Evaluating Discriminability of Vision-Language Models

Masayasu Muraoka[†], **Naoaki Okazaki**[† ‡]
[†]School of Computing, Institute of Science Tokyo, Tokyo 152-8550, Japan
[‡]Artificial Intelligence Research Center, AIST, Tokyo 135-0064, Japan
[‡]Research and Development Center for Large Language Models, NII, Tokyo 100-0003, Japan
masayasu.muraoka@nlp.comp.isct.ac.jp, okazaki@comp.isct.ac.jp

Abstract

We study the discriminative ability of vision-language models (VLMs). This ability refers to processing information by distinguishing key details from unnecessary or redundant parts to achieve specific goals. It is vital for the practical use of VLMs in applications like visual chatbots. Whereas recent VLMs have shown decent performance on various multimodal capabilities, their discriminative ability has not been thoroughly explored to date. To this end, we construct DiscriBench to evaluate the discriminability of VLMs in various daily life activities. We carefully design the dataset to require distinguishing information in both vision and language modalities, and semi-manually craft questions in English and Japanese, making them solvable without relying on external knowledge or expertise. Experimental results demonstrate a large performance gap (14.0 to 69.3 points) between humans and existing VLMs in discriminability, where humans can solve the task with an accuracy of 90% or higher. By reducing the difficulty of discriminability, our ablation studies elucidate that vision encoders cannot distinguish visual details well, given generally similar but partially different images. Besides, we observe that VLMs show inconsistent inference between modalities. We will publish DiscriBench (1,200 samples) to foster research in this direction.

Keywords: benchmark construction, evaluation, discriminability, vision-language model, VLM

1. Introduction

Discriminability is one of the ordinary abilities that humans use in daily life activities. As shown in Fig. 1, we decide trip plans such as destinations and routes through conversations while looking at photos provided in a travel guidebook. When we go shopping, we determine an item to buy from multiple candidates through a chat with shop staff, as shown on the left of Fig. 2. Such processes often involve multimodal information, and to achieve goals, we distinguish necessary information from other information regardless of modality, which we call discriminability in this work.

However, we ask: *Do current VLMs have enough discriminability to support us in our daily lives?* Whereas vision-language models (VLMs) have the potential to help us in various scenarios (Islam and Moushi, 2024; MIC, 2024), we observe that VLMs such as GPT-4o (OpenAI, 2023a) cannot solve visual question answering (VQA) problems that require discriminability, as shown at the bottom of Fig. 1. Also, the discriminative ability of VLMs has not been thoroughly explored to date. Typical multimodal tasks do not require distinguishing textual, visual, or both types of information from irrelevant information, since the input text or image in such tasks is sufficiently simple and does not contain redundant information. For example, an input text is "Describe the image in detail." in an image captioning task (Liu et al., 2023) or "What does the man have?" in a VQA task (Antol et al., 2015), both of which specify only the task instruction to VLMs without any other information. This differs from our daily-life conversation, as in Fig. 1, where additional information is included beyond the information that we finally need. More challenging benchmarks proposed in recent years (Gurari et al., 2018; Li et al., 2023; Fu et al., 2023; Liu et al., 2023; Yue et al., 2024a,b; Li et al., 2024b,a; Yu et al., 2024; Lu et al., 2024; Liu et al., 2024) sometimes contain longer contexts than conventional multimodal tasks. However, these benchmarks require additional skills such as multi-hop reasoning (Lu et al., 2024) or external knowledge (Yue et al., 2024a,b), which again impede the independent evaluation of the discriminative ability of VLMs.

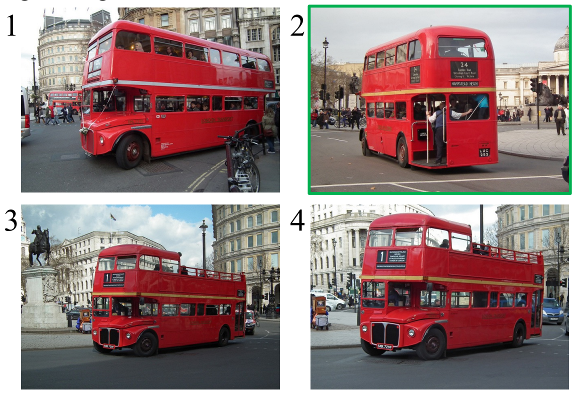
To this end, we propose DiscriBench, a new evaluation benchmark for the discriminability of VLMs. We aim to evaluate the discriminability of VLMs in both text and vision modalities through DiscriBench. As illustrated in Fig. 1, DiscriBench is formulated as a multiple-choice visual question answering task, where the choices are images. To solve this task, VLMs are required to distinguish necessary information from superfluities in both modalities and choose the most suitable image that matches the input text description from the image options given. Furthermore, to evaluate VLMs' discriminability on different image types, we include natural images and illustrations in DiscriBench. We obtain the actual examples used in university entrance exams, which include illustrations. As for natural images, we semi-manually create new examples by aligning their characteristics and difficulty level with the exams, which roughly correspond to CEFR A1-B1 levels (i.e., beginning to lower-intermediate language proficiencies). In addition, we meticu-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9368–9385
11–16 May 2026.

Situation: Two friends, Emma and Jack, are planning a sightseeing tour in London and discussing transportation options.

Conversation:
Emma: Have you seen those iconic red double-decker buses in London?
Jack: Yeah, they're everywhere! I heard some have open tops for sightseeing.
Emma: That sounds fun! And the yellow line in the middle of the bus goes well with the red body. But I'm worried about rain.
Jack: Let's check which routes have covered upper decks.

Question: Which bus would they choose for their sightseeing tour?



Human: 2 GPT-4o: 1 Claude 3.5 Sonnet: 1 InternVL-2.5-38B: 3

Figure 1: Example of DiscriBench and predictions by humans and VLMs (shown in gray). To answer the question, VLMs need to distinguish key details (shown in green) from unnecessary parts (shown in red) in the conversation while distinguishing visual details among the image options.

lously select image pairs that expose VLMs to a fine-grained comparison by using the erroneous agreement technique (Tong et al., 2024). We are also interested in evaluating the discriminability of VLMs in different languages, and thus, we support English and Japanese. We describe the detailed construction process of DiscriBench in Sec. 2.

We evaluate a number of both open and closed VLMs, including LLaVA-onevision (Li et al., 2024), Qwen2-VL (Wang et al., 2024), InternVL-2.5 (Chen et al., 2024a), Claude 3.5 Sonnet (Anthropic, 2024), and GPT-4o (OpenAI, 2023a), on our DiscriBench. We observe a large discrepancy in performance (14.0 to 69.3 points) between humans and current VLMs, whereas humans can easily solve the task with a high accuracy of 90%. We also demonstrate that the performance improves more when reducing the difficulty level of discriminability in the vision modality than in the language modality. Furthermore, we provide a case study showing that existing VLMs fail to identify or count visual objects, whereas they suggest valid reasoning.

Our contributions are three-fold:

- We propose DiscriBench to evaluate the discriminability of VLMs and make it available[1],
- We demonstrate a large performance gap between VLMs and humans in DiscriBench, and
- We analyze the extent of discriminability that VLMs have through ablation studies.

2. Construction of DiscriBench

Each sample in DiscriBench consists of a triplet of an input text, an image set providing four options, and an answer to the question included in the input text. The input text is further decomposed into a situation description, a conversation with four utterances on average between two people, and a question. The image set can be either four different images or a single image that provides four options. The answer is represented as an option number (1-4) of the correct image. We adopt two heterogeneous data sources: Exam and COCO. We detail each construction procedure in the subsections below. Fig. 2 illustrates an overview of the construction process.

2.1. Exam

We construct a subset of DiscriBench samples from a university entrance examination. We select the English listening subject in the Common Test for University Admissions[3] in Japan (NCUEE, 2024). This Common Test is conducted annually nationwide, and its questions, which are available online in PDF format, are created by subject experts. Specifically, questions in the English listening subject are created to align with A1 to B1 levels in CEFR (Common European Framework of Reference for Languages)[4] as a difficulty level to measure the language ability of examinees (Council of Europe, 2020; Kondo and Miyao, 2023). Topics handled in the English listening subject are related to daily lives and intentionally chosen such that most examinees are familiar with the topics. Hence, the questions are solvable from the information given without relying on external knowledge. The task format best matches DiscriBench, and thus, we adopt this Common Test in DiscriBench. Although the English reading subject in the Common Test was another candidate for DiscriBench, the reading subject contains fewer multimodal questions than the listening subject, and characteristics

[1]https://github.com/muraoka7/discribench

[2]The example used for Exam is adapted from Question 10 of the Regular Session in 2023.

