



valuable insights into basic compositional discrimination, they share three fundamental limitations: (1) they only test whether models can distinguish correct from incorrect captions for a single image, without requiring multi-step reasoning; (2) they do not evaluate counting, cross-image comparison, or object interaction; and (3) they use binary selection rather than multi-correct formats that probe the robustness of compositional judgment.

To address these gaps, we designed MMComposition’s taxonomy around three principles:

**Perception–Reasoning separation.** For each core semantic dimension (object, attribute, relation, counting), we distinguish perception tasks (recognizable “within a blink”) from reasoning tasks (requiring multi-step inference). This separation reveals that a model may perceive individual compositional facts correctly yet fail when sequential reasoning is needed—a distinction invisible to existing benchmarks.

**Counting as a distinct compositional skill.** Counting requires aggregating and individuating objects across a scene, often in the presence of occlusion, similar distractors, and irregular arrangements. No prior compositionality benchmark isolates this skill, yet it represents a major failure mode for current VLMs (see Table 2: GPT-5 achieves only 75.86% on Count-P vs. human 93.06%).

**Five novel categories.** We introduce Object Interaction (functional multi-object relations), Difference Spotting (cross-image compositional comparison), Visual Similarity (multi-image similarity judgment), Text Rendering (reading text within visual contexts), and Compositional Probing (multi-correct selection). These categories test compositional capabilities that are qualitatively different from the basic discrimination tasks in existing benchmarks. **Evaluation coverage comparison.** Table 9 compares MMComposition’s

Table 9: Evaluation coverage comparison across VL compositionality benchmarks. ✓ = directly evaluated as a dedicated category; △ = partially overlapping (the dataset touches this dimension but not as a targeted evaluation); ✗ = not covered.

Dataset	Task Format	Perception							Reasoning				Probing Prob	
		Obj-P	Attr-P	Count-P	Rel-P	Diff-S	TR	Vis-Sim	Attr-R	Obj-R	Count-R	Rel-R		Obj-Int
ARO	Binary I2T Retr.	△	△	✗	△	✗	✗	✗	✗	✗	✗	✗	✗	✗
SugarCrepes	Binary I2T Retr.	△	△	✗	△	✗	✗	✗	✗	✗	✗	✗	✗	✗
CREPE	I2T Retrieval	△	△	✗	△	✗	✗	✗	✗	✗	✗	✗	✗	✗
VL-Checklist	Binary I2T Retr.	△	△	✗	△	✗	✗	✗	✗	✗	✗	✗	✗	✗
Cola	T2I Retrieval	△	△	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Winoground	I2T + T2I Match	△	△	✗	△	✗	✗	✗	✗	△	✗	△	✗	✗
FineMatch	Mismatch Detec.	△	△	✗	△	✗	✗	✗	✗	✗	✗	✗	✗	✗
GQA	Compositional QA	△	△	△	△	✗	✗	✗	△	△	△	△	✗	✗
<b>MMComposition</b>	<b>Comp. QA</b>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

evaluation coverage with existing VL compositionality benchmarks. We use three levels of coverage: ✓ (directly and comprehensively evaluated), △ (partially overlapping—the dataset touches this dimension but not as a targeted evaluation), and ✗ (not covered). The comparison reveals that prior benchmarks predominantly evaluate along the object/attribute/relation perception axis, while MMComposition is the first compositionality benchmark to systematically cover counting, cross-image comparison, object interaction, and multi-correct probing.

## A.2 Definition of 13 distinct categories in MMComposition

- **Attribute Perception:** The specific attributes or properties of the object perception task that can be solved by humans “within a blink”.
- **Object Perception:** Identification or recognition of objects in the image.
- **Counting Perception:** Counting the number of objects or elements in the image.
- **Relation Perception:** Understanding the relationships between objects in the image.
- **Difference Spotting:** Identifying differences or changes between objects or scenes in two similar images.
- **Text Rendering:** Reading or interpreting text present in the image.

- **Visual Similarity:** Comparing similarities between objects or elements across multiple images.
- **Attribute Reasoning:** Identifying and reasoning about specific attributes or properties of objects in the image.
- **Object Reasoning:** Identifying and reasoning about objects in the image.
- **Counting Reasoning:** Identifying and reasoning about the number of objects or elements in the image.
- **Relation Reasoning:** Identifying and reasoning about the spatial arrangement or positioning of objects in the image.
- **Object Interaction:** Understanding interactions among multiple objects in the image.
- **VL Composition Probing:** Examining the composition or combination of visual and textual elements in images, where models are required to accurately find all the complex compositional descriptions about the image.

### A.3 Quantitive Results of MMComposition

In this section, we show statistical results for MMCOMPOSITION in Figure 5 through Figure 7.

### A.4 Comparison Analysis of Image Encoding in GPT-4o, Qwen2-VL, and InternVL-2

In GPT-4o, when the image detail parameters are set to “high”, images are first scaled to fit within a  $2048 \times 2048$  square while maintaining their aspect ratio. Then, the images are further scaled so that the shortest side is 768px long. Finally, GPT-4o calculates how many 512px squares the image contains, with each square costing 170 tokens. An additional 85 tokens for low resolution are always added to the final total. As a result, GPT-4o does not achieve true “any resolution” image processing.

In Qwen2-VL and InternVL-2, the image encoders adopt a dynamic “any resolution” encoding strategy. The images are first mapped to an optimal aspect ratio from predefined ratios, then divided into  $448 \times 448$  or  $28 \times 28$  pixel tiles, with each tile converted into 256 or 1 image tokens. A thumbnail is then generated to capture the global context. This allows the encoders to handle images of any resolution and aspect ratio. Furthermore, the image encoder in Qwen2-VL is a 675M ViT with a two-dimensional positional encoding mechanism, while InternVL-2 utilizes the more powerful InternViT with 6B parameters. This distinction contributes to the superior performance of the compositionality of Qwen2-VL and InternVL-2 in our benchmark. In Table 10, we provide a comparison of the properties of visual encoders for the aforementioned models.

Table 10: Visual encoder comparison of GPT-4o, InternVL2 and Qwen2-VL.

