

Supplementary Material for Blocks as Probes: Dissecting Categorization Ability of Large Multimodal Models

Bin Fu^{1,2}

bin.fu@vipl.ict.ac.cn

Qiyang Wan^{1,2}

qiyang.wan@vipl.ict.ac.cn

Jialin Li^{1,2}

jialin.li@vipl.ict.ac.cn

Ruiping Wang^{1,2}

wangruiping@ict.ac.cn

Xilin Chen^{1,2}

xlchen@ict.ac.cn

¹ Key Laboratory of AI Safety of CAS,
Institute of Computing Technology,
Chinese Academy of Sciences (CAS)
Beijing, China

² University of Chinese
Academy of Sciences
Beijing, China

1 Introduction

In this supplementary material, we elaborate on more details including dataset samples ([Sec.3.1 in the main paper¹](#)) and evaluation tasks ([Sec.3.2 in the main paper](#)) of our ComBo benchmark in [Sec.2](#), more details on evaluation settings including configuration of LMMs, prompts ([Sec.4.1 in the main paper](#)), and user studies in [Sec.3](#), more supplementary examples for experimental results ([Sec.4.2 in the main paper](#)) in [Sec.4](#), and extra discussion on task difficulty ([Sec.4.3 in the main paper](#)) in [Sec.5](#). As mentioned in the main paper, the full ComBo benchmark and all evaluation results will be released to the public upon acceptance.

2 The ComBo Benchmark

In this section, we first provide a detailed introduction to the construction process of composite objects and the dataset generation process ([Sec.2.1](#)). Subsequently, we elaborate on the specific details of three evaluation tasks ([Sec.2.2-2.4](#)).

2.1 Object Assembly and Image Rendering

Each object within **Composite Blocks** (ComBo) is composed of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The geometric primitives are

available in five shapes: cube, sphere, cylinder, cone, and conical frustum. During the assembly process, both the primary and secondary primitives select a contact point, denoted as the anchor and pivot, respectively, as shown in Fig.1. The secondary primitive’s pivot is then attached to the primary primitive’s anchor at a predefined angle.

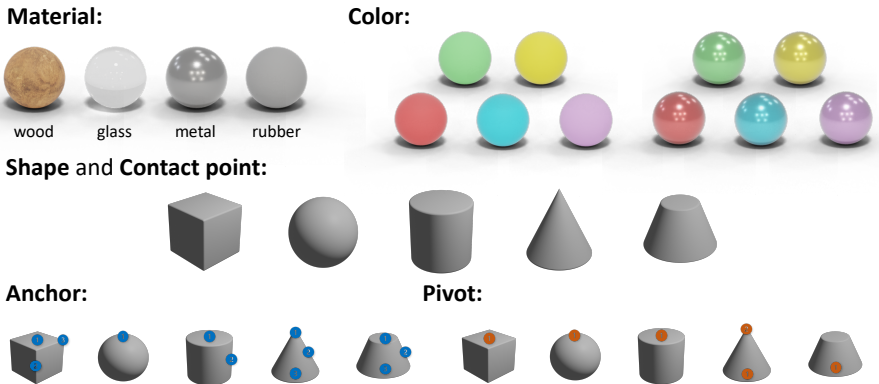


Figure 1: Types of patterns in ComBo.

It is noted that among secondary primitives, only cone has two selectable pivot options. Therefore, for other shapes, when examining the Large Multimodal Models’ (LMMs) perception of pivots through multiple-choice questions, there is only one possible answer, which solely assesses whether the LMMs can legally produce the unique correct answer. Although we have actually evaluated the LMMs’ perception of pivots, due to the limited practical significance of this metric and space limitations, we did not discuss pivots and related experiments in the main paper. We simplified the assembly process of primitives in the main paper to state that the primary and secondary primitives are connected through a contact point on the primary primitive. Here, the true meaning of the “contact point” is actually the anchor selected during the assembly process of the objects.

To enhance the visual diversity of ComBo, four different materials are assigned to the geometric primitives: rubber, metal, glass, and wood. Additionally, the rubber and metal materials are further differentiated by five colors: red, yellow, blue, green, and purple.

We utilize a ray tracing based rendering engine [24] to render each composite object, thereby obtaining more photorealistic images, inspired by CLEVR [9]. The composite objects are positioned at the center of the scene, with appropriate environment lighting and several point lights. Each object is rendered from 20 random viewpoints, culminating in the 190,080 images in ComBo. More composite objects are demonstrated in Fig.2.

2.2 Pattern Perception

Pattern Perception is the most fundamental visual perception task, requiring subjects to answer a set of questions about composite objects, including shape, color, material, and contact point. It is important to note that the contact point consists of the anchor on the primary primitive and the pivot on the secondary primitive, together forming the manner of object composition. However, for most geometric primitives, the choice of pivot when acting as a secondary primitive is unique (with only *cone* being the exception). Therefore, when we discuss the contact point in the main paper, we are primarily referring to the anchor on the



Figure 2: More exemplar images in ComBo.

primary primitive.

2.3 Abstraction Alignment

In this task, we invite cognitive science experts to select appropriate natural categories that can be abstracted by our ComBo objects.

Specifically, the cognitive science experts select objects from the dataset that correspond significantly to the natural categories in human cognition, such as “ice cream”. After filtering and voting, a final consensus of 24 categories, including image samples and category labels, are selected for evaluating, with specific categories and corresponding images are listed in Fig.3.

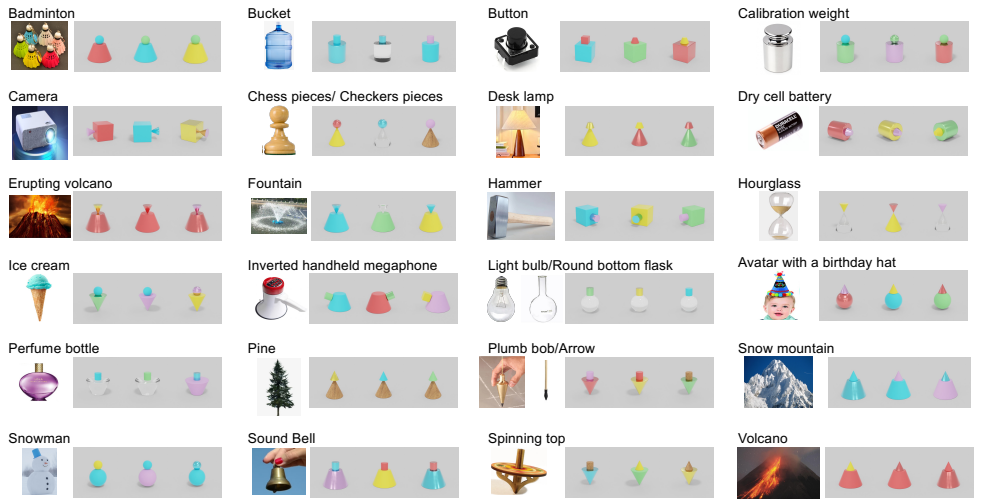


Figure 3: Category labels and corresponding images selected in **Abstraction Alignment**.

