

視覚言語モデルの識別性能に関する評価用ベンチマークの構築

村岡 雅康¹ 岡崎 直観^{1,2,3}

¹ 東京科学大学 ² 産業技術総合研究所 ³ NII LLMC

masayasu.muraoka@nlp.comp.isct.ac.jp okazaki@comp.isct.ac.jp

概要

本稿では、視覚言語モデル (VLM) の識別性能を評価するための新たな評価用ベンチマーク DiscrIBench を構築し、既存の VLM の識別性能を調査する。DiscrIBench は、提示される情報の中から必要な情報を識別し質問に回答するという、人間が日常生活において行なっているプロセスに焦点を当てたベンチマークである。全 100 問からなる選択肢形式の VQA であり、大学入試難易度を想定して作成している。GPT-4o を含む既存の VLM を DiscrIBench で評価した結果、人間の正解率は 89.2% だったのに対し、VLM はそれより 18.2 - 63.2 ポイント低く、識別性能において大きな差が認められた。

1 はじめに

人間は日常生活において物事を決める際、多くの関連する情報の中から必要な情報を識別し取捨選択するというプロセスを行なっている。例えば、図 1 から想像される旅行の道を決める際、会話の情報やガイドブック等にある画像情報の中から適切な情報を拾い、希望の道を決定する。また、買い物時には (図 3 左の会話やイラストを参照)、類似した候補選択肢の視覚的要素を比較したり、店員や友達との会話から得られた情報をもとに、購入品を決めることがある。これらのプロセスにはしばしばマルチモーダル情報が介在し、モダリティによらず必要な情報の識別・取捨選択が行われる。この情報を識別する能力を本研究では識別能力と呼ぶ。

視覚言語モデル (Vision-Language Model; VLM) は、コンテンツ制作や医療ドメインなど様々な実応用のプロセスにおいて、我々人間をサポートできる潜在能力があることが示されている [1, 2]。上述の日常生活においても VLM のサポートがあれば、より効率的で満足度の高い消費活動等につながると考えられる。しかしながら、現在多くのタスクで優れた性能を出している VLM [3, 4, 5, 6, 7, 8, 9, 10] はこれらの

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?

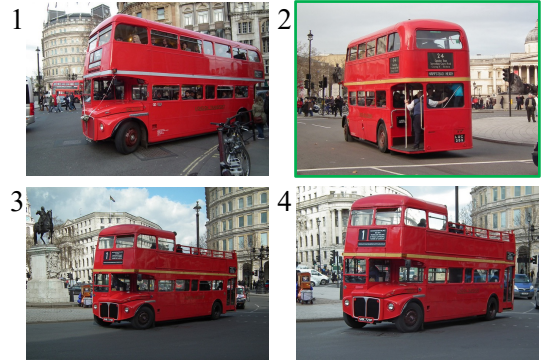


図 1 DiscrIBench の例と 人間および VLM の予測結果。

プロセスにおいて人間をサポートできるに足る識別能力があるかという点においてこれまで十分な調査が行われていない。実際、GPT-4o 等の VLM は図 1 のような比較的易しい問題に正しく答えられない。画像説明文生成 [11, 12] や画像質問応答 (VQA) [13] などの従来のマルチモーダルタスクでは入力テキスト、画像、またはその両方において情報の取捨選択を陽に求められず、近年提案されたより挑戦的なベンチマークにおいては複合的なスキルが求められ情報の識別能力を独立して評価する仕様になっていない [14, 15, 16, 17, 18, 19, 20, 21, 12, 22]。例えば、VLM の評価によく使われる LLaVA-Bench [12] の入力テキストは “Describe this photo in detail.” や “What type of fruit is this?” などの簡潔な指示や質問文が多く情報

の取捨選択は不要である。また、各事例 1 画像入力であるため類似画像を比較することも求められない。MMMU(-Pro) [14, 15] 等のベンチマークには比較的長い入力テキストと複数画像が含まれるが、解くためには大学で習う専門知識 (外部知識) を要し、情報の識別能力単体を評価することは難しい。