[3]https://www.dnc.ac.jp/kyotsu/ (Japanese site)

[4]https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

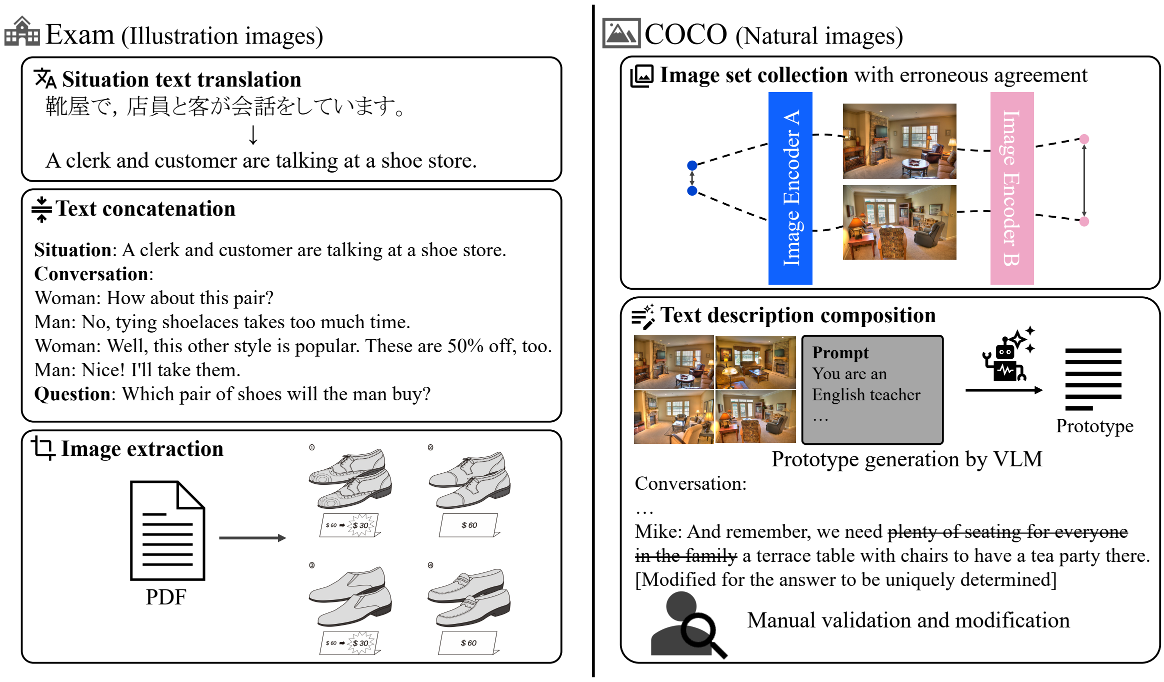


Figure 2: Overview of DiscriBench construction. Processes for Exam (left) and COCO data (right).[2]

of the reading subject are somewhat similar to existing work, such as MMMU (Yue et al., 2024a) and MMMU-Pro (Yue et al., 2024b). Therefore, we chose the listening portion to pursue the unique advantage of DiscriBench.

To match the data format in DiscriBench, we process the exam data (See Fig. 2:left). First, we machine-translate the situation description given in Japanese into English and then manually verify the translations. We next concatenate the English situation description with the associated conversation and question obtained from the ground-truth transcript data. For images, we convert the format from PDF into JPEG and resize them to 1280×1280.

2.2. COCO

Pre-training data of VLMs often involve natural images. To evaluate discriminability using natural images, we create another set of DiscriBench samples from COCO (Lin et al., 2014), incorporating the characteristics of the Exam. To do so, we first observe and analyze the Exam questions and identify the following characteristics:

- C1 Option images in Exam are generally similar to each other but differ in detail in each sample.
- C2 The conversation in Exam contains utterances that mislead examinees into a wrong answer.
- C3 Hints to a correct answer are scattered throughout the conversation.

We collect image sets from COCO and compose text descriptions based on these characteristics.

Image set collection. We use the erroneous agreement method (Tong et al., 2024) to collect images that share the common concept or view in general but differ in detail (C1). The erroneous agreement method enables us to find image pairs such that the image features generated by a vision encoder A are close to each other (e.g., a high cosine similarity), whereas those by another vision encoder B are distant from each other (e.g., a low cosine similarity). Following Tong et al. (2024), we use models trained in the CLIP manner (Radford et al., 2021) for the vision encoder A, and DINOv2 (Oquab et al., 2024) for the vision encoder B. It should be noted that CLIP-based models serve as a vision encoder in many existing VLMs. Hence, we can collect challenging image sets with which VLMs tend to get confused when distinguishing.

More formally, we collect four-image sets $\{S\}$, where S is an image set containing four images used as option images in a DiscriBench sample, to satisfy the following conditions: (i) any image pair taken from S has a cosine similarity $\geq \alpha$ in the feature space of the vision encoder A, and (ii) at least one image pair in S has a cosine similarity $< \beta$ in the feature space of the vision encoder B. α and β are thresholds of similarity. We apply a post-processing to blur faces to make it difficult to identify persons. See Appendix A.1 for more details on this processing.

Text description composition. We compose a text description, that is, a situation description, conversation, question, and answer, for each four-image set collected above with the help of a powerful VLM. In our case, we used Anthropic's Claude

Figure 3: Statistics and characteristics of DiscriBench.

3.5 Sonnet (Anthropic, 2024), given that we found that it can produce text of decent quality based on multiple input images. We let the VLM generate a prototype of a text description by providing a prompt along with a four-image set. We explicitly instruct the VLM to include the characteristics of C2 and C3 in the text description. One of the authors conducted manual validation of the generated text, carefully checking whether (a) the answer to the generated question can be uniquely determined given the input text and images, (b) the conversation has misleading information that can lead VLMs to a wrong answer, and (c) hints are spread over the conversation. We modify the generated text accordingly if needed and discard samples if and only if we find them difficult to meet all of the three criteria above ((a) to (c)), such as a sample with images of the same object from slightly different angles, which are so close to each other that we cannot create a meaningful conversation. After the manual validation is completed, we further ensured its validity by asking ten human annotators[5] to point out and fix any mistakes or ambiguities in the DiscriBench samples. See Appendix A.2 for more details on the text description composition.

2.3. 1-shot prototype generation and self-verification

To increase the number as well as diversity of DiscriBench samples, we devise a two-step prototypical generation process that allows us to reduce the manual validation effort while keeping its quality by making the best use of the DiscriBench samples already created. We generate a prototype by providing the same VLM as above with an already verified DiscriBench sample as a 1-shot demonstration, combined with a prompt.

Referring to self-verification approaches (Kadavath et al., 2022; Phute et al., 2024), we then ask the VLM to self-validate the generated prototype, whether it can be deterministically solved based on the information given, and to modify the prototype if the self-validation fails. The entire prompt we used for this self-validation is illustrated in Fig. 10. We can apply this generation process regardless of the image type.

[5]We used BAOBAB Inc., an annotation service company, for the validation: https://baobab-trees.com/en

2.4. Translation into Japanese

It is interesting to study how language affects VLMs' discriminability when using the same image options. To this end, we machine-translate the text description in DiscriBench into Japanese using Anthropic's Claude 3.5 Sonnet (Anthropic, 2024). We conducted a preliminary experiment to compare the quality of machine translation of the text description in our DiscriBench between GPT-4o and Claude 3.5 Sonnet, and found that Claude 3.5 Sonnet can produce better results since GPT-4o sometimes did not preserve the conversation format (e.g., removal of newline characters) and generated utterances in a non-colloquial style. Therefore, we adopt Claude 3.5 Sonnet. All the translated samples are validated for their correctness by ten human annotators, who were all Japanese native speakers, using the same annotation service mentioned in the previous subsection. We obtained a total of 400 DiscriBench samples throughout our construction process described so far.

