

Detect, Describe, Discriminate: Moving Beyond VQA for MLLM Evaluation

Manu Gaur Darshan Singh S Makarand Tapaswi
CVIT, IIT Hyderabad

<https://katha-ai.github.io/projects/detect-describe-discriminate/>

VQA Evaluation

Q: The Wii remote is positioned on which side of the girl's body?
(a) Left (b) Right

<ul style="list-style-type: none">🌀 The Wii remote is positioned on the left side of the girl's body.🌟 Left: As the Wii remote is in the girl's left hand.🌟 The Wii remote is positioned on the right side of the girl's body.		<ul style="list-style-type: none">🌀 A young girl in a pink floral dress stands in front of closed white blinds, holding controller in both hands as she looks to her left with a serious expression.🌟 A young girl in a pink floral dress holds a Wii remote in her left hand, arm slightly bent, and gazes to her left.🌟 Child in pink floral dress holding white object with both hands, looking to the side with a serious expression.
<ul style="list-style-type: none">🌀 The Wii remote is positioned on the girl's right side.🌟 Right: The Wii remote is positioned on the right side of the girl's body.🌟 The Wii remote is positioned on the right side of the girl's body.		<ul style="list-style-type: none">🌀 A young girl in pink floral dress stands in front of closed white blinds, holding controller in one hand as she looks slightly upward with a focused expression.🌟 A young girl in a pink floral dress swings a Wii remote towards right with both hands and her brow furrowed in concentration.🌟 Child in pink floral dress holding white object with one hand, looking directly at camera with a raised chin.

D3 Evaluation (Ours)

Detect the visual difference and
Describe the image uniquely such that the caption
Discriminates it from the distractor.

Figure 1. When prompted with a question and/or multiple choices (VQA evaluation), MLLMs show middling performance on identifying fine-grained differences between an image pair. Harder still, is when MLLMs need to independently detect and describe such differences (our evaluation). Our work finds that state-of-the-art MLLMs struggle to discern fine-grained difference with our detect-describe-discriminate evaluation framework, with open-source MLLMs failing to outperform random guess. The text highlighted in green represents the fine-grained differences captured by the MLLMs, while that marked in red represents erroneous descriptions (hallucinations). Results are presented for GPT-4o, Gemini-1.5-Pro, and Claude-Sonnet-3.5.

Abstract

Visual Question Answering (VQA) with multiple choice questions enables a vision-centric evaluation of Multimodal Large Language Models (MLLMs). Although it reliably checks the existence of specific visual abilities, it is easier for the model to select an answer from multiple choices (VQA evaluation) than to generate the answer itself. In this work, we offer a novel perspective: we evaluate how well an MLLM understands a specific visual concept by its ability to uniquely describe two extremely similar images that differ only in the targeted visual concept. Specifically, we assess the ability of MLLMs to capture specific points of

visual differences using self-retrieval [12], i.e. by retrieving the target image using its generated caption against the other image in the pair serving as the distractor. We curate 247 highly similar image pairs as part of the D_3 benchmark. For each image pair, the model is prompted to: (1) **Detect** a specific visual difference, and (2) **Describe** the target image uniquely such that it (3) **Discriminates** the target image from the distractor. Self-retrieval within D_3 enables white-box evaluation across six different visual patterns, revealing that current models struggle to independently discern fine-grained visual differences, with open-source models failing to outperform random guess.

1. Introduction

Multimodal Large Language Models (MLLMs) exhibit impressive capabilities in multimodal tasks such as image understanding, visual question answering, and instruction following [11, 16, 21]. These models leverage the strong reasoning abilities of LLMs [15, 17, 27], with advancements in MLLMs driven primarily by scaling up the language models [7, 11]. The rapid progress in model capabilities necessitates the development of more stronger benchmarks that are however missing. Recent works such as Cambrian-1 [21], OpenEQA [14], and MMStar [2] reveal that current benchmarks [6, 13, 24] exhibit a strong language bias and fail to accurately assess the visual understanding capabilities of MLLMs. This motivates us to further explore the vision-centric evaluation of these models.

Since MLLMs are conversational agents, users can evaluate the visual understanding of these models through Visual Question Answering (VQA), by examining their natural language responses. However, reliably evaluating natural language responses requires commonsense reasoning and language comprehension. Although LLMs can be used for this purpose, they may be inaccurate and slow when parsing a large number of responses (*e.g.* LLaMA-3-8B takes 3.1 s to generate 100 tokens on a single A6000 GPU). To circumvent this, recent works [4, 14, 21, 22] frame fine-grained visual tasks as VQA, where the model is required to *select an answer* from the multiple options.