In this task, the color and material of objects are usually ignored, while shape plays a crucial role (according to the human cognition). Additionally, although lighting and viewpoint are generally irrelevant, some specific categories are sampled under particular viewpoints if their pose greatly impacts category recognition.

Participants are asked two types of multiple-choice questions. *Img2Text* requires the participants to choose the label that best matches the given image out of four category labels. The four options include one correct answer and three distractors, which are selected from the remaining 23 categories and 21 categories introduced from the COCO dataset (shown in Fig.4). Since this task evaluates abstraction alignment capabilities of LMMs, we ask cognitive science experts to avoid choosing distractor categories that are very similar to the query image to prevent the problem from being overly difficult. *Text2Img* asks the participants to select the image that most resembles the given category from four image options.



Figure 4: Distractor categories from COCO.

2.4 Category Building

As discussed in the main paper, in this task, we require participants to simulate the category formation in human consensus by constructing abstract categories of different granularities, and then classify the test samples. Abstract categories are groups of object clusters defined based on rules, where all composite objects in an abstract category have the same constraint (e.g., “with a red cube as the primary primitive”). The more constraints imposed, the fewer the dimensions of patterns that can be freely chosen, resulting in fewer composite objects included within an abstract category and a finer granularity of the category.

It is worth noting that the difficulty of the task is related to the selected abstract categories. When the support samples from two categories are easily confused, summarizing the rules can become challenging. To further model the task difficulty, we consider extracting visual features of all composite objects in the abstract categories using a general visual model (e.g., CLIP) to quantitatively characterize the classification difficulty based on the distribution distances of the samples. Through cross-validation and user study, we finally adopt the solution of PCA dimensionality reduction, Gaussian parameter estimation, and Wasserstein distance as metrics. Under this metric, the difficulty of distinguishing between two abstract categories is close to the difficulty in human cognition.

3 Evaluation Settings

3.1 Details about LMMs

As mentioned in [Sec.4.1 in the main paper](#), we select the mainstream closed-source implementations of current LMMs (GPT-4V [5, 8], Gemini-1.5-Pro [9]) and open-source implementations (LLaVA-v1.5-13B [9], Qwen-VL-Chat [9]) as the subjects of our study.

To evaluate GPT-4V and Gemini, we utilize their official APIs. For LLaVA-v1.5-13B and Qwen-VL-Chat, we conduct local tests using a single NVIDIA A40 GPU. Considering the continuous updates to GPT-4V, we use different versions tailored to each task: for the pattern perception task, we use gpt-4-1106-vision-preview; for the abstraction alignment and category building tasks, we employ gpt-4-turbo-2024-04-09. Specifically, we set the temperature of GPT-4V as zero in all tasks. For all three evaluation tasks, we utilize the most recent version of Gemini, gemini-1.5-pro-latest, which was updated in April 2024.

LLaVA-v1.5-13B is an open-source chatbot trained by fine-tuning Vicuna v1.5 on GPT-generated multimodal instruction-following data. Qwen-VL-Chat is also a vision-language chatbot, based on Qwen-VL-7B, with enhanced capabilities in following instructions.

3.2 Prompts and Multiple Image Input

For all LMMs, our evaluation begins with the most comprehensive prompts. These prompts are intuitively designed to include system information about the ComBo benchmark, questions, multiple-choice options, and explicit instructions. In cases where some LMMs are incapable of processing multiple images, we merge these images into a single, larger composite image and make corresponding modifications to the prompts, as shown in Fig.5. For models with less robust instruction-following capability, excessively lengthy prompts and complex questions may hinder their ability to generate coherent responses. Consequently, we iteratively shorten or modify the prompts to facilitate normal output. These efforts are primarily aimed at ensuring that open-source LMMs can produce appropriate content, considering the different capabilities and preferences of various LMMs. It is important to emphasize that all adjustments to the images and prompts are made to ensure a fair evaluation, tailoring only the format of the problem to align with the capability of each model.

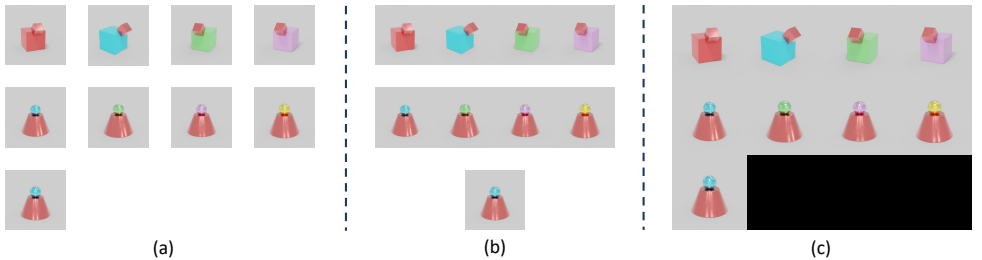


Figure 5: Taking **Category Building** task as an example, we illustrate three methods of presenting images to LMMs. In method (a), nine images are passed to the LMMs as separate files. In method (b), four images belonging to the same abstract category are concatenated, reducing the number of images to simplify the reasoning process. In method (c), all images are combined into one large image, which is useful for LMMs that can only accept a single image input.

3.3 User Study

We conducted multiple user studies to ensure the reasonableness of the questions and to provide human performance on different tasks for comparison with LMMs’ capabilities.

In the **Abstraction Alignment** task, we invited 20 participants to complete the user study, with each participant answering 20 questions for each of the two question types. After removing the highest and lowest total scores, we calculate the accuracy rates for each question type and the overall accuracy rate. The results of the user study indicate that the questions constructed by the cognitive science experts align with human perception of natural categories. Therefore, these data can be used to assess the **Abstraction Alignment** capabilities of LMMs compared to humans.

In the **Category Building** task, the user study was divided into two parts. First, we invited 8 participants to rank the difficulty of classifying abstract categories. Participants needed to complete 40 questions, each containing two options, with each option showing representative images of two abstract categories. Participants were asked to determine which pair of the two abstract categories is more difficult to separate. The aggregated answers were then compared with the Wasserstein distance between the two sets of categories, achieving a similarity of 95.4%, which demonstrates that this metric aligns with human cognition and can be used to measure the classification difficulty of abstract categories. Subsequently, we invited 23 participants to complete the **Category Building** task and recorded their scores across different task difficulties.

The webpage used for completing the user studies is shown in Fig.6.

4 Evaluation Results

In this section, we present the specific prompts used to evaluate LMMs across different tasks, as well as the responses from various LMMs.

As online LMMs continue to evolve and their outputs are non-deterministic, the actual output for a specific example cannot be guaranteed to be consistent with the results shown in the text. However, the overall score is derived from a large number of questions, which is statistically significant.

4.1 Pattern Perception

To save costs and fully utilize the batch inference capabilities of LMMs, we adopt the method described in [2], grouping multiple samples into one batch for a single query. In our experiments, we select the same batch size of 10 as recommended in [2], which has almost no negative impact on the inference results. In addition, based on the principle that an option should not be regarded as a test point when it is the only choice, in experiments concerning in-context learning where the anchor of the *sphere* has only one option, we do not consider cases where the primary primitive is a *sphere* in our experiments when evaluating the capability to recognize anchors.