2.5. Overview of DiscriBench

We analyze and briefly describe the characteristics of our DiscriBench in terms of activities, scenes, and objects using 100 samples. Fig. 3 illustrates that DiscriBench encompasses a wide range of daily life activities ranging from recreation, shopping, exploration, planning, learning, and social events to creative activities, travel, and professional services. DiscriBench also involves various scenes (e.g., indoor, outdoor, tourism venues, etc.) with

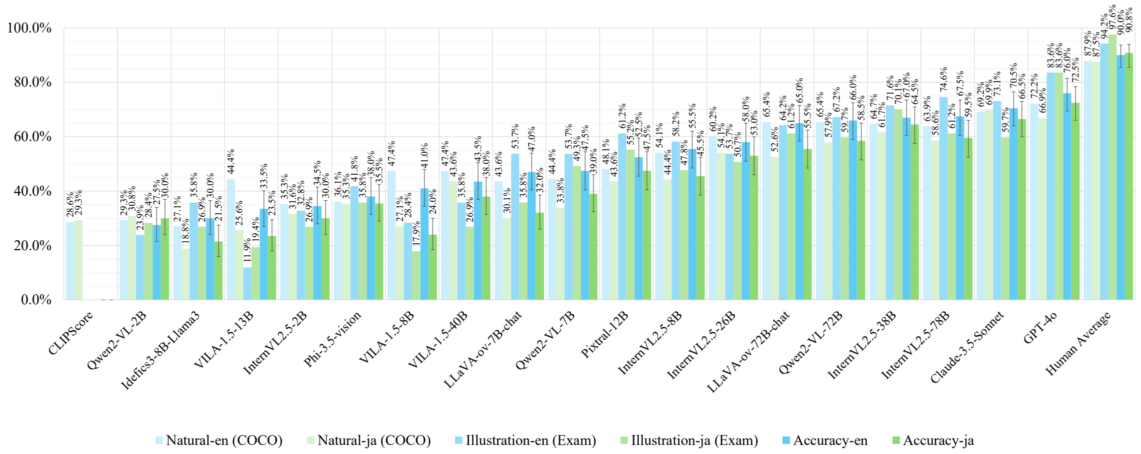


Figure 4: Accuracy of VLMs on DiscriBench. Whiskers indicate the 95% confidence intervals computed by the bootstrap method (Efron and Tibshirani, 1993) with 10,000 resampling.

many visual objects being discussed in the conversations, such as animals, home items, vehicles, food, and recreation items. Hence, DiscriBench allows us to evaluate VLMs' discriminability in a wide array of fine-grained daily activities. More examples can be found in Figs. 1, 2, 5, and 8.

3. Experiments

We now evaluate a number of VLMs, including both open-sourced and closed-sourced models, on DiscriBench. We also analyze the extent to which VLMs can (or cannot) distinguish necessary information from superfluities when the difficulty level of discriminability is relaxed.

3.1. Experimental setups

We evaluate Idefics3 (Laurençon et al., 2024), Phi-3.5-vision (Microsoft, 2024), VILA-1.5 (Lin et al., 2024), Pixtral (Mixtral AI, 2024), LLaVA-onevision (Li et al., 2024), Qwen2-VL (Wang et al., 2024), InternVL-2.5 (Chen et al., 2024a) as open-sourced VLMs, and Claude 3.5 Sonnet (Anthropic, 2024, `20241022`) and GPT-4o (OpenAI, 2023a, `2024-08-06`) as closed-sourced VLM. The selection criterion for the VLMs we test is that each VLM is trained on and thus can accept multiple images per sample, as required by DiscriBench. For a fair comparison, we use the same prompt for every VLM we evaluate, prepending images to the prompt while adapting the prompt format to align with each model's recommended usage. For example, the official usage for Claude 3.5 Sonnet recommends images preceding the instruction, whereas Pixtral encourages appending images after the instruction. We instruct VLMs to answer each question in DiscriBench with an image option number [1-4] and allow them to optionally append the reason for the answer. The full description of the prompt we use is found in Fig. 12. We extract the answer from VLM's response with a regular expression and compute accuracy as the model performance on 400 DiscriBench samples constructed in Sec. 2. We include CLIPScore (Hessel et al., 2021) as a baseline to see how well VLMs solve the task beyond simply matching images from input texts. CLIPScore solves DiscriBench via text-to-image matching: it encodes the input text and the image options separately into feature representations, and then selects the image option most similar to the text features as the answer. Note that we cannot apply CLIPScore to the illustration samples (Exam) since the visual input is a single image, and thus, we cannot perform text-to-image matching on these samples.

We use the official toolkit provided by OpenAI and Anthropic to obtain responses from GPT-4o (OpenAI, 2023) and Claude 3.5 Sonnet (Anthropic, 2023), respectively. For evaluating open-sourced VLMs, we use the official repository and pre-trained model checkpoints for VILA-1.5 (NVIDIA Research Projects, 2024), while we use `vllm` (Kwon et al., 2023) for the rest of the VLMs. We obtain the model weights of the VLMs from Hugging Face (Wolf et al., 2020). We employ greedy decoding during inference with the maximum number of tokens set to 512, and use the default values for the rest of the parameters. We use 4 NVIDIA A6000 GPUs with 48GB RAM each, and it takes up to 10 minutes to evaluate one VLM in a single run.

We also study human performance on DiscriBench to see how easily humans can solve this task. We asked ten participants and instructed them to solve 100 examples of DiscriBench in both languages. We randomly shuffled the examples of DiscriBench and presented them to the participants to avoid any potential bias. We report the average accuracy over ten participants.

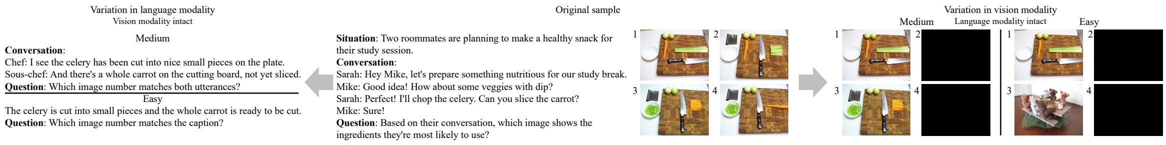


Figure 5: Samples used in our ablation study.

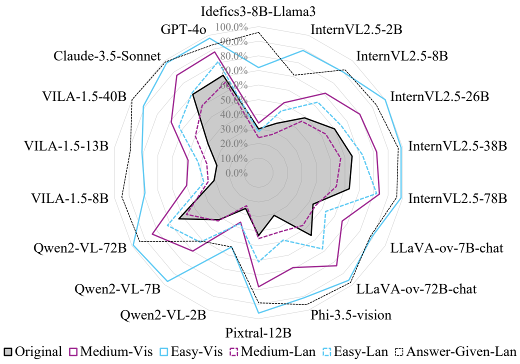


Figure 6: Accuracy of VLMs under relaxed difficulties in English.

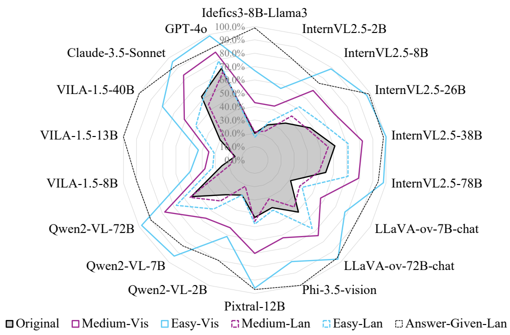


Figure 7: Accuracy of VLMs under relaxed difficulties in Japanese.

3.2. Main results

Fig. 4 shows the main results. Appendix B provides detailed accuracy of more VLMs. First, we verify that humans can solve this task with high accuracy: 90.0% in English and 90.8% in Japanese. This indicates that the task is easy for humans in general. However, we observe that some samples in DiscriBench are challenging even for humans, requiring careful attention to both modalities, such as prepositions or subtle visual differences, because it is constructed based on university entrance exams created by subject experts to measure the language ability of examinees. This could explain why it is difficult for humans to perfectly solve DiscriBench. In contrast, closed VLMs such as GPT-4o and Claude 3.5 Sonnet lag behind humans by 24.3+ points in overall accuracy (Accuracy-en&ja). Open VLMs achieve much lower accuracy than closed ones, ranging from 67.5% (InternVL-2.5 78B) to 27.5% (Qwen2-VL 2B), which is close to random guessing (25.0%). Roughly speaking, overall accuracy correlates with model size, with smaller VLMs tending to perform worse. We also verified that the gap between humans and top VLM, GPT-4o, is statistically significant since the confidence intervals of humans and GPT-4o do not overlap. Whereas we can expect certain performance gains with larger models that go beyond CLIPScore, the large gap from humans suggests room for improvement in the discriminability of VLMs.