VQA with multiple choices, provides a reliable method for checking the existence of specific facets of visual understanding. For instance, we can assess the MLLM’s ability to capture the *state* of the object in Fig. 2a by directly asking it if the cat’s eyes are *open* or *closed*. However, we find that it is *easier for the model to select an answer from multiple choices than to generate the answer* itself (see Fig. 1). Specifically, providing the answer along with task prompt (through multiple choice options or as part of the question) biases the MLLM’s output towards the visual concept that is being evaluated. This raises questions about whether the model actually understands this visual concept or is simply picking the more likely choice.

In this work, we offer an alternative perspective for evaluating fine-grained understanding exhibited by an MLLM. While looking at a pair of images, we ask the model to describe the target image such that a listener can distinguish the target image from an extremely similar distractor [5, 9]. We curate extremely similar image pairs, each having one prominent point of visual difference such that solving this task entails that the model captures a specific facet of visual understanding. With 247 such image pairs, we introduce the D_3 benchmark (examples in Fig. 2). For each image pair, we prompt the model to: (1) **Detect** the visual difference, and (2) **Describe** the target image uniquely such that it (3) **Discriminates** the target image from the distractor.

Unlike VQA, we do not constrain the output to multiple-choice answers and directly evaluate the model’s free-form natural language generation. Specifically, we assess the ability of the MLLM to capture subtle visual distinctions using self-retrieval [5, 12], *i.e.* retrieving the target image based on its generated description against the distractor image. Although Gaur et al. [5] also use self-retrieval for fine-grained evaluation, the captioner does not have access to the distractor images. In contrast, we evaluate the MLLM’s ability to discern fine-grained visual differences by showing both images simultaneously.

Inspired by MMVP [22], we annotate each image pair with a specific point of difference (POD), highlighting the visual concept that distinguishes both images. D_3 consists of image pairs that can be distinguished primarily using one of the following PODs: state, position, scene, orientation/direction, camera, or clutter (Sec. 2.2). Consequently, self-retrieval can be used as a white-box evaluation to assess different facets of visual understanding captured by the MLLM (Sec. 3.4). For example, successful self-retrieval for the image pair in Fig. 2d requires the model to have a fine-grained understanding of the dog’s *orientation*.

Although CounterCurate [26] and Spot-the-Diff [8] also elicit fine-grained visual discrimination with image pairs, CounterCurate [26] uses VQA with synthetic images, while Spot-the-Diff [8] leverages nearby frames from video-surveillance footage that often lack clear semantic differences beyond *negation* and *object position*. Most similar to us, MMVP [22] evaluates the ability of MLLMs to discern fine-grained visual distinctions across image pairs through VQA. While both tasks require the model to identify a specific visual difference, in the case of MMVP, the visual concept under scrutiny is presented to the MLLM within multiple choice options or as part of the question itself (see Fig. 1). In contrast, D_3 requires the model to *independently detect* the visual difference and incorporate it into captions that uniquely describe each image, making it a more challenging task than VQA (see Appendix A.2).

Unlike VQA, which offers flexibility through the phrasing of questions, self-retrieval provides flexibility through manual curation of image pairs. Specifically, one can assess a model’s ability to discern specific aspects of visual discrimination by curating image pairs that differ only in the targeted visual concept. As a result, image pairs in D_3 are significantly more challenging to distinguish. We verify this by using Gemini-1.5-Pro to perform self-retrieval evaluation on image pairs from both datasets under the same experimental settings – the model scores 87.3% in MMVP, compared to a mere 39.7% on our D_3 benchmark.

We evaluate various open- and closed-source MLLMs on our benchmark (Sec. 3.3). We find that current models struggle in capturing fine-grained visual differences, with open-source models fail to outperform random guess.

2. D₃ Benchmark

We present how we curate image pairs in the D₃ benchmark followed by an explanation of each point of difference.

2.1. Curating Image Pairs

Our benchmark creation process requires densely captioned images. We source them in two ways: (1) ShareGPT4V [1] that comprises 100K images with dense captions generated using GPT-4V [16], and (2) HolisticCaps [5] that blends multiple human annotated COCO captions and combines them with dense visual descriptions from Instruct-BLIP [3]. Lengthy captions are summarized to 65 tokens using Llama-3-70B [15] so they do not exceed SigLIP [25] text encoder’s token capacity. We briefly summarize the image pair curation process (please refer to [5] for a more detailed description): (1) The images and their corresponding (summarized) captions are encoded and concatenated using SigLIP’s image and text encoder respectively. (2) Image pairs are identified based on highest similarity in the multi-modal embedding space ensuring that each pair encodes a unique visual concept. (3) Finally, we perform manual filtering to get 247 image pairs, with each image pair having one prominent point of difference.

