

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}

jjialin.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

Abstract

Categorization, a core cognitive ability in humans that organizes objects based on common features, is essential to cognitive science as well as computer vision. To evaluate the categorization ability of visual AI models, various proxy tasks on recognition from datasets to open world scenarios have been proposed. Recent development of Large Multimodal Models (LMMs) has demonstrated impressive results in high-level visual tasks, such as visual question answering, video temporal reasoning, etc., utilizing the advanced architectures and large-scale multimodal instruction tuning. Previous researchers have developed holistic benchmarks to measure the high-level visual capability of LMMs, but there is still a lack of pure and in-depth quantitative evaluation of the most fundamental categorization ability. According to the research on human cognitive process, categorization can be seen as including two parts: category learning and category use. Inspired by this, we propose a novel, challenging, and efficient benchmark based on composite blocks, called **ComBo**, which provides a disentangled evaluation framework and covers the entire categorization process from learning to use. By analyzing the results of multiple evaluation tasks, we find that although LMMs exhibit acceptable generalization ability in learning new categories, there are still gaps compared to humans in many ways, such as fine-grained perception of spatial relationship and abstract category understanding. Through the study of categorization, we can provide inspiration for the further development of LMMs in terms of interpretability and generalization.

1 Introduction

Categorization is one of the most fundamental cognitive abilities of humans. As shown in Fig. 1, visual categorization involves the process of organizing objects into categories based on shared features or attributes (*category learning*), and using the mental representation to

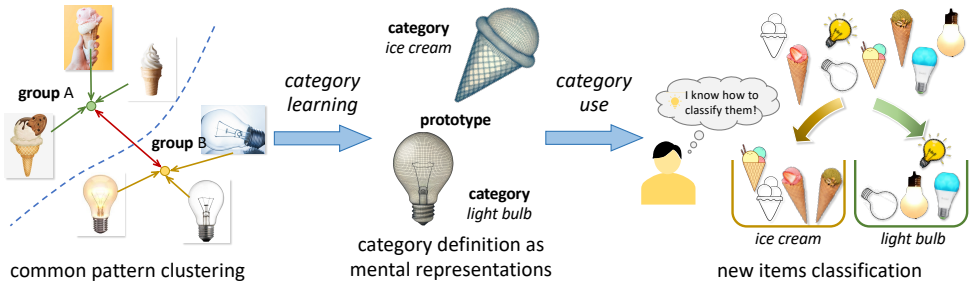


Figure 1: Human behavior in categorization. People can group objects together based on common patterns, form mental representation of categories, and classify novel items.

complete cognitive tasks, such as classifying new objects (*category use*) [26]. The learning and use of categories is not only a significant research topic in cognitive science but is also considered a critical feature of artificial intelligence [1].

With the progressive enhancement of computer vision models, there should be an implicit improvement in the capability for categorization, evaluated by the development of diverse proxy tasks, such as object recognition. In recent years, multimodal models like CLIP [50] align visual and textual modalities, thereby liberating categorization from the constraint of datasets and advancing towards open-world scenarios. Moreover, Large Multimodal Models (LMMs) have integrated modalities such as vision into language models with a large number of parameters, displaying remarkable performance on numerous high-level visual tasks [54, 48] and holistic benchmarks [51, 52]. Excellent understanding shown in image captioning [6] and visual question answering [13, 16, 25] implies that LMMs seem to possess sufficient categorization ability [4]. However, there is a lack of direct, objective, and decoupled evaluations of LMMs’ capabilities in the most fundamental tasks of visual perception.

To explore this question, a pure and in-depth benchmark is required to dissect categorization ability of LMMs. We argue that an effective benchmark should have the following characteristics: (1) **Avoiding data leakage.** Prevent not only data sample leakage [2, 24] but also leakage of evaluation categories. Similar to using abstract reasoning to test human intelligence in Wechsler Adult Intelligence Scale [41], some abstract and novel categories that are impossible to exist in the training set should be introduced. (2) **Establishing quantitative and discriminative tasks.** Select diverse and quantifiable evaluation tasks and questions to ensure objectivity and maximize dissection efficiency, allowing us to explore the boundaries of their capability through failure case analysis [2]. (3) **Performing unit tests and integration tests.** Design a diverse set of evaluation tasks that cover the entire cognitive process of category learning and use. These tasks should avoid unrestrained end-to-end questions to ensure a disentangled evaluation. To respond to these requirements, we will revisit the human cognitive process of categorization [26, 52, 59], and attempt to decouple the key evaluation points to probe the categorization process of LMMs, which will be detailed in Sec.2.