そこで本研究では、VLM の識別性能を評価するための新たな評価用ベンチマーク DiscrBench を構築し¹⁾、既存の VLM の識別性能評価を目指す。図 1 に示す通り、DiscrBench は選択肢形式の VQA として定式化され、状況説明文・会話文・質問文全てが入力テキストであり、画像が選択肢となっている。このタスクを解くためには、VLM は言語と視覚の両方のモダリティにおいて情報の識別が求められる。図 1 で言えば、会話文中の緑字の情報を拾い、それに対応する画像を選択肢の中から識別する必要がある。ただし、会話文中の赤字は質問の答えとは関係のないミスリードを誘う情報である。また、異なる画像タイプにおける性能を評価するため、自然画像とイラスト画像を取り扱う。イラスト画像に関しては大学共通テストで実際に使用された問題を使用し、自然画像に関しては COCO [11] から画像選択肢を収集し、Claude 3.5 Sonnet [8] を用いて入力テキストのプロトタイプを作成する。GPT-4o [9] を含む 9 つの VLM の識別性能を DiscrBench で評価した結果、正解率 89.2% と人間には容易に解けるタスクである一方、既存の VLM は人間の正解率から 18.2 から 63.2 ポイント低く、人間の識別能力に及ばないことが確認された。また、追加分析により、上位 3 つのモデルが共通して間違えた事例を改変し、識別に関する難易度を落とした場合に正解率の向上が見られたことから、誤りの原因は VLM の識別能力に起因することが示唆された。

2 DiscrBench の構築

DiscrBench の各事例は、入力テキストと 4 つの画像選択肢、答えの三つ組で構成される (図 1 参照)。入力テキストは状況説明文、会話文、および、質問文からなり、会話文は二者間で行われる平均 4 つの発話である。4 つの画像選択肢は、4 枚の異なる画像であるか、1 画像に 4 つの選択肢が提示されるかのいずれかである。異なる画像タイプを取り扱うため、試験問題 (Exam) と COCO の 2 つの異なる情報源を採用した。以下にそれぞれの情報源に対応する

構築手続きを詳述する。

2.1 Exam (イラスト画像)

大学共通テスト [23] の英語リスニング問題からイラスト画像に関する DiscrBench の事例を作成する。特に、上述の DiscrBench の形式に合う二者間の会話と質問文から正解画像を選ぶ選択肢問題を使用する。問題は専門家によって作成され、受験者の語学力を測るため、およそヨーロッパ言語共通参照枠 (CEFR) の A1 から B1 (語学初学者から標準程度) 相当の難易度となっている [24, 25, 26]。扱われるトピックは多くの受験者に馴染みのある日常生活の一場面である。したがって、特定の外部知識を必要とせず、与えられた情報のみから必要な情報を拾うことで解答可能という特徴を持つ。VLM が処理しやすいデータ形式かつ後述の COCO と表現形式をそろえるため、元の PDF 形式から以下の加工を行う (図 3 左)。状況説明文は機械翻訳により日英翻訳し人手で結果を確認・修正した。正解の会話文および質問文のトランスクリプトを PDF から抽出し、翻訳された状況説明文と結合した。選択肢画像は PDF から切り抜き、1280×1280 にリサイズし、JPEG 形式で保存した。結果として、公開されている試験問題 (試作問題含む) 5 年分から、計 33 問収集した。

2.2 COCO (自然画像)

多くの VLM の学習には自然画像が使われている。自然画像に関する識別性能評価のため、自然画像に関する事例を COCO [11] より作成する。また、難易度や特徴を可能な限り Exam と揃えるため、上記の試験問題を観察し以下の特徴を特定した。

- C1 選択肢内の画像は互いに全体が似ており細部が異なる。
- C2 会話文は受験者を間違った答えに導くようなミスリードさせる情報を含む。
- C3 正解のヒントは会話全体に分散している。

