

Multimodal Large Language Model Meets New Knowledge: A Preliminary Study

Junwen Mo Jiaxuan Li Duc Minh Vo Hideki Nakayama

The University of Tokyo

{mo, li, vmduc}@nlab.ci.i.u-tokyo.ac.jp nakayama@ci.i.u-tokyo.ac.jp

Abstract

Multimodal Large Language Models (MLLMs) have achieved impressive performance on established image understanding benchmarks. However, these benchmarks typically include images of objects that are already known, potentially not fully testing MLLMs' ability to understand unfamiliar objects. To address this, we assess the performance of MLLMs on images featuring synthesized novel objects. We use ChatGPT to create descriptions of novel objects by merging characteristics of existing objects and then employ a text-to-image generation model to generate synthesized objects. Using this dataset, we evaluate MLLMs in identifying and describing the elements of novel objects. Experiment results show that state-of-the-art MLLMs struggle to comprehensively understand images containing novel objects, often leading to hallucinated descriptions.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance in natural language generation and understanding [1, 2]. To advance the capabilities of intelligent systems, several works [3, 4, 5, 6] have extended LLMs to Multimodal Large Language Models (MLLMs) by incorporating visual elements. Although these MLLMs have achieved significant advancements in established benchmarks for image captioning and Visual Question Answering (VQA), their proficiency in recognizing and understanding novel entities in the open world remains unexplored. Moreover, existing benchmarks are not suitable anymore since MLLMs are trained on large amounts of data that include known objects. Consequently, developing more challenging benchmarks featuring novel objects is crucial for a thorough evaluation of MLLMs' capabilities.



Figure 1 A fruit that blends the characteristics of strawberries, kiwis, and pineapples.

In everyday life, we often come across unfamiliar objects. When describing these new objects, we typically refer to characteristics such as shape and color. To better understand and remember these novel objects, we instinctively draw comparisons to known objects, using our existing knowledge base instead of relying solely on the object's attributes. This necessitates the ability to understand known objects comprehensively.

Inspired by the aforementioned ideas, we construct a new image dataset representing fantasy objects that do not exist in reality. These novel objects are concocted by blending attributes of two or three known objects, utilizing image generation models, in order to evaluate whether current MLLM can identify that the presented objects are not real and can associate them to known concepts. Figure 1 provides an example.

Our contributions are as follows:

- We introduce a challenging new benchmark for MLLMs, comprised of novel objects that do not exist in the real world.
- We evaluate state-of-the-art MLLMs using our new benchmark and observe that these models struggle with novel objects, highlighting the importance of further developing MLLMs for open-world understanding.

2 Related Work

Multimodal Large Language Models Due to the difficulty in obtaining image-text pairs, most MLLMs are developed using existing visual models, such as CLIP [7], and language foundation models, like LLaMA [2] and Vicuna [8], rather than being trained from scratch. BLIP-2 [3] proposes a Querying Transformer (Q-former) to fuse vision and language module, which has been adopted by other works, including MiniGPT-4 [6]. For the development of instruction-following MLLMs, LLaVA [5] develops a multimodal instruction-following data, which is widely used by subsequent research [6, 4]. For our experiments, we test three well-known instruction-tuned MLLMs: LLaVA-1.5, MiniGPT-4, and InstructBLIP.

Traditional Benchmark Traditional Image captioning and VQA benchmarks contains large amounts of diverse images, providing robust evaluation for developing MLLMs. Image captioning requires models to recognize the information in an image and describe it accurately. One of the most famous image captioning dataset is MSCOCO [9]. VQA tasks require models to well understand the information in the image and sometimes need reasoning on the visual cues. A lot of VQA datasets are constructed to cover diverse scenarios, including COCO-QA [10], VQAv2 [11] and Visual Genome [12]. However, a limitation of these traditional benchmarks is that most images within these datasets contain existing knowledge, such as known objects and common sense, which may has already been included in the training data for MLLMs.

