オープンな VLM を活用した 日本語マルチモーダル指示データセットの構築

中尾 圭佑 [1] 栗田 修平 [2,3] 河原 大輔 [1,3]
[1] 早稲田大学 [2] 国立情報学研究所 [3] NII LLMC
{keisuke.nakao@akane., dkw@}waseda.jp skurita@nii.ac.jp

概要

日本語に特化した視覚言語モデル（VLM）の指示追従能力を高めるには，高品質な日本語のマルチモーダル指示データが不可欠である．既存の日本語指示データの多くは，ライセンスによる制限がある大規模言語モデル（LLM）を用いて合成されており，視覚情報の欠落による品質低下やライセンス制約の問題がある．本研究では，オープンな VLM を用いて自然画像を対象とした 2 種類の日本語マルチモーダル指示データセットを構築する．画像を参照して合成することで，視覚情報に基づく正確な記述を実現し，オープンな VLM を用いることで，ライセンスの問題を解決する．実験の結果，構築したデータセットで指示チューニングした日本語 VLM は，ライセンス制限付き指示データセットを学習に用いた既存の日本語 VLM と同等の性能を達成し，既存のオープンな日本語 VLM を上回る性能を示した．

1 はじめに

日本語に特化した視覚言語モデル（VLM）の指示追従能力を高めるには，高品質な日本語のマルチモーダル指示データが不可欠である．しかし，高品質な日本語のデータ量は英語と比べて限られており，その確保が性能向上の障壁となっている．

従来はクラウドソーシングなどを活用し，人手によるデータ作成が行われていた [1] が，コストと時間を要する問題があった．近年では，大規模言語モデル（LLM）の急速な発展を背景に，LLM を用いて指示データを合成する手法が主流になりつつある．代表的な手法として，LLaVA [2, 3] では，人手で付与されたテキスト情報（キャプション，物体ラベルと Bounding Box の組）を GPT-4 [4] に入力し，英語指示データを合成している．日本語指示データの構築においても，英語指示データの機械翻訳 [5] や LLaVA と同様の方法による合成 [6] が行われている．

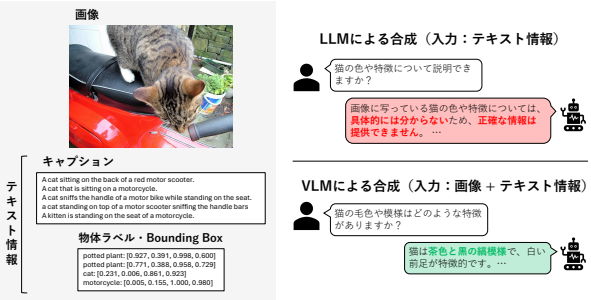

図 1 画像の有無による合成指示データの品質差．LLM による合成（右上）では，テキスト情報のみに依存するため品質が低下するのに対し，VLM による合成（右下）では，画像も参照することで正確に記述できている．

しかし，LLM による合成や機械翻訳などの既存手法には，二つの課題がある．第一に，視覚情報の欠落に起因する品質の限界である．生成時に画像を参照しないため，品質の低下やハルシネーションが生じる懸念がある [7]．第二に，ライセンスの制約である．公開されている既存の日本語指示データセットの多くは，ライセンスによる制限があるモデルを用いて構築されている．そのため，それらを学習に用いた VLM もライセンスの制約を受け，自由な活用が妨げられる可能性がある．

視覚情報の欠落の課題に対し，VLM を用いて画像から詳細な英語キャプションを合成する手法が提案されている [7]．日本語指示データの構築においても，VLM で合成した英語キャプションを LLM で日本語に翻訳する手法が報告されている [8] が，翻訳時に画像を参照しないことから，同様の課題が残る．また，VLM を用いて画像から直接日本語指示データを合成する取り組みもある [6] が，ライセンスによる制限がある VLM が用いられている．