これらの特徴に基づいて画像選択肢を COCO から収集し、入力テキストを作成する (図 3 右)。

画像選択肢収集 全体的には類似だが詳細が異なる 4 つ組画像を収集するため (C1)、erroneous agreement 法 [27] を使用する。この方法により、ある画像エンコーダ A での特徴量は互いに近く、別の画像エンコーダ B での特徴量が離れているような画像ペア (2 画像) を見つけることが可能である。Tong

1) huggingface.co/datasets/mmst004/dscribench-en

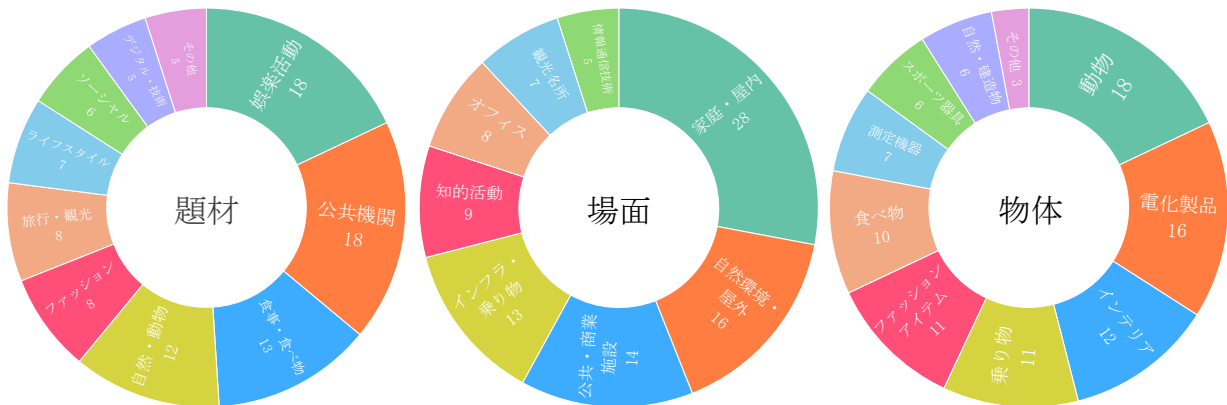


図2 DiscriBenchで扱われている題材，場面，および，物体に関する分布。

ら [27] に倣い，画像エンコーダ A と B にそれぞれ CLIP [28] と DINOv2 [29] を使用する。CLIP は既存の多くの VLM の画像エンコーダとしても使われているため，CLIP が似ていると判断している画像ペアは，VLM にとっても識別が難しいと期待される。

具体的には，以下のいずれも満たす 4 つ組画像 S_n を集める。(i) S_n 内の任意の画像ペアの CLIP 特徴量のコサイン類似度が 0.95 以上。(ii) S_n 内の少なくとも 1 つ以上の画像ペアの DINOv2 特徴量のコサイン類似度が 0.6 未満。事後処理として，個人が特定できるほど鮮明な人物画像に対して，特定を困難にするため顔部分にぼかし処理を施した。

入力テキスト作成 Claude 3.5 Sonnet (20240620) を使用し，上記で収集した各 4 つ組画像について，状況説明文，会話文，質問文，および，答えのプロトタイプを生成する。²⁾ その際与えるプロンプトには，明示的に Exam から観察された特徴 C2 および C3 を生成結果に含めるよう指示する。生成されたプロトタイプは次の項目を重点的に人手で確認し，必要があれば修正を行なった：(a) 入力テキストと画像選択肢から答えが一意に定まるか，(b) 会話には誤った答えにミスリードさせる言及は含まれているか，(c) 正解画像の情報は会話全体に散りばめられているか。一連の処理の結果，67 問作問できた。

2.3 DiscriBench の特徴

計 100 問からなる DiscriBench において扱われている題材，場面，および物体に関する分布を図 2 に示す。DiscriBench では日常生活の様々な場面における多様なやりとりが扱われていることが分かる。

2) GPT-4o と Claude 3.5 Sonnet を用いてどちらがより高品質なプロトタイプを生成できるか予備実験した結果，Claude 3.5 Sonnet だった。また，DiscriBench 構築時 20240620 が最新版のモデルだった。