In Fig.7, Fig.8, and Fig.9, we illustrate some Pattern Perception results from four LMMs. In Fig.10 and Fig.11, we illustrate in-context learning results of GPT-4V and Gemini.



Figure 6: Webpage screenshots of user studies.

4.2 Abstraction Alignment

In Fig.12 and Fig.13, we illustrate the prompts for *Img2Text* and *Text2Img* along with the outputs from all LMMs.

4.3 Category Building

We start by posing questions from the most natural prompts and gradually add hints or simplify the questions in the prompt when LMMs are completely unable to produce effective responses, until it is confirmed that LMMs can fully understand the question. Samples of abstract categories are provided in groups of category A and category B. Both GPT-4V and Gemini support multiple images as inputs. Therefore, we combine four samples from category A into a large composite image, as well as four samples from category B, along with one image to be tested, forming a set of three images (4+4+1) per question. The prompt is presented in Fig.14. LLaVA and Qwen do not support inference with multiple image inputs, therefore we design a manual chain of thought process, as shown in Fig.15-Fig.18.

Specifically, we require LLaVA to perform pattern perception on the nine input images, then concatenate the output results as input into the category building prompt, asking LLaVA to classify the test image according to the perceived patterns (in text) and the test image. However, Qwen cannot even handle the perception questions for eight patterns simultaneously. If we split the queries, it would require 72 queries to obtain the pattern perception results for all nine images, which is overly complex and cumbersome. Therefore, we further simplify the process by not requiring Qwen to output the perception results in a specific format. Instead, we allow it to perform free captioning on the input samples. Subsequently, we concatenate the generated captions into the category building prompt and perform the final classification task based on the text input and the test sample. Nevertheless, regardless of which instruction is chosen, Qwen is unable to produce valid results, or it answers all questions with the same option (for example, "A"). Therefore it is judged as unable to complete this task, and consequently, as described in [footnote 2 of the main paper](#), there is no corresponding bar in [Fig.7 in the main paper](#).

We also illustrate chain-of-thought (CoT) results of GPT-4V and Gemini in [Fig.19](#) and [Fig.20](#).

5 More Discussion on Task Difficulty

As discussed in [Sec.4.2 in the main paper](#), we continue to focus on the **difficulty** of the ComBo benchmark in this section. To verify that the questions in ComBo benchmark are much less difficult than these benchmarks in terms of logical reasoning, we use some smaller, commonly used computer vision models — *ResNet-50* and *ViT-B/16* — to complete the same tasks. By compared with LMMs by pre-trained and fine-tuned models, readers can better perceive the **fundamental** characteristics of the ComBo benchmark.

In the most challenging task of Category Building, for LMMs that can accept natural language input, we describe the problem and directly ask questions in the form of *instruction-following QA*, requiring them to respond in the specified format (usually in JSON format, as we do in the main paper). For CNN and ViT, we simply extract features from the input images by the pre-trained models, and then calculate the average features of all samples in each abstract category as its prototype, and classify the test image according to the nearest principle. This method is called *clustering & retrieval*. As shown in the upper part of [Table 1](#), through *clustering & retrieval*, ResNet-50 and ViT-B/16 both perform similarly to GPT-4V; in contrast, there is a significant gap between the pre-trained LLaVA and GPT-4V, which aligns with our prior.

To obtain a more appropriate feature extractor by fine-tuning CNN and ViT, we leverage the straightforward classification task among 9,504 categories as an auxiliary task. After a quick fine-tuning for one epoch, as shown in the lower part of [Table 1](#), applying the same *clustering & retrieval* methods can almost perfectly complete the task of Category Building, achieving an classification accuracy far higher than GPT-4V. As a comparison, we also fine-tune LLaVA. We construct a dataset containing 33,656 image descriptions and 40,000 task-specific question-answer pairs, and fine-tune LLaVA with LoRA. As shown in the last row of [Table 1](#), fine-tuned LLaVA can also achieve excellent performance on this task, but at the cost of greatly compromising its general visual capabilities.

In this section, the fine-tuned CNN, ViT, and LLaVA can all be regarded as **specialized models for completing specific tasks**. They perform excellently on the ComBo benchmark using pure visual capabilities, which is to be expected. The above experiment demonstrates

Model		Method	Easy	Medium	Hard	Expert
LMM	LLaVA	instruction-following QA	0.48	0.38	0.36	0.26
	GPT-4V		0.98	0.92	0.80	0.72
pre-trained	ResNet-50	clustering & retrieval	0.98	0.80	0.70	0.58
	ViT-B/16		1.00	0.98	0.86	0.66
fine-tuned	ResNet-50	clustering & retrieval	1.00	1.00	1.00	0.92
	ViT-B/16		1.00	1.00	0.96	0.94
fine-tuned	LLaVA	instruction-following QA*	1.00	1.00	0.98	0.92

Table 1: Category Building results for LMMs, pretrained models, and fine-tuned models. Instruction-following QA marked with an asterisk (*) indicates that the fine-tuned LLaVA has lost most of its general visual capabilities, only able to complete Pattern Perception and Category Building well. Moreover, it is extremely sensitive to prompts and cannot tolerate minor perturbations.

that the ComBo benchmark can effectively test the fundamental visual capabilities of LMMs in a Q&A manner. The experimental results indicate that current LMMs still have gaps compared to specialized models or humans when performing basic perception tasks. We believe that a powerful LMM should pay more attention to the ability of fundamental visual perception while accomplishing various high-level tasks, so that LMMs can develop more robustly.

I will be sharing 10 images with you for analysis. These images are sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive, and the contact points are denoted as the anchor and pivot respectively. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset. For the composite object in each image, I would like you to answer three specific questions. These questions are multiple-choice, and I will provide descriptions for each. Your task is to review the images in the order provided and give your answers accordingly. Here are the three questions.

Question 1: What are the shapes of primary and secondary primitives?

Description of Question 1: Each object is comprised of two geometric primitives, the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. As the secondary primitive is attached on the primary primitive, the primary primitive is always on the ground. Both of the primary and secondary primitives are available in five shapes given in the options.

Option of Question 1: A (cube), B (sphere), C (cylinder), D (cone), E (cone frustum)

Question 2: What are the contact points of primary and secondary primitives?

Description of Question 2: The contact point of primary primitive is called anchor, and the contact point of secondary primitive is called pivot. The options of this question are influenced by the shape of the primitives asked in Question 1, so you must maintain the consistency in the answers to both questions. For example, if you choose the primary primitive as (A) cube, in this question, the anchor of the primary primitive must be chosen from (A-1), (A-2), and (A-3). Please note that the anchor and pivot point refer to the relative position. You need to consider the rotation of the object when making judgments. In addition, we assume that the large circular surface of the cone frustum is the bottom base and the small circular surface is the top base.