本研究では，オープン[1)] な VLM を用いて，自然

1） 本研究では，Apache や Creative Commons など，利用規約による制約が最小限のライセンスの下で公開されていることを「オープン」と定義する．

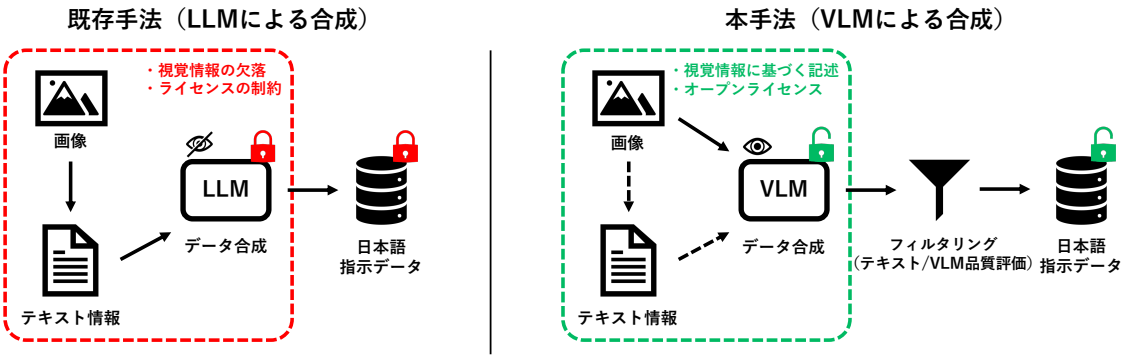

図 2　日本語マルチモーダル指示データセットの構築方法.

画像を対象とした 2 種類の日本語マルチモーダル指示データセットを構築する．画像を参照して合成することで，視覚情報に基づくより詳細かつ正確な記述が可能となる（図 1）．また，オープンな VLM を用いることで，ライセンス制約の問題を解決する．さらに，これらのデータセットを用いて指示チューニングを行い，オープンな日本語 VLM を構築する．

実験の結果，構築した日本語 VLM は，ライセンス制限付き指示データセットを学習に用いた既存の日本語 VLM と同等の性能を達成した．また，既存のオープンな日本語 VLM を上回る性能を示し，構築した日本語マルチモーダル指示データセットの有用性，ならびに，オープンかつ高性能な日本語 VLM の実現可能性を実証した．構築したデータセットは一般に公開している[2)]．

2　日本語指示データセット構築

自然画像を対象に，日本語版 LLaVA-Instruct，Japanese-photos 対話の 2 種類のオープンな日本語マルチモーダル指示データセットを構築する．データセットの各事例は，画像と，その内容に関する複数の質問と回答の組（QA ペア）で構成される．

図 2 右に，構築方法の概要を示す．まず，オープンかつ高い視覚言語能力を持つ VLM（Qwen3-VL-235B-A22B-Instruct [13]）を用いて，日本語指示データを合成する．次に，指示データの品質を高めるために，テキストまたは VLM による品質評価（VLM-as-a-judge [14]）に基づいたフィルタリングを行う．テキストフィルタリングでは，日本語以外のテキストが含まれる事例や，指定した書式を満たさない事例を除外する．VLM-as-a-judge フィルタリングでは，VLM を用いて QA ペアの品質評価を行い，基準を満たさないペアを除外する．

データの特性に応じて，これら 2 種類のフィルタリングを使い分ける．豊富な入力情報をもとに合成する日本語版 LLaVA-Instruct には，軽量なテキストフィルタリングを適用する．一方，画像のみを入力して合成する Japanese-photos 対話には，より高度な VLM-as-a-judge フィルタリングを適用する．