また，言及されている物体も豊富である。したがって，DiscriBench を用いて多種多様な日常生活における VLM の識別性能の評価が可能となる。³⁾ また，単に評価用としてだけでなく，VLM の挙動を分析・診断するための使用も認める。具体例については，図 1 や 4，公開ページ⁴⁾を参照されたい。

3 実験

3.1 実験設定

DiscriBench を用いて，1 事例につき複数画像の入力を受け付ける以下の VLM の識別性能を評価する：Idefics3 [31]，Phi-3.5-vision [3]，VILA-1.5 [32]，Pixtral [33]，LLaVA-onevision [4]，Qwen2-VL [6]，InternVL-2.5 [7]，Claude 3.5 Sonnet (20241022) [8]，GPT-4o (2024-08-06) [9]。全てのモデルで同一のプロンプトを用い，答えを画像番号 1-4 で出力させ，任意で回答理由も生成させることを許可した。正規表現を用いて画像番号を抽出し，正解率を算出した。また，人間の正解率を求めるため，学部生および院生含む 5 名に DiscriBench 100 問を解くように有償で依頼した。人間の正解率として 5 人の平均を求めた。

3.2 実験結果

表 1 に結果を示す。人間は約 9 割の正解率を達成したことから，人間にとって簡単なタスクであることが分かる。一方，評価した VLM の中で好成績を収めた非公開モデル (GPT-4o および Claude 3.5 Sonnet) でも約 7 割の正解率であり，18.2 ポイント以上人間の正解率より低い結果となった。公開モデル

3) データセットサイズについて，近年よく使われている評価用ベンチマークにおいても 100 から 200 件程度のものがあり [12, 21, 30]，DiscriBench もそれらと同規模サイズである。

4) huggingface.co/datasets/mmst004/discribench.en

表 1 DiscriBench における各 VLM の識別性能 (正解率).
黄色背景 は非公開のモデル.

モデル	サイズ	自然画像	イラスト	全体
無作為回答	-	25.0%	25.0%	25.0%
Idefics3	8B	29.9% (20 / 67)	30.3% (10 / 33)	30.0%
Phi-3.5-vision	4B	26.9% (18 / 67)	39.4% (13 / 33)	31.0%
VILA-1.5	8B	37.3% (25 / 67)	18.2% (6 / 33)	31.0%
VILA-1.5	13B	37.3% (25 / 67)	12.1% (4 / 33)	29.0%
VILA-1.5	40B	44.8% (30 / 67)	30.3% (10 / 33)	40.0%
Pixtral	12B	40.3% (27 / 67)	48.5% (16 / 33)	43.0%
LLaVA-ov-chat	7B	38.8% (26 / 67)	51.5% (17 / 33)	43.0%
LLaVA-ov-chat	72B	56.7% (38 / 67)	54.5% (18 / 33)	56.0%
Qwen2-VL	2B	28.4% (19 / 67)	21.2% (7 / 33)	26.0%
Qwen2-VL	7B	38.8% (26 / 67)	48.5% (16 / 33)	42.0%
Qwen2-VL	72B	64.2% (43 / 67)	60.6% (20 / 33)	63.0%
InternVL-2.5	2B	35.8% (24 / 67)	36.4% (12 / 33)	36.0%
InternVL-2.5	8B	47.8% (32 / 67)	51.5% (17 / 33)	49.0%
InternVL-2.5	26B	65.7% (44 / 67)	48.5% (16 / 33)	60.0%
InternVL-2.5	38B	59.7% (40 / 67)	75.8% (25 / 33)	65.0%
InternVL-2.5	78B	62.7% (42 / 67)	63.6% (21 / 33)	63.0%
Claude 3.5	N/A	70.1% (47 / 67)	69.7% (23 / 33)	70.0%
GPT-4o	N/A	65.7% (44 / 67)	81.8% (27 / 33)	71.0%
人間の平均値	-	85.1%	97.6%	89.2%

はさらに低く、最高でも InternVL-2.5 38B の 65.0% で、Qwen2-VL 2B が最も正解率が低い 26.0% だった。これはほぼ乱択と同じ結果である。これらの結果から、現在の VLM は識別性能に関して人間には及ばず、改善の余地があると言える。