Option of Question 2 (anchor):

If you choose the primary primitive as cube, the options of this question is: A-1 (center of a cube face), A-2 (midpoint of a cube edge), A-3 (vertex of a cube)

If you choose the primary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the primary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's base), C-2 (midpoint of a cylinder's generatrix)

If you choose the primary primitive as cone, the options of this question is: D-1 (apex of cone), D-2 (midpoint of the cone's slant height), D-3 (center of the cone's base)

If you choose the primary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's top base), E-2 (midpoint of the cone frustum's slant height), E-3 (center of the cone frustum's bottom base)

Option of Question 2 (pivot):

If you choose the secondary primitive as cube, the options of this question is: A-1 (center of a cube face)

If you choose the secondary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the secondary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's base)

If you choose the secondary primitive as cone, the options of this question is: D-1 (center of the cone's base), D-2 (apex of cone)

If you choose the secondary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's bottom base)

Question 3: What are the materials and colors of primary and secondary primitives?

Description of Question 3: The primary and secondary primitives are each made from one of the following materials: rubber, metal, glass, or wood. (1) The rubber material has a matte surface, and the color of rubber surface is almost uniform and consistent. (2) The metal material exhibits a metallic luster. Compared to rubber, metal has a sheen and sharper edges. (3) The glass material is transparent. (4) The wood material is brownish yellow with brown texture on the surface. Moreover, primitives made from glass and wood have unique appearance, if a primitive's material is glass or wood, the answer of its' color should be F (NA). On the other hand, the rubber and metal materials come in five different colors: red, yellow, blue, green, and purple. Do Not choose F (NA) for a rubber or metal primitive.

Option of Question 3 (material): A (rubber), B (metal), C (glass), D (wood)

Option of Question 3 (color): A (red), B (yellow), C (blue), D (green), E (purple), F (NA)

Requirement: You should answer questions based solely on your vision ability, not writing and running code. And your answer should only contain JSON and strictly in the format below, you are not allowed to output anything else. { "image 1": { "shape of primary primitive": "", "shape of secondary primitive": "", "anchor of primary primitive": "", "pivot of secondary primitive": "", "material of primary primitive": "", "material of secondary primitive": "" }, "image 2": {}, "image 3": {}, ... "image 10": {} }

Please ensure your responses are accurate and based on the content of each image.



GPT-4V


```
...
"image 2": {
  "shape of primary primitive": "D",
  "shape of secondary primitive": "D",
  "anchor of primary primitive": "D-3",
  "pivot of secondary primitive": "D-1",
  "material of primary primitive": "B",
  "color of primary primitive": "B",
  "material of secondary primitive": "B",
  "color of secondary primitive": "A"
}, ...
"image 6": {
  "shape of primary primitive": "C",
  "shape of secondary primitive": "A",
  "anchor of primary primitive": "C-1",
  "pivot of secondary primitive": "A-1",
  "material of primary primitive": "A",
  "color of primary primitive": "A",
  "material of secondary primitive": "D",
  "color of secondary primitive": "E"
}, ...
```

Gemini

```
...
"image 2": {
  "shape of primary primitive": "E",
  "shape of secondary primitive": "D",
  "anchor of primary primitive": "E-1",
  "pivot of secondary primitive": "D-1",
  "material of primary primitive": "A",
  "color of primary primitive": "B",
  "material of secondary primitive": "B",
  "color of secondary primitive": "B"
}, ...
"image 6": {
  "shape of primary primitive": "C",
  "shape of secondary primitive": "A",
  "anchor of primary primitive": "C-2",
  "pivot of secondary primitive": "A-1",
  "material of primary primitive": "A",
  "color of primary primitive": "A",
  "material of secondary primitive": "D",
  "color of secondary primitive": "F"
}, ...
```


Figure 7: Examples of the **Pattern Perception** task, using standard prompts, showcasing partial outputs from GPT-4V and Gemini.

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the shape of the large primitive?
Options: A.cube, B.sphere, C.cylinder, D.cone, E.cone frustum
Requirements: Choose one from options, you can **ONLY** output the identifier.




Answers: Qwen: A LLaVA: A

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the shape of the small primitive?
Options: A.cube, B.sphere, C.cylinder, D.cone, E.cone frustum
Requirements: Choose one from options, you can **ONLY** output the identifier.




Answers: Qwen: A LLaVA: C

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the material of the large primitive?
Options: A.rubber, B.metal, C.transparent glass, D.wood
Requirements: Choose one from options, you can **ONLY** output the identifier.




Answers: Qwen: D LLaVA: D

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the color of the large primitive?
Options: A.red, B.yellow, C.blue, D.green, E.purple, F.NA(color is transparent or wood color)
Requirements: Choose one from options, you can **ONLY** output the identifier.




Answers: Qwen: C LLaVA: F

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the material of the small primitive?
Options: A.rubber, B.metal, C.transparent glass, D.wood
Requirements: Choose one from options, you can **ONLY** output the identifier.




Answers: Qwen: D LLaVA: D

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the color of the small primitive?
Options: A.red, B.yellow, C.blue, D.green, E.purple, F.NA(color is transparent or wood color)
Requirements: Choose one from options, you can **ONLY** output the identifier.



Answers: Qwen: C LLaVA: C

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
Where is the junction of the large and small primitives located in the large primitive?
Options: A.center of a cube face, B.midpoint of a cube edge, C.vertex of a cube
Requirements: Choose one from options, you can **ONLY** output the identifier.



Answers: Qwen: B LLaVA: C

Figure 8: An example of the **Pattern Perception** task, we use simpler prompts for LLaVA and Qwen.

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the shape of the large primitive?
Options: A.cube, B.sphere, C.cylinder, D.cone, E.cone frustum
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: D LLaVA: E

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the shape of the small primitive?
Options: A.cube, B.sphere, C.cylinder, D.cone, E.cone frustum
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: D LLaVA: E

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the material of the large primitive?
Options: A.rubber, B.metal, C.transparent glass, D.wood
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: A LLaVA: C

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the color of the large primitive?
Options: A.red, B.yellow, C.blue, D.green, E.purple, F.NA(color is transparent or wood color)
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: C LLaVA: C

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the material of the small primitive?
Options: A.rubber, B.metal, C.transparent glass, D.wood
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: C LLaVA: C

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
What is the color of the small primitive?
Options: A.red, B.yellow, C.blue, D.green, E.purple, F.NA(color is transparent or wood color)
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: D LLaVA: D

There is an object in the image I gave you. This object is composed of two primitives, one large and one small.
Where is the junction of the large and small primitives located in the large primitive?
Options: A.apex of cone, B.midpoint of a cone's slant height, C.center of the cone's base
Requirements: Choose one from options, you can **ONLY** output the identifier.

Answers: Qwen: A LLaVA: C

Figure 9: An example of the **Pattern Perception** task, we use simpler prompts for LLaVA and Qwen.

