日本語 VLM 構築に向けた合成データのフィルタリングの検討

大島 遼祐 [1,2*] 小澤 圭右 [2] 品川 政太朗 [2] 鈴木 哲平 [2]
[1] 早稲田大学 [2] SB Intuitions 株式会社
{ryosuke.oshima,keisuke.ozawa,seitaro.shinagawa,teppei.suzuki}@sbintuitions.co.jp

概要

Vision-Language Model の学習に必要な Visual Instruction Tuning データセットは、コストの観点から主に人手ではなくデータ合成によって作成される。合成データの利用における課題は、合成時に発生する不適切な合成データを取り除く上での正確性と、時間的効率性である。本稿では、CLIP および VLM as a judge を用いたフィルタリングの正確性と効率性について検証する。実験の結果、VLM as a judge は正確性が 17.3%高いが CLIP の方が 62 倍高速に動作すること、これら手法の併用により VLM as a judge 単体に比べ正確性を損なうことなく 36%高速にフィルタリング可能なことがわかった。

1 はじめに

Vision-Language Models (VLM) とは、画像とテキストの指示プロンプトを紐づけ、適切な回答を出力し問題解決を行うモデルの総称である。多くの VLM は画像エンコーダと大規模言語モデル（LLM）が接続された構造を持ち、Visual Instruction Tuning [1, 2] と呼ばれる「画像」、「指示（質問）」、「回答」の 3 つ組で構成されるデータセットを用いて学習される。

この Visual Instruction Tuning 用データセットの構築を人手で行うのは多大なコストと時間がかかるため、LLM や VLM による合成データが一般的に用いられる [1, 3, 4, 5]。特に、日本語を始めとした低資源の言語では、合成データ生成はより重要となる [6]。

しかし、合成されたサンプルには、学習に不適切なサンプルが一定数混入する。例えば、画像と質問に対する誤った回答が含まれているサンプルは、VLM の学習に悪影響を及ぼし、ハルシネーションを助長する恐れがある。より良い VLM の構築には、不適切なサンプルのみを正確にフィルタリングし、できるだけ適切なサンプルで構成された高品質なデータセットを用意することが不可欠である [7, 8]。さらに、大規模な合成データ生成とフィルタリングを実現するためには、不適切なサンプルのみを取り除く正確性だけでなく、金銭的に低コストかつ時間的に高効率な手法が要求される [9, 10]。

そこで本研究では、合成された Visual Instruction Tuning 用データセットの正確かつ効率的なフィルタリングの実現に向け、既存のマルチモーダルなフィルタリング手法を検証する。具体的には、日本語能力を持つ VLM で日本語合成 VQA データセットを作成し、CLIP [11] を用いたフィルタリングと、近年 VQA の評価方法として注目を集めている VLM as a judge [12] の 2 つのフィルタリング方法の有効性・特徴を分析する。また、上記手法を組み合わせることによる効率性の向上についても報告する。

2 日本語合成 VQA の生成

まず、本研究で検証対象とする、日本語による合成 VQA データセットの構築方法について説明する。はじめに、日本各地で撮影された画像データセットである Japan Diversse Images Dataset [1)] の画像から、商用利用が可能かつ軽量で高い日本語能力を持つ Qwen2-VL-7B-Instruct [13] を用いて、直接質問と回答ペア（QA ペア）の生成を行った。具体的には、サンプル画像 126 枚を用い、各画像について 200 個ずつ QA ペアを出力させ、そのうち「質問:{question}」かつ「回答:{answer}」の形式として抽出できたサンプルからランダムに 25 個ずつ選択し、合計 3150 個の VQA サンプルからなる合成データセットを構築した（詳細は、付録 B を参照）。この時、Qwen2-VL-7B-Instruct の温度パラメータは $T = 1.0$ に設定した [2)]。