2.2. Annotating Points of Difference (POD)

Each image pair is essentially identical except for one prominent visual difference. We annotate the differences among 6 visual concepts. The number of image pairs differentiated by each concept is mentioned in parentheses.

- **State** (72): relates to different states of the same object, *e.g.* the toilet seat is *up/down*. This also includes entities performing different actions, *e.g.*, in Fig. 2a, we see a cat *rests* with eyes *open/closed*.
- **Camera** (55): relates to the camera’s position or different properties such as *perspective*, *viewpoint*, *depth*, or *zoom*. Fig. 2b shows an example of a different viewpoint.
- **Position** (26): relates to different relative positions of objects in an image. *E.g.* the black cow is on the *left/right* of the brown cow in Fig. 2c. This also includes cases where a *single* object has different relative positioning with respect to the background.
- **Orientation/Direction** (63): relates to the entity or object facing a different direction or having dissimilar orientation with respect to the camera. Fig. 2d shows an example of a dog looking left or at the camera.
- **Scene** (18): relates to fine-grained difference in attributes or background characteristics that span the image or some area around an object of interest. *E.g.*, Fig. 2e shows two zebras standing in a *lush green/muddy* environment.
- **Clutter** (13): relates to a pair of images with extremely

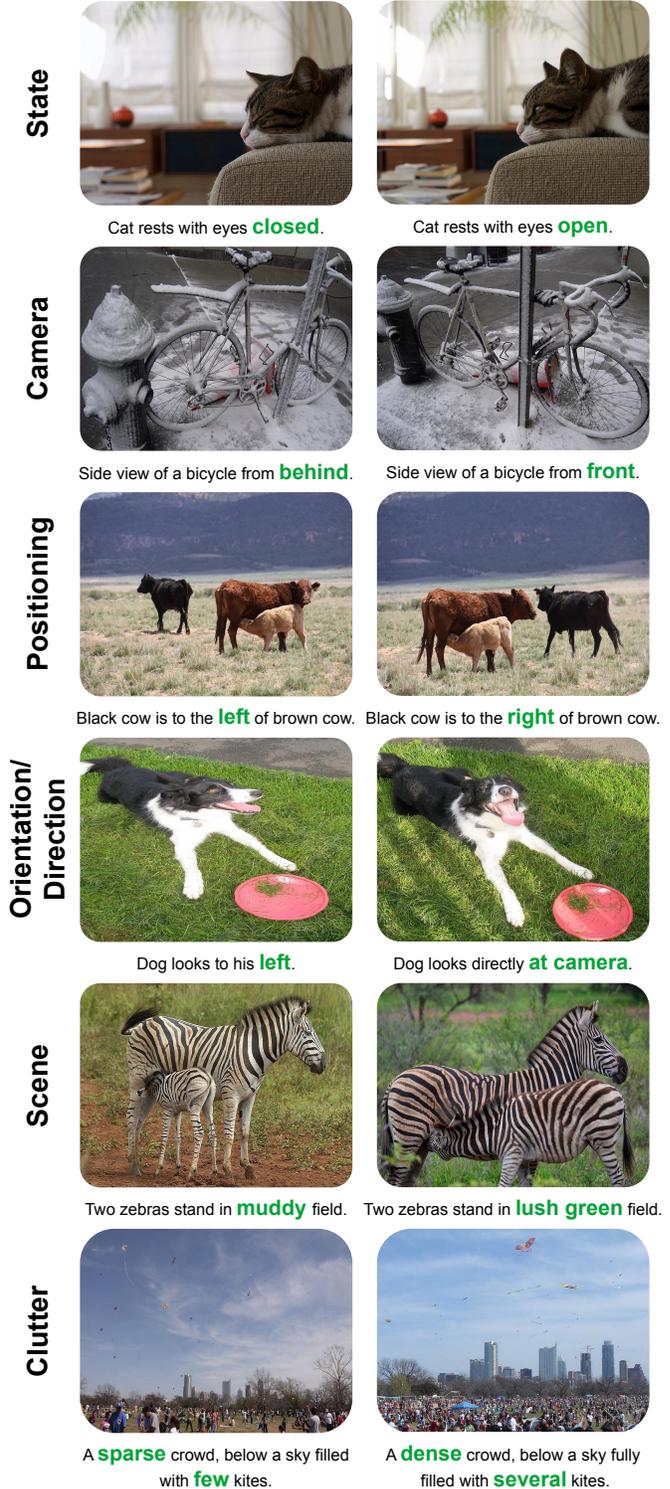


Figure 2. Each row illustrates one of the six Points of Difference (PODs) present in D₃ benchmark: (a) State, (b) Camera, (c) Positioning, (d) Orientation/Direction, (e) Scene, and (f) Clutter. We provide exemplar discriminative captions for each image, highlighting the fine-grained point of difference in **green**.

similar scenes, *e.g.* a fridge or a desk, cluttered with too many objects. The image pair often contains fine-grained differences in the characteristics of the non-prominent objects or scene. An example of this is the *dense/sparse* crowd of people in Fig. 2f.

Inter-annotator agreement is studied by randomly sampling 100 image pairs from the benchmark and asking two annotators to pick the most relevant POD based on the original instructions. After removing some invalid annotations, we obtain an agreement accuracy of 71.3% (67/94), indicating that the PODs are reasonably distinct.

3. Experiments

3.1. Self-Retrieval Setup