Benchmark with Novel Elements The WHOOPS benchmark introduces images that defy common sense, with the objective of evaluating whether the generation by MLLMs exhibits bias towards common sense and assesses their reasoning ability to identify unusual parts. Additionally, the ISEKAI dataset [13], featuring images with fantasy objects, aims to assess the few-shot learning capabilities of MLLMs. To assess the ability of LLMs when meeting new knowledge, KnowGen [14] is proposed. It is a knowledge generation method by altering entity attributes and relationships to create entities not existing in the world. A Question Answering (QA) benchmark named ALCUNA is constructed using KnowGen. Our work is inspired by KnowGen. The main distinction lies in our focus on MLLMs. Compared to the ISEKAI dataset, we aim to

synthesize novel entities with a more complex mixture of objects.

3 Image Synthesis

In this section, we introduce how to design new objects and synthesize the corresponding images.

3.1 Prompt construction

In our preliminary study, we employed a straightforward procedure to synthesize images. Our approach involved prompting ChatGPT to design novel objects through the combination of existing entities, with a specific focus on animals and fruits. It is instructed to provide detailed descriptions of combining these objects to generate novel ones. This process yielded 100 novel animal descriptions and 50 fruit descriptions.

```
###User: Generate visually rich prompts for Text-to-Image models, envisioning novel fruits by amalgamating physical attributes of diverse fruits. Ensure the following:
```

1. Keep the prompts non-threatening.
2. Avoid duplicating existing examples.
3. Cultivate creativity in the examples you generate.
4. PLEASE eliminate redundancy and background details unrelated to amalgamating appearance features, like taste and impression. The content should include shape, surface, and flesh.
5. Steer clear of repetition within your examples.
6. Aim for maximum diversity in the generated creature designs.
7. Each created fruit combines the characteristics of 2 or 3 kinds of fruits.

```
Example:
```

```
Melonkiwipear: a pear-shaped fruit that combines the fuzzy skin of a kiwi with the juicy, red, transparent flesh of a watermelon.
```

```
###ChatGPT:
```

3.2 Data construction

In the second step, we prompt the text-to-image model to synthesize images, utilizing OpenDalleV1.1 [15]. Due to the potential mismatch between the descriptions generated by ChatGPT and the requirements of the image generation model, we employed Promptist [16] to generate refined prompts. Both the original and refined descriptions were fed into the model, resulting in the generation of 9 images for each description.

Finally, to ensure the quality of images for our preliminary experiment, we manually selected 84 satisfactory images based on our judgment of high quality.

4 Experiment

4.1 Implementation

We employed beam search for inference in our evaluations. For each model, the beam width is set to 5, and the length penalty is set to 1. The implementation is based on the HuggingFace Transformers library.

4.2 Compared methods

We conducted tests on three well-known MLLMs: LLaVA-1.5-7b [5], InstructBLIP [4], and MiniGPT-4 [6], using our selected images.

4.3 Evaluation metrics

The evaluation focused on two main aspects:

Entity Identification. The first test aimed to assess the models' capability to identify whether the entities within the synthesized images were artificially generated.

Recognition of Unreal Features. We evaluated whether the models could recognize the source or origin of the features if we inform the models that the entities within the images were unreal.

To achieve these goals, we design 3 questions/instructions:

- 1) What is the {ENT} in this image?
- 2) Describe this image.
- 3) The {ENT} in the image is a combination of multiple {ENT}s. Please respond with only the names of the {ENT}s, without additional information. {ENT} will be replaced to be either "animal" or "fruit" depending on the image being evaluated.

4.4 Results

What is the {ENT} in this image? The objective of this question was to assess whether the models could identify that the entities in the images did not exist in the world. The responses are checked manually. We found that all models provided accurate but incomplete answers by correctly naming the main entity. For instance, if presented with an image featuring an elephant with eagle wings, the models will answer "elephant" without expressing doubt about the synthetic nature of the entity. It is observed that the responses were limited to naming the primary entity and did not explicitly acknowledge the artificial or unreal characteristics of the image. One plausible explanation for this limitation is the lack of diversity in the current instruction-tuning dataset, leading the models to overlook expressing doubts about the authenticity of the entities, as discussed in a recent work [17].

Among the analyzed images, responses for 52 out of 64 animal images were consistently aligned across the chosen models, whereas responses for 7 out of 20 fruit images were consistent. One of the inconsistent examples is shown in Table 1 (col 2). The result suggests that our synthesized images for fruits are more confusing for the models.