Method	Visual Encoder	Image Tile Size	Maximum Number of Tiles	Maximum Aspect Ratio	# of Tokens for One Tile
GPT-4o	-	512 x 512	8	any	170
InternVL2	InternViT-6B	448 x 448	12	1:6	256
Qwen2-VL	ViT-675M	28 x 28	dynamic	any	1

### A.5 Characteristics of Questions Where Models Underperform

We define the comprehensive performance value (CPV) for each question as the average score across 54 representative VLMs. By comparing each question’s CPV with the score of a random choice within its class, we find that 1,159 questions have a CPV lower than that of random chance. We show statistical results questions with low CPV in Figure 9 through Figure 12.



Figure 4: Interpretable analysis of different VLMs. Green letters indicate correct answers, while red letters represent wrong (predicted) answers.

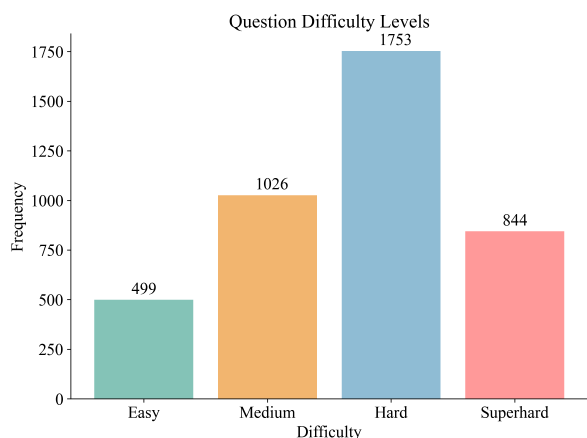


Figure 5: Distribution of difficulty levels across the question set, illustrating the challenging nature of tasks.

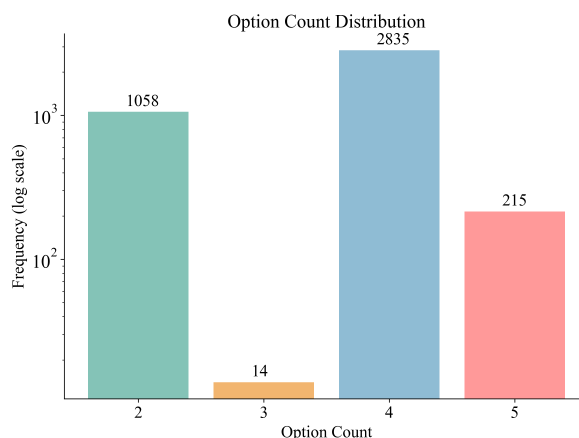


Figure 6: Distribution of option counts per question, showing the variety in answer choices provided to evaluate VLMs.

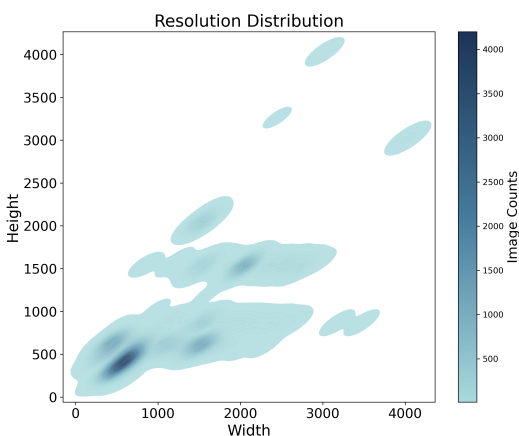


Figure 7: Resolution distribution of images in our benchmark, reflecting the portion of high-quality images in MMCOMPOSITION.

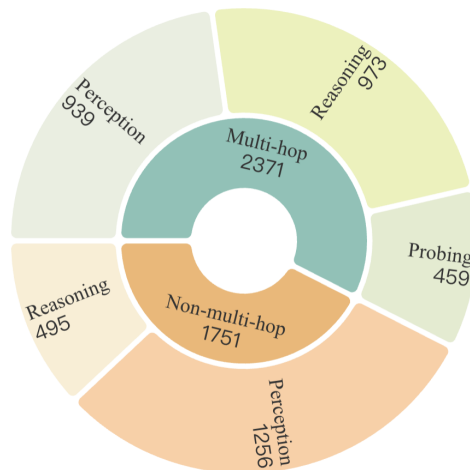


Figure 8: Distribution of multi-hop and non-multi-hop questions across categories.

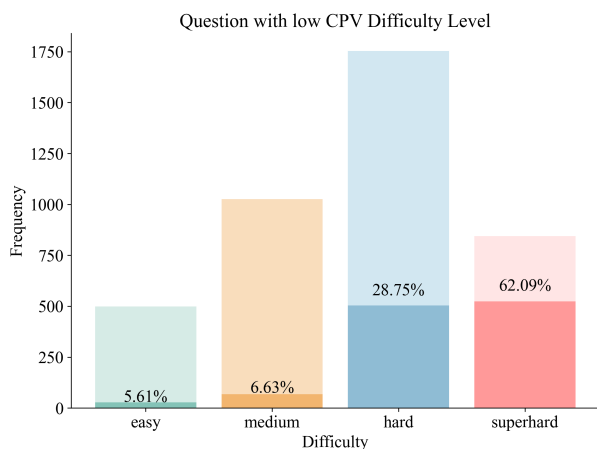


Figure 9: Distribution of difficulty levels of questions with low CPV, illustrating the authenticity of the difficulty distribution in MMCOMPOSITION.

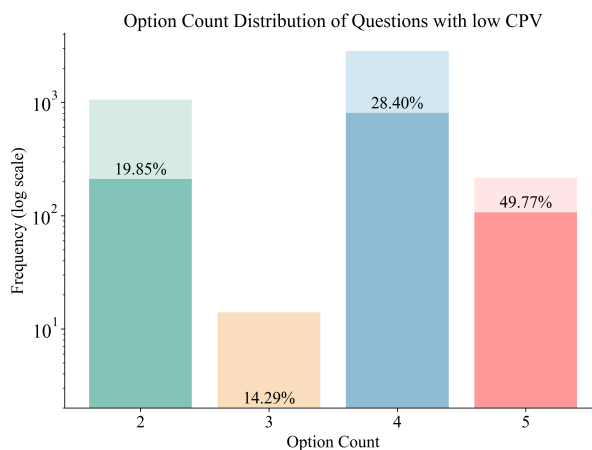


Figure 10: Distribution of option counts for questions with low CPV.