Given an image pair I_0 and I_1 , we prompt an MLLM to generate image captions C_0 and C_1 that allow a listener to distinguish between them. To evaluate the MLLM, we first encode both the images and the generated captions using siglip-so400m-patch14-384. Next, similar to Winoground [20], we compute the *self-retrieval score* that checks whether the scorer is able to pair the generated caption to the correct target image:

$$f(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } \text{sim}(C_0, I_0) > \text{sim}(C_0, I_1) \\ & \text{and } \text{sim}(C_1, I_1) > \text{sim}(C_1, I_0), \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where $\text{sim}(C_i, I_j)$ corresponds to the cosine similarity between the encoded representations of the caption and image respectively.

3.2. Prompting MLLMs

Both images are simultaneously given to the MLLM with the prompt shown in Fig. 3. Notably, since Cambrian [21] and LLaVA-NeXT [11] do not support multi-image prompting, we concatenate the image pair horizontally (without resizing). Extensive prompt tuning is done for open-source models such as Cambrian-34B, Chameleon-30B [19], and LLaVA-NeXT-34B to improve performance on the self-retrieval task (see Appendix A.1).

3.3. MLLMs struggle with fine-grained Differences

The performance of MLLMs on D_3 benchmark is shown in Fig. 4. MLLMs struggle to incorporate fine-grained visual details in their captions, as evidenced by their performance on the benchmark. State-of-the-Art (SotA) open-source models such as Cambrian-34B and LLaVA-NeXT-34B fail to even outperform random guess. Although, closed-source models outscore random guess, they still struggle to generate discriminant captions for images in D_3 , with GPT-4o achieving only 33.2%. We find Claude Sonnet 3.5 to be

Prompt: Identify fine-grained visual differences between both images and generate a discriminant caption for each image. Each caption should uniquely describe the image and highlight the distinct features that set the image apart from the other. Output in JSON format with 'left' and 'right' as keys, and their captions as string values.

Figure 3. The prompt given to GPT 4o, Claude Sonnet 3.5, Gemini Pro 1.5, and Gemini Flash 1.5 to uniquely describe images within our benchmark. Prompts for the open-source models are presented in Appendix A.1.

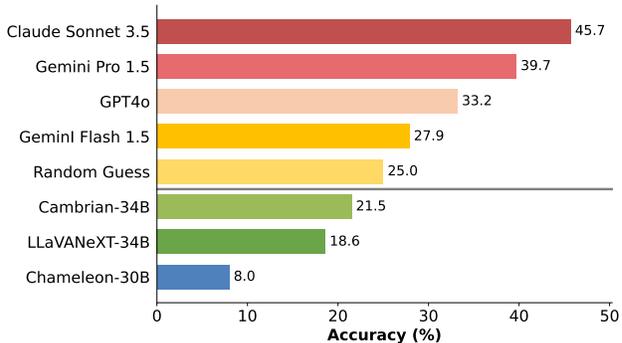


Figure 4. Benchmarking different open- and closed-source MLLMs on D_3 benchmark. The scores are averaged across all image pairs.

most capable in discerning fine-grained visual differences, achieving the highest score of 45.7% on our benchmark.