Image Captioning We prompt models with "Describe this image" for image captioning, aiming to assess their ability to comprehend synthesized entities and recognize unusual parts within them. Due to the lack of ground truth captions for synthesized images, automatic evaluation of this task is unfeasible. Upon observation, we noted a significant issue of hallucination in this test. Most responses do not mention the unusual parts present in the synthesized entities. Even in cases where a model recognized these distinctive features from different entities, it often generated incorrect captions that include hallucinations, shown in Table 1 (col 3). This




			
	Q: What is the fruit in this image?	Q: Describe this image.	Q: The animal in the image is a combination of multiple animals. Please respond with only the names of the animals, without additional information.
Components	apricot, passion fruit, citrus	parrot, leopard	cheetah, kangaroo
LLaVA-1.5-7b	The fruit in this image is a peach.	... a jungle scene with a green parrot perched on a tree branch	Cheetah
MiniGPT-4	The fruit in this image is an orange.	... a jaguar sitting on a tree branch with a green parrot perched on its back ...	The animal in the image is a cheetah.
InstructBLIP	The fruit in this image is an apricot.	... a leopard sitting on a tree branch, with a colorful parrot perched on its back ...	The animal in the image is a combination of a cheetah and a leopard.

Table 1 Col 2 shows a case where models have varying opinions on the fruit name. Col 3 demonstrates that the models successfully recognize elements of a parrot and a leopard, while they do not notice that these elements constitute a single entity in the captions. Col 4 showcases a component analysis. The features of the cheetah are obvious, while the features of a kangaroo are less distinct, with its characteristics mainly manifested in its movements and paws.

highlights a challenge in the models’ ability to accurately capture and describe the unique elements of the synthesized objects.

Component The final test aims to assess whether models could correctly identify the components of synthesized entities when explicitly informed that the entities depicted in the images were synthesized. The results for prediction accuracy are presented in Table 2. Notably, InstructBLIP displayed the highest recall and F1 score, while LLaVA-1.5-7b achieved the best precision. It is noteworthy that recalls for all models were close to 50%, indicating that one of the components could be correctly recognized in most cases. Table 1 (col 4) shows an example.

Model	Precision	Recall	F1
LLaVA-1.5-7b	0.7292	0.5327	0.5850
MiniGPT-4	0.6877	0.4990	0.5372
InstructBLIP	0.6607	0.6518	0.6495

Table 2 Average precision, recall and f1 score on the prediction of components

5 Conclusion

This work is a preliminary study on the ability of three well-known MLLMs on images with synthesized objects. We conducted experiments to assess the ability of MLLMs in understanding novel objects combined with familiar entities. Through the experiments, we found that 1) these MLLMs failed to express doubt about the presented questions. 2) The observed prevalence of hallucination indicates that these MLLMs can identify known elements but often lack a comprehensive understanding of the overall entity.

Limitation The dataset size in our experiments is relatively small, leading to a less convincing conclusion. In our future work, we aim to devise methods to expand both the size and diversity, enabling a more comprehensive analysis. Additionally, most evaluation processes are conducted manually, incurring both time costs and subjective judgments. To enhance the evaluation process, we are considering the utilization of well-developed LLMs, such as ChatGPT, which will be explored in our future experiments.

References

- [1] OpenAI. ChatGPT [large language model], 2024. <https://chat.openai.com>.
- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, Vol. abs/2302.13971, , 2023.
- [3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.
- [4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, Vol. abs/2305.06500, , 2023.
- [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *CoRR*, Vol. abs/2304.08485, , 2023.
- [6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, Vol. abs/2304.10592, , 2023.
- [7] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16399–16408. IEEE, 2022.
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- [9] 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.
- [10] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2953–2961, 2015.
- [11] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Evaluating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017.
- [12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, Vol. 123, No. 1, pp. 32–73, 2017.
- [13] Yan Tai, Weichen Fan, Zhao Zhang, Feng Zhu, Rui Zhao, and Ziwei Liu. Link-context learning for multimodal llms. *CoRR*, Vol. abs/2308.07891, , 2023.
- [14] Xunjian Yin, Baizhou Huang, and Xiaojun Wan. AL-CUNA: large language models meet new knowledge. In Houada Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 1397–1414. Association for Computational Linguistics, 2023.
- [15] <https://huggingface.co/dataautogpt3/OpenDalleV1.1>.
- [16] Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *CoRR*, Vol. abs/2212.09611, , 2022.
- [17] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, Vol. 1, , 2023.