### Object Perception

**Which caption accurately describes the image?**




A: A big plate with some potatoes, broccoli, and meat on it.  
 B: A white plate with some broccoli, carrots, and meat.  
 C: A big plate with some potatoes and meat on it.   
 D: A plate containing potatoes, broccoli, carrots and chicken. 

 **D.** The image shows a plate containing potatoes, broccoli, carrots, and chicken, which matches the description in option D. The other options either miss some items or inaccurately describe the contents on the plate.

### Object Perception

**Which caption accurately describes the image?**



A: A tree underneath a green park bench in a park.  
 B: A sidewalk next to the outdoor sitting area of a restaurant.   
 C: A self service parking lot with an intersection behind it and a building in front of it.  
 D: An outdoor sitting area next to the sidewalk of a restaurant. 

 **D.** The image shows an outdoor sitting area next to a sidewalk with tables, chairs, and umbrellas, typical of a restaurant setting. People are seated and there is a sidewalk running alongside the sitting area, fitting the description in option D. This is in contrast to the other options which describe different settings not evident in the image.

### Counting Perception

**How many different colored pillows are stacked on top of each other in one of the images?**



A: One.  
 B: Three   
 C: Five   
 D: Two  
 E: Four 

 **C.** In Image 2, there are a total of five pillows of different colors stacked on top of each other, including purple, yellow, green, blue-green, and orange.

### Counting Perception

**How many large light fixtures are there above the area in the formal dining room?**



A: One.  
 B: Four   
 C: Three   
 D: Two   
 E: Zero

 **C.** In the formal dining room (Image 1), there are three large light fixtures hanging above the area. These fixtures can be seen evenly spaced out on the ceiling. This matches option C, which is "Three."

### Attribute Perception

**Which caption accurately describes the image?**



A: A black toilet with a white seat on top of a tiled floor.   
 B: A black toilet with a white seat sitting in a small stall.  
 C: Toilet with black splatter paint on a white seat in a restroom stall.   
 D: Toilet with white splatter paint on a black seat in a restroom stall. 

 **B.** The image shows a toilet with a white seat and lid that appears to be glittery or metallic. The toilet is situated in a confined space that resembles a restroom stall. The flooring appears to be tiled with a reddish or maroon color. The most accurate description based on these observations is option B: "A black toilet with a white seat sitting in a small stall."

### Attribute Perception

**Which caption accurately describes the image?**




A: A tall building with four double-decker buses driving along a parking lot.   
 B: Four tall buses driving along a parking lot next to a double-decker building.   
 C: Old buses sit at the curb in front of a double-decker building.   
 D: A tall bus parked in a lot beside a colorful building with many windows.

 **C.** The image depicts old buses parked at the curb in front of a double-decker building. The description accurately matches the scene in the image.

### Counting Reasoning

**In an outdoor medium view of the light blue sky, how many cirrus clouds are visible?**



A: Three.  
 B: Four   
 C: Five   
 D: Six

 **A.** In the image, there are exactly three cirrus clouds visible against the light blue sky.

### Counting Reasoning

**How many objects are tiny matte objects in front of the yellow metallic cylinder or tiny things left of the gray thing?**



A: 1  
 B: 2   
 C: 3   
 D: 4

 **B.** There are two objects that fit the criteria. The tiny matte red sphere is in front of the yellow metallic cylinder, and the tiny matte green cylinder is to the left of the gray metallic sphere.

### Relation Reasoning

**Which image, the left or the right, shows a pink scoop of ice cream on the bottom and a blue scoop on top?**





A: Left.  B: Right. 

 **A.** The image on the left shows an ice cream cone with a pink scoop of ice cream on the bottom and a blue scoop on top, whereas the image on the right shows a blue scoop on top and a white/pinkish scoop on the bottom.

### Relation Reasoning

**The pizza is \_\_\_\_\_ the person. (Choose the most appropriate option)**



A: attached to   
 B: surrounding  
 C: along  
 D: on 

 **D.** The pizza is positioned on the table in front of the person, making "on" the most appropriate choice.

Figure 13: GPT-4o Weak Category Analysis. The logos of the models or human displayed to the right of the option(s) indicate that the model or human has selected the option(s) as the correct answer(s).