画像タイプごとの結果に着目すると、イラスト画像に関して人間は 97.6% と満点に近い一方、VLM は全体的に低い正解率となっており、特に VILA-1.5 8/13B と Qwen2-VL 2B は乱択よりも低い結果になった。また、全体傾向としてどちら画像タイプを得意とするかはモデルごとに異なっているが、70B 級のモデルでは性能差は小さく、どちらの画像タイプでも同程度の正解率を達成できることが分かる。

3.3 追加分析

エラー分析 VLM の誤りの傾向を掴むため、また、言語と視覚どちらのモダリティによるエラーかを調査するためエラー分析を行った。表 1 において上位 3 つの VLM (GPT-4o, Claude 3.5 Sonnet, InternVL-2.5 38B) が共通して間違えた事例を調べたところ 12 件あり、その誤った要因のモダリティをモデルが出力した回答理由から特定した。結果は、12 件のうち 5 件は言語におけるエラーであり、7 件が視覚におけるエラーであった。図 4 に具体例を示す。分析した VLM は物体の数え上げや会話文にお

表 2 識別難易度易化実験における VLM の正解率.

モデル	言語のみ	視覚のみ	両方改変
InternVL-2.5 38B	25.0%	41.7%	58.3%
Claude 3.5 Sonnet	16.7%	25.0%	50.0%
GPT-4o	50.0%	50.0%	83.3%

ける正解画像に含まれるべき物体情報の識別が求められるような一見簡単に見える事例を間違っていた。しかしながら、これら 12 件の人間の正解率は 75% と全件の正解率より少し低く、人間にとっても難しい問題を含み、特にそれらの問題で上位の VLM は一貫して不正解だったと言える。

識別難易度易化実験 次に、エラー分析において確認されたエラーが実際に識別能力に起因するものかを調べる実験を行う。上記 12 件の問題を改変し、識別能力に関する難易度を下げ、同じモデルに解かせることで正解率が向上するか調査する。具体的には、入力テキストにおいて状況説明文および会話文を削除し、正解画像にのみ合うキャプションを手動で作成した (言語モダリティ改変)。また、画像選択肢に対しては、CLIP 特徴量のコサイン類似度で正解画像と最も似ていない (= 識別しにくい) 画像を残し、それ以外を削除した (視覚モダリティ改変)。

表 2 の結果から、もともと全く解けなかった問題が両モダリティを改変し難易度を易化させると 5 割から最大 8 割以上解けるようになることが確認できた。これより、元の 12 問中 5-8 割は、エラーの原因が分析したモデルの識別能力の欠如によるものであることが示唆される。また、特に InternVL-2.5 38B と Claude 3.5 Sonnet においては、視覚モダリティのみ改変した時の方が言語モダリティのみ改変した時より正解率が高い。このことから、少なくともこの 2 つのモデルに関しては、視覚モダリティにおける識別能力に改善の余地が多いことが示唆される。

4 おわりに

本稿では、VLM における、提示された情報の中から重要な情報を抜き出し質問に回答するという識別性能を調査した。この目的のため、新たな評価分析用ベンチマーク DiscriBench を構築した。DiscriBench は、専門的な外部知識は必要としない日常生活を題材とし、語学初学者から標準相当の難易度を意図して作成されている。実験結果から、既存の VLM は識別性能において人間と差があり、改善の余地が多く残されていることが確認された。

謝辞

本研究成果は、国立研究開発法人情報通信研究機構（NICT）の委託研究（22501）により得られたものです。また、独立行政法人大学入試センターより大学共通テスト問題の利用許諾をいただきました。