To meet the aforementioned requirements of categorization evaluation, we construct a synthetic dataset consisting of **Composite Blocks (ComBo)**, which will be elaborated in Sec.3. The objects and categories in ComBo are entirely unseen to LMMs, meeting the need of preventing data leakage. Since the synthetic data is completely controllable, we can easily control the difficulty of the tasks and inexpensively generate a large number of questions with ground truth for quantitative evaluation. Inspired by the analysis of the categorization process, we design a series of tasks covering the entire cognitive process, aiming to compre-

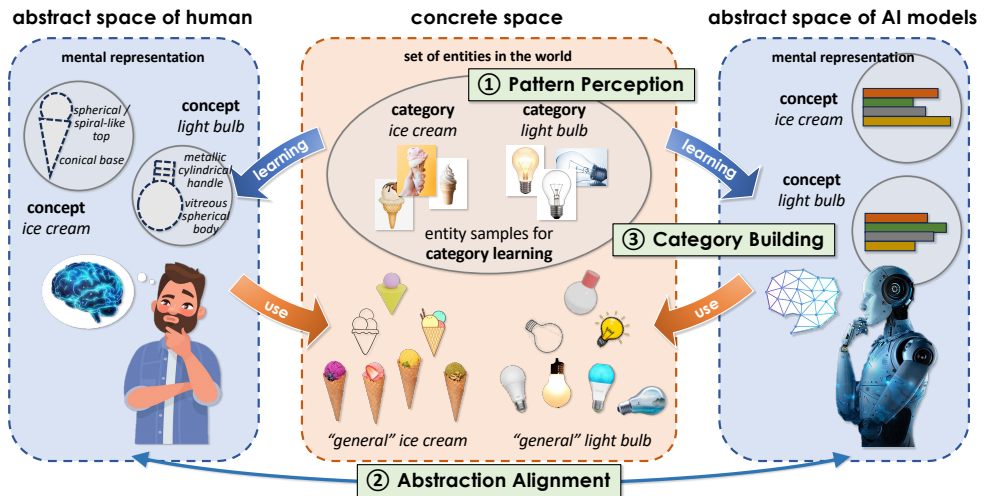


Figure 2: The cognitive processes of humans and LMMs in categorization. Categorization can be modeled as a process of category learning and category use between concrete and abstract spaces. The proposed evaluation tasks are shown in green blocks.

hensively evaluate the categorization capability of LMMs. First, we evaluate LMMs’ ability to perceive low-level patterns that is critical for accurate object recognition. Next, we explore their capability to align abstract category representations with human mental concepts by predefined semantic categories, which verifies the consistency of the learned concepts. Finally, we challenge the models with unseen abstract categories to examine their generalizability of categorization ability. These experiments are designed to illuminate the strengths and limitations of LMMs in replicating human-like category cognition, thereby pushing the boundaries of LMMs in understanding and interacting with the real-world objects.

The experimental results and analyses in Sec.4 reveal that while LMMs demonstrate enhanced categorization capability over traditional CV models, they continue to be stuck in spatial detail recognition, abstract conceptual reasoning, and learning unseen categories in some scenarios. The corresponding discussions and related work are presented in Sec.4.3 and Sec.5. Analyzing these failure cases allow us to explore LMMs’ shortcomings from a more basic level and make effective promotions. We believe that studying the low-level visual capability such as categorization of LMMs will contribute to the further development of generalizability and interpretability in AI models.

2 Categorization Dissection

In this section, we will break down the design of categorization dissection, starting with an introduction to the cognitive processes of categorization of humans and LMMs in Sec.2.1, and based on this, we will introduce the design philosophy of our evaluation in Sec.2.2.