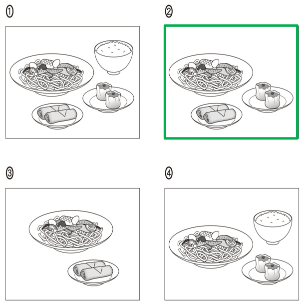
Comparing results between image types (Natural-*, and Illustration-*), we notice that humans and large VLMs perform better on illustrations than on natural images, whereas smaller VLMs such as Qwen2-VL 2B and VILA-1.5 8B/13B show the opposite trend. Thus, the competence across image types varies across VLMs. More interestingly, though we see performance drops from the English results (*-en) in almost all VLMs, the results in Japanese (*-ja) are encouraging, considering that the training data used to train VLMs mainly consist of English. We expect these promising Japanese results to be attributable to the effective use of powerful pre-trained LLMs trained on multilingual data as the language encoder of the VLMs. Overall, we hope that these findings on image types, languages, and various VLMs are beneficial to the community.

3.3. Analysis

Experiment with relaxed difficulty: It is interesting to see how accuracy changes as the difficulty level of discriminability decreases in vision and language modalities, and which modality affects discriminability more, if such effects exist. To verify this, we conduct an ablation study. We extend 100 samples of DiscriBench in both English and Japanese by stepwise reductions in discriminative difficulty in the respective modalities. Fig. 5 illustrates the samples used in this ablation study.

More specifically, to reduce discriminability in language, we create two variations with shorter and easier conversations than the original ones. We use the text composition technique introduced in Sec. 2.2, while keeping image options unchanged.

[6]The above example is adapted from Question 9 of the Regular Session in 2022.



Situation: A clerk is taking an order from a customer.
Conversation:
Woman: Are you ready to order, sir?
Man: Yes, I'd like the fried noodle set.
Woman: Certainly. Would you like rice with that?
Man: Well.... It comes with two side dishes, so that's enough.
Question: What did the man order?

Human: 2 (average accuracy: 100%)
GPT-4o: Answer: ③
Reason: The man ordered the fried noodle set with two side dishes, but he declined the rice. Image 3 shows the noodles with two side dishes and no rice.
Claude 3.5 Sonnet (20241022): Answer: 3
Reason: Based on the conversation, the man ordered the fried noodle set but declined the rice when asked by the server. Looking at image 3, it shows exactly what he ordered - a plate of fried noodles and spring rolls as a side dish, without the rice and other side dishes shown in the other images. (redacted)
InternVL-2.5-38B: Answer: 3
Reason: The man ordered the fried noodle set, which includes two side dishes but no rice. Image 3 shows the fried noodle set with two side dishes and no rice.

Situation: A couple discusses their weekly grocery shopping and meal planning while looking at fresh produce.
Conversation:
Sarah: We need to plan our meals for the week. What vegetables should we buy?
Mark: How about some broccoli and carrots? They're always good staples.
Sarah: Good idea. I'm also thinking of making a corn dish. Should we get some?
Mark: Actually, let's focus on leafy greens this week. I saw some nice lettuce varieties.
Question: Based on their conversation, which image best represents their likely shopping cart?

Human: 4 (average accuracy: 100%)
GPT-4o: Answer: Image 2
Reason: The conversation mentions broccoli, corn, and lettuce. Image 2 includes these vegetables.
Claude 3.5 Sonnet (20241022): Answer: 1
Reason: Image 1 best matches their conversation because it contains broccoli, which Mark specifically suggested, leafy greens (both red and green varieties) which Mark wanted to focus on, and some root vegetables. The image also shows a simple, focused selection of vegetables that would be typical for basic weekly meal planning, unlike images 2, 3, and 4 which show much larger and more diverse produce hauls. Additionally, Image 1 doesn't include corn (which Sarah considered but Mark redirected toward leafy greens instead), making it more consistent with their final conversation direction.
InternVL-2.5-38B: Answer: 3
Reason: Image 3 includes a variety of leafy greens, such as lettuce and kale, which aligns with Mark's suggestion to focus on leafy greens for the week. It also contains broccoli and carrots, which were mentioned by Mark as good staples. The presence of corn in Image 2 and the absence of leafy greens in Image 1 and Image 4 make them less suitable choices based on the conversation.

Figure 8: Examples of errors made by top VLMs.[6]

One variant is of medium difficulty and contains two utterances, each of which matches multiple image options but determines the single ground-truth answer image in combination. Another variation is easier and consists of a single caption that directly indicates only the ground-truth answer image, without any misleading information. To reduce the difficulty in vision, we replace two to three images in each image option set. We create an image option set of medium difficulty by retaining the most dissimilar image to the ground-truth image in the original four-image set as an incorrect option, while dropping the remaining two. We use cosine similarity, as computed in Sec. 2.2, to measure image similarity. An easier image option set consists of a ground-truth image and an irrelevant image randomly taken from other samples in DiscriBench. While keeping the text part of the samples unchanged, we add two black images to each modified image option set in both the medium and easy settings to keep the number of image tokens given to VLMs constant relative to the original samples. In total, we obtain 800 new samples for this ablation.

Figs. 6 and 7 show the results. Original indicates accuracy in the original, unmodified DiscriBench; Medium-* and Easy-* correspond to accuracy at the medium and easy difficulty levels, respectively; and *-Vis and *-Lan indicate the modalities (vision and language, respectively) in which the difficulty is relaxed. We observe that performance indeed improves in almost all VLMs and in both languages as the difficulty level of discriminability decreases. These results indicate that current VLMs lack discriminability because they fail to distinguish necessary information from unnecessary information when more and more unnecessary information is given to the VLMs (i.e., in the Original and Medium settings). It is notable that the performance gain is much larger when the vision modality is altered (*-Vis) than when the language modality (*-Lan) is altered, as shown by half of the VLMs achieving $\geq 94\%$ accuracy in Easy-Vis, compared to Easy-Lan, where only two VLMs achieve $\geq 80\%$. These trends may suggest that the language encoders of VLMs have more limited discriminability than the vision encoders. However, there might be another possibility that could explain the lower accuracy in Easy-Lan: the vision encoders fail to select the correct image despite sufficient discriminability of the language encoders. To exclude this possibility, we conduct an additional experiment where the answer image option is provided in the text description, such as "... Question: {question} *The answer is third image.* Please choose the number of the image that matches the answer to the question." This setting allows VLMs to solve the task without looking at the images. The results are shown by Answer-Given-Lan in Figs. 6 and 7, which clearly indicate that the performance is significantly improved from Easy-Lan, reaching the same or higher accuracy than Easy-Vis. This demonstrates that the lower accuracy in Easy-Lan is due to the lack of discriminability in the vision encoders.

Case study: We provide a case study to see how and when VLMs are confused in DiscriBench. Fig. 8 shows examples of errors commonly made by three VLMs that performed well in Fig. 4. As shown in Fig. 8, the VLMs made mistakes even in intuitively easy cases that required only object recognition or counting. Considering the reasons for the answers provided by the VLMs (see the reasons in the top example and those provided by by InternVL-2.5-38B in the bottom example in Fig. 8), it seems that the VLMs can often distinguish key information for the correct answer in the conversation (i.e., the language modality), whereas they fail to choose the correct image This could explain the lack of discriminability in the vision modality observed in the ablation study above. Humans can answer these cases correctly, as the conversation itself is easy to follow and the image options are sufficiently different from each other to distinguish. More examples can be found in Appendix C.

4. Related work

4.1. Vision-language models (VLMs)

Following the emergence of large language models (LLMs) with instruction-following capabilities (Brown et al., 2020), the development of instruction-following VLMs has been an active research area (Alayrac et al., 2022; Li et al., 2022, 2023; Liu et al., 2023; Bai et al., 2023b; OpenAI, 2023a,b; Laurençon et al., 2024; Lin et al., 2024; Microsoft, 2024; Li et al., 2024; Liu et al., 2024; Wang et al., 2024; Chen et al., 2024a; Anthropic, 2024). Such VLMs usually consist of pre-trained vision and language encoders, connected by modality adapters such as multi-layer perceptrons (MLPs) and attention layers to process both modalities. Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have been the default choice as the vision encoder and are often taken from pre-trained CLIP or its derivative models (e.g., SigLIP (Zhai et al., 2023)), which were trained on image-caption paired data with contrastive learning (Radford et al., 2021). In contrast, different LLMs have been used as the language encoder in different VLMs, such as Llama (Touvron et al., 2023a,b; Dubey et al., 2024), Vicuna (Zheng et al., 2024), Qwen (Bai et al., 2023a; Yang et al., 2024), etc. Although training of VLMs involves interleaved multimodal data (Byeon et al., 2022) or high-quality instruction-tuning data (Chen et al., 2023), it is unclear how these data contribute to VLMs' discriminability. To evaluate the discriminative ability of these VLMs, we aim to construct DiscriBench.

