

# 大規模画像言語モデルは物体の裏側を認識できるか？ 物体の見えない部分の認識を問うタスクの提案

竹中 誠<sup>1</sup> 谷中 瞳<sup>2</sup>

<sup>1</sup> 三菱電機株式会社 <sup>2</sup> 東京大学

Takenaka.Makoto@bc.MitsubishiElectric.co.jp hyanaka@is.s.u-tokyo.ac.jp

## 概要

人間は視覚的な情報と言語的な情報を受け取ると、それに関連する事前知識を用いて物事の“見えない部分”を想像することができる。物体の見えない部分を予測するタスクはコンピュータビジョンの分野では精力的に研究されているが、多くは対象物の一部が隠れた物体を対象とした物体認識であり、裏側や側面など見えない部分には焦点は当てられていない。そこで本研究では、物体の裏側や側面の認識能力を評価するためのタスクを提案し、線画と写実画像を用いてデータセットを構築した。

実験では、提案タスクを用いて人間と代表的な大規模画像言語モデルの性能を比較した。その結果、人間にとって簡単なタスクであるにもかかわらず、最先端の商用モデルでも人間の性能には及ばないことを確認した。

## 1 はじめに

最近の大規模画像言語モデルの著しい進化[1, 2, 3]に伴い、モデルの能力を評価することの重要性も増している。これまでに、さまざまな観点からVLMの能力を評価するための包括的なベンチマークが提案されている。[4, 5, 6, 7, 8, 9]。これらのタスクはいずれも、物の見える部分を認識する能力を評価するために設計されている。一方、本研究の興味は、物の見えない側面を認識する能力である。最新のVLMがウェブスケールの大規模データで事前学習されていることを考えると、物の見えない部分を想像する能力をVLMに期待することは不思議ではない。見えない部分を分類すると、(i) 一部が他の物体に隠されている場合、(ii) 物体の裏側や側面のように視覚的に無情報な場合、に大別できる。前者はアモーダル補完(Amodal completion)に関連しておりコンピュータビジョンの分野では勢力的に研究

されている。代表的なタスクとしては、オクルージョンを伴う物体認識やセグメンテーションタスク[10, 11, 12]などがある。

本研究が対象とするのは後者である。提案タスクに関連する先行研究には、PerspectiveTaking[13]やCLEVR-MRT[14]がある。これらのタスクは物の見えない部分に焦点は当てておらず、視点獲得[15, 16]や心的回転[17, 18]を重視している。

そこで我々は、裏側や側面などの見えない部分に焦点を当てた二つのタスクを提案する。提案タスクの入出力の例を図1に示す。図1aの反対側予測タスクは、与えられた画像を反対側から見たときの見え方に対応する画像を選択するタスクである。図1bの視点推論タスクは、与えられた画像と、その画像に関する視点の説明文の整合性を問う二値分類タスクである。

本論文の主な貢献は以下の通りである。

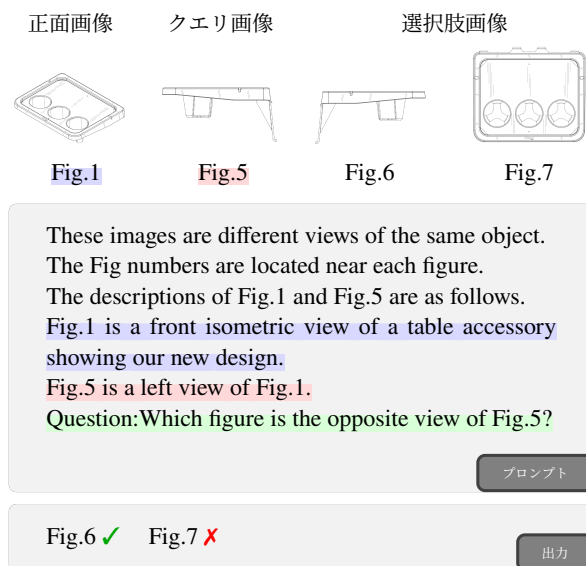
- ✓ 物体の見えない部分を認識する能力を評価する2つのタスクを提案する。
- ✓ 線画と写実画像を使用して、提案タスクのデータセットを構築する。データセットは公開予定である。
- ✓ 提案タスクを用いてプロプライエタリモデルおよびオープンソースモデルを評価し、最新のモデルでも人間に及ばないことを示す。