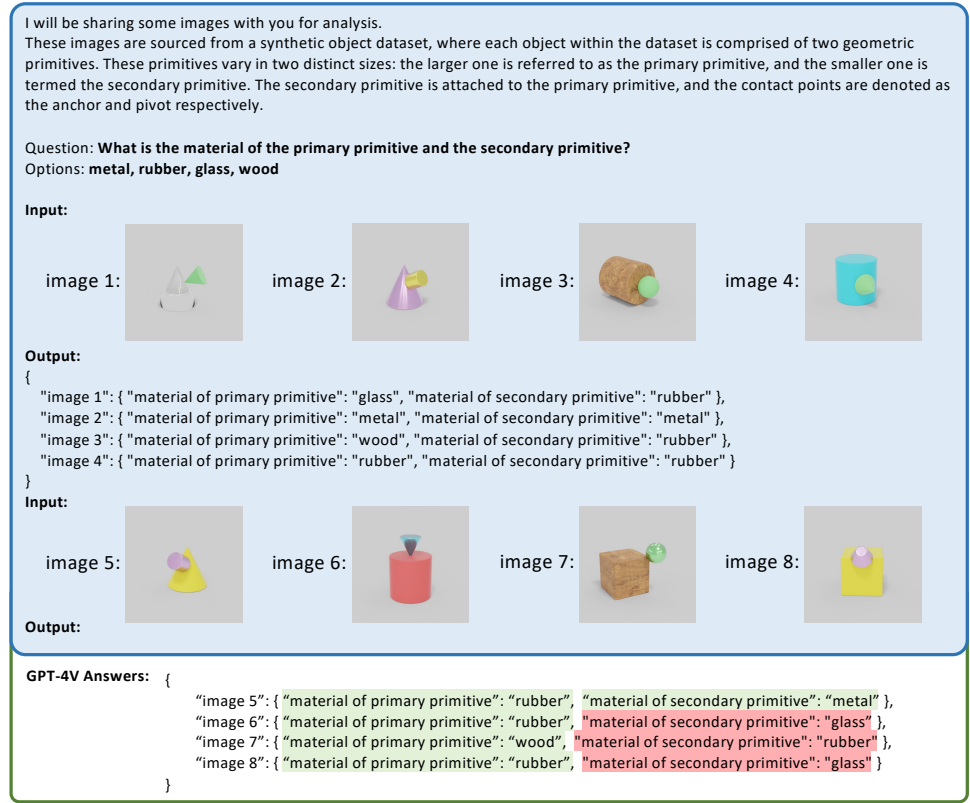


Figure 10: The results of GPT-4V recognizing material when providing in-context examples.

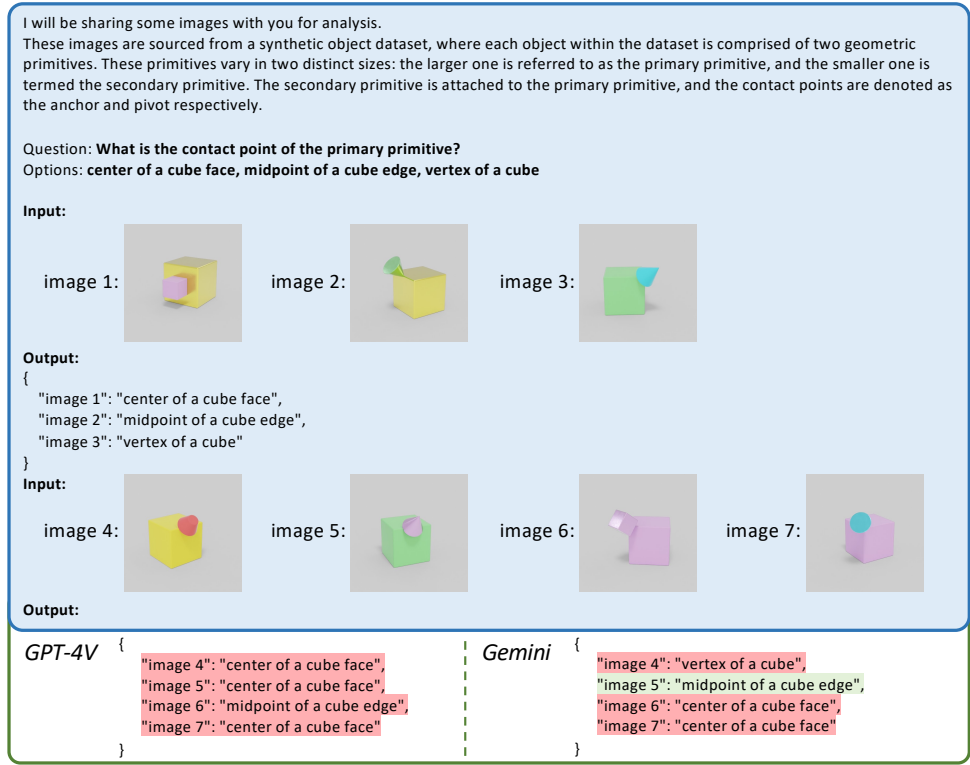


Figure 11: The results of GPT-4V and Gemini recognizing contact point when providing in-context examples.

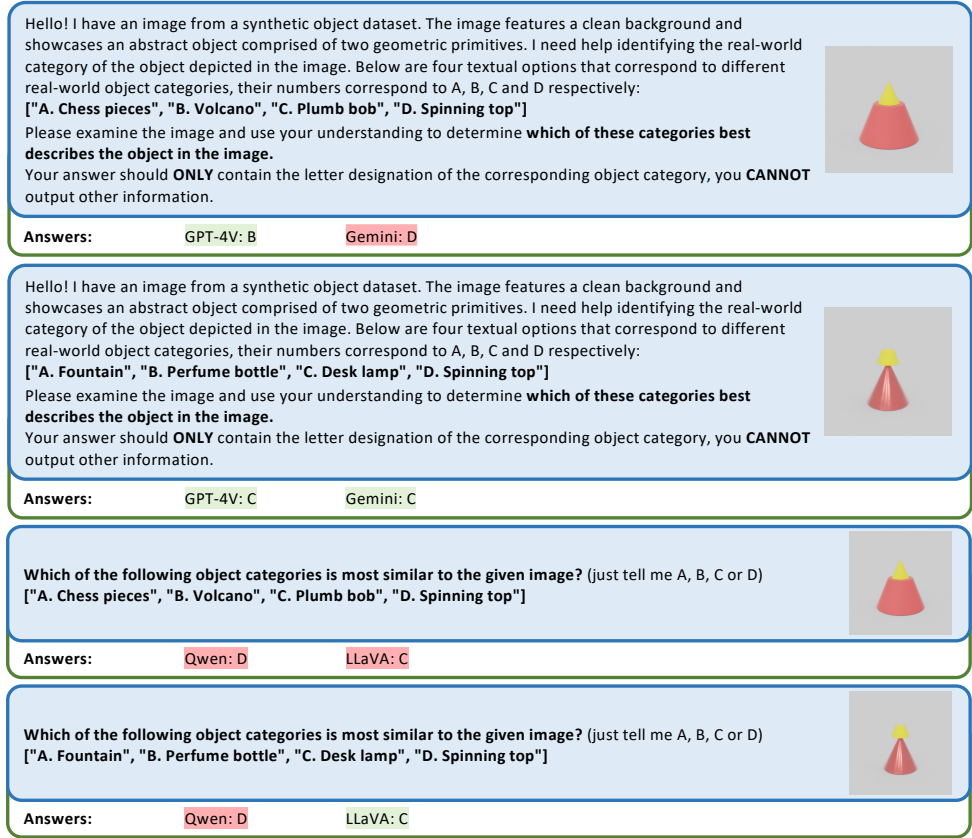



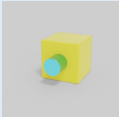
Figure 12: Prompts for *Img2Text* and outputs from all LMMs.