4.2. Evaluation benchmarks for VLMs

To evaluate VLMs, a variety of evaluation benchmarks have been proposed to date (Gurari et al., 2018; Li et al., 2023; Fu et al., 2023; Liu et al., 2023; Yue et al., 2024a,b; Li et al., 2024b,a; Yu et al., 2024; Lu et al., 2024; Liu et al., 2024). Similar to DiscriBench, the SEED benchmark (Li et al., 2024b) is constructed with the help of GPT-4 to generate a question, multiple-choice options, and a ground-truth answer. Most of these benchmarks were not created to evaluate discriminability, as they do not contain irrelevant or unnecessary information in both modalities from which VLMs should distinguish key information. Some benchmarks contain examples that require distinguishing relevant parts from irrelevant ones, but these benchmarks usually require additional skills to solve the tasks. MMMU(-Pro) (Yue et al., 2024a,b) are composed of college-level examples in 30 subjects, which, however, require external knowledge in these subjects. MathVista (Lu et al., 2024) contains samples derived from an IQ test, which requires multi-hop reasoning to reach answers. Unlike these benchmarks, DiscriBench evaluates VLMs' discriminability without relying on such additional expertise.

4.3. Multi-image multimodal tasks

Our work is also relevant to other multimodal work that proposes datasets that contain multiple images in a sample (Suhr et al., 2019; Li et al., 2024c; Kil et al., 2024; Fu et al., 2024; Wang et al., 2025). BLINK (Fu et al., 2024) includes visual perception tasks that require multi-image comparison. MuirBench (Wang et al., 2025) offers 12 multi-image understanding tasks with 10 multi-image relations. Although these datasets consider multiple images per sample, the language modality lacks complexity because the input text is limited to a few short sentences. Therefore, they do not pose substantial challenges to the language encoder in VLMs. Moreover, none of these datasets aim to evaluate the discriminability of VLMs addressed in this work. In particular, they do not assess VLMs' capability to distinguish key information from superfluities in both modalities simultaneously, as opposed to within either modality individually.

5. Conclusion

We study the discriminative ability of VLMs, namely, the ability to distinguish key details from unnecessary or irrelevant information when exposed to a mixture of such information. To this end, we propose a new evaluation benchmark, DiscriBench, which is challenging because it requires discriminability in both vision and language modalities. A unique feature of DiscriBench is that it involves multiple images that are generally similar yet different in detail, coupled with a multi-turn conversation. This feature is practically important for VLMs to support us in our daily lives, such as recommending items among similar items. Experimental results demonstrate that recent VLMs struggle to achieve high accuracy, although the task is easy enough for humans. We also conduct an ablation study and case studies to take a step further in investigating the low accuracy of VLMs on DiscriBench, whereas identifying the root causes is challenging due to the limited transparency of model details for some VLMs (e.g., undisclosed training data or model architecture). Despite exceptional performance on existing tasks, we observe that the vision encoders in VLMs can be further improved to distinguish details of images with subtle differences. We believe the insights obtained in our study are beneficial for guiding future improvements of VLMs. We are interested in expanding our analysis of VLMs on DiscriBench in the future to study how the output formats impact overall accuracy (i.e., comparison between a multiple-choice setting and free-form generation) and how VLM performance on DiscriBench on DiscriBench could be improved; for instance, visual prompting (Wu et al., 2024) and dialogue annotation (Bunt et al., 2012) could help the vision and language encoders in VLMs interpret their inputs more appropriately.

6. Ethics Statement

The evaluation results obtained in this work are applicable to English, Japanese, and the daily life activities covered in DiscriBench, but not to other languages or activities. We also made our best effort to mask and remove information that can identify individuals during manual validation, but there is a possibility of mistakenly passing personal information through manual validation. However, we set the intended use of our benchmark for evaluation purposes only, which we believe can minimize the potential risk of unintended use.

7. Limitations

The development of VLMs in the community is unprecedentedly rapid, and new VLMs have been released on a monthly or even weekly basis. We evaluated newer models such as Qwen3-VL (QwenTeam, 2025) and GPT-5 (OpenAI, 2025), which have been released near or after the time of submission, and added the results in Appendix B. In contrast, we could not evaluate very large models such as Llama 4 Maverick (Meta AI, 2025) due to the lack of computing resources. Based on the results we obtained, we anticipate that the gap will close further as more powerful VLMs are developed. However, this direction does not conflict with our work; rather, it aligns with our work because we hope that our DiscriBench allows future VLMs to address discriminability.

One might also be concerned about the dataset size. However, we emphasize that DiscriBench is carefully curated to target a specific and underexplored capability, discriminative ability in multimodal contexts, which has not been the focus of existing datasets. We believe that the focused nature of our benchmark can provide meaningful insights, as specific capabilities like discriminability are often underrepresented in large-scale general-purpose datasets. Moreover, it is comparable in size to other focused benchmarks such as LLaVA-Bench (Liu et al., 2023) (60 questions), MMVet (Yu et al., 2024) (218 questions), and MM-UDP Bench (Miyai et al., 2025) (350–920 questions per subtask).

Acknowledgment

These research results were obtained from the commissioned research (No.22501) by the National Institute of Information and Communications Technology (NICT), Japan. This work was also supported by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology. We thank the National Center for University Entrance Examinations for granting permission to use the data in the English listening subject in the Common Test for University Admissions to construct our dataset. We are also grateful to BAOBAB Inc. for their human annotation, which helped us produce a high-quality DiscriBench.

8. Bibliographical References

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2023. Getting vit in shape: Scaling laws for compute-optimal model design. In *NeurIPS*, volume 36, pages 16406–16425.

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, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pages 23716–23736.

Anthropic. 2023. Anthropic python api library. Accessed on September 15, 2024.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic.

Anthropic. 2025. System card:claude sonnet 4.5. Technical report, Anthropic.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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. 2020. Language models are few-shot learners. In *NeurIPS*, volume 33, pages 1877–1901.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proc. 2012 LREC*, pages 430–437.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proc. IEEE/CVF CVPR*, pages 24185–24198.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, Strasbourg.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *20th ICLR*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Bradley Efron and R. J. Tibshirani. 1993. *An introduction to the bootstrap*. 57. Chapman & Hall.

Flickr. 2004. Flickr. Accessed on December 14, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. 2021 EMNLP*, pages 7514–7528, Online and Punta Cana, Dominican Republic.

Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. *TechRxiv*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova

DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Mutsumi Kondo and Mari Miyao. 2023. Does the common test measure the thinking skills necessary for the 21st century? *ARELE*, 34:81–96.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proc. 29th SOSP*, page 611–626, New York, NY, USA.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. 40th ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proc. 39th ICML*, volume 162, pages 12888–12900.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proc. IEEE/CVF CVPR*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proc. IEEE/CVF CVPR*, pages 26296–26306.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2026-03-04.

MIC. 2024. *WHITE PAPER Information and Communications in Japan*, pages 75–88. Ministry of Internal Affairs and Communications, Japan.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. Technical report.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Technical report, Microsoft.

Mixtral AI. 2024. Pixtral 12b. Technical report, Mixtral AI.

NCUEE. 2024. About the national center for university entrance examinations. Accessed on December 14, 2024.

NVIDIA Research Projects. 2024. Vila: Optimized vision language models. Accessed on August 23, 2024.

OpenAI. 2023a. Gpt-4 technical report. Technical report, OpenAI.

OpenAI. 2023b. Gpt-4v(ision) system card. Technical report, OpenAI.

OpenAI. 2023. Openai python api library. Accessed on October 11, 2024.

OpenAI. 2025. Gpt-5 system card. Technical report, OpenAI.

OpenAI. 2025. Introducing gpt-4.1 in the api. Accessed on March 4, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning robust visual features without supervision. *TMLR*.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. LLM self defense: By self examination, LLMs know they are being tricked. In *2nd Tiny Papers Track at ICLR 2024*.