2.1　日本語版 LLaVA-Instruct

日本語版 LLaVA-Instruct は，一般画像を用いた日本語指示データセットであり，155,657 事例で構成される．

指示データの合成　LLaVA の指示データの構築方法を参考に，VLM を用いて日本語指示データを合成する．具体的には，COCO [15] の画像と，人手で付与されたテキスト情報（キャプション，物体ラベルと Bounding Box の組）を VLM に入力し，指示データを生成する．LLaVA の構築方法との相違点は，テキスト情報だけでなく，画像も入力として利用する点である．画像も参照することで，視覚情報がより正確に反映され，指示データの品質向上が期待される．推論パラメータは，モデルのデフォルト設定を用いる．

テキストフィルタリング　まず，日本語以外のテキストが含まれる事例や，指定された書式を満たさない事例を抽出する．次に，抽出された事例について，指示データを再度合成する．これらの手順を数回繰り返し，それでも条件を満たさない事例は除外する．フィルタリングの結果，156,564 事例のうち，907 事例が除外された．

2) • https://huggingface.co/datasets/llm-jp/llava-instruct-ja-qwen3vl
• https://huggingface.co/datasets/llm-jp/japanese-photos-conversation-qwen3vl

表 1　モデルの日本語ベンチマーク性能比較．本モデル以外のスコアは，llm-jp-eval-mm の Leaderboard の値を引用した．“LLM” は gpt-4o-2024-11-20 を用いた LLM-as-a-Judge による評価結果を表す．† 印は，ライセンス制限付きデータを学習に用いていないオープンなモデルを示す．**太字**は GPT-4o 以外のモデルの中で最高のスコアを，下線はオープンなモデルの中で最高のスコアを意味する．

		Heron-Bench	JA-VLM-Bench-ItW		JA-VG-VQA-500	
モデル	LM サイズ	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
Japanese InstructBLIP Alpha† [9]	7B	23.5	14.2	2.3	–	–
Japanese Stable VLM† [10]	7B	48.4	23.2	3.3	–	–
LLaVA-CALM2-SigLIP† [11]	7B	54.1	46.3	3.7	17.7	3.6
Asagi 14B† [8]	13B	41.9	30.9	2.9	9.3	2.0
LLM-jp-3 VILA [6]	13B	68.0	**52.4**	**4.1**	16.2	3.9
本モデル †	13B	**73.4**	45.5	4.0	**19.8**	**4.0**
GPT-4o [12]	–	93.7	32.2	4.4	11.8	3.9

表 2　QA ペアの品質評価基準．

質問の評価基準	
流暢性	質問が自然な文章であるか。
簡潔性	質問が簡潔で適切な長さであるか。
正確性	質問が画像に基づいて正しく、回答可能か。
明瞭性	質問の内容が明瞭であるか。
画像依存性	質問が画像を参照しなければ答えられない内容か。
回答の評価基準	
流暢性	回答が自然な文章であるか。
簡潔性	回答が簡潔で適切な長さであるか。
正確性	回答が画像と質問に基づいて正しいか。
整合性	回答が質問に対して整合性のある内容か。
一般知識と画像依存性	回答が一般常識と画像から得られる情報を基に導き出せる内容か。

2.2　Japanese-photos 対話

Japanese-photos 対話は，日本で撮影された画像を用いた対話形式の日本語指示データセットであり，11,809 事例で構成される．

指示データの合成　日本で撮影された画像データセット japanese-photos [16] を用いて，日本語指示データを合成する．画像のみを VLM に入力し，zero-shot で QA ペアを 3-5 個生成する．推論パラメータは，モデルのデフォルト設定を用いる．合成に使用したプロンプトは，付録 D.1 に示す．

VLM-as-a-judge フィルタリング　評価用 VLM に画像と QA ペアを与え，表 2 に示す計 10 個の品質基準それぞれについて，基準を満たしているか否かを 2 値分類させる．すべての基準を満たす QA ペアのみを指示データとして採用し，それ以外を除外する．評価用 VLM には，高い視覚言語能力を持つことから，データ合成と同一のモデルを採用し，自己評価を行う．出力の多様性を抑制するため，温度パラメータは 0 に設定する．VLM-as-a-judge に用いたプロンプトは，付録 D.2 に示す．フィルタリングの結果，58,832 個の QA ペアのうち，201 個が除外された．VLM-as-a-judge フィルタリングの性能を評価した結果，低品質 QA ペアの検出において一定の Precision を示す一方で，Recall は極めて低く，改善の余地がある（付録 C）．