参考文献

- [1] Raisa Islam and Owana Marzia Moushi. GPT-4o: The cutting-edge advancement in multimodal LLM. July 2024.
- [2] 総務省. 令和6年版情報通信白書, pp. 75–88. 2024.
- [3] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone. arXiv:2404.14219, 2024.
- [4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. arXiv:2408.03326, 2024.
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024.
- [6] 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. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. arXiv:2409.12191, 2024.
- [7] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv:2412.05271, 2024.
- [8] Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku, 2024.
- [9] OpenAI. GPT-4 technical report, 2023.
- [10] OpenAI. GPT-4V(ision) system card, 2023.
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.
- [12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pp. 34892–34916, 2023.
- [13] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: Visual question answering. *IJCV*, pp. 4–31, 2017.
- [14] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pp. 9556–9567, 2024.
- [15] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. arXiv:2409.02813, 2024.
- [16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pp. 292–305, 2023.
- [17] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *CVPR*, pp. 13299–13308, 2024.
- [18] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: Stronger llms supercharge multimodal capabilities in the wild, 2024.
- [19] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394, 2023.
- [20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, pp. 216–233, 2025.
- [21] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *ICMR*, pp. 57730–57754, 2024.
- [22] 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 *ICLR*, 2024.
- [23] 大学入試センター. 大学入学共通テスト, 2024年12月14日閲覧. <https://www.dnc.ac.jp/kyotsu/>.
- [24] 大学入試センター. 令和6年度大学入学者選抜に係る大学入学共通テスト問題作成方針, 2024年12月28日閲覧. <https://www.dnc.ac.jp/albums/abm.php?d=677&f=abm00004546.pdf>, p.13.
- [25] Council of Europe. Common european framework of reference for languages: Learning, teaching, assessment – companion volume, 2020. <https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>.
- [26] Mutsumi Kondo and Mari Miyao. Does the common test measure the thinking skills necessary for the 21st century? *ARELE*, pp. 81–96, 2023.
- [27] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *CVPR*, pp. 9568–9578, 2024.
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.
- [29] 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. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.
- [30] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhenhao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NIPS*, 2024.
- [31] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. arxiv:2408.12637, 2024.
- [32] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *CVPR*, pp. 26689–26699, 2024.
- [33] Mixtral AI. Pixtral 12b. arXiv:2410.07073, 2024.

Exam (イラスト画像)

状況説明文日英翻訳
靴屋で、店員と客が会話をしています。
↓
A clerk and customer are talking at a shoe store.

入力テキスト結合
Situation: A clerk and customer are talking at a shoe store.
Conversation:
Woman: How about this pair?
Man: No, tying shoelaces takes too much time.
Woman: Well, this other style is popular. These are 50% off, too.
Man: Nice! I'll take them.
Question: Which pair of shoes will the man buy?

画像抽出・変換・加工
PDF → JPEGに変換 (1280×1280) → にリサイズ
[Image showing shoe pairs with price tags: \$110, \$30, \$60]

COCO (自然画像)

画像選択肢収集 (erroneous agreement法)
[Diagram showing image selection process]

入力テキスト作成
プロンプト: You are an English teacher ...
Claude 3.5 Sonnetによるプロトタイプ生成
Conversation:
Mike: And remember, we need plenty of seating for everyone in the family a terrace table with chairs to have a tea party there.
[答えが一意に定まるように修正]
人手による確認および修正

図3 DiscriBench 構築方法の概要。Exam に使われている例は令和5年度本試験問10より作成。

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.

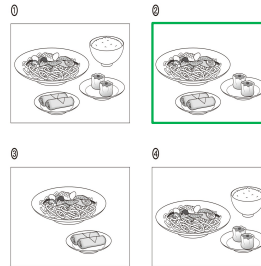
Claude 3.5 Sonnet (20241022): Answer: 3

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.

エラー理由: 視覚モダリティによるエラー。VLMは会話から必要な情報(the fried noodle set with two side dishes and no rice)を特定している一方、間違った画像を選んでいる。'two side dishes'のtwoに引っ張られたと考えられる。



Situation: A couple discusses their weekly grocery shopping and meal planning 1 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.

エラー理由: 言語モダリティによるエラー。VLMは会話文から必要とそうでない物体の識別ができていない。今回の場合、ブロッコリー、ニンジン、レタスが必要で、とうもろこしは不要。



図4 DiscriBench 上位 VLM によるエラーの例。上の問題は令和4年度本試験問9より作成。