QwenTeam. 2025. Qwen3-vl: Sharper vision, deeper thought, broader action. Technical report, QwenTeam.

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. 2021. Learning transferable visual models from natural language supervision. In *Proc. 38th ICML*, volume 139, pages 8748–8763.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proc. IEEE/CVF CVPR*, pages 9568–9578.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proc. 2020 EMNLP: Demo*, pages 38–45, Online.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. 2024. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proc. IEEE/CVF ICCV*, pages 11975–11986.

9. Language Resource References

Antol, Stanislaw and Agrawal, Aishwarya and Lu, Jiasen and Mitchell, Margaret and Batra, Dhruv and Lawrence Zitnick, C and Parikh, Devi. 2015. *Vqa: Visual question answering*.

Byeon, Minwoo and Park, Beomhee and Kim, Haecheon and Lee, Sungjun and Baek, Woonhyuk and Kim, Saehoon. 2022. *COYO-700M: Image-Text Pair Dataset*.

Chen, Lin and Li, Jisong and Dong, Xiaoyi and Zhang, Pan and He, Conghui and Wang, Jiaqi and Zhao, Feng and Lin, Dahua. 2023. *ShareGPT4V: Improving Large Multi-Modal Models with Better Captions*.

Fu, Chaoyou and Chen, Peixian and Shen, Yunhang and Qin, Yulei and Zhang, Mengdan and Lin, Xu and Yang, Jinrui and Zheng, Xiawu and Li, Ke and Sun, Xing and others. 2023. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*.

Fu, Xingyu and Hu, Yushi and Li, Bangzheng and Feng, Yu and Wang, Haoyu and Lin, Xudong and Roth, Dan and Smith, Noah A. and Ma, Wei-Chiu and Krishna, Ranjay. 2024. *BLINK: Multimodal Large Language Models Can See but Not Perceive*.

Gurari, Danna and Li, Qing and Stangl, Abigale J and Guo, Anhong and Lin, Chi and Grauman, Kristen and Luo, Jiebo and Bigham, Jeffrey P. 2018. *Vizwiz grand challenge: Answering visual questions from blind people*.

Kil, Jihyung and Mai, Zheda and Lee, Justin and Chowdhury, Arpita and Wang, Zihe and Cheng, Kerrie and Wang, Lemeng and Liu, Ye and Chao, Wei-Lun. 2024. *MLLM-CompBench: A Comparative Reasoning Benchmark for Multimodal LLMs*.

Li, Bo and Zhang, Kaichen and Zhang, Hao and Guo, Dong and Zhang, Renrui and Li, Feng and Zhang, Yuanhan and Liu, Ziwei and Li, Chunyuan. 2024a. *LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild*.

Li, Bohao and Ge, Yuying and Ge, Yixiao and Wang, Guangzhi and Wang, Rui and Zhang, Ruimao and Shan, Ying. 2024b. *SEED-Bench: Benchmarking Multimodal Large Language Models*.

Li, Feng and Zhang, Renrui and Zhang, Hao and Zhang, Yuanhan and Li, Bo and Li, Wei and Ma, Zejun and Li, Chunyuan. 2024c. *LLaVA-NeXT: Tackling Multi-image, Video, and 3D in Large Multimodal Models*.

Li, Yifan and Du, Yifan and Zhou, Kun and Wang, Jinpeng and Zhao, Xin and Wen, Ji-Rong. 2023. *Evaluating Object Hallucination in Large Vision-Language Models*.

Lin, Tsung-Yi and Maire, Michael and Belongie, Serge and Hays, James and Perona, Pietro and Ramanan, Deva and Dollár, Piotr and Zitnick, C Lawrence. 2014. *Microsoft coco: Common objects in context*.

Liu, Haotian and Li, Chunyuan and Wu, Qingyang and Lee, Yong Jae. 2023. *Visual Instruction Tuning*.

Liu, Yuan and Duan, Haodong and Zhang, Yuanhan and Li, Bo and Zhang, Songyang and Zhao, Wangbo and Yuan, Yike and Wang, Jiaqi and He, Conghui and Liu, Ziwei and others. 2024. *Mmbench: Is your multi-modal model an all-around player?*

Pan Lu and Hritik Bansal and Tony Xia and Jiacheng Liu and Chunyuan Li and Hannaneh Hajishirzi and Hao Cheng and Kai-Wei Chang and Michel Galley and Jianfeng Gao. 2024. *MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts*.

Miyai, Atsuyuki and Yang, Jingkang and Zhang, Jingyang and Ming, Yifei and Yu, Qing and Irie, Go and Li, Yixuan and Li, Hai Helen and Liu, Ziwei and Aizawa, Kiyoharu. 2025. *Unsolvable Problem Detection: Robust Understanding Evaluation for Large Multimodal Models*.

Suhr, Alane and Zhou, Stephanie and Zhang, Ally and Zhang, Iris and Bai, Huajun and Artzi, Yoav. 2019. *A Corpus for Reasoning about Natural Language Grounded in Photographs*.

Fei Wang and Xingyu Fu and James Y. Huang and Zekun Li and Qin Liu and Xiaogeng Liu and Mingyu Derek Ma and Nan Xu and Wenxuan Zhou and Kai Zhang and Tianyi Lorena Yan and Wenjie Jacky Mo and Hsiang-Hui Liu and Pan Lu and Chunyuan Li and Chaowei Xiao and Kai-Wei Chang and Dan Roth and Sheng Zhang and Hoifung Poon and Muhao Chen. 2025. *MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding*.

Weihao Yu and Zhengyuan Yang and Linjie Li and Jianfeng Wang and Kevin Lin and Zicheng Liu and Xinchao Wang and Lijuan Wang. 2024. *MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities*.

Yue, Xiang and Ni, Yuansheng and Zhang, Kai and Zheng, Tianyu and Liu, Ruoqi and Zhang, Ge and Stevens, Samuel and Jiang, Dongfu and Ren, Weiming and Sun, Yuxuan and others. 2024a. *Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi*.

Yue, Xiang and Zheng, Tianyu and Ni, Yuansheng and Wang, Yubo and Zhang, Kai and Tong, Shengbang and Sun, Yuxuan and Yu, Botao and Zhang, Ge and Sun, Huan and others. 2024b. *Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark*.

Zheng, Lianmin and Chiang, Wei-Lin and Sheng, Ying and Zhuang, Siyuan and Wu, Zhanghao and Zhuang, Yonghao and Lin, Zi and Li, Zhuohan and Li, Dacheng and Xing, Eric P. and Zhang, Hao and Gonzalez, Joseph E. and Stoica, Ion. 2024. *Judging LLM-as-a-judge with MT-bench and Chatbot Arena*.

Vision encoder A	α
OpenAI's CLIP-L/14-336	0.95
SigLIP-SO400M-14/384	0.95
InternViT-6B-448-V1-2	0.988
LLaVA-1.5's CLIP-L/14-336	0.91
VILA-13B's SigLIP-SO400M-14/384	0.94
VILA-40B's InternViT-6B-448-V1-2	0.99

Table 1: Threshold α used for each vision encoder in four-image set collection.

Appendix

A. Construction of DiscriBench

A.1. Four-image set collection

To ensure the diversity of four-image sets, we used OpenAI's CLIP-L/14-336 (Radford et al., 2021), SigLIP-SO400M-14/384 (Zhai et al., 2023; Alabdulmohsin et al., 2023), InternViT-6B-448-V1-2 (Chen et al., 2024b), LLaVA-1.5's CLIP-L/14-336 (Liu et al., 2024), VILA-13B's SigLIP-S0440M-14/384, VILA-40B's InternViT-6B-448-V1-2 (Lin et al., 2024) as vision encoders A, and DINOv2 with registers (`dinov2_vitl14_reg`) (Darcet et al., 2024) as a vision encoder B. While we use the threshold $\beta = 0.6$ following (Tong et al., 2024), we set the threshold α for each vision encoder A as shown in Table 1, where we empirically find suitable threshold values through manual exploration.