3.4. Whitebox Evaluation

Each image pair within D_3 contains a prominent visual concept or point of difference that. Generating discriminative descriptions for both images requires the caption to incorporate the targeted visual difference. When an MLLM fails to uniquely describe and retrieve both images within a pair in D_3 , we are able to accurately identify the visual concept that the MLLM is unable to pick up (see Appendix A.3 for qualitative examples).

Fig 5 illustrates the self-retrieval performance of various MLLMs on image pairs of D_3 across all six points of difference. Trends indicate that SOTA MLLMs, both open and closed, struggle to perceive fine-grained changes in orientation/direction, camera angle, object’s state, or positioning. These findings are similar to those established in MMVP [22]. In contrast, uniquely describing similar images with differing scenes appears to be easier for these models. This may be because differences in the scene often span the entire image which is easier for current models to identify. V* [23] finds that MLLMs struggle to focus on fine-grained details in visually crowded images. Inter-

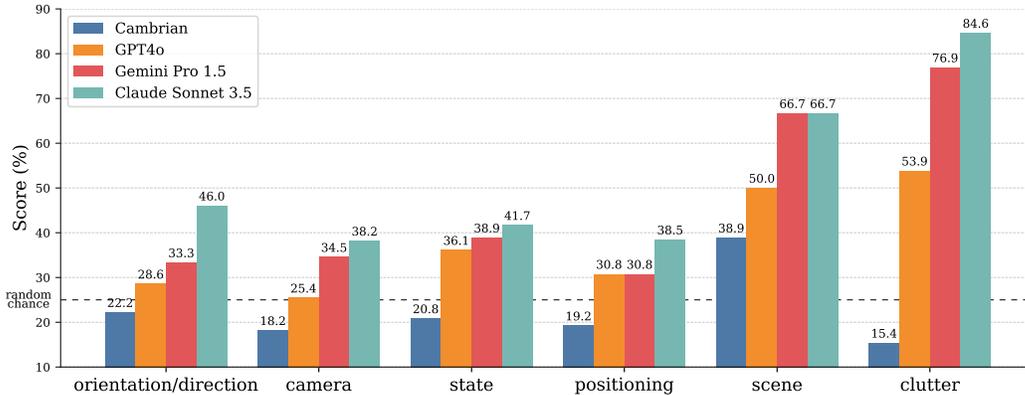


Figure 5. **Whitebox evaluation on D_3 .** We compute self-retrieval scores on individual *POD* subsets within our benchmark for Cambrian-34B, GPT 4o, Gemini Pro 1.5, and Claude Sonnet 3.5. The random guess is 25%.

estingly, while we find this to be true for Cambrian-34B, Fig 5 demonstrates that closed-source MLLMs, specifically Claude 3.5 Sonnet and Gemini Pro 1.5, are capable of identifying characteristics of non-prominent objects to distinguish cluttered images.

3.5. Validity of Self-Retrieval Scorer

While self-retrieval is able to assess an MLLM’s ability to detect and describe subtle visual distinctions between images, quantifying its success depends on the ability of the scorer function used for retrieval. Even if an MLLM successfully captures and incorporates visual differences in its captions, an inferior scorer may fail to capture these fine-grained details, leading to unreliable pairings between generated captions and images. To address this challenge as effectively as possible, we adopt SigLIP, one of the most fine-grained open-source VLM, as our scorer.

We conduct a study to evaluate the reliability of SigLIP as a scorer for the self-retrieval task by comparing three scorers: an average human, an expert human, and SigLIP. Each scorer is presented 100 image pairs sampled from the benchmark. We use captions generated by GPT-4o and pick one (among two) randomly as the caption of the target image. From the 100 captions generated using GPT-4o, 23 captions are deemed non-discriminant by the expert human scorer and are filtered out.

Among the remaining 77 captions, a text-to-image retrieval task is set up: within each image pair, all three scorers are asked to retrieve the target image using the given caption. We find an agreement of 94.8% between the average human and expert human scorer, with the average human picking the same image as the expert for 73/77 samples. SigLIP shows a 79.2% agreement with the expert human scorer, matching the image selection for 61/77 image pairs. This suggests that although SigLIP is not a perfect scorer, it is good enough to be used for self-retrieval evaluation on our benchmark.