次に、先行研究 [8] にならい、CLIP や VLM as a

* SB Intuitions 株式会社でのインターンシップ中の成果。

1） https://huggingface.co/datasets/ThePioneer/japanese-photos
2） 先行研究 [14] を参考に、多様性を重視しつつも日本語としてある程度自然な文を生成できる値として経験的に設定したもので、検証の余地は残されている。

judge によるフィルタリングの前処理として、テキストフィルタリングを行った。本研究では、2 つの条件「1. 単一サンプルに 2 組以上の QA ペアが含まれる」「2. 日本語テキスト以外（例：中国語や英語）の QA ペアである」のうち、少なくとも一方を満たすサンプルを除外し、残った 2453 サンプルを用いてフィルタリングの検証実験を行なった。

3 フィルタリング

本章では、2 章で作成した合成データセットに適用した、フィルタリング手法について説明する。

3.1 CLIP フィルタリング

画像テキストペアの整合性を評価する有名な方法として、CLIP モデルによる画像と QA ペアの埋め込み ($\mathbf{v}$ と $\mathbf{t}$) を用いて類似度スコア $\cos(\mathbf{v}, \mathbf{t})$ を計算し、閾値以下のサンプルを除外する方法が知られている [15, 16]。本研究では、日本語性能の優れた CLIP モデルとして clip-japanese-base[3] を採用した。

3.2 VLM as a judge フィルタリング

VLM as a judge [12] とは、VLM に画像と質問と評価対象の回答をプロンプトとして与え、回答の品質を評価する手法である。本研究ではこれをフィルタリング手法として採用する。具体的には、質問と回答の品質に関する合計 12 個の基準をプロンプトとして与え、すべての基準を満たすか否かを VLM に 2 値分類問題として出力させる（具体的な品質の基準・モデルに与えたプロンプトは、付録 C 参照）。

4 実験設定

4.1 フィルタリングの評価方法

参照値ラベルの作成 本研究では、フィルタリングが適切に作用しているかどうかを評価するために、GPT-4o により画像と QA ペアが整合するか否かをラベル付けして、これを参照値（仮の真値）として扱うこととした。GPT-4o によるラベル付けは現状最高性能の VLM as a judge フィルタリングに相当する[4]。著者一名が、ランダムに 100 サンプルを抽出して人手で付与したラベルと GPT-4o で付与したラベルを比較したところ、90%程度が一致したため、この参照値ラベルは正確性と効率性の議論に十分であると判断した。

まず、2 章で作成した合成 VQA に対して、GPT-4o（gpt-4o-2024-08-06）を用いた VLM as a judge フィルタリング操作（3.2 節）により、それぞれ 2 値ラベルを付与した。結果、全 2453 サンプル中 936 個が整合サンプル、1513 個が不整合サンプルとなった。残りの 4 サンプルは、GPT-4o の出力からのラベルの抽出ができなかった。本研究では、ラベルの抽出ができた 2453 サンプルを実験に用いた。

評価指標 上記ラベルを参照値とし、F 値・Precision・Recall の評価指標をもって、3 章で説明した 2 つのフィルタリング手法を評価する。Precision が高いほど抽出したデータプールに不整合サンプルが含まれる割合が少なく、Recall が高いほど整合サンプルをフィルタリングで排除せず、うまくデータプールに残せていることを意味する。また、正解率ではなくこれらの評価指標を用いたのは、本合成データセットの整合サンプルが全体の 38% 程度であった不均衡さを考慮してのことである。

4.2 VLM as a judge で用いたモデル

1 章で述べたように、合成データの生成・フィルタリングでは、低コストかつ効率的な方法が求められる。そこで本研究では、オープンソースかつ軽量な VLM を用いた。具体的には、Qwen2-VL-Instruct [13] の 2B・7B サイズのモデル、InternVL-2.5 [5] の 2B・4B・8B サイズのモデル、LLaVA-OneVision [17] の 0.5B, 7B サイズのモデルを用いた。また、日本語 VLM として、日本語 VQA ベンチマーク [18] で高いスコアを示す Llama-3-EvoVLM-JP-v2[6] [19]、llava-calm2-siglip [7] も用いた（表 1）。LLM-jp-3 VILA 14B [8] は、本研究の合成データの元となる Japan Diverse Images Dataset を使用して学習しているため、検証対象に含めなかった。