## 2 提案タスク

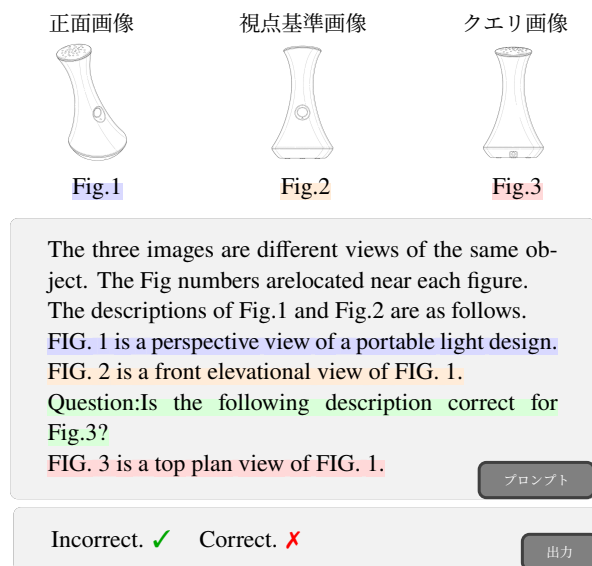
本章では、提案タスクとデータセット構築の流れを述べる。

### 反対側予測タスク

入出力の例を図1aに示す。このタスクは、クエリ画像の反対側から見た見え方に対応する画像を選択肢から選ぶ問題である。入力する画像は、対象物の正面画像(図1a Fig.1), クエリ画像



(a) 反対側予測タスクの入出力



(b) 視点推論タスクの入出力

図 1: 意匠データセットの入出力の一例. スペースの都合上プロンプトは一部のみ表示している. プロンプトの全文は Appendix A を参照されたい.

(図 1a Fig.5), および回答の選択肢画像として 2 つの画像 (図 1a Fig.6, Fig.7) である. プロンプトとして, 正面画像の説明文, クエリ画像の説明文, および質問文を入力する. 正面画像とクエリ画像の説明には対象物に関する情報と, 視点情報が述べられている. なお本稿の実験では, 入力画像枚数の増加による問題の難化を防ぐため入力画像数は 4 枚に制限する.

## 視点推論タスク

入出力の例を図 1b に示す. このタスクは, クエリ画像と, その画像の視点説明の整合性を問う Yes or No の質問応答タスクである. 入力する画像は, 対象物の正面画像 (図 1b Fig.1), クエリ画像 (図 1b Fig.3), および, 左右や前後の視点の基準を与えるための視点基準図 (図 1b Fig.2) である. また, プロンプトとして, 正面画像の説明文, 視点基準画像の説明文, および質問文とクエリ画像の説明文を入力する. なお, 実験では, 視点基準画像を入力する場合としない場合とで性能を比較する.

## 3 データセット

異なるタイプの画像での性能を比較するため, 線画と写実画像を用いてデータセットを構築する. 線画は意匠データの図を用い, 写実画像は一般物体の 3D スキャンデータの画像を用いる.

### 3.1 意匠データセット

画像は DeepPatent データセット [19]<sup>1)</sup> の一部を使用する. DeepPatent データセットは, 米国特許商標庁 (USPTO<sup>2)</sup>) から収集された大規模な意匠画像データセットであり, USPTO のライセンス条件<sup>3)</sup> に則り CC0 1.0<sup>4)</sup> に基づいて公開されている. このデータセットには 45,000 件の意匠から収集された 350,000 の画像で構成されている. 各画像には 1 つ以上の図が含まれており, 各図の説明文は意匠の公開番号を手掛かりに Web から収集可能である. 意匠の図の特徴は, (i) 線画であり, (ii) 同じ物体について, 複数の異なる視点から描かれており, (iii) 各図がどこから見た図なのかという視点情報が自然言語で記述されているというところにある. 本稿の実験では, CPC クラスの D, E, F から 147 をサンプリングしてデータセットを構築した. これらのクラスに限定した理由は, 容易に変形する物体や極端に薄い物体などを可能な限り排除するためである. 各画像には一つ以上の図が含まれ, 各図には “Fig. N” という形式で図番号が付与されている. 提案タスクの入力単位は画像ではなく図であるため, 画像から図をトリミングする必要がある. この作業は手作業で実施し, このとき同時に図番号を付与する. 図と説

1) <https://github.com/GoFigure-LANL/DeepPatent-dataset>

2) [www.uspto.gov](http://www.uspto.gov)

3) <https://www.uspto.gov/terms-use-uspto-websites>

4) <https://creativecommons.org/publicdomain/zero/1.0/>

画像ソース	タスク	トークン数			
		size	mean	min	max
意匠	反対側予測	335	80	67	98
意匠	視点推論 (基準有)	294	85	68	106
意匠	視点推論 (基準無)	294	73	60	91
GSO	反対側予測	150	69	63	74
GSO	視点推論 (基準有)	300	71	71	71
GSO	視点推論 (基準無)	300	61	61	61

表 1: 画像ソース別の各タスクのデータセットのサイズとプロンプトのスペース分割したときのトークン数.