4. Conclusion

Our study sheds light on the capabilities as well as limitations of current Multimodal LLMs (MLLMs) in perceiving and describing subtle visual differences. We propose a novel benchmark, D_3 , consisting of 247 image pairs, each comprised of highly similar images that differ in one fine-grained visual concept. By using self-retrieval, we directly evaluate the natural language outputs requiring the model to independently identify the visual difference and incorporate it into unique captions. Our study reveals that while MLLMs excel at distinguishing scene changes and visually crowded images, they struggle with nuanced aspects of visual understanding such as camera angle or an object’s state, positioning, and orientation. Among all the MLLMs tested, Claude Sonnet 3.5 performs relatively well on D_3 , while other models, especially open-source ones, struggle. We also conduct a human study to investigate the validity of the retrieval scorer. Our results reveal SigLIP to be a reliable scorer for self-retrieval evaluation on D_3 .

Future work could focus on scaling up the benchmark by adopting large image datasets with accompanying dense captions such as PixelProse [18], VeCap [10]. This could result in a larger number of fine-grained image pairs, enhancing diversity and reliability of evaluation.

Acknowledgements. This project was supported in part by funding from SERB SRG/2023/002544 and a research gift by Adobe Research India.

References

- [1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. In *European Conference on Computer Vision (ECCV)*, 2024. 3
- [2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang

- Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*, 2024. 2
- [3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3
- [4] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024. 2
- [5] Manu Gaur, Darshan Singh S, and Makarand Tapaswi. No Detail Left Behind: Revisiting Self-Retrieval for Fine-Grained Image Captioning. *arXiv preprint arXiv:2409.03025*, 2024. 2, 3
- [6] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kallioikoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A Bateman. AI2D-RST: A Multimodal Corpus of 1000 Primary School Science Diagrams. *Language Resources and Evaluation*, 2021. 2
- [7] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, et al. Training Compute-optimal Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [8] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to Describe Differences Between Pairs of Similar Images. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2
- [9] Elisa Kreiss, Fei Fang, Noah D Goodman, and Christopher Potts. Concadia: Towards Image-based Text Generation with a Purpose. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 2
- [10] Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, et al. VeCLIP: Improving CLIP Training via Visual-enriched Captions. *arXiv preprint arXiv:2310.07699*, 2023. 5
- [11] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-Next-Interleave: Tackling Multi-Image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*, 2024. 2, 4
- [12] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, Tell and Discriminate: Image Captioning by Self-Retrieval with Partially Labeled Data. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2
- [13] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*, 2024. 2
- [14] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, et al. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [15] Meta. The Llama 3 Herd of Models. *arXiv preprint, arXiv:2407.21783*, 2024. 2, 3
- [16] OpenAI. GPT-4V(ision) System Card. <https://openai.com/index/gpt-4v-system-card/>, 2023. 2, 3
- [17] OpenAI. GPT-4 Technical Report. *arXiv preprint, arXiv:2303.08774*, 2023. 2
- [18] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From Pixels to Prose: A Large Dataset of Dense Image Captions. *arXiv preprint arXiv:2406.10328*, 2024. 5
- [19] Chameleon Team. Chameleon: Mixed-modal Early-Fusion Foundation Models. *arXiv preprint arXiv:2405.09818*, 2024. 4
- [20] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4
- [21] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. *arXiv preprint arXiv:2406.16860*, 2024. 2, 4
- [22] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4
- [23] Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4
- [24] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoji Liu, Ge Zhang, et al. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [25] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-training. In *International Conference on Computer Vision (ICCV)*, 2023. 3
- [26] Jianrui Zhang, Mu Cai, Tengyang Xie, and Yong Jae Lee. CounterCurate: Enhancing Physical and Semantic Visio-Linguistic Compositional Reasoning via Counterfactual Examples. In *Association of Computational Linguistics (ACL)*, 2024. 2
- [27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

A. Appendix

A.1. Prompt Tuning for Open-source MLLMs

Open-source MLLMs required further prompt tuning compared to closed-source MLLMs. The following prompt was used for benchmarking Cambrian-34B and LLaVA-NeXT-34B on D_3 :

Given two images, generate a discriminant description for each image that uniquely describes it. The description for each image should highlight its key visual differences compared to the other distractor image. Only focus on distinct visual features that make each image unique. Output in JSON format, with a key 'left' and a key 'right' with string values corresponding to descriptions for the left and right image.

Chameleon-30B struggled to follow the above instructions, and its prompt was tuned further. We found the following prompt to perform best on our benchmark:

You are given two images. Identify finegrained visual differences between both images <image><image>. Using these visual differences, uniquely describe each image. The output should be in JSON format, with 'image1' and 'image2' as keys, and their respective captions as string values.