5 実験結果

5.1 CLIP フィルタリング

図 1 に、CLIP による画像と QA ペアの類似度スコアの分布を示す。整合サンプルの分布の大部分が不

3） https://huggingface.co/line-corporation/clip-japanese-base
4） GPT-4o による直接的な VLM as a judge フィルタリングを適用した合成データセット構築は金銭的コストや商用利用上ライセンスの問題が生じるため、本研究の研究対象としては扱わなかった。
5） https://internvl.github.io/blog/2024-12-05-InternVL-2.5/
6） https://huggingface.co/SakanaAI/Llama-3-EvoVLM-JP-v2
7） https://huggingface.co/cyberagent/llava-calm2-siglip
8） https://huggingface.co/llm-jp/llm-jp-3-vila-14b

フィルタリング方法	F 値	Precision	Recall
CLIP フィルタリング	0.606	0.478	0.826
VLM (Qwen2-VL-2B-Instruct)	0.625	0.469	0.936
VLM (Qwen2-VL-7B-Instruct)	0.626	0.457	0.990
VLM (InternVL2.5-2B)	0.584	0.413	0.995
VLM (InternVL2.5-4B)	0.627	0.461	0.981
VLM (InternVL2.5-8B)	**0.711**	**0.572**	0.938
VLM (LLaVA-OneVision-0.5B)	0.554	0.386	0.985
VLM (LLaVA-OneVision-7B)	0.575	0.403	**0.998**
VLM (Llama-3-EvoVLM-JP-v2)	0.561	0.391	0.997
VLM (llava-calm2-siglip)	0.000	0.000	0.000

表 1　CLIP または VLM as a judge によるフィルタリング結果のスコア。VLM は、VLM as a judge フィルタリングを意味する。CLIP フィルタリングによるスコアは、F 値が最高となる閾値 $\theta = 0.5536$ でのスコアを示す。

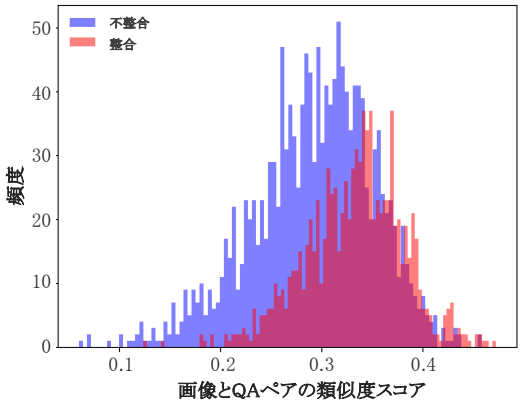


図 1　CLIP の類似度スコアのヒストグラム

整合サンプルの分布に含まれていることから、CLIP フィルタリングでは、整合・不整合サンプルの分類を行う事が困難であることが示唆される。実際、表 1 からわかるように、CLIP フィルタリングのスコアは F 値が最高で 0.606、Precision も 0.478 と低いことから、多くの不整合サンプルを除外できていないことがわかる。

図 2 に、参照値は不整合であるにも関わらず、CLIP による類似度スコアが高いサンプルの例を示す。生成された回答では、「虹色の石鉄」という理解困難な単語が含まれているのにも関わらず、類似度は高くなっている。これは、回答に含まれるその他の「電車」「鉄道」「富士山」といった単語が画像と類似性が高いためであると考えられる。このように、テキスト文の一部分のみに内容の間違いがあるサンプルは CLIP では除外しきれないことがわかった。しかし、図 1 から、類似度スコアが約 0.20 以下のサンプルの大部分が不整合であることから、CLIP は足切りの役割としては利用可能なことがわかる。