3　実験

構築した日本語指示データセットの有用性を検証するために，これらのデータセットを用いて VLM を指示チューニングし，その性能を日本語ベンチマークにより評価する．

3.1　モデルの指示チューニング

LLM-jp-3 VILA [6] をベースとし，ライセンス制限のない学習データのみを用いてオープンな日本語 VLM を構築した．モデル構成および学習手順は，LLM-jp-3 VILA に準拠する．モデル構成は，LLaVA のアーキテクチャに倣い，画像エンコーダ（SigLIP [17]）と LLM（llm-jp-3-13b-instruct [18]）をプロジェクタ（2 層の MLP）で接続する方式である．学習手順は，VILA [19] に倣い 3 段階で構成され，最終段階で指示チューニングを行う．LLM-jp-3 VILA では，この段階において GPT-4o [12] により合成したライセンス制限付き指示データセット（LLaVA-Instruct-150K [2, 3]，llava-instruct-ja [20]，japanese-photos-conversation [21]）を用いている．これらを新たに構築したオープンな指示データセットに置き換え [3]，学習を行った．

3）　英語指示データセットの LLaVA-Instruct-150K も，日本語版 LLaVA-Instruct と同様の方法で構築し，置き換えた．

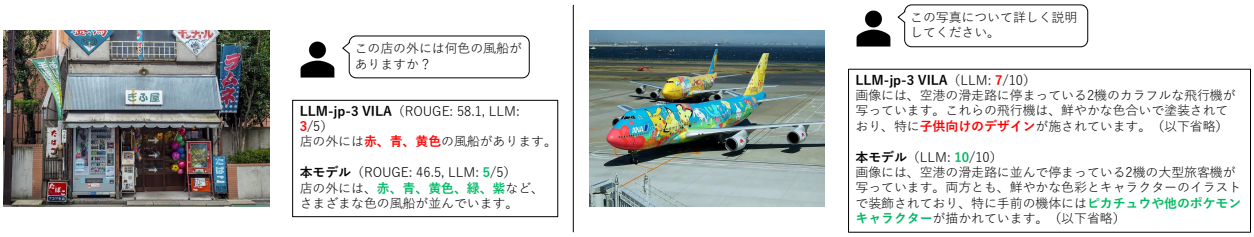

図 3　生成結果の比較例（左: JA-VLM-Bench-ItW，右: Heron-Bench）．

表 3　指示データのアブレーション結果．“LLM” は gpt-4o-2024-11-20 を用いた LLM-as-a-Judge による評価結果を表す．Heron-Bench における Complex，Conv，Detail は，それぞれ複雑な推論，単純な会話形式，詳細説明の質問カテゴリを指す．**太字**は最高のスコアを意味する．

	Heron-Bench				JA-VLM-Bench-ItW	JA-VG-VQA-500
モデル	Complex (%)	Conv (%)	Detail (%)	Overall (%)	LLM (/5.0)	LLM (/5.0)
本モデル w/o 日本語版 LLaVA-Instruct	58.9	73.9	59.3	64.1	**4.0**	**4.1**
本モデル w/o Japanese-photos 対話	67.4	75.4	**74.4**	72.4	3.7	4.0
本モデル	**67.6**	**78.7**	73.7	**73.4**	**4.0**	4.0