明文の対応付けは、図番号の完全一致によるルールベースによって自動で行う.

### 3.2 GSO データセット

画像は、Google Scanned Objects データセット (GSO) [20, 21] に含まれる画像を使用し、意匠データセットと同様のデータセットを構築する. GSO は 1030 の物体の 3D スキャン画像で構成され、CC BY 4.0<sup>5)</sup>で公開されている. GSO には、各オブジェクトにつき前面、裏面、左側、右側からの画像が含まれている. 前面などの視点ラベルは明示的に付与されていないが、ファイル名の整数が視点ラベルに対応しているため自動的にラベリング可能である. ただし視点情報が記述された説明文は含まれないため、“Fig.2 is a front view of the object.” のような説明文を人間が作成する.

作成したデータセットの統計は表 1 にまとめる.

## 4 実験

### 4.1 実験設定

評価対象のモデルは、代表的なプロプライエタリモデルとオープンソースモデルを用いる. プロプライエタリモデルは実験時点の最新モデルである GPT-4o[1], Claude 3.5 Sonnet<sup>6)</sup>, および Gemini 1.5 Pro[22] である. オープンソースモデルは DeepSeek-VL [23], Mantis [24], MiniCPMv2.6 [25], Pixtral-12B<sup>7)</sup> を使用する. top-p, top-k, 温度などの推論時のパラメータはすべてデフォルト値を指定する. また, “Let’s think step by step.” [26] をプロンプト末尾に付与したプロンプトでも評価する. 実験は各モデル各タスクにつき 3 回行い正解率の平均を最終スコアとする. 人間による評価は 2 名のアノテータによって

実施する. 2 名の評価結果のカップ係数は、表 2 のカラム順に左から 0.760, 0.656, 0.571, 0.759, 0.625, 0.625 であった.

### 4.2 実験結果

各モデルの各タスクの正答率を表 2 に示す.

反対側予測および視点基準を入力した場合の視点推論の人間の結果によると、いずれも正解率は 95% を超えており人間にとっては非常に簡単なタスクであることがわかる. 視点基準を入力しない場合は、視点基準が定まらないため左右や前後に不定性があることから人間の性能は悪化している.

ほとんどのオープンソースモデルの性能はランダム以下である. 一方、プロプライエタリモデルのスコアはばらつきはあるものの、概ね Gemini の性能が低く、Claude と GPT-4o は比較的高性能である. また、CoT プロンプトを用いた場合、一部のモデルでスコアの改善がみられたものの、その効果は小さい.

以下では、Claude の出力を例にいくつかの観点から出力を分析する.

### 言語情報の効果

反対側予測の結果から、人間では言語情報が無くとも容易に正解できるような事例でも、モデルはまずテキストプロンプトから反対側の言語表現を推測し、その結果を手がかりに反対側から見た画像を探す傾向があることがわかった. そこで、言語情報を極力排除するために、クエリ画像の説明文を入力せずに実験を行った. その結果、Claude の正解率は 0.500 であり、人間を大幅に下回る結果となった. この結果は、最先端のモデルであって言語情報に強く依存して推論を行なっていることを示唆する.

### エラー分析

モデルと人間の両方が誤答した事例を図 2 に示す. この例は、製品パッケージの前面画像の反対側を予測するタスクである. 前後左右の側面がほぼ同じサイズの正方形であるため、視覚情報だけでは人間にとっても難しい事例である. Claude (CoT) の振る舞いは、「裏面には栄養成分表示があるだろう」という事前知識を手掛かりにして箱の各側面にプリントされたテキスト情報を読み取り、回答を試みていることが読み取れる. 結果的には不正解だが、こ