A.2. VQA is easier than D_3 evaluation

Given an image pair, we find that it is easier for the MLLM to discern fine-grained differences during VQA than D_3 evaluation. Although providing the visual concept as part of the question aids VQA in assessing the existence of specific visual abilities, it makes the task easier. Figs. 6, 7, 8, 9 provide qualitative examples where Gemini-1.5-Pro, despite achieving perfect VQA scores, struggles to independently identify the fine-grained difference and incorporate them into discriminant captions. In the figures, the text highlighted in **green** represents the fine-grained differences captured by the MLLMs, while that marked in **red** represents erroneous descriptions (hallucinations).



Q: The tree with brown leaves is to the left or right of the polar bear?
(a) Left (b) Right

◆ The tree with brown leaves is to the **left of the polar bear**. So the answer is (a).

◆ The tree with brown leaves is to the **right of the polar bear**. So the answer is (b).

Detect, Describe, Discriminate

◆ A majestic polar bear with a thick coat of creamy white fur strides confidently across a rocky outcrop, its dark eyes focused intently on something beyond the frame. The bear's right paw is slightly raised, capturing a moment of dynamic movement.

◆ A powerful polar bear with thick, creamy white fur carefully navigates a rocky terrain, its head lowered and gaze fixed on the ground ahead. The bear's front left paw extends forward, highlighting the precision of its movement as it traverses the uneven surface.

Figure 6. VQA vs. D_3 evaluation, 1/4. Example from D_3 , where model fails during D_3 evaluation despite achieving perfect score in VQA.



Q: Is the orange in front of the apple? (a) Yes (b) No

◆ (b) **No**. The orange is **behind** the apple.

◆ (a) **Yes**. The orange is **in front** of the apple.

Detect, Describe, Discriminate

◆ A closeup of a bowl of fruit, featuring a vibrant orange in the center, flanked by a green pear and a red and yellow apple.

◆ A close-up shot of a bowl overflowing with fruits, highlighting a large, red apple centered in the frame, surrounded by pears, oranges.

Figure 7. **VQA vs. D₃ evaluation, 2/4**. Example from D₃, where model fails during D₃ evaluation despite achieving perfect score in VQA.



Based on the image, is the following statement correct: The surfer's face is visible.
(a) True (b) False

◆ (b) **False**. The surfer's face is not visible.

◆ (a) **True**. Surfer's face is visible in profile.

Detect, Describe, Discriminate

◆ A surfer carves a high line on a steep wave, his upper body tucked low within the barrel.

◆ A surfer pumps his legs for speed, carving across the face of a glassy wave.

Figure 8. **VQA vs. D₃ evaluation, 3/4**. Example from D₃, where model fails during D₃ evaluation despite achieving perfect score in VQA.



Q: How many clock faces are visible in this image of Big Ben?

(a) One (b) Two

◆ (a) **Correct.**

◆ (b) **Incorrect.**

Detect, Describe, Discriminate

◆ The clock face of Big Ben, shows the time just past quarter after 1, set against a vibrant blue sky with scattered white clouds.

◆ Big Ben, shows the time just past **7 o'clock**. The sky behind the tower is **paler blue**, with a few faint clouds.

Figure 9. **VQA vs. D₃ evaluation, 4/4.** Example from D₃, where model fails during D₃ evaluation despite achieving perfect score in VQA.

A.3. Qualitative Analysis of MLLMs on Different Points of Difference

We present qualitative examples in Figs. 10, 11, 12, 13, 14, 15, corresponding to each point of difference. For each image pair, we also include the captions generated by GPT-4o, Gemini-1.5-Pro, and Claude-Sonnet-3.5.

State



A tennis player in a maroon outfit is captured right before executing a serve, with the ball in mid-air directly above her head. Her knees are bent in preparation, and is surrounded by greenery and fences on a sunny day.



A female tennis player with her body angled to the left, eyes on the ball, winds up for a powerful serve with her **racket held high**.



Tennis player in purple dress completing serve motion, with **racket extended upward** and both arms raised.



Tennis player in a maroon outfit is about to make contact with the ball during her serve. The **ball is slightly lower and closer** to her racket. The outdoor court features similar surroundings with greenery and fences.



A female tennis player with her body angled to the right prepares to return a high shot, her **racket drawn back** and her left hand raised.



Tennis player in purple dress preparing to serve, with racket raised high **behind head** and left arm extended forward.

Figure 10. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **State** as POD.