3.2　評価

ベンチマーク評価　Heron-Bench [5]，JA-VLM-Bench-In-the-Wild [22]（以下 JA-VLM-Bench-ItW と略記），JA-VG-VQA-500 [1] の 3 つの日本語ベンチマークを用いて，構築したモデルの性能を評価した．評価には，自動評価ツール llm-jp-eval-mm [23] を利用した．表 1 に，モデルの日本語ベンチマーク性能比較を示す．構築したモデルは，ベースとした LLM-jp-3 VILA と遜色ない性能を示した．また，既存のオープンな日本語 VLM を上回る性能を示した．

事例分析　図 3 に，構築したモデルと LLM-jp-3 VILA の生成結果の比較例を示す．図 3 左の例では，LLM-jp-3 VILA は赤，青，黄の 3 色のみを挙げているのに対し，本モデルは緑や紫など，その他の風船の色も認識できている．図 3 右の画像の詳細説明を求める例では，LLM-jp-3 VILA は「子供向けのデザイン」という抽象的な表現に留まっている一方で，本モデルは「ピカチュウ」や「ポケモン」といった具体的な固有名詞を特定しており，より詳細かつ正確に描写できている．

指示データのアブレーション分析　構築した 2 種類の日本語指示データセットが最終的なモデルの性能に与える影響を明らかにするために，アブレーション分析を行った．各データセットを指示データから除いた際のモデル性能の変化を調べた．その結果を表 3 に示す．

日本語版 LLaVA-Instruct を指示データから除いた場合，Heron-Bench の全てのカテゴリでスコアが低下した．特に，Complex と Detail カテゴリの性能低下が顕著であった．このことから，日本語版 LLaVA-Instruct は，汎用的な推論・描写能力の強化に寄与していると考えられる．JA-VLM-Bench-ItW や JA-VG-VQA-500 のスコアに大きな影響がなかった原因として，これらのベンチマークが短答形式であることが挙げられる．

Japanese-photos 対話を指示データから除いた場合，Heron-Bench の Conv カテゴリや，JA-VLM-Bench-ItW のスコアが低下した．これらはいずれも日本ドメインの画像を対象としたベンチマークであることから，Japanese-photos 対話は日本ドメインへの適応に寄与していると考えられる．

4　おわりに

本研究では，オープンな VLM を用いて，自然画像を対象とした 2 種類の日本語マルチモーダル指示データセットを構築した．実験の結果，構築したデータセットで指示チューニングした日本語 VLM は，ライセンス制限付き指示データセットを学習に用いた既存の日本語 VLM と同等の性能を達成した．また，既存のオープンな日本語 VLM を上回る性能を示し，構築した日本語マルチモーダル指示データセットの有用性，ならびに，オープンかつ高性能な日本語 VLM の実現可能性を実証した．今後，VLM-as-a-judge フィルタリングの性能向上や，日本ドメインの画像を用いたより多様で大規模な日本語マルチモーダル指示データセットの構築に取り組みたい．

謝辞

本研究は，文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けた．また，本研究の実施にあたっては，データ活用社会創成プラットフォーム mdx [24] を利用するとともに，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」の支援を受けて利用した．

参考文献

[1] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In **COLING**, pp. 1918–1928. Association for Computational Linguistics, 2018.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, Vol. 36, pp. 34892–34916, 2023.

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **CVPR**, pp. 26296–26306, 2024.

[4] OpenAI. GPT-4 technical report, 2024.

[5] Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. Heron-Bench: A benchmark for evaluating vision language models in japanese, 2024.

[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. In **NAACL (System Demonstrations)**, pp. 470–484, 2025.

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. In **ECCV**, pp. 370–387. Springer, 2024.

[8] 上原康平, 黒瀬優介, 安道健一郎, Chen Jiali, Gao Fan, 金澤爽太郎, 坂本拓彌, 竹田悠哉, Yang Boming, Zhao Xinjie, 村尾晃平, 吉田浩, 田村孝之, 合田憲人, 喜連川優, 原田達也. Asagi: 合成データセットを活用した大規模日本語 vlm. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[9] Makoto Shing and Takuya Akiba. Japanese instruct-blip alpha, 2023. https://huggingface.co/stabilityai/japanese-instructblip-alpha.