5) <https://creativecommons.org/licenses/by/4.0/>

6) <https://www.anthropic.com/news/claude-3-5-sonnet>

7) <https://mistral.ai/news/pixtral-12b/>

		意匠データセット			GSO データセット		
		反対側予測	視点推論	視点推論 (w/o ref.)	反対側予測	視点推論	視点推論 (w/o ref.)
Random		0.333	0.500	0.500	0.333	0.500	0.500
Human		0.960	0.969	0.816	0.957	0.958	0.923
	#params	Open-source models					
deepseek-vl-7b-chat	7B	0.040 (0.051)	0.623 (0.431)	0.506 (0.458)	0.260 (0.207)	0.517 (0.477)	0.500 (0.503)
Mantis-8B-siglip-llama3	8B	0.026 (0.023)	0.500 (0.461)	0.488 (0.446)	0.413 (0.460)	0.543 (0.557)	0.543 (0.460)
MiniCPM-V-2.6	8B	0.382 (0.346)	0.624 (0.595)	0.547 (0.549)	0.467 (0.451)	0.506 (0.577)	0.513 (0.570)
Pixtral-12B-2409	12B	0.704 (0.755)	0.578 (0.560)	0.497 (0.545)	0.487 (0.673)	0.591 (0.577)	0.605 (0.620)
		Proprietary models					
Gemini 1.5 Pro (Accessed in early Oct. 2024)		0.632 (0.567)	0.502 (0.528)	0.414 (0.441)	0.842 (0.807)	0.582 (0.509)	0.582 (0.517)
Claude 3.5 Sonnet (claude-3-5-sonnet-20240620)		0.823 (0.830)	0.761 (0.775)	0.687 (0.691)	0.747 (0.753)	0.647 (0.643)	0.583 (0.586)
GPT-4o (gpt-4o-2024-08-06)		0.888 (0.866)	0.916 (0.901)	0.738 (0.708)	0.893 (0.873)	0.637 (0.727)	0.697 (0.720)

表 2: 各モデル各タスクの正解率。数値は 3 回の試行の正解率の平均を表し ( ) 内の数値は CoT プロンプトの結果であり、赤字は CoT でスコアが向上したことを示し、青字は悪化したことを示す。“視点推論 (w/o ref.)” は視点基準画像とその説明文を入力しない場合を意味する。

のような推論過程は人間によるそれに近い振る舞いである。

## 5 関連研究

提案タスクの類似タスクとして BLINK [4] のサブタスクである multi-view reasoning が近い。このタスクは、同じ物体を 2 つ異なる視点から撮影したときの 2 つの画像の回転方向 (時計周り/反時計周り) を予測するタスクである。また、CLEVR-MRT [14] は、CLEVR データセット [27] に基づき、心的回転能力 [17] を評価する VQA タスクである。PerspectiveTaking タスク [13] は、物の位置や文字 (M および W, 6 および 9) が他者視点からどのように見えるかを問うタスクとして提案されている。このタスクは、別視点からの見え方を問うという意味で我々のタスクに近いが、視点獲得 [18] という心理学的な現象の検証を重視しており、見えない部分の認識能力を測るタスクとは異なる。

## 6 おわりに

我々は、物の見えない部分を想像する能力を評価する新しいタスクを提案し、線画と写真画像を用いてデータセットを作成した。実験では、提案タスクで代表的な VLM を評価し、人間の性能と比較した。実験の結果、提案タスクは人間にとって非常に簡単であるが、最も優れた性能である GPT-4o でさえも人間には劣ることがわかった。我々の提案タスクが、これまで明示的に評価されていなかった VLM の能力の一つの側面を浮き彫りにすることで、将来の VLM の開発を促進することを期待する。



Fig.1

Fig.2

Fig.3

Fig.4

プロンプト:

These images are different views of the same object. The Fig numbers are located near each figure.

The descriptions of Fig.2 and Fig.1 are as follows.

Fig.2 is a front view of the object.

Fig.1 is an image of Fig.2 viewed from the left.

Question: Which figure is the opposite view of Fig.2?

Let's think step-by-step. Please state your conclusion in the first line of your response.

Your answer:

モデル出力:

The opposite view of Fig.2 is Fig.3. ✗

Here's the step-by-step reasoning:

- Fig.2 shows the front view of a Nescafe Dolce Gusto Cappuccino box.
- In packaging, the opposite side of the front is typically the back, which usually contains nutritional information and other details.
- Looking at the other figures:
  - Fig.1 shows a side of the box with some product information and a small image.
  - Fig.3 displays detailed nutritional facts and information in multiple languages.
  - Fig.4 shows an array of colorful capsules, likely representing different flavors.
- Among these, Fig.3 is the only one that matches what we'd expect to see on the back of a food product packaging - comprehensive nutritional information.
- Therefore, Fig.3, containing the nutritional facts, is most likely the opposite view (back) of the front view shown in Fig.2.