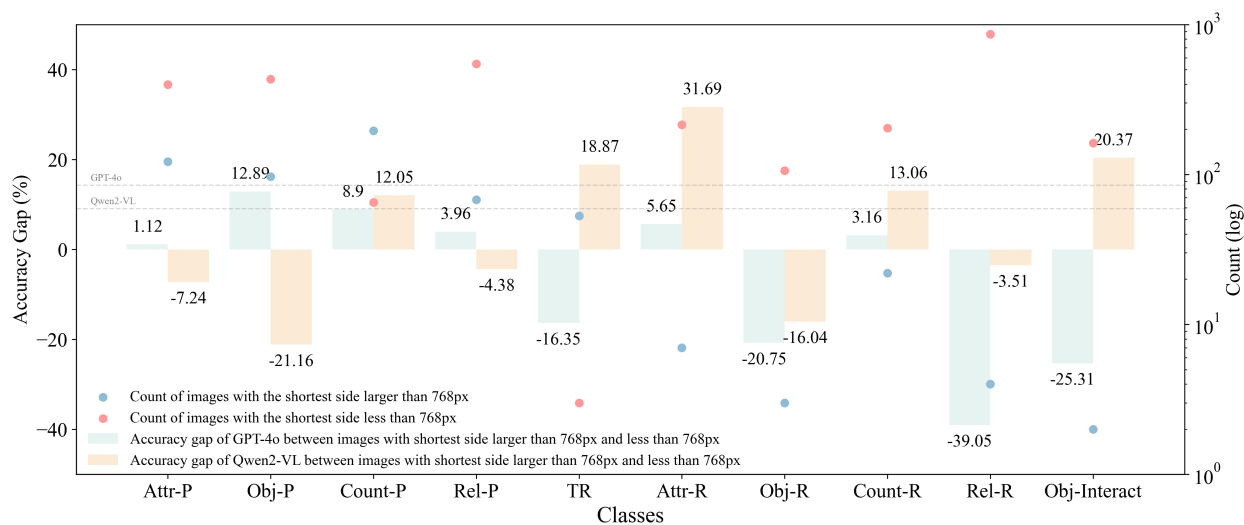



Figure 14: Performance gap between images whose shortest side  $> 768px$  and those  $\leq 768px$ , defined as  $gap = Acc_{>768px} - Acc_{\leq 768px}$ . The histogram shows the distribution of performance gaps across 13 tasks. The average performance gap for GPT-4o is 14.26, while for Qwen2-VL, it is 9.05. The smaller gap for Qwen2-VL indicates its greater effectiveness in processing high-resolution images. Additionally, Qwen2-VL’s performance gaps are more consistently positive across different tasks, further highlighting its robustness in handling high-resolution images.

### Counting Perception

How many people are wearing caps?

A: 5 B: 3  
C: 4 D: 6



**B.** There are three people in the image wearing caps. The first person is wearing a black cap, the second person is wearing a white straw hat, and the third person is wearing a green cap.


**B.** There are three people wearing caps: the man on the left, the person in the middle back, and the boy on the right.

**B.** In the image, there are three people wearing caps. One person is wearing a black cap, another is wearing a white cap, and the third is wearing a green cap.

**B.** The image shows 3 people wearing caps.

### VL Composition Probing

Select the false statement(s) about the image.



A: The man is walking on the sidewalk.

B: It is lawn.

C: The man is wearing sneakers.

D: The man is caucasian.

**A.** The man is not walking on the sidewalk. Instead, he is walking on a grass surface, indicating that option (A) is false.

**A, B, C.** The small sign appears to be made of a material similar to brick or clay, which are typically brown in color.

**A, B, C.** (A) The man is not walking on the sidewalk; he is on a grassy lawn. The statement is false because there is no sidewalk present in the image. (B) The image does not show a lawn; it shows a grassy field. The statement is false because the term "lawn" typically refers to a well-maintained area of grass, often found in residential or recreational settings, which is not the case here. (C) The man is not wearing sneakers; he is wearing sandals. The statement is false because sneakers are a type of closed-toe footwear, which is not what the man is wearing.

**A.** The man is not walking on the sidewalk. He is walking on the grass.

Figure 15: More interpretable analysis of different VLMs. Green indicates correct answers, while red represents the predicted wrong answers.

### Object Interaction Which caption accurately describes the image?



A: A person on a skateboard in front of a building on a street.  
 B: A person on a skateboard on a street.  
 C: A young man doing a jump with his scooter and skateboard in a skate park.  
 D: A person on a bike and a skateboard are riding on the street.

 Human  
 GPT-4o  
 Gemini-1.5-Pro  
 Qwen-VL-Max  
 ViLA  
 VILA-40B

**Human:** C. The image shows a young person doing a jump using a skateboard in an environment that appears to be a skate park, which is consistent with option (C). The other options either mention scenarios not depicted (like a building on a street, or a bike) or incorrectly describe the activity (person on a skateboard on a street).

**GPT-4o:** C. The image shows a young person performing a jump on a skateboard at what appears to be a skate park. Options A and B are not specific enough and D is incorrect as there is no bike in the image.

**Gemini-1.5-Pro:** C. The image shows a person performing a jump with a skateboard in a skate park. The other options are incorrect because they mention a building, a street, a scooter, or a bike, none of which are present in the image.

**Qwen-VL-Max:** C. The image shows a young man doing a jump with his scooter and skateboard in a skate park. The person is not on a bike, and the skateboard is not on a street. The person is not on a skateboard in front of a building on a street. The person is not on a bike and a skateboard riding on the street.

**ViLA:** C. The image shows a young man doing a jump with his scooter and skateboard in a skate park. The person is not on a bike, and the skateboard is not on a street. The person is not on a skateboard in front of a building on a street. The person is not on a bike and a skateboard riding on the street.

---

### Attribute Perception

Based on the overhead view, what is the color of the small sign?



A. Red  
B. Maroon  
C. Brown  
D. Purple

**Human:** A. The overhead view clearly shows that the sign is a vibrant red color, matching the color specified in option (A). The sign has a consistent red hue, rather than any shades of maroon, brown, or purple, confirming that the correct answer is "Red."

**GPT-4o:** C. The small sign appears to be made of a material similar to brick or clay, which are typically brown in color.

**Gemini-1.5-Pro:** C. The small sign in the image is a brownish-red color, which is best described by option (C) Brown. The other options do not accurately represent the color of the sign.

**Qwen-VL-Max:** A, B. The small sign is red and maroon in color.

**ViLA:** A, B. The small sign is red and maroon in color.

### Attribute Reasoning

What is the color of the sphere that is the same material as the big blue thing?



A: None of the choices provided  
B: Image 3  
C: Image 4  
D: Image 2

**Human:** D. Image 2 shows the same building structure and surrounding features as Image 1, confirming it as a building from the same university.

**GPT-4o:** A. None of the images provided show the same building as in Image 1.

**Gemini-1.5-Pro:** A. Image 1 depicts a modern building with a blue and white color scheme, surrounded by a parking lot and a roundabout. None of the other images (Images 2, 3, and 4) show a building that matches the description of the university building in Image 1. Therefore, the correct answer is A, as none of the choices provided show a building from the university depicted in Image 1.

**Qwen-VL-Max:** A. The university is not visible in any of the images.

**ViLA:** A. The university is not visible in any of the images.

---

### Relation Reasoning

The dining table is \_\_\_\_ the airplane.  
(Choose the most appropriate option)



A: three  
B: four  
C: above  
D: below

**Human:** A. The image shows a dining room with a table, chairs, shelves, and a cabinet. There is no visible airplane in the image. Given this information, and the possible inclusion of objects "not in the scene" as per usual interpretations, it is most reasonable to conclude that the dining table is "away from" the airplane (which is not in the scene). Thus, Option (A) "away from" is the most appropriate.