[10] Makoto Shing and Takuya Akiba. Japanese stable vlm, 2024. https://huggingface.co/stabilityai/japanese-stable-vlm.

[11] Aozora Inagaki. llava-calm2-siglip, 2024. https://huggingface.co/cyberagent/llava-calm2-siglip.

[12] OpenAI. GPT-4o system card, 2024.

[13] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. **arXiv preprint arXiv:2511.21631**, 2025.

[14] 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.

[15] 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. Springer, 2014.

[16] ThePioneer. Japan diverse images dataset, 2024. https://huggingface.co/datasets/ThePioneer/japanese-photos.

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **ICCV**, pp. 11975–11986, 2023.

[18] LLM-jp. llm-jp-3-13b-instruct, 2024. https://huggingface.co/llm-jp/llm-jp-3-13b-instruct.

[19] 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.

[20] LLM-jp. llava-instruct-ja, 2024. https://huggingface.co/datasets/llm-jp/llava-instruct-ja.

[21] LLM-jp. japanese-photos-conversation, 2024. https://huggingface.co/datasets/llm-jp/japanese-photos-conversation.

[22] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. **Nature Machine Intelligence**, Vol. 7, No. 2, pp. 195–204, 2025.

[23] 前田航希, 杉浦一瑳, 小田悠介, 栗田修平, 岡崎直観. llm-jp-eval-mm: 日本語視覚言語モデルの自動評価基盤. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[24] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[25] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. **arXiv preprint arXiv:2508.18265**, 2025.

表 4　構築したデータセットの統計.

データセット	ユニーク画像数	事例数	QA ペア数
日本語版 LLaVA-Instruct	80,958	155,657	386,229
Japanese-photos 対話	11,809	11,809	58,631

表 5　評価に用いた日本語ベンチマークの特性比較.

ベンチマーク	画像ドメイン	正解文の長さ
Heron-Bench	日本ドメイン	1 文〜複数文
JA-VLM-Bench-ItW	日本ドメイン	1 文
JA-VG-VQA-500	一般ドメイン	1 フレーズ程度

A　構築したデータセットの統計

表 4 に，構築した日本語マルチモーダル指示データセットの統計を示す.

B　日本語ベンチマークの特性比較

表 5 に，評価に用いた 3 つの日本語ベンチマークの特性比較を示す．画像ドメインや，正解文の長さが異なる.

C　オープンな VLM による VLM-as-a-judge フィルタリングの性能

Japanese-photos 対話の構築時に合成した 58,832 個の QA ペアを対象に，オープンな VLM による VLM-as-a-judge フィルタリングの性能を評価した．本実験では，Qwen3-VL-235B-A22B-Instruct および InternVL3.5-241B-A28B-Instruct [25] の 2 つのモデルを評価対象とした.

真値ラベルとして，gpt-4o-2024-11-20 の品質評価結果を用いた．同モデルにより低品質と判定された QA ペアは 2,616 個（全体の約 4.45%）であった．低品質 QA ペアが少数である不均衡なデータであることを考慮し，「低品質」を正例として Precision，Recall および F 値を算出した.

評価結果を表 6 に示す．オープンな VLM による VLM-as-a-judge フィルタリングは，低品質 QA ペアの検出において一定の Precision を示す一方で，Recall は極めて低く，低品質 QA ペアの多くを見逃していることがわかる.