Hello! I have four images from a synthetic object dataset. Each image features a clean background and showcases an abstract object comprised of two geometric primitives. Your task is to find the image that best matches the real-world object category I give you. Please review each image carefully and determine **which one of the four contains the object that best matches the category "Bucket"**.


Options: A, B, C and D, which are correspond to four input images respectively.

Your answer should **ONLY** contain the letter designation of the corresponding image, you **CANNOT** output other information.

A. 

B. 

C. 


D. 


Answers: GPT-4V: A Gemini: C


Hello! I have four images from a synthetic object dataset. Each image features a clean background and showcases an abstract object comprised of two geometric primitives. Your task is to find the image that best matches the real-world object category I give you. Please review each image carefully and determine **which one of the four contains the object that best matches the category "Perfume bottle"**.

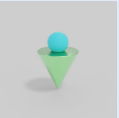
Options: A, B, C and D, which are correspond to four input images respectively.

Your answer should **ONLY** contain the letter designation of the corresponding image, you **CANNOT** output other information.

A. 

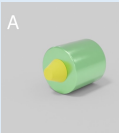
B. 

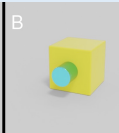
C. 

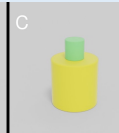
D. 

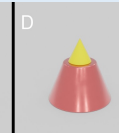
Answers: GPT-4V: A Gemini: C

Which object best matches the category "Bucket"? (Just tell me A, B, C or D)

A. 


B. 

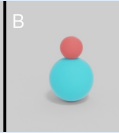
C. 

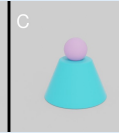
D. 

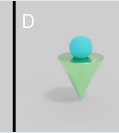
Answers: Qwen: A LLaVA: C

Which object best matches the category "Perfume bottle"? (Just tell me A, B, C or D)

A. 

B. 

C. 

D. 

Answers: Qwen: A LLaVA: B

Figure 13: Prompts for *Text2Img* and outputs from all LMMs.

I will be sharing some images with you for analysis. These images are sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

I will provide you with 3 images from the synthetic dataset described above. The four objects in image 1 belong to Category A, and the four objects in image 2 belong to Category B. The third image contains a single object that needs to be classified.


Based on the characteristics you observe in the sample images for each category, your task is to analyze the object in the third image and determine its classification.

Question: Which category does the object in the third image belong to?


Options: A or B (indicating category A and category B, respectively)

Please respond with the option identifier only.


Category A:



Category B:



Query:



Answers:

GPT-4V: A

Gemini: A

I will be sharing some images with you for analysis. These images are sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

I will provide you with 3 images from the synthetic dataset described above. The four objects in image 1 belong to Category A, and the four objects in image 2 belong to Category B. The third image contains a single object that needs to be classified.


Based on the characteristics you observe in the sample images for each category, your task is to analyze the object in the third image and determine its classification.

Question: Which category does the object in the third image belong to?


Options: A or B (indicating category A and category B, respectively)

Please respond with the option identifier only.


Category A:



Category B:



Query:



Answers:

GPT-4V: A

Gemini: B

Figure 14: Examples of the **Category Building** task, using standard prompts, showcasing outputs from GPT-4V and Gemini.

Stage 1

I have given you an image, this image is sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive, and the contact points are denoted as the anchor and pivot respectively. Also, different materials and colors are assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

For the image I gave you, I would like you to answer three specific questions. These questions are multiple-choice, and I will provide descriptions for each.

Your task is to review the image and give your answers accordingly.

Here are the three questions.

Question 1: What are the shapes of primary and secondary primitives?

Option of Question 1: A (cube), B (sphere), C (cylinder), D (cone), E (cone frustum)

Question 2: What are the contact points of primary and secondary primitives?

Option of Question 2 (anchor):

If you choose the primary primitive as cube, the options of this question is: A-1 (center of a cube face), A-2 (midpoint of a cube edge), A-3 (vertex of a cube)

If you choose the primary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the primary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's face), C-2 (midpoint of a cylinder's generatrix)

If you choose the primary primitive as cone, the options of this question is: D-1 (apex of cone), D-2 (midpoint of the cone's slant height), D-3 (center of the cone's base)

If you choose the primary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's top base), E-2 (midpoint of the cone frustum's slant height), E-3 (center of the cone frustum's bottom base)

Option of Question 2 (pivot):

If you choose the secondary primitive as cube, the options of this question is: A-1 (center of a cube face)

If you choose the secondary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the secondary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's base)

If you choose the secondary primitive as cone, the options of this question is: D-1 (center of the cone's base), D-2 (apex of cone)

If you choose the secondary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's bottom base)

Question 3: What are the materials and colors of primary and secondary primitives?

Option of Question 3 (material): A (rubber), B (metal), C (glass), D (wood)

Option of Question 3 (color): A (red), B (yellow), C (blue), D (green), E (purple), F (NA)

Requirement: You should answer questions based on the characteristics of the object. And your answer should strictly in the format below, you should use the letter designation as the answer, and you are not allowed to output anything else.

("shape of primary primitive": "", "shape of secondary primitive": "", "anchor of primary primitive": "", "pivot of secondary primitive": "", "material of primary primitive": "", "color of primary primitive": "", "material of secondary primitive": "", "color of secondary primitive": "")

Results from LLaVA:






Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 
("shape of primary primitive": "C", "shape of secondary primitive": "D", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "D-1", "material of primary primitive": "B", "color of primary primitive": "D", "material of secondary primitive": "B", "color of secondary primitive": "E")	("shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "C", "color of primary primitive": "C", "material of secondary primitive": "C", "color of secondary primitive": "C")	("shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "C", "color of primary primitive": "D", "material of secondary primitive": "C", "color of secondary primitive": "B")	("shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "C", "color of primary primitive": "D", "material of secondary primitive": "C", "color of secondary primitive": "E")
Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 
("shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "B", "color of primary primitive": "Yellow", "material of secondary primitive": "B", "color of secondary primitive": "Purple")	("shape of primary primitive": "C", "shape of secondary primitive": "D", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "D-1", "material of primary primitive": "A", "color of primary primitive": "E", "material of secondary primitive": "A", "color of secondary primitive": "E")	("shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "A", "color of primary primitive": "E", "material of secondary primitive": "A", "color of secondary primitive": "B")	("shape of primary primitive": "D", "shape of secondary primitive": "A", "anchor of primary primitive": "D-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "A", "color of primary primitive": "E", "material of secondary primitive": "A", "color of secondary primitive": "E")
Stage 1 prompt + query image 		("shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "D", "material of secondary primitive": "A", "color of secondary primitive": "B")	

Figure 15: An example of the **Category Building** task. In Stage 1, LLaVA answers eight questions regarding pattern perception for each image.

Stage 2

I will be sharing an image with you for analysis. This image is sourced from a synthetic object dataset. Each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints, resulting in the final image dataset.