A.2. Text description composition

We used Anthropic's Claude 3.5 Sonnet (Anthropic, 2024) to generate prototypes. Specifically, we used the model version `20240620` to create the first 100 samples since it was the latest model at that time, whereas we used `20241022` for 1-shot prototype generation and self-verification later. Figs. 9 and 10 present the prompts we used to generate prototypes of the text descriptions. It cost 0.008 USD per sample to generate a prototype and 0.778 USD for 100 samples. We manually ensure that the composed text does not include personally identifiable information. Fig. 11 provides samples that we needed to discard due to difficulty in creating meaningful conversations.

A.3. Use of DiscriBench

We publish our DiscriBench dataset and the evaluation scripts.[7] When you use our dataset, especially the Exam portion, please refer to the instructions described in our release site above. Regarding licensing, Apache 2.0 applies to the evaluation scripts, while the DiscriBench dataset has different licenses or requirements. In particular, for the Exam portion derived from the English listening subject in the Common Test for University Admissions, all rights are reserved by its original creator, the National Center for University Entrance Examinations (NCUEE).[8] We obtained written permission from NCUEE for their use. The COCO images originate from Flickr (Flickr, 2004) and are subject to their own licenses, which we follow as well. The text part of the COCO portion in DiscriBench is a modified output from Claude 3.5 Sonnet, which is also subject to the Terms of Services of Anthropic.[9] We define the intended use of DiscriBench as an evaluation of machine learning models.

[7] https://github.com/muraoka7/discribench

B. Experiment details

Table 2 shows detailed accuracy of VLMs, including more recent ones, i.e., Phi-4-multimodal (Microsoft et al., 2025), Llama 4 Scout (Meta AI, 2025), InternVL3.5 (Wang et al., 2025), Qwen3-VL (QwenTeam, 2025) as open VLMs, and Claude Sonnet 4.5 (Anthropic, 2025), GPT-4.1 (OpenAI, 2025), and GPT-5 Chat (OpenAI, 2025) as closed VLMs. For the recent open VLMs, although we tried inference using `vllm` as in the main experiments, all models except Qwen3 failed to generate outputs reliably. Consequently, we conducted inference for these models using the `transformers` library instead. Recent models generally tend to exhibit performance improvements over their predecessors of the same model size. For example, Phi-4-multimodal-instruct outperforms Phi-3.5-vision by approximately 3-3.5 points, and GPT-5 Chat and GPT-4.1 show gains of about 4.5-6 points over GPT-4o, respectively. Interestingly, large open VLMs such as InternVL3.5-241B-A28B and Qwen3-VL-235B-A22B-Instruct achieve performance comparable to closed VLMs, including Claude Sonnet 3.5/4.5 and GPT-4o. A similar trend of narrowing the performance gap between open and closed VLMs has also been observed in other VL tasks (QwenTeam, 2025), indicating that these large open VLMs have acquired strong general-purpose VL capabilities.

C. More examples in our case study

Figs. 13 and 14 show additional examples where VLMs made errors. Interestingly, we see some valid reasons provided by the VLMs, indicating that they can correctly distinguish key information from superfluities in the conversation, but choose wrong image options, possibly due to the lack of discriminability

[8] https://www.dnc.ac.jp/

[9] https://www.anthropic.com/legal/consumer-terms

Image 1: {image_1}
Image 2: {image_2}
Image 3: {image_3}
Image 4: {image_4}
You are an English teacher in a high school and creating a final exam to evaluate the reading comprehension skills of your students.
Based on the given four images, please create a quiz by generating (1) a chat with three or four conversations between two people, such as a man and woman or a boy and girl, (2) a situation that fits the aforementioned chat and the images, (3) a question whose answer must match only one of the given four images (i.e., never match the rest of three images), (4) the answer to the question, and (5) its explanation (up to 50 words) that justifies the answer. In doing so, please meet the requirements below.
1. Generate an output text following the next template:
"""Situation: <situation>
Conversation:
<person_1>: <utterance_1>
<person_2>: <utterance_2>
<person_1>: <utterance_3>
<person_2>: <utterance_4>
Question: <question>
Answer: <answer>
Reason: <explanation>""",
in which "<string>" is a placeholder and should be replaced with the corresponding actual text generated.
2. The length of the situation, the question, and each utterance should be less than 20 words.
3. The answer should be just a number of the image that matches the answer.
4. Scatter the clues or hints for answering the question you generate over different utterances so that you can evaluate the reading comprehension skills of the examinees of this quiz.
5. Include information that can mislead the examinees to a wrong answer.

Figure 9: Prompt provided to a VLM to generate a prototype of a text description. {image_i} is a placeholder and is replaced with an actual image.

Image 1: {image_1}
Image 2: {image_2}
Image 3: {image_3}
Image 4: {image_4}
=== Prototypical text ===
{unverified_prototype}
=== End of prototypical text ===
Given the images and the prototypical text, which includes a situation description, conversation, question, and its ground truth answer (image number), please validate whether or not the question can be deterministically solved based on the context, that is, the conversation, situation description, and images. This means that the context must specify and match only the ground truth image, not multiple images.
If the above validation fails, assuming that you are a quiz creator to assess the reading comprehension of examinees, please correct the prototypical text to ensure that the question can be deterministically solved at last, while the conversation should contain information that could match multiple images in the middle. Follow the following output format:
"""Validation result: <validation_result>
--- Modified text ---
<modified_text>
"""

Figure 10: Prompt used for self-validation of a prototype text generated by a VLM. {·} is a placeholder and is replaced with corresponding data, an image or text.

in the vision modality. For example, GPT-4o seems to correctly capture that the man sees a bank and a flower shop across from the bank in the bottom example in Fig. 13, but fails to choose the appropriate image option (4). Similar failure patterns can be seen in the reasons provided by GPT-4o and Claude 3.5 Sonnet in the top example in Fig. 13, by Claude 3.5 Sonnet and InternVL-2.5-38B in the top example in Fig. 14, and by Claude 3.5 Sonnet in the bottom example in the same figure.

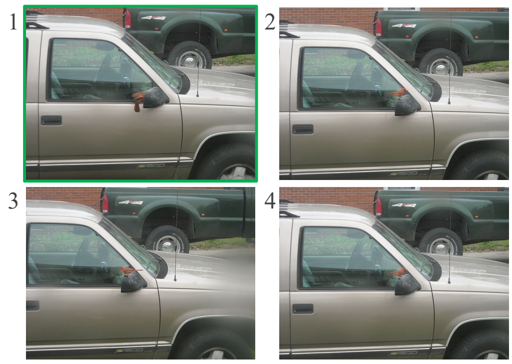
Figure 11: Examples of discarded images.

