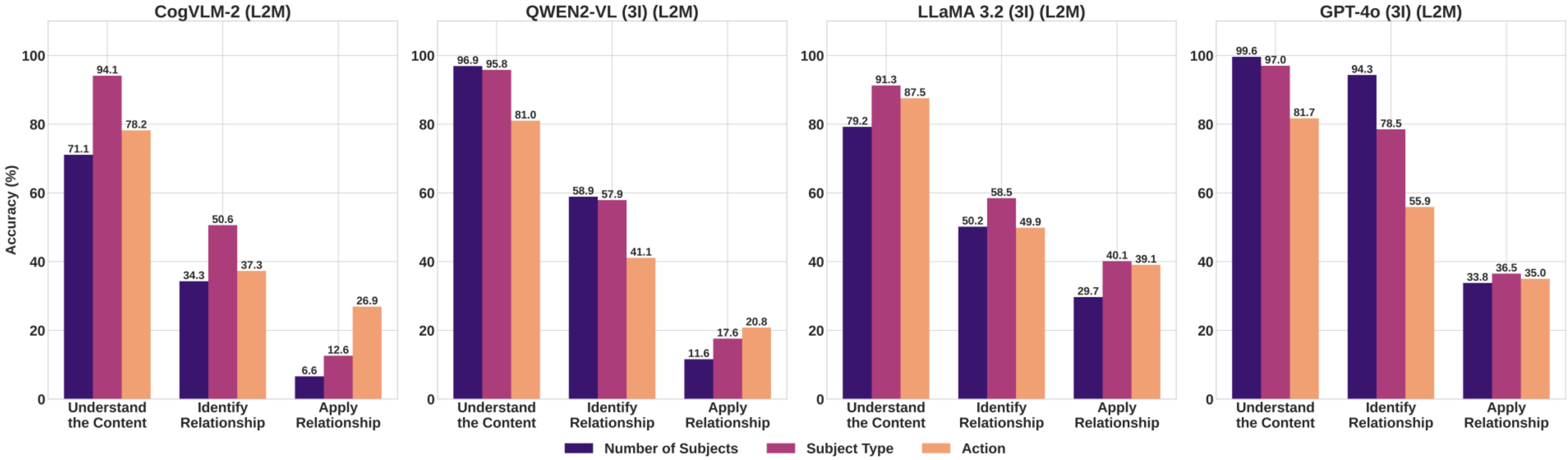
2. Is there {number} subject?

3. Is subject {action}?

How Good are Current Multimodal Large Language Models?



Property Success of Models at Each Step

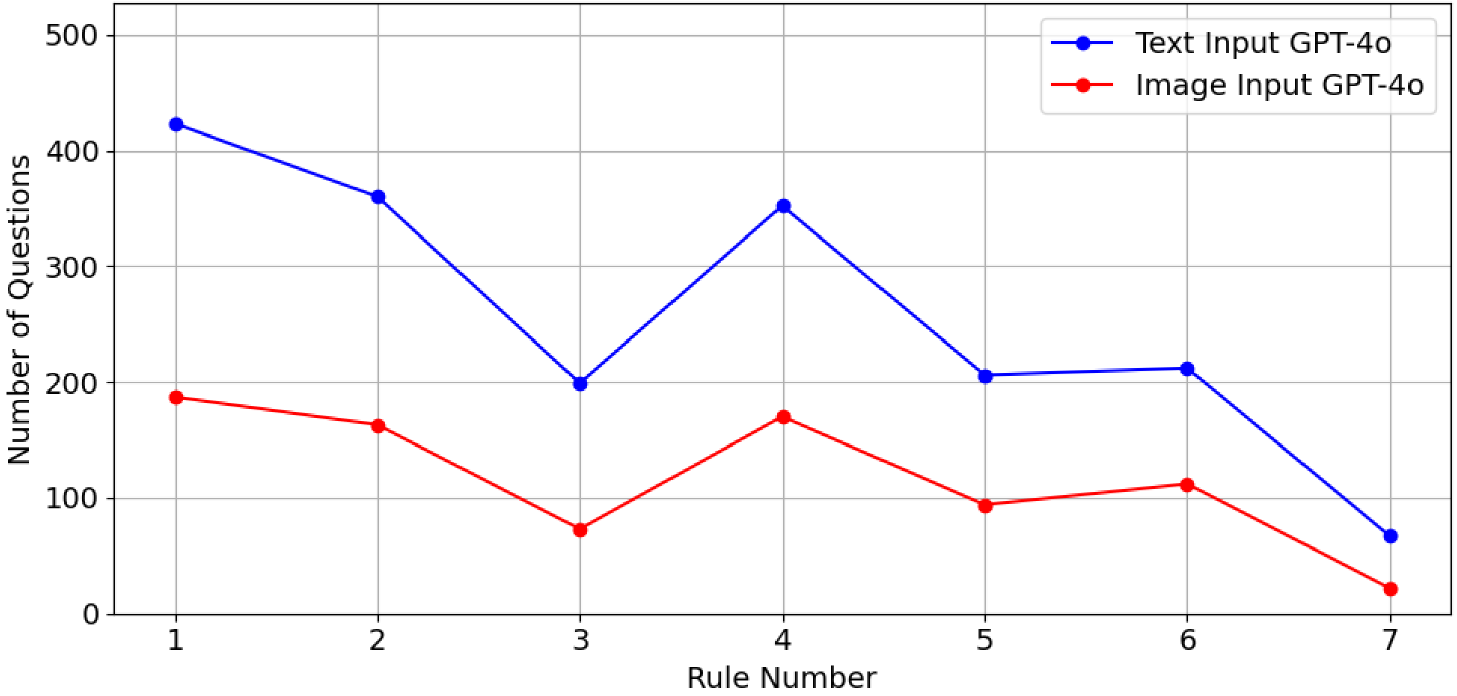


Model Performance With Access To Ground Truth Information

- **Phase 2:** GPT-4o with ground truth image descriptions → **97%**
- **Phase 3:** GPT-4o with ground truth relationships → **17%**
- **Human performance (Phase 3): 71%**

How Does Visual Information Affect Performance?

- Three sequential image input : 22%
- Text description of images: 49%



Summary

- New benchmark for visual analogical reasoning
- New dataset creation pipeline using text-to-image models
- Comprehensive evaluation of MLLMs and analysis of factors influencing performance.

Code and data: https://github.com/nlylmz/Voila