In addition to the object in the image, I also took out 8 objects from this dataset, which are labeled 1 to 8 respectively. It is now known that objects 1 to 4 belong to category A, objects 5 to 8 belong to category B, and the object in the image is labeled 9. The attribute information of these 9 objects is as follows.

```
{'object_id': 'object 1', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cone', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': "center of the cone's base", 'material of primary primitive': 'metal', 'color of primary primitive': 'green', 'material of secondary primitive': 'metal', 'color of secondary primitive': 'purple'},
{'object_id': 'object 2', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'glass', 'color of primary primitive': 'blue', 'material of secondary primitive': 'glass', 'color of secondary primitive': 'blue'},
{'object_id': 'object 3', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'glass', 'color of primary primitive': 'green', 'material of secondary primitive': 'glass', 'color of secondary primitive': 'yellow'},
{'object_id': 'object 4', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'glass', 'color of primary primitive': 'green', 'material of secondary primitive': 'glass', 'color of secondary primitive': 'purple'},
{'object_id': 'object 5', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'metal', 'color of primary primitive': 'Yellow', 'material of secondary primitive': 'metal', 'color of secondary primitive': 'Purple'},
{'object_id': 'object 6', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cone', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': "center of the cone's base", 'material of primary primitive': 'rubber', 'color of primary primitive': 'purple', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'purple'},
{'object_id': 'object 7', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'rubber', 'color of primary primitive': 'purple', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'yellow'},
{'object_id': 'object 8', 'shape of primary primitive': 'cone', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': 'apex of cone', 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'rubber', 'color of primary primitive': 'purple', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'purple'},
{'object_id': 'object 9', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'green', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'yellow'}
```

Now you need to summarize the characteristics of categories A and B and classify object 9 based on the image and the information I gave you.

Question: **Which category does object 9 belong to?**

Options: A or B (indicating category A and category B, respectively)

Please respond with the **option identifier only**.

+ query image



Answers:

LLaVA: B

Figure 16: An example of the **Category Building** task. In Stage 2, the analyses of all images are concatenated into the prompt, assisting LLaVA in classifying the query image.

Stage 1

I have given you an image, this image is sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive, and the contact points are denoted as the anchor and pivot respectively. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

For the image I gave you, I would like you to answer three specific questions. These questions are multiple-choice, and I will provide descriptions for each. Your task is to review the image and give your answers accordingly.
Here are the three questions.

Question 1: What are the shapes of primary and secondary primitives?

Option of Question 1: A (cube), B (sphere), C (cylinder), D (cone), E (cone frustum)

Question 2: What are the contact points of primary and secondary primitives?

Option of Question 2 (anchor):

If you choose the primary primitive as cube, the options of this question is: A-1 (center of a cube face), A-2 (midpoint of a cube edge), A-3 (vertex of a cube)
If you choose the primary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the primary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's base), C-2 (midpoint of a cylinder's generatrix)

If you choose the primary primitive as cone, the options of this question is: D-1 (apex of cone), D-2 (midpoint of the cone's slant height), D-3 (center of the cone's base)

If you choose the primary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's top base), E-2 (midpoint of the cone frustum's slant height), E-3 (center of the cone frustum's bottom base)

Option of Question 2 (pivot):

If you choose the secondary primitive as cube, the options of this question is: A-1 (center of a cube face)

If you choose the secondary primitive as sphere, the options of this question is: B-1 (point on the sphere)

If you choose the secondary primitive as cylinder, the options of this question is: C-1 (center of a cylinder's base)

If you choose the secondary primitive as cone, the options of this question is: D-1 (center of the cone's base), D-2 (apex of cone)

If you choose the secondary primitive as cone frustum, the options of this question is: E-1 (center of the cone frustum's bottom base)

Question 3: What are the materials and colors of primary and secondary primitives?

Option of Question 3 (material): A (rubber), B (metal), C (glass), D (wood)

Option of Question 3 (color): A (red), B (yellow), C (blue), D (green), E (purple), F (NA)

Requirement: You should answer questions based on the characteristics of the object. And your answer should strictly in the format below, you should use the letter designation as the answer, and you are not allowed to output anything else.

{"shape of primary primitive": "", "shape of secondary primitive": "", "anchor of primary primitive": "", "pivot of secondary primitive": "", "material of primary primitive": "", "color of primary primitive": "", "material of secondary primitive": "", "color of secondary primitive": ""}

Results from LLaVA:

Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 
{"shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "B", "color of secondary primitive": "C"}	{"shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "B", "color of secondary primitive": "B"}	{"shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "A", "color of secondary primitive": "D"}	{"shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "C", "color of primary primitive": "D", "material of secondary primitive": "B", "color of secondary primitive": "C"}
Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 	Stage 1 prompt + 
{"shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "A", "color of secondary primitive": "D"}	{"shape of primary primitive": "C", "shape of secondary primitive": "B", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "C", "color of primary primitive": "C", "material of secondary primitive": "C", "color of secondary primitive": "B"}	{"shape of primary primitive": "A", "shape of secondary primitive": "B", "anchor of primary primitive": "A-1", "pivot of secondary primitive": "B-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "B", "color of secondary primitive": "C"}	{"shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "A", "color of secondary primitive": "D"}
Stage 1 prompt + query image 		{"shape of primary primitive": "C", "shape of secondary primitive": "A", "anchor of primary primitive": "C-1", "pivot of secondary primitive": "A-1", "material of primary primitive": "B", "color of primary primitive": "C", "material of secondary primitive": "A", "color of secondary primitive": "D"}	

Figure 17: An example of the **Category Building** task. In Stage 1, LLaVA answers eight questions regarding pattern perception for each image.

Stage 2

I will be sharing an image with you for analysis. This image is sourced from a synthetic object dataset. Each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints, resulting in the final image dataset.

In addition to the object in the image, I also took out 8 objects from this dataset, which are labeled 1 to 8 respectively. It is now known that objects 1 to 4 belong to category A, objects 5 to 8 belong to category B, and the object in the image is labeled 9. The attribute information of these 9 objects is as follows.

```

{object_id: 'object 1', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'metal', 'color of secondary primitive': 'blue'},
{object_id: 'object 2', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'metal', 'color of secondary primitive': 'yellow'},
{object_id: 'object 3', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'green'},
{object_id: 'object 4', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'glass', 'color of primary primitive': 'blue', 'material of secondary primitive': 'glass', 'color of secondary primitive': 'blue'},
{object_id: 'object 5', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'green'},
{object_id: 'object 6', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'glass', 'color of primary primitive': 'blue', 'material of secondary primitive': 'glass', 'color of secondary primitive': 'yellow'},
{object_id: 'object 7', 'shape of primary primitive': 'cube', 'shape of secondary primitive': 'sphere', 'anchor of primary primitive': 'center of a cube face', 'pivot of secondary primitive': 'point on the sphere', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'metal', 'color of secondary primitive': 'blue'},
{object_id: 'object 8', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'green'},
{object_id: 'object 9', 'shape of primary primitive': 'cylinder', 'shape of secondary primitive': 'cube', 'anchor of primary primitive': "center of a cylinder's base", 'pivot of secondary primitive': 'center of a cube face', 'material of primary primitive': 'metal', 'color of primary primitive': 'blue', 'material of secondary primitive': 'rubber', 'color of secondary primitive': 'green'}

```


Now you need to summarize the characteristics of categories A and B and classify object 9 based on the image and the information I gave you.