2.1 Overview of Categorization

As shown in Fig.2, the cognitive process of categorization involves information transmission between *concrete space* and *abstract space* [26]. The concrete space consists of perceivable

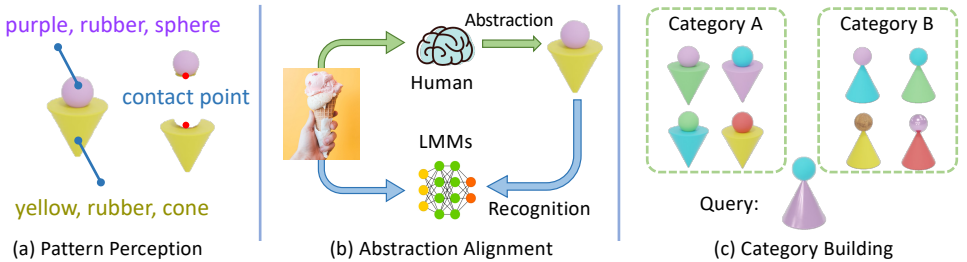


Figure 3: Three progressive tasks on categorization evaluation. (a) **Pattern Perception:** Evaluating LMMs’ low-level pattern recognition ability. (b) **Abstraction Alignment:** Comparing the category abstract representations between humans and LMMs. (c) **Category Building:** Examining LMMs’ categorization ability on abstract unseen categories.

visual entities in the real world, including various data forms of the object categories, such as a photorealistic ice cream, an ice cream sketch, and toy blocks like an ice cream. The abstract space is where both humans and LMMs store categorization rules about these categories respectively, such as the shape of the category “ice cream” (typically consisting of a cone and ice cream balls) and some attributes (a cold dessert). Humans use mental representations to encode key aspects about category members [76], while LMMs store knowledge about entities in their internal implicit representation spaces, such as feature vectors.

Thus, the cognitive process of categorization can be represented as follows. (1) *category learning*: humans and LMMs perceive data in concrete space, gathering some items with common features together. They then abstract and summarize the commonalities of these items to form a concept representation of the category in abstract space. (2) *category use*: humans and LMMs utilize the concepts in abstract space to construct various cognitive functions. For example, classifying a newly encountered object as ice cream, inferring the cold taste and other attributes a new ice cream should have, and even implicitly applying it to tasks such as image captioning and visual question answering about a dessert shop.

2.2 Design Philosophy of Evaluation

Based on the cognitive process of object categorization described above, we design three evaluation tasks (green blocks in Fig.2) corresponding to different stages of the categorization process, in order to conduct a comprehensive evaluation of the categorization capability of LMMs.

Pattern Perception: pre-CL evaluation. When perceiving entities in the concrete space, patterns are the direct perceptual targets for humans and LMMs. The ability to accurately identify low-level patterns is a prerequisite for category learning (pre-CL). As shown in Fig.3 (a), we evaluate the ability of LMMs to recognize patterns in multiple dimensions, such as shape, material, color, etc., in a fully disentangled manner.

Abstraction Alignment: post-CL evaluation. Alignment between the abstract spaces is one of the important topics in eXplainability AI (XAI) research [4]. We further explore whether LMMs’ learned representations of category learning (post-CL) are aligned with human mental representations. As shown in Fig.3 (b), LMMs are asked to recognize abstract visual stimuli agreed upon by humans and align them with the correct semantic labels.

Category Building: full-chain evaluation. As shown in Fig.3 (c), to examine the categorization capability of LMMs from learning to use, we define several groups of abstract



Figure 4: Overview of **Composite Blocks (ComBo)** dataset: exemplar images and attributes. Each object can be represented by four-dimension fully-disentangled attributes as shape, color, material, and contact point between the primary primitive and the secondary primitive.

unseen categories in ComBo, requiring LMMs to observe exemplar objects, induce the definitions of two categories and classification boundaries, and finally classify new objects. This task is considered complex as it involves both perception and cognition, closely resembling many laboratory tasks designed for human participants.