**GPT-4o:** A. The dining table is situated at a distance from the airplane (which is a miniature model placed on the shelf). Therefore, "away from" is the most suitable option.

**Gemini-1.5-Pro:** A. The dining table is positioned away from the airplane, as the airplane is located on a shelf above the table, not directly above or beneath it.

**Qwen-VL-Max:** A. The dining table is away from the airplane.

**ViLA:** A. The dining table is away from the airplane.

### Counting Perception

How many dogs are there?



A: Zero  
B: None of the choices provided  
C: Three  
D: Four  
E: One

**Human:** E. There is only one dog shown in the two images provided. Both images depict the same dog with similar markings and physical characteristics. Therefore, the correct answer is that there is one dog.

**GPT-4o:** E. There is only one dog present in the image. It is the same dog in both Image 1 and Image 2, just in different locations.

**Gemini-1.5-Pro:** E. There is one dog in each image, making a total of two dogs. However, the question asks for the number of dogs in the images, and since there is one dog in each image, the correct answer is "E" (One).

**Qwen-VL-Max:** E. There is one dog in the image.

**ViLA:** E. There is one dog in the image.

Figure 16: More interpretable analysis of different VLMs. Green indicates correct answers, while red represents the predicted wrong answers.

slightly higher than that of a random choice, indicating that our questions cannot be answered using common sense.

**V2L Adapters Comparison.** We also compare different V2L adapters. As shown in Table 13, models that utilize an MLP adapter (e.g., LLaVA1.5-13B) generally outperform those with a Q-Former adapter in overall performance. Specifically, LLaVA1.5-13B achieves an overall score of **41.02**, surpassing InstructBLIP-13B’s score of **36.76**. This suggests that the choice of adapter architecture significantly influences a model’s ability to effectively integrate visual features.

**Multi-hop/Non-multi-hop Question Setting.** We analyze the performance on multi-hop versus non-multi-hop question settings in Table 14 and observe that some models perform better on multi-hop questions, indicating strength in complex reasoning tasks. For example, InternVL2-40B achieved an overall score of **62.95** on multi-hop questions, compared to **73.35** on non multi-hop ones. This demonstrates the model’s enhanced capability to handle questions requiring multiple reasoning steps.

**Multi-image Setting.** As shown in Table 15, models respond differently to multiple images. Qwen2-VL-72B improves its overall score by **4.57**, indicating effective use of additional visual evidence, whereas InternVL2-40B drops by **1.63**, suggesting challenges in integrating information across images. These additional experiments underscore the difficulty of MMCOMPOSITION and highlight the role of visual evidence, adapter design, and the challenges of multi-hop reasoning in multimodal models. They also offer guidance for future work on improving vision–language model capabilities.

Table 12: Results for Image-Blind Setting using VLMs and LLMs.

Model	Perception	Reasoning	Probing	Overall
Qwen2-VL-72B	59.67	76.77	69.57	66.86
Qwen2-VL-72B-blind	44.72 <sub>-14.95</sub>	47.41 <sub>-29.36</sub>	30.76 <sub>-38.81</sub>	44.12 <sub>-22.74</sub>
InternVL2-26B	59.51	69.48	52.43	62.27
InternVL2-26B-blind	33.97 <sub>-25.54</sub>	42.30 <sub>-27.18</sub>	32.17 <sub>-20.26</sub>	36.74 <sub>-25.53</sub>
InternVL2-40B	64.55	74.66	59.59	67.60
InternVL2-40B-blind	37.46 <sub>-27.09</sub>	43.66 <sub>-31.00</sub>	34.28 <sub>-25.31</sub>	39.32 <sub>-28.28</sub>
InternVL2-76B	62.56	75.34	58.46	66.65
InternVL2-76B-blind	33.24 <sub>-29.32</sub>	43.26 <sub>-32.08</sub>	32.68 <sub>-25.78</sub>	36.74 <sub>-29.91</sub>
GPT-3.5	26.87	41.08	32.93	32.61
LLaMA-3.1-70B	36.11	35.35	26.58	34.78

Table 13: Comparison of Different Adapters for Model’s Performance.

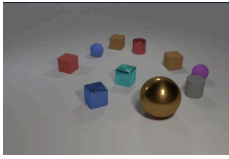
Model	Visual Encoder	LLM	V2L Adapter	Perception	Reasoning	Probing	Overall
mPLUG-Owl2 [Ye et al. (2024)]	ViT-L/14	LLaMA2-7B	Q-Former	36.26	45.16	30.36	38.77
InstructBLIP-7B [Dai et al. (2023)]	ViT-G/14	Vicuna-7B	Q-Former	33.23	43.05	31.41	36.53
LLaVA1.5-7B [Liu et al. (2024b)]	ViT-L/14	Vicuna-7B	MLP	35.76	44.28	30.32	38.19
InstructBLIP-13B [Dai et al. (2023)]	ViT-G/14	Vicuna-13B	Q-Former	35.51	42.23	25.24	36.76
LLaVA1.5-13B [Liu et al. (2024b)]	ViT-L/14	Vicuna-13B	MLP	36.79	47.89	39.32	41.02

Table 14: Comparison of models’ performance on multi-hop and non multi-hop questions.


Model	Perception	Reasoning	Probing	Overall
InternVL2-40B-non-multi-hop	72.56	75.35	-	73.35
InternVL2-40B-multi-hop	52.35	74.31	59.59	62.95
Qwen2-VL-72B-non-multi-hop	59.49	76.97	-	64.43
Qwen2-VL-72B-multi-hop	58.81	76.46	69.57	68.31
VILA-40B-non-multi-hop	65.54	69.29	-	66.60
VILA-40B-multi-hop	44.40	68.14	62.16	57.82
GPT-4o-non-multi-hop	62.30	63.03	-	62.51
GPT-4o-multi-hop	48.95	63.00	54.65	55.94
LLaVA-1.6-34B-non-multi-hop	65.27	68.69	-	66.24
LLaVA-1.6-34B-multi-hop	44.06	51.90	58.17	50.12
Gemini-1.5-Pro-non-multi-hop	54.71	53.54	-	54.38
Gemini-1.5-Pro-multi-hop	42.48	58.99	49.60	50.78

Table 15: Results for Multi-image Setting.