Model	Natural-en (COCO)	Natural-ja (COCO)	Illustration-en (Exam)	Illustration-ja (Exam)	Accuracy-en	Accuracy-ja
CLIPScore	28.6%	29.3%	N/A	N/A	N/A	N/A
Qwen2-VL-2B	29.3%	30.8%	23.9%	28.4%	27.5%	30.0%
Idefics3-8B-Llama3	27.1%	18.8%	35.8%	26.9%	30.0%	21.5%
InternVL3.5-1B	33.8%	32.3%	23.9%	20.9%	30.5%	28.5%
VILA-1.5-13B	44.4%	25.6%	11.9%	19.4%	33.5%	23.5%
InternVL2.5-2B	35.3%	31.6%	32.8%	26.9%	34.5%	30.0%
InternVL3.5-2B	36.1%	30.1%	37.3%	29.9%	36.5%	30.0%
Phi-3.5-vision	36.1%	35.3%	41.8%	35.8%	38.0%	35.5%
VILA-1.5-8B	47.4%	27.1%	28.4%	17.9%	41.0%	24.0%
Phi-4-multimodal-instruct	41.4%	42.9%	44.8%	29.9%	42.5%	38.5%
VILA-1.5-40B	47.4%	43.6%	35.8%	26.9%	43.5%	38.0%
LLaVA-ov-7B-chat	43.6%	30.1%	53.7%	35.8%	47.0%	32.0%
Qwen2-VL-7B	44.4%	33.8%	53.7%	49.3%	47.5%	39.0%
InternVL3.5-8B	50.0%	45.1%	53.7%	44.8%	48.1%	45.0%
Pixtral-12B	48.1%	43.6%	61.2%	55.2%	52.5%	47.5%
InternVL3.5-4B	54.9%	39.8%	50.7%	37.3%	53.5%	39.0%
InternVL2.5-8B	54.1%	44.4%	58.2%	47.8%	55.5%	45.5%
InternVL2.5-26B	60.2%	54.1%	53.7%	50.7%	58.0%	53.0%
InternVL3.5-38B	63.2%	57.9%	49.3%	62.7%	58.5%	59.5%
InternVL3.5-30B-A3B	60.9%	54.1%	56.7%	52.2%	59.5%	53.5%
InternVL3.5-14B	62.4%	47.4%	58.2%	55.2%	61.0%	50.0%
Qwen3-VL-4B-Instruct	59.4%	54.9%	65.7%	61.2%	61.5%	57.0%
Llama-4-Scout-17B-16E-Instruct	69.2%	68.4%	50.7%	67.2%	63.0%	68.0%
LLaVA-ov-72B-chat	65.4%	52.6%	64.2%	61.2%	65.0%	55.5%
Qwen3-VL-30B-A3B-Instruct	62.4%	51.9%	73.1%	65.7%	66.0%	56.5%
Qwen2-VL-72B	65.4%	57.9%	67.2%	59.7%	66.0%	58.5%
Qwen3-VL-8B-Instruct	64.7%	60.2%	68.7%	56.7%	66.0%	59.0%
InternVL2.5-38B	64.7%	61.7%	71.6%	70.1%	67.0%	64.5%
InternVL2.5-78B	63.9%	58.6%	74.6%	61.2%	67.5%	59.5%
Claude-3.5-Sonnet	69.2%	69.9%	73.1%	59.7%	70.5%	66.5%
InternVL3.5-241B-A28B	69.9%	66.2%	74.6%	65.7%	71.5%	66.0%
Claude Sonnet 4.5	70.7%	64.7%	80.6%	61.2%	74.0%	63.5%
GPT-4o	72.2%	66.9%	83.6%	83.6%	76.0%	72.5%
Qwen3-VL-235B-A22B-Instruct	78.2%	71.4%	79.1%	73.1%	78.5%	72.0%
GPT-4.1	76.7%	73.7%	88.1%	86.6%	80.5%	78.0%
GPT-5 Chat	78.9%	72.9%	88.1%	88.1%	82.0%	78.0%
Human Average	87.9%	87.5%	94.2%	97.6%	90.0%	90.8%

Table 2: Detailed accuracy of VLMs on DiscriBench in ascending order of Accuracy-en.

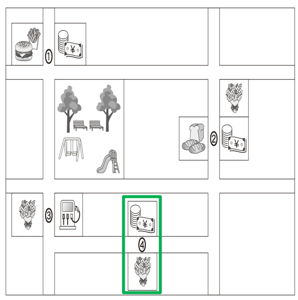
```
Image 1: {image_1} # {image_i} is replaced with an actual image
Image 2: {image_2}
Image 3: {image_3}
Image 4: {image_4}
Situation: {situation}
Conversation:
{conversation}
Question: {question} Please choose the number of the image that matches the answer to the question, given
the situation and conversation. The output format is ‘Answer: <image_number>
Reason: <reason>’, in which the reason can be optionally added to explain your answer.
```

Figure 12: Prompt given to VLMs to answer DiscriBench. The actual text is inserted in {}.



Situation: Two friends discuss a strange sight they witnessed in a parking lot involving a parked car.
Conversation:
Tom: Hey Sarah, did you see that weird thing with the car earlier?
Sarah: Yeah, I did! A small red bird stayed on the side mirror.
Tom: Right. Then, the bird checked the front and back a few times.
Sarah: Finally, the bird looked into the mirror to check itself. It was so interesting!
Question: Which image shows the state of the bird in the end?

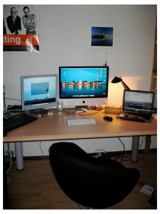
Human: 1 (average accuracy: 100%)
GPT-4o: Answer: Image 4
Reason: Image 4 displays the bird analyzing itself in the mirror, consistent with the description from Tom and Sarah.
Claude 3.5 Sonnet (20241022): Answer: 2
Reason: Looking at the sequence of images, Image 2 shows the red bird appearing to look directly at the side mirror of the car, which matches the conversation's conclusion where Sarah mentions "the bird looked into the mirror to check itself." The positioning of the bird in this image is consistent with a bird examining its reflection in the side mirror, making it the final state described in the conversation.
InternVL-2.5-38B: Answer: 4
Reason: In the conversation, Sarah mentions that the bird looked into the mirror to check itself. Image 4 shows the bird on the side mirror, which aligns with the description of the bird checking itself in the mirror.



Situation: They are talking on the phone with a friend who is lost.
Conversation:
Woman: What can you see there?
Man: Well, there's a bank.
Woman: Is there a bakery across from the bank?
Man: No, I don't see one, but there is a flower shop.
Question: Where is the man standing?

Human: 4 (average accuracy: 100%)
GPT-4o: Answer: 3
Reason: The man sees a bank and a flower shop. In image 3, there is a flower shop across from a bank (represented by money).
Claude 3.5 Sonnet (20241022): Answer: 4
Reason: Based on the conversation, the man is standing at a location where he can see both a bank (indicated by the money/coins symbol) and a flower shop. Looking at the images, location #4 is the only one that shows both a bank (money/yen symbol) and a flower shop next to each other. (redacted)
InternVL-2.5-38B: Answer: 3
Reason: The man mentions seeing a flower shop, which corresponds to the flower shop in image 3.

Figure 13: Examples of errors made by top VLMs. The bottom example is adapted from Question 11 of the Make-up/Retake Examination in 2024.

1 
2 
3 
4 

Situation: Two roommates discuss their home office setups and plan to upgrade their workspaces.
Conversation:
Sarah: Hey Mike, I'm thinking of upgrading my desk setup. Any suggestions?
Mike: Sure! I recently added a second monitor. It's great for productivity.
Sarah: Nice! Do you think a third monitor is too much? I already have a second monitor.
Mike: That'd be enough, then. Oh, don't forget about proper lighting. It makes a big difference.
Question: Based on their conversation, which image best represents Mike's suggestion of the desk setup?

Human: 3 (average accuracy: 60%)
GPT-4o: Answer: Image 4
Reason: Image 4 shows a desk with two monitors, which matches Mike's suggestion of having a second monitor for increased productivity.
Claude 3.5 Sonnet (20241022): Answer: 1
Reason: Image 1 best represents Mike's suggestion as it shows a dual monitor setup with proper lighting (there's a desk lamp providing targeted illumination). The setup includes an iMac as the main display and a secondary monitor, which aligns with Mike's recommendation about adding a second monitor for productivity. The lighting element he mentioned is also clearly visible in this setup, making it the most suitable match for his suggestions in the conversation.
InternVL-2.5-38B: Answer: 1
Reason: Image 1 shows a desk setup with two monitors, which aligns with Mike's suggestion of adding a second monitor for productivity. Additionally, the desk has proper lighting, as indicated by the presence of a desk lamp.

---

1 
2 
3 
4 

Situation: A mother and daughter discuss a teddy bear picnic party the daughter attended at a friend's house.
Conversation:
Daughter: Mom, I went to Sarah's teddy bear picnic today. It was so cute!
Mother: That sounds fun! How many teddy bears were there?
Daughter: Six bears total. They sat around a wooden table with tiny cups.
Mother: Did you take a photo of them?
Daughter: Yes, here it is. Some had their backs to us, facing the table. Among these three, I liked the right white bear.
Mother: So cute! Did they have any snacks?
Daughter: Yes! Little jars of honey and what looked like tiny sandwiches. It was adorable!
Question: Which image best represents the daughter's photo of the teddy bear picnic?

Human: 4 (average accuracy: 100%)
GPT-4o: Answer: Image 3
Reason: Image 3 shows three teddy bears with their backs to us, facing a wooden table with tiny cups and jars, closely matching the daughter's description.
Claude 3.5 Sonnet (20241022): Answer: 1
Reason: Image 1 best matches the daughter's description because:
1. It shows exactly 6 teddy bears
2. Three bears are facing away from the camera, sitting with their backs to the viewer, just as the daughter mentioned
(redacted)
5. Of the three bears facing away, the rightmost one is white/light colored, matching the daughter's comment about liking "the right white bear"
The other images show the same scene from different angles, but Image 1 specifically matches the daughter's perspective and description of how she took the photo, with three bears facing away from the camera.
InternVL-2.5-38B: Answer: 3
Reason: The third image shows six teddy bears sitting around a wooden table with tiny cups, jars of honey, and sandwiches, which matches the daughter's description of the teddy bear picnic.

Figure 14: Examples of errors made by top VLMs.