3 The ComBo Benchmark

3.1 Overview of ComBo

We construct a large-scale repository of **Composite Blocks** for categorization (ComBo), where each object within the dataset is composed of two geometric primitives, named primary primitive and secondary primitive according to the size of the primitives. The primary and secondary primitives are contacted through a contact point on the primary primitive. The optional shapes of the primitives and the optional contact points on the primary primitive are all displayed in Fig.4. To enhance the visual diversity of ComBo, four different materials are assigned to the primitives. Additionally, the rubber and metal materials are further differentiated by five colors.

By enumerating all the values across the four disentangled dimensions of shape, material, color, and contact point, a total of 9,504 objects can be obtained, with each pair of objects differing in at least one dimension. We utilize a ray tracing based rendering engine [9] to render each composite object from 20 random viewpoints, culminating in the 190,080 images in ComBo, inspired by CLEVR [10]. More details are shown in supplementary materials.

The benchmark content and evaluation results are publicly available at: <https://fubin29.github.io/Blocks-as-Probes/>.

3.2 Tasks

As mentioned in Sec.2.2, to evaluate LMMs’ categorization capability, we start with three tasks: **Pattern Perception**, **Abstraction Alignment**, and **Category Building**. Examples for three tasks and the corresponding answers by LMMs are illustrated in Fig.5, and some statistics about the benchmark are shown in Tab.1.

Pattern Perception. We randomly select 5,000 objects from ComBo as evaluation subjects, and sample one rendered image for each object. Participants are required to sequentially answer seven questions about the low-level patterns present in the object, as shown in Tab.2. All questions are multiple-choice, and a brief description of ComBo along with all the options are provided.

Abstraction Alignment. In this task, we invite cognitive science experts to select appropriate natural categories that can be abstracted by our ComBo objects. Following a fil-

Pattern Perception

Question: What is the shape of the primary primitive?
Options: A. Cube; B. Sphere; C. Cylinder; D. Cone; E. Conical frustum
 ...

Question: What is the color of the secondary primitive?
Options: A. Red; B. Yellow; C. Blue; D. Green; E. Purple
 ...

GPT Shape of the primary primitive: E ...
 Color of the secondary primitive: D ...

Gemini Shape of the primary primitive: E ...
 Color of the secondary primitive: D ...

Abstraction Alignment

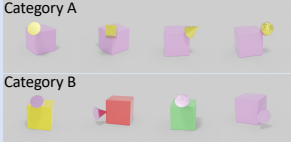
Question: Which option that best matches the object in the image?
Options: A. Light bulb; B. Hourglass; C. Plumb bob; D. Fountain

GPT A

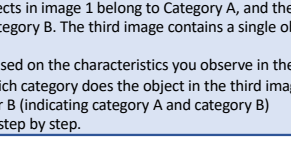
Gemini B

Category Building


Category A



Category B



Query



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.

Question: Based on the characteristics you observe in the sample images for each category, which category does the object in the third image belong to?
Options: A or B (indicating category A and category B)
 Please think step by step.

GPT: Analyzing image 1: Objects consistently feature a primary primitive that is a consistent color (light purple). The secondary primitives are a **different color**. ... **Analyzing image 2:** The secondary primitives' **shapes vary** as in Category A. ... **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. ... **Conclusion:** The object belongs to **Category A**.

Gemini: Reviewing Image 1: All objects have a pink cube as the primary primitive. ... **Reviewing Image 2:** The primary primitives vary in color and **shape**. ... All objects have a **flat, circular secondary primitive** with a consistent lavender material. ... **Classification:** However, the secondary primitive is a cone, which is **not observed** in any of the Category B objects. the third image is more likely to belong to **Category A**.

Figure 5: Examples of the QA pairs for three evaluation tasks. Due to space constraints, prompts and answers are abbreviated. Refer to supplementary materials for details.