Model	Perception	Reasoning	Probing	Overall
Qwen2-VL-72B	55.36	74.68	89.86	70.09
Qwen2-VL-72B-multi	63.01 <sub>+7.65</sub>	78.55 <sub>+3.87</sub>	89.19 <sub>-0.67</sub>	74.66 <sub>+4.57</sub>
InternVL2-40B	42.35	71.45	88.51	63.79
InternVL2-40B-multi	39.29 <sub>-3.06</sub>	70.81 <sub>-0.64</sub>	86.49 <sub>-2.02</sub>	62.16 <sub>-1.63</sub>



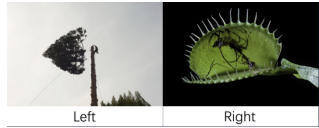
(Relation Reasoning) What size is the shiny thing that is behind the tiny gray thing and in front of the small red cylinder?  
A. large  
B. small  
**Answer: B**



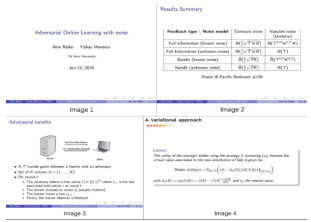
(Attribute Perception) What is the color of the line that horizontally intersects the centered 'X' in the image?  
A. Black B. White C. Red  
D. Blue  
**Answer: B**



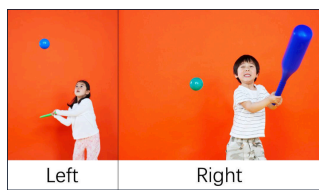
(Attribute Reasoning) Which image, left or right, shows the person with hair to their shoulders who has blue eyes, while the other person has brown eyes?  
A. Left  
B. Right  
**Answer: A**




(Object Interaction) Which image (left or right) shows a plant that was harmed by another organism, resulting in the plant being broken into pieces?  
A. Left  
B. Right  
**Answer: A**



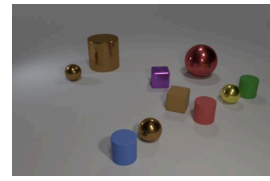
(Difference Spotting) Three of the following four slides are from the same presentation, but one is from a different one. Please identify the outlier.  
A. Image 4  
B. Image 3  
C. None of the choices provided  
D. Image 2  
**Answer: A**



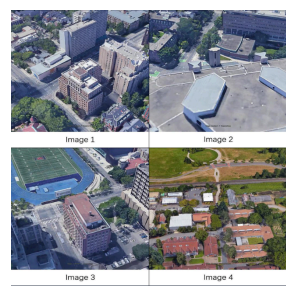
(VL Composition Probing) Which image, the left or the right, depicts a young person playing baseball with a blue bat and a green ball?  
A. Left  
B. Right  
**Answer: B**




(Counting Perception) How many different colored pillows are stacked on top of each other in one of the images?  
A. One  
B. Three  
C. Five  
D. Two  
E. None of the choices provided  
**Answer: E**




(Counting Reasoning) There is a matte cylinder to the left of the tiny purple shiny object; how many red metal things are behind it?  
A. 2 B. 3 C. 1 D. 4  
**Answer: C**




(Visual Similarity) Is it possible for you to unearth images containing the identical building as portrayed in Image 1?  
A. Image 2  
B. Image 4  
C. Image 3  
D. None of the choices provided  
**Answer: D**




(Text Rendering) Is there a red and black sticker covering any part of the text on the severely weathered white sign that reads "SKATEBOARDS OR BICYCLES ALLOWED ON SIDEWALK"?  
A. Yes, covering the word "ALLOWED"  
B. No, there is no sticker  
C. Yes, covering the word "SKATEBOARDS"  
D. Yes, covering the word "NO"  
**Answer: C**



(Object Perception) Which of the following statements is true?  
A. The first sign has a white arrow pointing up diagonal and the sign next to it has a white arrow pointing down.  
B. The first sign has a white arrow pointing down and the sign next to it has a white arrow pointing up diagonal.  
C. Both signs have white arrows pointing down.  
D. Both signs have white arrows pointing up diagonal.  
**Answer: B**



(Relation Perception) Is the gray street light directly behind the arm or the utility pole in the bottom right corner of the image?  
A. The street light is behind the arm.  
B. The street light is behind the utility pole.  
C. The utility pole is in front of the arm.  
D. The utility pole is in front of the street light.  
**Answer: A**



(Object Reasoning) One image shows a ferret standing on all fours on dirt, with its body in profile and its head turned.  
A. True  
B. False  
**Answer: B**

Figure 17: Examples of multi-hop questions: The ratio of multi-hop to non-multi-hop questions in our dataset is 2,371 to 1,751.

Table 16: The comprehensive performance of 93 VLMs on Acc, including open source models and API-based models . The **best** and second best results are in bold and underlined, respectively.