D　プロンプト

D.1　Japanese-photos 対話の合成

```
あなたは、画像理解に優れた AI アシスタントです。
以下の情報が与えられます:
- 画像

これらの情報をヒントとして利用しながら、**画像の内容と一般知識で正しく答えられる質問と答え(QAペア)**を3〜5組作成してください。

質問の条件:
1. 第三者が画像および一般知識だけで答えられる内容にしてください。
2. 以下のタイプの質問をバランスよく含めてください:
   - 画像の主題を問う
   - 画像中の物体・人物・建造物などを問う
   - 色・形・構造・数・位置・関係などの特徴を問う
   - 文字や記号が含まれていれば、それを読み取る質問
3. 質問と答えは自然な日本語にしてください。
4. 推測や意見を含めないでください。

出力形式は JSON 構造:

{{
"conversations": [
    {{"from":"human","value":"質問 1"}},
    {{"from":"gpt","value":"答え 1"}},
    ...
]
}}

出力:
```

表 6　VLM-as-a-judge フィルタリングの評価結果.

モデル	Precision	Recall	F 値
Qwen3-VL-235B-A22B-Instruct [13]	0.602	0.046	0.086
InternVL3.5-241B-A28B-Instruct [25]	0.521	0.047	0.086

D.2　VLM-as-a-judge

```
[指示]
あなたは、VQA(Visual Question Answering) のサンプルの品質の評価者です。
「画像」とその画像に基づいた「質問と回答の組」が与えられるので、以下の評価基準に基づいてそれらの質問と回答が適切であるかどうかを評価してください。

まず、質問の評価基準は以下のとおりです。
流暢性: 質問が自然な文章であるか評価してください。文法的に誤っている質問には低い評価をつけてください。
簡潔性: 質問が簡潔で適切な長さであるか評価してください。不必要に長い質問には低い評価をつけてください。
正確性: 質問が画像に基づいて正しく、回答可能な内容であるか評価してください。画像から得られる情報と矛盾する質問には低い評価をつけてください。
明瞭性: 質問の内容が明瞭であるか評価してください。複数の解釈が可能で、解釈によって答えが変わるような曖昧な質問には低い評価をつけてください。
画像依存性: 質問が画像を参照しなければ答えられない内容であるか評価してください。画像を参照せずとも答えられる質問には低い評価をつけてください。

次に、回答の評価基準は以下のとおりです。
流暢性: 回答が自然な文章であるか評価してください。文法的に誤っている回答には低い評価をつけてください。
簡潔性: 回答が簡潔で適切な長さであるか評価してください。不必要に長い回答には低い評価をつけてください。
正確性: 回答が画像と質問に基づいて正しいか評価してください。誤っている回答には低い評価をつけてください。
整合性: 回答が質問に正しく対応し、整合性のある内容になっているか評価してください。質問と関係のない内容を含む回答には低い評価をつけてください。
一般常識と画像依存性: 回答が一般常識と画像から得られる情報を基に導き出せる内容であるか評価してください。画像や一般常識から推論できない余分な内容を含む回答には低い評価をつけてください。

評価値は 0 または 1 の 2 値です。
初めに評価の理由を述べ、その後に評価値を記入してください。
評価値を二重角括弧で囲み (例: [[1]])、以下の形式で評価結果を記述してください。

質問の評価
流暢性 (評価理由): 評価理由を記入
流暢性: [[評価値を記入]]

簡潔性 (評価理由): 評価理由を記入
簡潔性: [[評価値を記入]]

正確性 (評価理由): 評価理由を記入
正確性: [[評価値を記入]]

明瞭性 (評価理由): 評価理由を記入
明瞭性: [[評価値を記入]]

画像依存性 (評価理由): 評価理由を記入
画像依存性: [[評価値を記入]]

回答の評価
流暢性 (評価理由): 評価理由を記入
流暢性: [[評価値を記入]]

簡潔性 (評価理由): 評価理由を記入
簡潔性: [[評価値を記入]]

正確性 (評価理由): 評価理由を記入
正確性: [[評価値を記入]]

整合性 (評価理由): 評価理由を記入
整合性: [[評価値を記入]]

一般知識と画像依存性 (評価理由): 評価理由を記入
一般知識と画像依存性: [[評価値を記入]]

[入力]
質問:
{question}

回答:
{answer}
```