Task Name	Evaluation Goal	# Questions	Question Type	Metric
Pattern Perception	pre-CL	5000	multiple-choice	accuracy
Abstraction Alignment	post-CL	240	multiple-choice	accuracy
Category Building	full-chain	400	multiple-choice	accuracy
		400	CoT	manual scoring

Table 1: Summary of three evaluation tasks. For multiple-choice questions, we use the accuracy of correct options as evaluation metric. For answers by Chain-of-Thought (CoT), they are manually scored by experts to obtain accuracy, process score, and error attribution.

tering and voting process, a consensus is reached on 24 categories, which include image samples and category labels, for evaluation purposes. Subsequently, we generate two types of multiple-choice questions, each comprising 120 questions, by incorporating distractors among the matched abstract objects and category labels. *Img2Text* requires the participants to choose the label that best matches the given image out of four category labels. *Text2Img* asks the participants to select the image that most resembles the given category from four image options. The detailed process of question generation is described in supplementary materials, and the validity of the questions is verified by user study discussed in Sec.4.2.

Category Building. 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”). We present multiple samples from two abstract categories to the participants, requiring them to observe and summarize the rules for building both abstract categories. Furthermore, we randomly show test samples belonging to the two categories to the participants multiple times. Participants should be able to classify all samples correctly when they understand the categories. We also design a similarity measurement method to calculate the evaluation task difficulty for classifying composite objects

	Primary Primitive			Secondary Primitive			Contact Point
	Shape	Material	Color	Shape	Material	Color	
LLaVA	55.7	50.7	40.6	48.6	38.1	29.5	47.2
Qwen	70.7	70.2	65.8	37.2	43.1	25.4	43.8
GPT-4V	89.4	75.8	94.8	64.6	68.4	87.9	43.7
Gemini	95.9	96.0	99.5	79.6	86.6	94.8	64.9

Table 2: Pattern Perception Results (%). For GPT-4V and Gemini, all seven low-level patterns are queried simultaneously in a single question, whereas each pattern is addressed in a separate question for LLaVA and Qwen.

within abstract categories. See supplementary materials for more details.

4 Experiments

4.1 Evaluation Settings

In this study, we select the mainstream closed-source implementations of current LMMs (GPT-4V [28, 47], Gemini-1.5-Pro [30]) and open-source implementations (LLaVA-v1.5-13B [20], Qwen-VL-Chat [3]) as the subjects of our analyses. To evaluate GPT-4V and Gemini, we utilize their official APIs. For LLaVA and Qwen, we conduct local tests using a single NVIDIA A40 GPU. Additionally, we include other implementations as comparative references in different experiments, such as representation classification based on CLIP [30] pre-trained models, evaluations from human users, etc. More details and examples about image input and prompt are presented in supplementary materials.

4.2 Evaluation Results

Pattern Perception. Tab.2 demonstrates the pattern perception capability of various LMMs, without any fine-tuning or in-context prompting. Gemini and GPT-4V exhibit significantly stronger low-level pattern recognition and instruction-following capability, compared to open-source LMMs (even in simpler separate questions). Notably, Gemini and GPT-4V generally achieve the accuracy of larger than 90% in recognizing the primary primitive’s shape and colors of both primitives. We also find that the recognition of patterns in smaller secondary primitives presents greater challenges, resulting in performance declines across all the LMMs. Overall, Gemini achieves the best results in all metrics, and especially excels in predicting the contact points, indicating its advanced spatial perception capability.

Considering the domain transfer challenges posed by the ComBo dataset, we conduct additional in-context learning experiments [10, 5, 38] on GPT-4V and Gemini, focusing on contact points and materials of both primitives. As shown in Fig.6, the results indicate that GPT-4V can significantly improve its performance through in-context learning. We speculate that Gemini’s performance advantage over GPT-4V might stem from Gemini’s exposure to similar block data during training and specialized training on spatial relationships.

Abstraction Alignment. We invite 20 human participants to complete the user study to validate the reasonableness of our questions. The results indicate that the abstract objects in the images bear a good resemblance to the mental representation of natural categories held by humans. Tab.3 presents the alignment between the category concepts learned by LMMs and

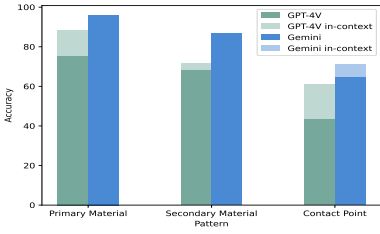


Figure 6: Comparison of Accuracy (%) of Pattern Perception Task: GPT-4V and Gemini with and without in-context examples.