Method	Perception <sup>†</sup>							Reasoning <sup>†</sup>					Probing <sup>†</sup>	Overall <sup>†</sup>
	Attr-P	Obj-P	Count-P	Rel-P	Diff-S	TR	Visual-Sim	Attr-R	Obj-R	Count-R	Rel-R	Obj-Interact	Prob	
Human	97.94	98.04	93.06	92.00	79.02	85.71	86.54	91.20	78.83	100.00	77.35	88.00	91.84	90.31
Qwen2.5-VL-72B-Instruct (team, 2024)	<b>74.05</b>	<b>77.39</b>	<b>55.56</b>	68.01	39.00	50.00	63.51	<b>89.19</b>	<b>86.24</b>	<b>87.61</b>	<b>73.90</b>	<u>74.39</u>	42.92	<b>68.16</b>
InternVL2-40B (Chen et al., 2024b)	72.22	<u>75.99</u>	45.21	<b>72.53</b>	31.12	73.21	48.65	83.78	82.57	84.51	69.75	65.85	59.59	<u>67.60</u>
Qwen2-VL-72B-Instruct (team, 2024)	59.57	63.87	52.49	62.52	<b>45.23</b>	<b>82.14</b>	<u>67.57</u>	<u>87.84</u>	84.40	84.51	<u>71.49</u>	70.12	<b>69.57</b>	66.86
InternVL2-76B (Chen et al., 2024b)	70.65	75.52	48.28	<u>70.00</u>	19.09	78.57	48.65	85.14	83.49	<u>85.40</u>	70.01	67.07	58.46	66.65
InternVL2.5-78B-MPO (Chen et al., 2024b)	<u>73.28</u>	73.43	<u>53.64</u>	67.25	34.85	50.00	<b>74.32</b>	87.39	84.40	<u>85.40</u>	69.61	70.12	37.98	65.61
InternVL2.5-38B-MPO (Chen et al., 2024b)	71.55	72.03	51.34	66.99	<u>39.83</u>	44.64	<b>74.32</b>	84.68	<u>85.32</u>	86.73	68.41	70.12	41.00	65.34
InternVL-Chat-V1.2-Plus (Chen et al., 2024b)	69.81	65.73	43.68	69.02	31.12	78.57	28.38	78.83	77.98	80.53	66.27	60.98	<u>65.80</u>	64.58
InternVL2.5-26B-MPO (Chen et al., 2024b)	69.33	71.10	47.13	64.15	27.39	50.00	51.35	85.14	81.65	84.96	70.95	64.63	37.58	62.90
InternVL2.5-78B (Chen et al., 2024b)	70.07	66.90	47.13	64.23	32.37	48.21	60.81	85.59	82.57	80.09	68.27	73.17	37.58	62.64
68.46	67.13	40.23	66.96	22.82	<u>80.36</u>	62.16	79.28	79.82	81.86	62.65	63.41	52.43	62.27	
Qwen2-VL-7B-Instruct (team, 2024)	68.30	71.79	41.38	64.63	32.37	39.29	52.70	81.08	76.15	80.53	67.34	69.51	41.43	62.09
VILA-40B (Lin et al., 2024)	65.70	64.10	45.21	63.65	23.65	75.00	44.59	70.72	77.06	67.26	69.08	59.15	62.16	61.83
InternVL2.5-38B (Chen et al., 2024b)	66.51	67.60	46.74	60.28	30.29	53.57	59.46	84.23	83.49	80.97	65.19	71.95	41.68	61.43
InternVL2.5-26B (Chen et al., 2024b)	67.44	68.53	43.68	61.99	21.58	41.07	51.35	84.68	81.65	82.30	66.13	61.59	39.58	60.45
Ovis1.6-Gemma2-27B (Lu et al., 2024)	66.25	61.07	49.04	60.13	28.22	42.86	54.05	81.53	80.73	80.53	68.81	57.93	41.14	60.27
Qwen2.5-VL-7B-Instruct (team, 2024)	69.91	66.90	43.30	59.74	22.41	41.07	48.65	82.88	81.65	80.09	64.39	68.29	40.41	60.06
Molmo-7B (Deitke et al., 2024)	76.08	68.53	48.28	65.20	25.31	51.79	43.24	65.77	75.23	62.39	58.63	73.78	41.47	59.59
InternVL-Chat-V1.2 (Chen et al., 2024b)	64.58	62.00	41.38	62.98	25.73	76.79	29.73	63.06	71.56	61.06	63.19	65.24	60.71	59.13
GPT-4o (Achiam et al., 2023)	63.97	<u>57.58</u>	<u>37.93</u>	<u>66.76</u>	<u>32.37</u>	<b>82.14</b>	60.81	62.61	79.82	61.95	<u>58.37</u>	<b>75.00</b>	<u>54.65</u>	<u>59.03</u>
POINTS1.5-7B-Chat (Liu et al., 2024c)	70.13	61.54	39.46	60.39	24.90	46.43	44.59	76.13	77.06	76.11	60.24	69.51	45.21	58.66
InternVL-Chat-V1.5 (Chen et al., 2024b)	59.44	58.97	38.31	60.47	21.58	76.79	51.35	77.93	83.49	78.32	59.57	63.41	57.01	58.64
InternVL2.5-8B-MPO (Chen et al., 2024b)	65.64	66.20	45.21	58.12	21.99	41.07	56.76	78.83	80.73	76.55	65.19	59.76	37.44	58.49
Qwen2.5-VL-3B-Instruct (team, 2024)	64.80	65.03	40.61	58.98	29.46	46.43	35.14	76.58	77.06	78.76	61.45	60.37	39.58	57.71
Ovis1.6-Gemma2-9B (Lu et al., 2024)	65.16	61.54	40.61	55.20	23.65	41.07	54.05	80.18	77.06	76.99	66.13	67.68	38.63	57.88
InternVL2-8B (Chen et al., 2024b)	62.68	59.21	31.80	59.54	25.31	73.21	33.78	78.83	75.23	73.89	60.37	62.20	54.10	57.76
LLaVA-V1.6-34B (Liu et al., 2024a)	67.24	66.90	44.06	61.31	25.73	76.79	21.62	53.15	67.89	53.10	59.44	54.27	58.17	57.28
Llama-3.2-90B-Vision-Instruct	68.85	69.46	39.85	62.87	23.65	53.57	41.89	64.86	69.72	54.87	57.56	64.63	46.84	57.23
InternVL2.5-8B (Chen et al., 2024b)	64.10	61.07	39.85	59.23	19.09	39.29	36.49	71.62	73.39	76.11	63.45	62.20	40.41	56.54
MiniCPM-V2.6 (Yao et al., 2024)	65.19	58.04	41.00	61.80	21.99	73.21	37.84	63.96	73.39	68.14	52.07	60.98	54.43	56.07