5.2　VLM as a judge フィルタリング

表 1 の 2 行目以降に、各モデルの VLM as a judge フィルタリング結果のスコアを示す。この表から、全てのモデルの中で、InternVL2.5-8B が最も F 値が高くフィルタリング性能が良いことがわかる。また Recall は 0.938 であり、整合サンプルのほとんどを整合と正しく予測できている。これは、多種多様なデータで学習するために学習に適切なサンプルはできるだけ残したい合成データフィルタリングにとって望ましい特徴である。一方で、Precision は 0.572 であり、フィルタリング後のデータの約半数が不整合サンプルであることがわかる。

InternVL2.5-8B の偽陽性サンプル（すなわち不整合サンプルを誤って整合と判断したサンプル）を人手で確認したところ、主に「画像から類推不可能な情報を持つサンプルの指摘失敗」「画像認識能力の欠如」「日本語テキスト単体の認識能力の欠如」「日本に関する知識不足による見過ごし」の 4 種類が確認された。図 3 に、「画像認識能力の欠如」が原因と思われる、偽陽性の例を示す（他の種類の原因については、付録 A を参照。）。生成された回答には、「大きなシンボルマーク」や「デジタルサイネージ」といった画像に含まれない物体について言及されているが、InternVL2.5-8B はこの間違いを指摘せず、整合と予測してしまっている。

また、表 1 から日本語 VLM はその他の VLM に比べて F 値が低いことがわかる。特に llava-calm2-siglip は、全てのサンプルを不整合と予測したため、F 値、Recall、Precision が全て 0.0 となっている。これらの原因は、日本語 VLM はその他の VLM に比べて複雑な指示に対する追従能力が欠けているためだと考えられる。

5.3　CLIP と VLM as a judge の併用

本節では、CLIP と VLM as a judge を併用することを考える。表 1 から、InternVL2.5-8B を用いた VLM as a judge は、CLIP より高い正確性を持つことがわかるが、モデルサイズ、エンコーダ・デコーダモデルの違いのため実行時間は長くなる。実際に、InternVL2.5-8B を用いた VLM as a judge フィルタリングは 1 つのサンプルあたり 3.82 秒かかるが、CLIP は 0.0614 秒ですむ。従って、効率性の観点から、まず CLIP フィルタリングを行った後に除外されなかったサンプルに対して VLM as a judge を行う、と

Q: 画像には何が映っていますか？
A: 画像には電車と虹色の石鉄が鉄道に並んでおり、その背後に富士山が見えます。
参照値: 不整合　CLIPによる類似度スコア: 0.4355

図 2　参照値は不整合であるが、CLIP の類似度スコアが高いサンプル例。画像と矛盾する部分を赤字にしている。

Q: この写真は何を示していますか
A: これは夜に撮影された狭い通りを示しています。
通りの左側には大きなシンボルマークが点在しています。
途中に映っているデジタルサイネージもあります。
参照値: 不整合　InternVL2.5-8B: 整合

図 3　InternVL2.5-8B を用いた VLM as a judge によるフィルタリングにおける、偽陽性のサンプル例。画像と矛盾する部分を赤字にしている。

いった多段フィルタリングが望ましい。

そこで本研究では、 5.1 節の CLIP フィルタリングの結果と 5.2 節の InternVL2.5-8B の結果を利用し、多段フィルタリングのシミュレーション実験を以下の手順で行った。

1. CLIP フィルタリングにおける類似度の閾値 θ を決める。θ 未満のサンプルを不整合、θ 以上のサンプルを整合とする。
2. 1. の結果と 5.2 節の VLM as a judge フィルタリング結果が両者とも整合の場合、多段フィルタリングによる結果を整合サンプルとし、それ以外のものを不整合サンプルとする。
3. 1. で設定する θ を 0 から 0.472（類似度の最高値）まで変化させ、各閾値における多段フィルタリングの結果を模倣する。