	<i>Img2Text</i>	<i>Text2Img</i>	Overall
CLIP	47.5	75.8	61.7
LLaVA	55.0	29.2	42.1
Qwen	48.3	32.5	40.4
GPT-4V	55.0	76.7	65.9
Gemini	52.5	74.2	63.4
Human	80.8 ± 9.5	94.4 ± 5.2	87.6 ± 10.2

Table 3: Abstraction Alignment Results (%). Specifically, by calculating the similarity between CLIP features of the query and the options, we provide CLIP’s alignment results.

human mental representations for natural categories. For *Img2Text*, which involves matching one image to four semantic labels, all the LMMs exhibit similar performance with an accuracy rate of around 50%. For *Text2Img*, matching one semantic label to four images, GPT-4V and Gemini outperform the open-source LMMs. However, considering both types of questions, the abstract reasoning ability of LMMs still falls short of humans.

Additionally, two conclusions can be drawn from the experimental results: (1) The poor performance of open-source LMMs on *Text2Img* stems from the need to reason based on multiple input images, a relative weakness for these models compared to others like GPT-4V. (2) *Img2Text* is harder than *Text2Img*. In the semantic space, a single image corresponds to a relatively definite feature representation, while the feature representation for a label is actually the centroid of a group of similar images’ features. Therefore, matching images to multiple semantic labels in the question introduces greater uncertainty.

Category Building. In this task, we invite 8 human participants to assess whether the proposed similarity measurement method correlates with human cognition and whether it is applicable for evaluating the classification difficulty of abstract categories. Based on this similarity measurement, we design four difficulty levels: easy, medium, hard, and expert. We also invite another 23 human participants to complete this task, providing a human reference score (the red line in Fig.7) for comparative analysis. Fig.7 illustrates the complete categorization capability of different LMMs across varying difficulties.¹ As the difficulty increases, represented by the diminishing differences between two categories, the challenge for LMMs to construct accurate category representations and classify query objects also escalates, and all LMMs demonstrate varying degrees of performance decline. In contrast, human participants maintain high classification accuracy across different difficulty levels. Even in the expert level, human participants can perceive the increase in difficulty but still manage to cope effortlessly. However, when the differences between two categories become minimal, and they must rely on the shape of secondary primitives or the combination of two primitives for distinction, the weaknesses of LMMs in pattern perception are further magnified in this experiment.

We further employ the Chain-of-Thought (CoT) approach [9, 42]. Both GPT-4V and Gemini possess robust reasoning ability; they can effectively perform category building and application when objects are accurately perceived. The performance differences displayed by the two models in more difficult questions primarily stem from Gemini’s advantage in low-level pattern recognition. From Fig.7, requiring LMMs to explicitly output their decision-

¹Qwen failed to produce a valid output in this task, hence there is no corresponding entry in Fig.7

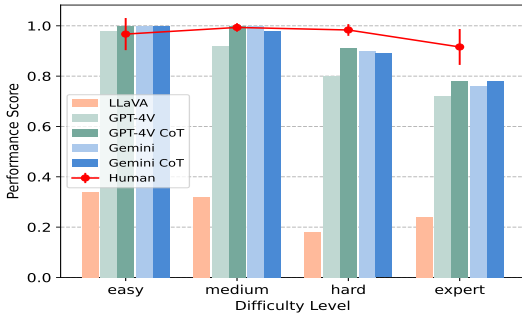


Figure 7: Category Building Results. GPT-4V and Gemini both demonstrate strong performance on this task, with GPT-4V achieving greater improvements when employing the CoT technique.

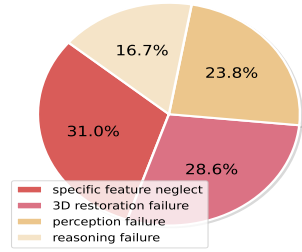


Figure 8: Error Analysis of LMMs Using CoT Technology: distribution of error types in LMMs' decision-making processes.

making processes in CoT format indeed enhances their performance in this task. Moreover, CoT technology helps us understand the causes of misclassification, as shown in Fig.8.