Positioning



Four riders on horseback advancing on a broad, sandy shore with **deeper water near the shoreline**, marked by distant foamy waves and a **slightly cloudier sky**.



Four horseback riders traverse a windswept beach, their horses leaving a trail of imprints in the damp sand. The horses are closely grouped together.



Four riders on horseback walking in a line along the beach, with the leftmost horse's legs visible and a thin strip of water in the foreground.



Four riders on horseback leisurely walk along a wide, sandy beach with calm, shallow water pooling at the edge, showcasing the gentle waves in the background under a **clear sky**.



Four horseback riders enjoy a leisurely ride along the shoreline, their horses' hooves kicking up puffs of sand with each stride. The **horses are spaced further apart**.



Four riders on horseback side-by-side on the beach, with the **horses' legs more visible** and a **larger pool of water in the foreground**.

Figure 11. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **Positioning** as POD.

Camera




 Donuts passing under a glossy white glaze curtain on a production line. The donuts left to of the conveyor are glazed, while those on the right are about to enter the glaze waterfall, creating a smooth and uniform coating.

 A row of plain donuts on a conveyor belt are about to move under a machine that will pour icing on top of them.

 Donut production line with golden-brown donuts on a metal conveyor belt, featuring a large white icing dispenser in the foreground.

 Donuts on a production line as they undergo the glazing process. The donuts left to the conveyor belt are covered with a shiny layer of glaze, while the ones on the right are freely moving towards the glaze curtain.

 A row of glazed donuts on a conveyor belt have just moved out from under a machine that has poured icing on top of them.

 Close-up of donut glazing process, showing both glazed and unglazed donuts side by side, with a visible divider and nozzle for applying glaze

Figure 12. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **Camera** as POD.

Scene




 Close-up of vibrant oranges hanging from a tree with a **clear, blurred pastel background**, emphasizing the smooth texture of the fruit and the glossy, light green leaves.

 Three vibrant, ripe oranges are the focal point of this image, hanging prominently from a branch with lush green leaves against a **bright, washed-out background**.

 Bright orange citrus fruits with smooth skin hanging from thin branches, set against a **soft focus background** with vibrant green leaves in the foreground.

 Close-up of oranges clustered on tree branch against a **lush green background**, with some visible blemishes around oranges and the darker, denser foliage around them.

 A cluster of oranges, some partially obscured by leaves, with a single drop of water clinging to the bottom of the most prominent fruit, set against a **backdrop of leafy green foliage**.

 Close-up of textured orange citrus fruits among dark green leaves, visible water droplets on the fruit and a **more densely leaved background**.

Figure 13. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **Scene** as POD.

Clutter





-  A modern purple train stationed at platform with passengers boarding on the left side. The front of the train is completely intact, including the extended front bumper, and platform tiles are aligned with tactile paving.
-  A maroon train with a gold stripe, distinguished by **two passengers in suits** walking away from the train on the left side of the platform.
-  Purple train with a sleek, rounded front at station platform, passengers visible boarding a green train in the background

-  Modern purple train at the platform, with passengers boarding from the right side. The front bumper of the train is missing, revealing internal parts, distinguishing it from the other image.
-  A maroon train with a gold stripe, identifiable by the green train visible in the background on the right side of the platform.
-  Purple train with a more prominent nose and visible undercarriage, **red-striped pillar in foreground**, fewer passengers on platform.

Figure 14. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **Clutter** as POD.

Orientation/Direction





-  A close-up view of a giraffe standing on an open plain with dense vegetation in the background. The giraffe's body fills most of the frame, showcasing its detailed patterned coat in front of a slightly cloudy sky.
-  A giraffe with bold reticulated pattern stands tall against cloudy sky. Its long legs and neck are fully extended, giving a sense of its impressive height.
-  Giraffe with distinct, intricate pattern of dark spots stands tall amidst dense, green vegetation under a cloudy sky

-  A giraffe standing in a more sparse landscape with scattered trees and bushes. The giraffe is **positioned slightly to the left**, with space around it, revealing a wide expanse of the plain under a clearer sky.
-  A giraffe with a **more dispersed spot pattern** blends into the savanna landscape. Its **neck slightly bent**, and the focus on surrounding trees and grasslands creates a sense of its natural habitat.
-  Giraffe with **more uniform, less intricate spot pattern** peers over bushes in a drier, more open savanna landscape

Figure 15. Outputs of different closed-source MLLMs during D_3 evaluation on an image pair having **Orientation** as POD.