図 2: 箱の裏側の予測に失敗した例。CoT により推論過程を出力したところ、モデルは箱側面の記載内容を手掛かりに裏側の予測を試みていることがわかる (ハイライト部分)。



## 謝辞

本研究の一部は JST さきがけ JPMJPR21C8 の支援を受けたものである。

## 参考文献

- [1] OpenAI et al. GPT-4 Technical Report, 2024.
- [2] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.
- [3] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal Foundation Models: From Specialists to General-Purpose Assistants. *arXiv e-prints*, page arXiv:2309.10020, 2023.
- [4] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal Large Language Models Can See but Not Perceive, 2024.
- [5] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pages 9556–9567, 2024.
- [6] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *arXiv e-prints*, page arXiv:2409.02813, 2024.
- [7] Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in Long Context. In **First Conference on Language Modeling**, 2024.
- [8] Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. *arXiv e-prints*, page arXiv:2406.09411, 2024.
- [9] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024.
- [10] Kaziwa Saleh, Sandor Szenasi, and Zoltan Vamossy. Occlusion Handling in Generic Object Detection: A Review. In **2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMII)**. IEEE, 2021.
- [11] Dan Zeng, Raymond N. J. Veldhuis, and Luuk J. Spreeuwens. A survey of face recognition techniques under occlusion. *IET Biom.*, 10:581–606, 2020.
- [12] Rohit Mohan and Abhinav Valada. Amodal panoptic segmentation. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pages 21023–21032, 2022.
- [13] Bridget Leonard, Kristin Woodard, and Scott O. Murray. Failures in perspective-taking of multimodal ai systems, 2024.
- [14] Christopher Beckham, Martin Weiss, Florian Golemo, Sina Honari, Derek Nowrouzezahrai, and Christopher Pal. Visual Question Answering From Another Perspective: CLEVR Mental Rotation Tests, 2022.
- [15] J. Piaget and B. Inhelder. **Psychology Of The Child**. The Psychology of the Child. Basic Books, 1969.
- [16] John H. Flavell. The development of knowledge about visual perception. **Nebraska Symposium on Motivation. Nebraska Symposium on Motivation**, 25:43–76, 1977.
- [17] Roger N. Shepard and Metzler Jacqueline. Mental rotation of three-dimensional objects. **Science**, 171(3972):701–703, 1971.
- [18] Andrew Surtees, Ian Apperly, and Dana Samson. Similarities and differences in visual and spatial perspective-taking processes. **Cognition**, 129(2):426–438, 2013.
- [19] Michal Kucer, Diane Oyen, Juan Castorena, and Jian Wu. DeepPatent: Large scale patent drawing recognition and retrieval. In **2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pages 557–566, 2022.
- [20] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A High-Quality Dataset of 3D Scanned Household Items. In **2022 International Conference on Robotics and Automation (ICRA)**, pages 2553–2560, 2022.
- [21] GoogleResearch. Scanned objects by google research.
- [22] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024.
- [23] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World Vision-Language Understanding, 2024.
- [24] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. MANTIS: Interleaved Multi-Image Instruction Tuning, 2024.
- [25] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM-V: A GPT-4V Level MLLM on Your Phone, 2024.
- [26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, **Advances in Neural Information Processing Systems**, 2022.
- [27] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In **2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, pages 1988–1997, 2017.

## A 参考情報

実験で使ったプロンプトは「複数の入力画像は同じ対象物について異なる視点から見た画像である」ことをモデルに教えることを意図して実験的に決定した。プロンプトの全体を Appendix A.1, Appendix A.2 に示す。太字の部分は事例毎に異なり、それ以外はすべての事例で共通である。一部のオープンソースモデルは仕様によってプロンプトの先頭に `< image >` などの image token を挿入する。

### A.1 反対側予測タスクのプロンプト全体

These images are different views of the same object.  
The Fig numbers are located near each figure.  
The descriptions of Fig.1 and **Fig.6** are as follows.  
**FIG. 1 is a top perspective view of a of a wall-mounted laundry bin cabinet showing my new design. FIG. 6 is a front elevation view of FIG. 1.**  
Question: Which figure is the opposite view of **Fig.6**?  
Please answer in a straightforward manner. Do not include a question in your response.  
(Let's think step-by-step. Please state your conclusion in the first line of your response. )  
Your answer :

### A.2 視点推論タスクのプロンプト全体

The three images are different views of the same object.  
The fig numbers are located near each figure.  
The descriptions of Fig.1 and **Fig.6** are as follows.  
**FIG. 1 is a top perspective view of a of a wall-mounted laundry bin cabinet showing my new design. FIG. 6 is a front elevation view of FIG. 1.**  
Question: Is the following description correct for **Fig.7**?  
**FIG. 7 is a rear elevation view of FIG. 1.**  
If this statement is correct, answer "correct" otherwise, answer "incorrect".  
Please answer in a straightforward manner. Do not include a question in your response.  
(Let's think step-by-step. Please state your conclusion in the first line of your response. )  
Your answer :