4.3 Discussion

4.3.1 Prompt Design and Instruction Following

During the evaluation process, we find significant differences in the instruction following [29] capability among different LMMs. GPT-4V and Gemini can understand the most complete description of the questions, and can also receive multiple image inputs, allowing them to handle more complex problems that require reasoning between multiple images. In pattern perception evaluation, querying different patterns separately can improve the performance of LLaVA. Qwen's capability to follow complex instructions is slightly inferior to other LMMs, requiring more adjustments to the form of the questions.

4.3.2 Task Difficulty

In this section, we discuss the **difficulty** of our benchmark. Although ComBo, as a dataset rendered based on geometric primitives, defines problems that are formally similar to abstract reasoning problems [12] in other benchmarks, it is much less difficult than those benchmarks in terms of logical reasoning. To verify this, we use some smaller, commonly used computer vision models to complete the same tasks. By retrieval, models pre-trained on ImageNet-1k [83] can achieve similar performance to GPT-4V, and almost are able to completely solve the task after fine-tuning. See supplementary materials for more details. The fine-tuned small model is capable of completing the task, implying that it does not necessitate a complex reasoning process. However, current LMMs still show a significant gap compared to humans in such simple visual tasks, indicating that LMMs are **far from being able to claim** that the fundamental tasks of visual categorization **have been completely solved**.

4.3.3 Limitation and Future Work

We believe that evaluating both the lower and upper bounds of LMMs' capabilities is equally important. Compared to other comprehensive evaluation benchmarks, our evaluation bench-

mark leans more towards the in-depth evaluation of **categorization capability**, which is considered one of the most fundamental visual cognitive abilities. Notably, we utilize brand new synthetic data rather than real 2D images to completely **prevent data leakage** and facilitate **decoupled, controllable** evaluation. However, it is still necessary to use more complex, even real images to further evaluate the capabilities of LMMs in real-world application scenarios. In future work, we will involve more complex composite objects and controllable 3D models, and incorporate a wider array of cognitive tasks to further explore the current conclusion. Moreover, we are committed to developing new methodologies and datasets that enhance LMMs’ performance in perceiving spatial details, reasoning about abstract concepts, and learning new categories. We believe these efforts will enhance the applicability and reliability of LMMs in various high-level tasks.

5 Related Work

Large Multimodal Models. LMMs [2, 9, 49] integrate visual [22] or other modalities [13, 43] into Large Language Models (LLMs), enabling them to handle a variety of multimodal tasks. High-performance and closed-source LMMs like PaLM-E [10], GPT-4V [47], and Gemini [65] represent a critical branch of development. These models benefit from substantial investments in proprietary datasets and computing resources, achieving superior performance across a range of complex tasks. Another branch consists of open-source models such as LLaMA-Adapter [63], LLaVA [22], MiniGPT-4 [64], Otter [18], and Qwen [8]. These LMMs are typically developed by modularly integrating other modalities into open-source LLMs [67]. Both branches have demonstrated strong capability in various applications, such as medical image understanding [20, 27] and embodied agents [40, 46]. Consequently, we select two models from each branch for evaluation in our study.

LMM Benchmarks. Due to the more generalized multimodal perception and reasoning capability of LMMs, traditional vision-language benchmarks are inadequate for providing a comprehensive and sufficient evaluation. Consequently, recent developments in the evaluation of LMMs have primarily focused on several key aspects [49]: (1) addressing specific common issues such as visual shortcomings [66] and hallucinations [8, 24]; (2) comprehensive benchmarks that entail complex tasks and diverse capability [19, 23, 45, 60, 61]; (3) expert-level domain knowledge and advanced reasoning [62]. In contrast to these evaluation efforts, our study concentrates on assessing the fundamental categorization ability of LMMs.

6 Conclusion

In this work, we introduce the ComBo benchmark, focusing on evaluating the categorization capability of Large Multimodal Models (LMMs). Inspired by research on categorization in cognitive science, we design three evaluation tasks from different perspectives, comprehensively assessing the LMMs’ ability in pattern perception, abstract concept alignment, and generalization of categorization. The evaluation results reveal that LMMs still exhibit deficiencies in spatial detail perception, abstract concept reasoning, and learning of new categories. Although in-context learning or Chain-of-Thought (CoT) techniques can further improve the performance of LMMs, there remains a gap compared to human categorization capability, providing recommendations for future improvements in LMMs.