Question: **Which category does object 9 belong to?**

Options: A or B (indicating category A and category B, respectively)

Please respond with the **option identifier only**.

+ query image



Answers:

LLaVA: B

Figure 18: An example of the **Category Building** task. In Stage 2, the analyses of all images are concatenated into the prompt, assisting LLaVA in classifying the query image.

I will be sharing some images with you for analysis. These images are sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

I will provide you with 3 images from the synthetic dataset described above. The four objects in image 1 belong to Category A, and the four objects in image 2 belong to Category B. The third image contains a single object that needs to be classified. Based on the characteristics you observe in the sample images for each category, your task is to analyze the object in the third image and determine its classification.

z

Question: Which category does the object in the third image belong to?

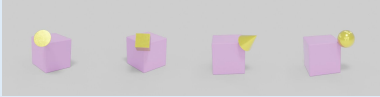
Options: A or B (indicating category A and category B, respectively)

Instructions:

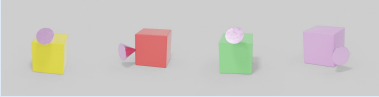
- Review image 1 and image 2 to understand the distinguishing features of Categories A and B.
- Analyze the third image, comparing its features to those observed in the previous images.
- Classify the object in the third image based on your analysis.

Please think step by step.


Category A:



Category B:



Query:



GPT-4V:

Step 1: Review of Image 1 and Image 2

Image 1 (Category A) Description:

- All objects consist of a conical primary primitive and an inverted conical secondary primitive.
- The primary cones are consistently in pastel shades: pale pink, red, purple, and yellow.
- The secondary cones are all a consistent bright yellow.

Image 2 (Category B) Description:

- Similar to Category A, all objects feature a conical primary primitive and an inverted conical secondary primitive.
- However, the primary cones differ in their coloration, they are in bright yellow, red, violet, and pastel green.
- The secondary cones are in a consistent soft green.

Step 2: Analyze the third image

Third Image Description:

- The object features a primary cone in pastel purple and a secondary cone in soft green.

Step 3: Classification

Comparison:

- The secondary primitive color (soft green) in the third image matches the secondary primitive color observed in Category B images.
- The appearance and configuration of the object in terms of primitive shapes (cone with inverted cone on top) are consistent across both categories, leaving the color as the distinguishing characteristic.

Conclusion:

- Based on the color similarity of the secondary primitive, the object in the third image belongs to Category B.
- To classify the object in the third image, I will first analyze the distinguishing features of objects in Categories A and B using the first two images provided:

1. Analyzing image 1 (Category A):
 - Objects consistently feature a primary primitive that is a solid, consistent color (light purple).
 - The secondary primitives are a different color (yellow) and are consistent in color across all objects.
 - Geometric shapes of the secondary primitives vary (circle, square, triangle, sphere).
 - The color contrast between the primary and the secondary primitive is noticeable but follows a consistent scheme.
2. Analyzing image 2 (Category B):
 - Objects each have primary primitives of different colors.
 - Secondary primitives are also of various colors but tend to blend more subtly with the primary primitive or are the same color.
 - The secondary primitives' shapes vary as in Category A.
 - The overall color scheme is more varied compared to Category A.
3. Analyzing the third image:
 - The primary primitive is a cube with a solid green color.
 - The secondary primitive is a cone with a light purple color.
 - The contrast and color scheme follow a pattern similar to the objects in Category A, where there is a stark color contrast between the primary and secondary primitives.

Based on these observations:

- The green primary primitive and light purple secondary primitive in the third image create a clear contrast, following the trend seen in Category A.
- The object's composition uses color contrast similarly to how all objects in Category A are structured, rather than the blending or matching scheme observed in Category B.

Conclusion: The object in the third image belongs to **Category A**.

Figure 19: An example from the task **Category Building** demonstrates GPT-4V’s analytical process and classification result under the Chain of Thought (CoT) setting.

I will be sharing some images with you for analysis. These images are sourced from a synthetic object dataset, where each object within the dataset is comprised of two geometric primitives. These primitives vary in two distinct sizes: the larger one is referred to as the primary primitive, and the smaller one is termed the secondary primitive. The secondary primitive is attached to the primary primitive. Also, different materials and colors were assigned to the geometric primitives. Each object is then rendered from random viewpoints (this means that you need to take into account the rotation of the combined object when you answer the following questions), resulting in the final image dataset.

I will provide you with 3 images from the synthetic dataset described above. The four objects in image 1 belong to Category A, and the four objects in image 2 belong to Category B. The third image contains a single object that needs to be classified. Based on the characteristics you observe in the sample images for each category, your task is to analyze the object in the third image and determine its classification.

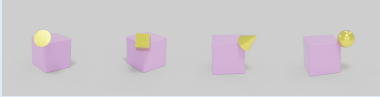
Question: Which category does the object in the third image belong to?
Options: A or B (indicating category A and category B, respectively)

Instructions:

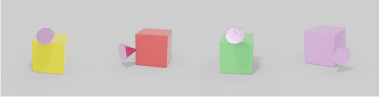
1. Review image 1 and image 2 to understand the distinguishing features of Categories A and B.
2. Analyze the third image, comparing its features to those observed in the previous images.
3. Classify the object in the third image based on your analysis.

Please think step by step.


Category A:



Category B:



Query:



Gemini:
Step-by-Step Analysis of Object Classification

1. Reviewing Image 1 (Category A):

- Primary Primitive: All objects have a pink cube as the primary primitive.
- Secondary Primitive: The secondary primitives vary in shape, including a flat circle, a flat square, a cone, and a sphere. They all share a gold material.

2. Reviewing Image 2 (Category B):

- Primary Primitive: The primary primitives vary in color (yellow, red, green, and pink) and shape, including cubes and a rectangular prism.
- Secondary Primitive: All objects have a flat, circular secondary primitive with a consistent lavender material.

3. Analyzing Image 3:

- Primary Primitive: The object has a green cube as the primary primitive.
- Secondary Primitive: The secondary primitive is a cone with a silver material.

4. Classification:

- The object in the third image shares the characteristic of a green cube primary primitive with one object in Category B. However, the secondary primitive is a cone, which is not observed in any of the Category B objects.
- The silver material of the cone is also closer to the gold material observed in Category A than the lavender material in Category B.

Conclusion:
Based on the analysis, the object in the third image is more likely to belong to **Category A**.

Figure 20: An example from the task **Category Building** demonstrates Gemini’s analytical process and classification result under the Chain of Thought (CoT) setting.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- [2] Blender Online Community. Blender: a 3d modelling and rendering package. <http://www.blender.org>, 2024.
- [3] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, 2017.
- [4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024.
- [5] OpenAI. GPT-4V(ision) technical work and authors. <https://openai.com/contributions/gpt-4v/>, 2023.
- [6] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- [7] Wenhao Wu, Huanjin Yao, Mengxi Zhang, Yuxin Song, Wanli Ouyang, and Jingdong Wang. GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? *arXiv preprint arXiv:2311.15732*, 2024.
- [8] 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). *arXiv preprint arXiv:2309.17421*, 2023.