InternVL2.5-4B-MPO (Chen et al., 2024b)	59.67	64.80	42.91	58.40	19.92	33.93	40.54	72.52	74.31	69.47	60.78	59.76	41.58	55.74
InternLM-XComposer2-4KHD-7B (Dong et al., 2024b)	62.24	55.24	39.08	58.36	23.65	67.86	27.03	70.72	74.31	60.18	55.82	59.15	60.02	55.35
Qwen-VL-Max (Bai et al., 2023)	53.76	<u>53.15</u>	<u>36.40</u>	<u>58.67</u>	<u>22.82</u>	<u>80.36</u>	41.89	53.60	65.14	<u>53.98</u>	60.91	62.80	63.87	<u>54.75</u>
InternVL2.5-4B (Chen et al., 2024b)	61.43	61.54	36.40	58.41	21.99	32.14	32.43	73.87	71.56	69.47	57.56	59.15	41.21	54.52
InternLM-XComposer2.5-7B (Zhang et al., 2024a)	56.68	56.64	37.93	56.82	21.58	71.43	28.38	71.17	75.23	61.06	58.50	60.98	49.64	54.38
InternLM-XComposer2-VL-7B (Dong et al., 2024a)	59.18	52.45	40.23	56.91	25.31	66.07	31.08	67.57	73.39	61.06	52.34	53.66	57.15	53.80
Hunyuan-Vision	61.95	61.31	37.16	58.58	26.97	76.79	36.49	61.26	72.48	56.19	52.21	59.15	45.03	53.67
Gemini-1.5-Pro (Reid et al., 2024)	55.30	53.50	39.46	57.11	24.48	67.86	55.41	59.91	74.31	50.44	56.29	65.24	49.60	53.27
Qwen2-VL-2B-Instruct (team, 2024)	58.86	59.21	36.40	54.77	20.33	44.64	20.27	71.62	70.64	69.91	53.15	60.98	44.81	52.81
InternVL2.5-2B-MPO (Chen et al., 2024b)	60.95	62.47	37.16	53.40	36.93	30.36	32.43	68.02	68.81	63.72	53.28	58.54	36.31	52.65
Mini-Gemini-34B (Li et al., 2023b)	58.35	55.01	37.93	53.70	25.31	73.21	39.19	54.50	73.39	58.41	55.82	61.59	41.79	51.96
Ovis1.6-Llama3.2-3B (Lu et al., 2024)	59.79	51.98	41.00	53.14	24.07	39.29	45.95	74.32	77.06	68.58	52.21	62.80	32.79	51.64
Molmo-7B-D (Deitke et al., 2024)	68.02	55.71	37.16	52.40	24.90	48.21	40.54	56.76	67.89	46.02	53.41	60.98	42.70	51.61
InternVL2-4B (Chen et al., 2024b)	53.82	51.05	31.42	52.17	18.26	73.21	25.68	77.03	71.56	72.57	54.08	55.49	41.18	51.00
MiniCPM-Llama3-V2.5 (Yao et al., 2024)	51.93	50.12	36.40	49.88	19.92	76.79	20.27	69.37	77.06	68.14	56.49	62.20	41.79	50.95
Mini-Gemini-34B-HD (Li et al., 2023b)	54.95	47.09	37.55	48.35	27.80	73.21	40.54	59.91	72.48	58.85	58.37	66.46	35.90	50.35
InternVL2.5-2B (Chen et al., 2024b)	58.61	58.74	36.78	54.26	22.82	28.57	24.32	66.67	66.06	62.83	48.73	55.49	39.80	50.93
Bunny-Llama-3B-V (He et al., 2024)	58.16	51.05	34.87	54.07	21.58	50.00	12.16	45.95	66.06	53.10	48.73	57.32	59.44	49.33
Mini-Monkey (Huang et al., 2024)	52.25	56.64	26.82	52.53	26.56	73.21	18.92	68.92	65.14	59.29	50.60	50.00	42.37	49.46
Phi3.5-Vision-Instruct (Abdin et al., 2024)	55.01	45.69	30.27	52.61	21.16	66.07	31.08	45.05	63.30	53.10	53.95	53.66	54.65	49.12
ColgVLM2-Llama3-Chat-19B (Hong et al., 2024)	57.67	51.05	34.48	51.69	38.17	57.14	48.65	50.90	65.14	47.35	40.96	59.15	50.69	48.83
Phi3-Vision-128K-Instruct (Abdin et al., 2024)	55.30	39.86	30.27	51.61	25.31	69.64	40.54	45.05	65.14	47.79	45.25	60.37	56.75	47.57
Yi-VL-34B (AI et al., 2024)	53.02	38.23	30.27	50.33	26.14	64.29	17.57	50.45	56.88	55.31	51.00	52.44	53.88	47.38
Step-1V-32K	46.11	39.86	26.44	46.25	25.31	67.86	43.24	66.67	66.97	62.83	50.60	59.76	45.46	47.12
Yi-VL-6B (AI et al., 2024)	51.99	43.82	30.27	49.34	25.73	60.71	20.27	45.05	51.38	52.21	50.33	51.83	48.76	46.34
ConvLLaVA-1024-7B (Ge et al., 2024)	51.73	44.29	32.57	44.96	28.22	69.64	21.62	55.41	65.14	53.10	49.53	54.88	40.89	46.21
ConvLLaVA-1536-7B (Ge et al., 2024)	50.03	45.69	28.35	41.25	27.39	69.64	29.73	51.35	64.22	52.21	48.19	64.02	34.20	45.52
Bunny-3B (He et al., 2024)	49.97	49.65	26.82	48.79	25.73	50.00	12.16	46.40	61.47	47.79	42.17	51.22	55.08	45.40
Bunny-4B-V1.0 (He et al., 2024)	52.50	46.85	39.08	46.00	21.16	51.79	17.57	43.69	62.39	52.21	46.59	52.44	42.66	45.23
LLaVA-HR-13B (Luo et al., 2024)	50.32	41.26	35.25	39.81	32.37	66.07	27.03	45.50	60.55	45.58	48.46	57.32	48.80	45.12
InternVL2-2B (Chen et al., 2024b)	43.32	53.61	26.82	45.79	22.82	67.86	17.57	63.06	58.72	49.56	44.31	53.66	38.16	44.22
Monkey-Chat (Li et al., 2024)	49.20	47.55	24.14	47.13	16.60	69.64	13.51	51.35	58.72	44.25	44.18	51.22	48.91	44.10
InternVL2.5-1B (Chen et al., 2024b)	47.43	52.45	28.35	49.43	25.31	35.71	14.86	49.55	67.89	52.65	41.63	53.66	36.93	43.98
InternVL2.5-1B-MPO (Chen et al., 2024b)	45.05	51.05	28.35	46.03	29.46	25.00	35.14	48.65	64.22	48.67	42.30			