Acknowledgement

This work is partially supported by National Key R&D Program of China No. 2021ZD0111901, 2023YFF1105104, and Natural Science Foundation of China under contract No. U21B2025. Bin Fu and Qiyang Wan contributed equally to this work. Bin Fu was primarily responsible for the construction of the dataset and also participated in the design and implementation of the evaluation benchmark. Qiyang Wan mainly provided theoretical formulation in categorization within cognitive science and similarly contributed to the design and implementation of the evaluation benchmark.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Miłkoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 23716–23736, 2022.
- [2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*, 2023.
- [3] 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.
- [4] Blender Online Community. Blender: a 3d modelling and rendering package. <http://www.blender.org>, 2024.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, 2020.
- [6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015.
- [7] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, 2020.

- [8] Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges. *arXiv preprint arXiv:2311.03287*, 2023.
- [9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 49250–49267, 2023.
- [10] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 8469–8488, 2023.
- [11] Li Fei-Fei and Ranjay Krishna. Searching for Computer Vision North Stars. *Daedalus*, 151(2):85–99, 2022.
- [12] Chaoyou Fu, Renrui Zhang, Zihan Wang, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, Yunhang Shen, Mengdan Zhang, Peixian Chen, Sirui Zhao, Shaohui Lin, Deqiang Jiang, Di Yin, Peng Gao, Ke Li, Hongsheng Li, and Xing Sun. A Challenger to GPT-4V? Early Explorations of Gemini in Visual Expertise. *arXiv preprint arXiv:2312.12436*, 2023.
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.
- [14] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. *arXiv preprint arXiv:2310.14566*, 2024.
- [15] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-Bind & Point-LLM: Aligning Point Cloud with Multi-modality for 3D Understanding, Generation, and Instruction Following. *arXiv preprint arXiv:2309.00615*, 2023.
- [16] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019.
- [17] 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.
- [18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: Multi-Modal In-Context Instruction Tuning. *arXiv preprint arXiv:2306.05425*, 2023.
- [19] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024.
- [20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 28541–28564, 2023.
- [21] 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.
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 34892–34916, 2024.
- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281*, 2024.
- [24] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, 2024.
- [25] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204, 2019.
- [26] Arthur B Markman and Brian H Ross. Category Use and Category Learning. *Psychological Bulletin*, 129(4):592–613, 2003.
- [27] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: a Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225, pages 353–367, 2023.
- [28] OpenAI. GPT-4V(ision) technical work and authors. <https://openai.com/contributions/gpt-4v/>, 2023.

- [29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 27730–27744, 2022.
- [30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- [31] 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.
- [32] Eleanor Rosch. Principles of Categorization. In *Cognition and Categorization*, pages 27–48. Routledge, 1978.
- [33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
- [34] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326, 2019.
- [35] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*, 2023.
- [36] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, 2024.
- [37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.
- [38] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 200–212, 2021.

- [39] Amos Tversky. Features of Similarity. *Psychological Review*, 84(4):327–352, 1977.
- [40] Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-Agent: Autonomous Multi-Modal Mobile Device Agent with Visual Perception. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.
- [41] David Wechsler. *Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV)*. Psychological Corporation, San Antonio, TX, 2008.
- [42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.
- [43] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [44] 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.
- [45] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. LVLm-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *arXiv preprint arXiv:2306.09265*, 2023.
- [46] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Hao-ran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied Vision-Language Programmer from Environmental Feedback. *arXiv preprint arXiv:2310.08588*, 2023.
- [47] 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.
- [48] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint arXiv:2304.14178*, 2024.
- [49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*, 2024.
- [50] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, LEI BAI, Jing Shao, and Wanli Ouyang. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 26650–26685, 2023.

-
- [51] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- [52] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024.
- [53] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*, 2023.
- [54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The 12th International Conference on Learning Representations (ICLR)*, 2024.