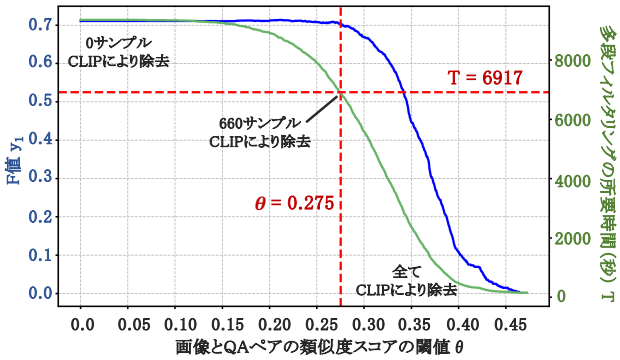


図 4　多段フィルタリングのシミュレーション結果

図 4 に多段フィルタリングのシミュレーション結果を示す。注意点として、閾値 θ は大きい方がより多くのサンプルを CLIP フィルタリングで除外でき、VLM as a judge を適用するサンプル数が減るため、実行時間は短くなる。しかし、InternVL2.5-8B による VLM as a judge の方が CLIP フィルタリングよりもフィルタリングの正確性が高いため、偽陰性のサンプル数が増え多段フィルタリングの性能が落ちる。従って、できるだけ大きい閾値 θ で F 値（正確性）を維持できるのが望ましい。

図 4 から、閾値 θ が 0.0 から約 0.275 の範囲では F 値は横ばいだが、その値を超えると急激に減少することがわかる。これは、本研究の問題設定において、閾値 $\theta = 0.275$ が正確性を損なわずに効率化できる限界であることを意味する。図 4 に示すように、限界値である $\theta = 0.275$ では、660 個（26.9%）のサンプルを CLIP フィルタリングで除去できることで、VLM as a judge 単体でのフィルタリング所要時間に比べて 1.36 倍短縮できており、フィルタリングの一定の効率化を実現できていることがわかる。

6　まとめ

本研究では、既存のマルチモーダルなフィルタリング手法について、日本語合成 VQA データセットのフィルタリングにおける正確性と効率性の観点から分析した。CLIP または VLM as a judge によるフィルタリング手法について検討し、これらの手法は偽陽性サンプルが多く発生してしまうことを確認した。また、正確性を維持しつつ時間的効率を上げるため、CLIP と VLM as a judge を併用した多段フィルタリングを検討した。今後の展望として、日本語 LLM を用いたフィルタリングなど異なるフィルタリング手法の検証を考えている。

参考文献

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.

[2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 26296–26306, June 2024.

[3] Zhuowan Li, Bhavan Jasani, Peng Tang, and Peng Tang. Synthesize step-by-step: Tools, templates and llms as data generators for reasoning-based chart vqa. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2024.

[4] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In **Proceedings of the European Conference on Computer Vision (ECCV)**, 2024.

[5] Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning, 2023.

[6] Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. Constructing multimodal datasets from scratch for rapid development of a japanese visual language model, 2024.

[7] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[8] Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. Synthvlm: High-efficiency and high-quality synthetic data for vision language models, 2024.

[9] Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. Finetuned multimodal language models are high-quality image-text data filters. 2024.

[10] Patrick Amadeus Irawan, Genta Indra Winata, Samuel Cahyawijaya, and Ayu Purwarianti. Towards efficient and robust vqa-nle data generation with large vision-language models, 2024.

[11] 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, 2021.

[12] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In **ICML**, 2024.

[13] 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 preprint arXiv:2409.12191**, 2024.

[14] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing, 2024.

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In **Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track**, 2023.

[16] Anonymous. A simple baseline for zero-shot visual question answering via synthetic data generation. In **Submitted to ACL Rolling Review - June 2024**, 2024. under review.

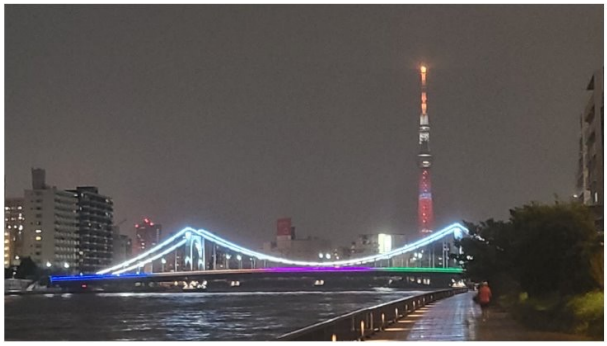
[17] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. **arXiv preprint arXiv:2408.03326**, 2024.

[18] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. **arXiv preprint arXiv:2410.17250**, 2024.

[19] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes, 2024.

A　VLM as a judge の偽陽性の例

InternVL2.5-8B の「日本に関する知識不足による見過ごし」による偽陽性の例を示す。スカイツリーは築地地区にはないが、InternVL2.5-8B は指摘する事ができず整合と出力してしまっている。



Q: この写真のスカイツリーはどこですか？
A: この写真のスカイツリーは、日本のトーキョー築地地区にある、東京スカイツリーです。
参照値: 不整合　InternVL2.5-8B:整合

図 5　日本に関する知識不足による見過ごしの例

B　VQA の合成におけるプロンプト

Magpie [14] を参考に、以下のシステムプロンプトのみを指定し、空のユーザプロンプトから質問と回答を含む文を生成した。

システムプロンプト

あなたは画像についての指示に正確に回答する AI アシスタントです。与えられた画像に基づいて、日本語で質問と回答のペアを作成してください。

質問は、画像の内容に関連し、多様である必要があります。それぞれの質問に対する回答は、画像から導き出せる正確なものである必要があります。回答は簡潔にしてください。

質問と回答のフォーマットは以下のようにしてください：
質問: { ここに質問 } 回答: { ここに回答 }

C　VLM as a judge でのプロンプト

日本語 VLM に与えたプロンプトを以下に示す。GPT-4o を含めた日本語 VLM 以外のモデルには、これを英訳したプロンプトを与えた。

ユーザプロンプト

あなたは、VQA（Visual Question Answering）のサンプルの品質の評価者です。「画像」と「画像に基づいた質問と回答のペア」が与えられた場合、次の評価基準に基づいてそれらが適切であるかどうかを評価してください。
評価基準:
質問の妥当性:
　文法エラー: 日本語の文法に間違いはあるか。
　繰り返し: 不必要に言葉やフレーズが繰り返されているか。
　冗長性: 質問が不必要に長くないか。簡潔で適切な長さか。
　二重サンプル: 質問と回答のペアが複数含まれていないか。
　矛盾: 質問が画像と矛盾していないか。また、質問は答えが存在する形で作成されているか。
　画像依存性: 質問は、画像を見なければ答えられないものになっているか。
回答の妥当性:
　文法エラー: 日本語の文法に間違いはあるか。
　繰り返し: 不必要に言葉やフレーズが繰り返されていないか。
　冗長性: 質問が不必要に長くないか。簡潔で適切な長さか。
　二重サンプル: 質問と回答のペアが複数含まれていないか。
　正確性: 回答は、画像と質問に基づいて正しいか。
　一般知識と画像依存性: 回答は、一般知識と画像から得られる情報を基に導き出せる内容か。画像や一般知識から類推できない余計な情報が含まれていないか。

評価対象サンプル:
質問: {question}
回答: {answer}
では、はい/いいえで評価してください。
このサンプルは基準を満たしていますか？
はい/いいえ:
評価した後に、そう考える理由も述べてください。
理由: