LLM-jp-3 VILA: 日本語マルチモーダルデータセット及び強力な日本語マルチモーダルモデルの構築

笹川 慶人 ♣,‡, 前田 航希 ◇,‡*, 杉浦 一瑳 ♠,‡*
栗田 修平 †,‡, 岡崎 直観 ◇,‡, 河原 大輔 ♣,‡
♣ 早稲田大学, ◇ 東京科学大学, ♠ 京都大学,
† 国立情報学研究所, ‡ 国立情報学研究所 大規模言語モデル研究開発センター
{kate@fuji.,dkw@}waseda.jp sugiura.issa.q29@kyoto-u.jp
{koki.maeda@nlp.,okazaki@}comp.isct.ac.jp skurita@nii.jp

概要

強力な視覚言語モデル (VLM) を構築するためには、画像・テキスト対データや交互配置 (interleaved) データ、指示データなどのマルチモーダルデータが必要である。しかし、英語のデータは豊富にあるが、日本語のデータは限られている。この問題に対処するため、我々はゼロから日本語のマルチモーダルデータセットを構築する。我々はウェブ上の画像・テキスト対データと interleaved データを構築し、さらに既存の大規模言語モデル (LLM) や VLM を利用して日本語のマルチモーダル指示データを構築する。これらのデータセットを利用して学習したモデルは、先行研究の機械翻訳を用いて構築したデータによるモデルよりも高い性能を示した。

1 はじめに

日本語特化の強力な視覚言語モデル (VLM) を構築するためには、日本語のマルチモーダルデータが必要である。英語のマルチモーダルデータは先行研究において多く提案されているが、日本語のデータセットは不足している。この問題に対する一つの解決策には、英語のマルチモーダルデータを機械翻訳によって日本語に翻訳することが挙げられる [1]。しかし、この方法では英語のテキストの翻訳時に画像の内容を考慮できないこと、翻訳誤りが生じることなどの問題がある。さらに、英語のマルチモーダルデータの画像のソースは一般的には英語のウェブサイトであるため、日本の文化を反映した画像はほとんど含まれない問題がある。

このような言語間の差に対応するため、本研究では

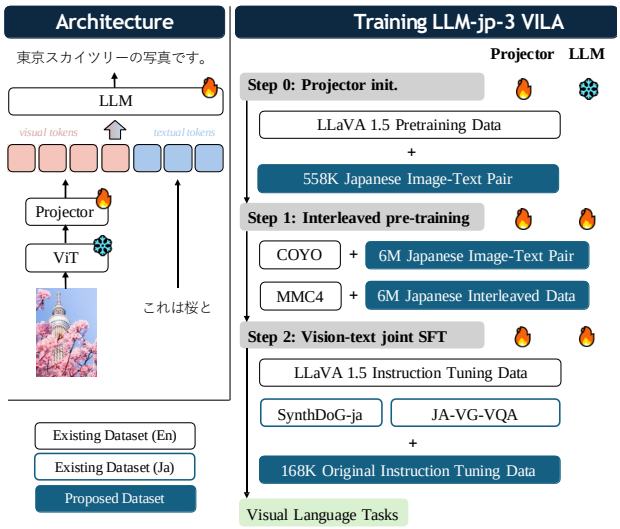

図 1: **LLM-jp-3 VILA** のアーキテクチャと学習方法。■ 色は本研究で構築したデータを示す。

はゼロから日本語マルチモーダルデータセットを構築する手法を提案する。我々は事前学習データと指示データの 2 種類のデータを日本語で構築する。事前学習データとしては、ウェブ上からデータを収集することで画像・テキスト対データと交互配置 (interleaved) データを構築する。指示データとしては、LLaVA [2] のデータ構築方法にならって画像のキャプション等を既存の大規模言語モデル (LLM) に入力して画像に関する指示データを日本語で生成する。また、日本に関連する画像を既存の VLM に入力し指示データを生成することによって、より日本の文化を反映したデータセットを構築する。

実験において、提案したデータセットで事前学習と指示学習を行ったモデルは、機械翻訳されたデータセットで学習したモデルよりも高い性能を持つことを示す。モデルのアーキテクチャとしては VILA [3] を採用し、英語データセットと提案した日

* Equally contributed.

本語データセットを用いてモデルを学習する。この手法は日本語に限らず、あらゆる言語に適用できる。本研究により、日本語 VLM 研究に利用できるリソースが大幅に強化され、VLM におけるより効果的なローカライゼーションと文化理解が可能になると考えられる。モデルの重み、データセット、学習コードは一般に公開している[1)]。

2 関連研究

視覚言語モデルの学習データ VLM の学習には事前学習と指示学習の両方において大量の画像・テキスト対データが必要である。LLaVA [2, 4] の視覚指示学習では OpenAI API[2)] による合成データを主に使用している。この合成データは InstructBLIP [5] や VILA [3] などのさまざまな VLM の開発に使用されている。これらの成功にならって、我々は指示学習のための合成データを日本語で構築する。

日本語視覚言語モデルの言語資源 図やテキストを含む文書に対する質問に回答するタスクは、文書質問応答タスクとして注目されている [6, 7, 8, 9, 10, 11]。ただし、これらのデータセットは主に英語で開発されており、他言語における知識は反映されていない。日本語における文書質問応答データセットとしては、JDocQA [12] が提案されている。また、日本語 VLM の評価ベンチマークとして、Heron Bench [1] も提案された。しかし、英語のデータセットと比較すると日本語のデータセットは非常に不足しており、既存のデータセットは主にモデルのファインチューニングを想定している。本研究では、日本語 VLM 構築のための大規模な画像・テキストのデータセットを構築する。

3 日本語データセットの構築

Interleaved データ MMC4 [14] のデータ構築方法に基づいて日本語の interleaved データを構築した。まず llm-jp-corpus [15] における 2020 年から 2022 年の Common Crawl のダンプから日本語テキストを抽出した。このテキストに対して、`bunkai` [16] を利用して文章を文に分解した。この操作の後、数字やアルファベット、日本語の文字が含まれない文を一つ前の文と結合した。

表 1: LLM-jp-3 VILA の学習データ量。“Full” 列は最終的なモデルの学習に使ったデータを示す。

Data	# Images	# Step	Full
English			
LLaVA-1.5 Pretrain Data	558K	0	✓
LLaVA-1.5 Instruction Data (subset)	358K	2	✓
COYO [13] (subset)	6M	1	✓
mmc4-core [14] (subset)	6M	1	✓
Japanese			
(Proposed) 画像・テキスト対データ	6.6M	0 & 1	✓
(Proposed) 交互配置データ	6M	1	✓
(Proposed) 指示データ	369K	2	✓
翻訳データ [3)]	620K	2	✗

次に、テキストのソースのウェブページから画像の URL を抽出し、画像をダウンロードした。特定のサーバーへの負荷を避けるため、高頻度ドメインの URL はスキップした。ダウンロードした画像について、同一文書内の画像の重複除去を行った。`ImageHash`[4)] ライブラリを利用して、画像の phash 値を計算し、そのハミング距離が 5 以下の画像集合について、最も解像度の高いものを残した。また、複数文書を跨いだ重複除去も行った。それぞれの年のデータについて、サンプリングされた 6 万画像のうち 10 以上の重複がある画像を除去した。この操作をサンプリングした画像の枚数の合計が全体の画像の枚数となるまで繰り返した。アプリのアイコンや広告、画像のリンクが切れているときに挿入される画像などが除去された。その後、`dataset2metadata`[5)] [17] を用いて NSFW 画像を除去した。

上記のフィルタリングを通過した画像について、LAION5B [18] で学習された OpenCLIP[6)] [19] を用いて全ての画像と文のペアの類似度を計算した。文書内の全ての文について、画像との類似度が 0.20 未満ならばその画像は除去した。MMC4 と同様に、割り当てられた画像と文の類似度を文書全体で最大化するように割り当て問題を解くことで、文書内の画像を文に割り当てた。割り当て問題は `lapjv`[7)] ライブラリを使用して解いた。最後に llm-jp-corpus において使われている有害文書フィルタリングを行った。事前学習に利用するため、画像の枚数が 2 から 5 のもので、文数が 10 から 100、トークン数が LLM

1) https://huggingface.co/llm-jp/llm-jp-3-vila-14b
2) https://openai.com/index/openai-api
3) https://huggingface.co/datasets/turing-motors/LLaVA-v1.5-Instruct-620K-JA
4) https://github.com/JohannesBuchner/imagehash
5) https://github.com/mlfoundations/dataset2metadata
6) https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k
7) https://pypi.org/project/lapjv

表 2: 日本語ベンチマークにおける既存の VLM と **LLM-jp-3 VILA** の比較結果。”–”はベンチマークデータが学習に使われており、評価ができないことを示す。**太字**は GPT-4o 以外の最高スコアを示す。“LLM” は LLM-as-a-Judge を示す。ベースラインモデルの詳細は付録 A に示す。

	Heron-Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
Models	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
Japanese InstructBLIP Alpha	14.0	20.8	2.42	–	–
Japanese Stable VLM	24.2	23.3	2.47	–	–
Llama-3-EvoVLM-JP-v2	39.3	41.4	2.92	**23.5**	2.96
LLaVA-CALM2-SigLIP	43.3	47.2	3.15	17.4	3.21
LLaVA-1.6 7B	25.8	28.6	2.40	11.7	2.67
LLaVA-1.5 7B	34.8	40.6	2.48	13.9	2.66
Llama 3.2 11B Vision	36.5	27.4	2.77	13.8	2.95
InternVL2 8B	45.2	33.7	2.98	11.6	3.13
Qwen2-VL 7B Instruct	54.8	45.3	3.53	16.2	3.48
LLM-jp-3 VILA (Ours)	**57.2**	**52.3**	**3.69**	16.2	**3.62**
GPT-4o	87.6	37.6	3.85	12.1	3.58

の最大長に収まり、さらに全ての画像と割り当てられた文のペアの類似度が 0.20 以上の文書のみを残した。結果としてデータセット内の画像の枚数は 9.9M となった。画像・テキスト対データの量とバランスをとるため、学習にはサブセットを利用した。

画像・テキスト対データ　interleaved データから NSFW 画像を除去したものについて代替テキストを Web から収集した。COYO [13] の手法を参考に、代替テキストのフィルタリングを行った。まず、日本語データのみを抽出するために、正規表現を用いてひらがなやカタカナ、漢字が含まれていないものを除去した。また、文字数が少なすぎるものやファイル名をルールベースで除去した。次に `Hojichar`[8] の `DiscardAdultContentJa` フィルタを用いて、NSFW テキストを除去した。続いて、文頭・文末の連続する空白文字を除去し、連続する複数の空白文字を 1 つの半角スペースに置き換えた。その後、データの重複除去を行った。10 個以上存在する代替テキストを除去し、(画像の phash 値, 代替テキスト) のペアを一つだけ残した。

最後に各画像・代替テキストペアの類似度を LAION-5B で学習した OpenCLIP と Japanese CLIP[9] を用いて算出し、類似度が下位 30%以下のペアを除去した。類似度は 2 つの CLIP のスコアをそれぞれの中央値で重みづけしたものを利用した。結果的に 6.6M の画像・テキスト対データセットが得られた。

指示データ　本研究では LLaVA の指示学習データセットと同様に、COCO [20] の画像から OpenAI API を用いて日本語の指示データを構築した。具体的には、画像に付与されているキャプションやバウンディングボックスを GPT-4o-mini に入力し、画像に関する質問と回答のペアを複数生成させた。結果として 156K 事例のデータとなった。

さらに、日本で撮影された画像データセット[10] を用いて、画像を GPT-4o に入力しマルチターンの会話データを生成させ、**Japanese-photos-conv** データセットを構築した。このデータは 12K 事例を含む。

以上の方法で構築したデータセットを表 1 にまとめる。これらのデータセットをモデルの学習に用いた。

4　実験

4.1　モデルの学習

我々が提案する **LLM-jp-3 VILA** のアーキテクチャは、図 1 にも示したように、画像エンコーダと LLM を MLP 層によって接続する構造である。画像エンコーダには SigLIP[11] [21] を、LLM には `llm-jp-3-13b-instruct`[12] を使う。

図 1 に示すように、VILA と同様の三段階の学習ステップによりモデルを構築する。Step 0 で日本語と英語の画像・テキスト対データを 558K 事例ずつ用いて MLP 層のパラメータをチューニングする。Step 1 において画像・テキスト対データと interleaved

8) https://github.com/HojiChar/HojiChar
9) https://huggingface.co/line-corporation/clip-japanese-base
10) https://huggingface.co/datasets/ThePioneer/japanese-photos
11) https://huggingface.co/google/siglip-so400m-patch14-384
12) https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

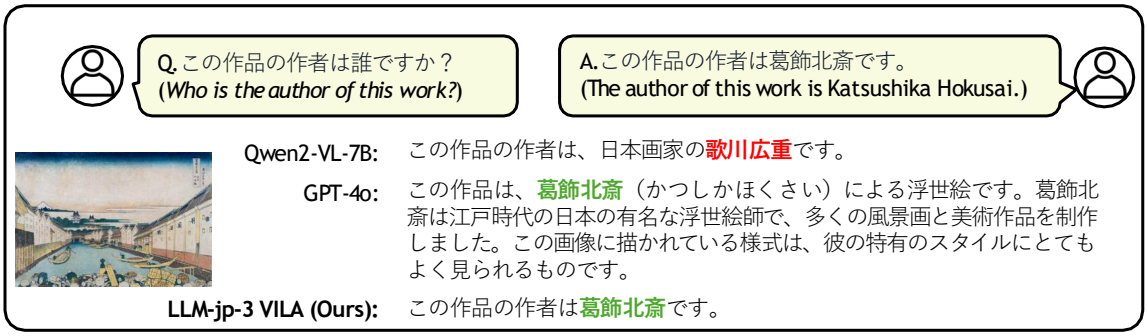

図 2: それぞれのモデルの Heron Bench の質問に対する出力例。緑は正しい回答を、赤は間違った回答を示す。

表 3: **LLM-jp-3 VILA** のアブレーション結果。**太字**は最高スコアを示す。“LLM” は LLM-as-a-Judge を表す。

			Heron-Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
Step-0	Step-1	Step2	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
✓	✓	translated	47.2	45.6	3.19	15.7	3.33
✓	✗	✓	56.5	**57.3**	3.47	16.1	3.54
✓	w/o interleaved	✓	**58.6**	52.2	3.50	**16.7**	3.61
✓	✓	✓	57.2	52.3	**3.69**	16.2	**3.62**

データを用いてマルチモーダル継続事前学習を行うことで MLP 層と LLM のパラメータをチューニングする。英語のデータセットには、mmc4-core の 6M 画像のサブセットと COYO の 6M 画像のサブセットを使用する。日本語のデータセットには我々が提案するデータを使用する。Step 2 において、モデルが指示に従うように MLP 層と LLM のパラメータをチューニングするマルチモーダル指示学習を行う。日本語のデータとして本研究で構築した指示データと JA-VG-VQA [22] と synthdog-ja [23] を使用する。英語のデータには LLaVA-Instruct のサブセットを使用する。日本語のデータセットは 369K 事例、英語のデータセットは 358K 事例を含む。

4.2 評価

LLM-jp-3 VILA の性能を検証するために、Heron Bench [1]、JA-VLM-Bench-In-the-Wild [24]、および JA-VG-VQA500[13] の 3 つのベンチマークを使用する。評価ツールには llm-jp-eval-mm [25] を用いた。

ベンチマーク結果　表 2 に 3 つのベンチマークにおけるスコアを示す。既存の日本語 VLM と比較して、LLM-jp-3 VILA は一貫して最高性能であった。また、JA-VG-VQA-500 ベンチマークでは、我々のモデルは GPT-4o の性能を上回った。

事例分析　図 2 に LLM-jp-3 VILA とベースラインモデルの出力の比較を示す。Qwen2-VL-7B は質問に対する答えを「葛飾北斎」と答えるべきところを「歌川広重」と回答したが、我々のモデルは正しく答え、日本文化に特化した知識を持っていることを示した。GPT-4o も詳細な説明を含む良い答えを出力した。

学習データセットのアブレーション　構築したデータセットの有効性を検証するために、アブレーション実験を行った。指示データ、Step 1 の学習、interleaved データのそれぞれを省略したときのモデルの性能を比較した。表 3 に結果を示す。提案した指示データを機械翻訳したデータに置き換えると性能がかなり悪化した。また、Step 1 の学習ではある程度の改善が見られるが、いくつかの評価指標では interleaved データによる性能向上は限定的である。その理由の一つは、我々のデータセットの画像数は、VILA の Step 1 のデータセットの画像数の約半分であることである。本研究の今後の課題として、データの量を増やすことが挙げられる。

5 おわりに

本研究では日本語のマルチモーダルデータを構築し、日本語に強い VLM である LLM-jp-3 VILA を提案した。実験により、このモデルは日本語におけるさまざまなマルチモーダルタスクにおいて良い性能を示した。今後の課題として、文書画像など、さまざまなデータセットを用いて学習したモデルの構築に取り組みたい。

13) ベンチマークの詳細については付録 B を、評価方法については付録 C を参照

謝辞

この研究プロジェクトは、GPT-4o を利用したモデルの性能比較に Microsoft Accelerate Foundation Models Research (AFMR) プログラムの助成により Microsoft Azure を使用しました。

参考文献

[1] 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.
[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
[3] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. **2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 26679–26689, 2023.
[4] 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.
[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
[6] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In **Computer Vision – ECCV 2016**, pp. 235–251, Cham, 2016. Springer International Publishing.
[7] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, July 2017.
[8] Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. Semantic parsing to probabilistic programs for situated question answering. In Jian Su, Kevin Duh, and Xavier Carreras, editors, **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**, pp. 160–170, Austin, Texas, November 2016. Association for Computational Linguistics.
[9] Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In **CVPR**, 2018.
[10] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. Visuo-linguistic question answering (VLQA) challenge. In Trevor Cohn, Yulan He, and Yang Liu, editors, **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 4606–4616, Online, November 2020. Association for Computational Linguistics.
[11] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In **AAAI**, 2023.
[12] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese document question answering dataset for generative language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9503–9514, Torino, Italy, May 2024. ELRA and ICCL.
[13] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.
[14] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 8958–8974. Curran Associates, Inc., 2023.
[15] LLM-jp. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **CoRR**, Vol. abs/2407.03963, , 2024.
[16] Yuta Hayashibe and Kensuke Mitsuzawa. Sentence boundary detection on line breaks in Japanese. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, **Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)**, pp. 71–75, Online, November 2020. Association for Computational Linguistics.
[17] 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 Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J 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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 27092–27112. Curran Associates, Inc., 2023.
[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 25278–25294. Curran Associates, Inc., 2022.
[19] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.
[20] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, **Computer Vision – ECCV 2014**, pp. 740–755, Cham, 2014. Springer International Publishing.
[21] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 11975–11986, 2023.
[22] 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 **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1918–1928. Association for Computational Linguistics, 2018.
[23] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In **European Conference on Computer Vision**, pp. 498–517. Springer, 2022.
[24] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes, 2024.
[25] 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), March 2025.
[26] Meta. Llama-3.2-11b-vision, 2024.
[27] 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.
[28] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. **arXiv preprint arXiv:2404.16821**, 2024.
[29] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
[30] Aozora Inagaki. llava-calm2-siglip, 2024.
[31] Makoto Shing and Takuya Akiba. Japanese stable vlm, 2024.
[32] Makoto Shing and Takuya Akiba. Japanese instructblip alpha, 2023.
[33] OpenAI. Gpt-4 technical report, 2024.
[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. **International Journal of Computer Vision**, Vol. 123, pp. 32–73, 2017.

表 4: ベースラインモデルの詳細.

VLM	Reference	Base LM	LM Size	Hugging Face / API
Llama 3.2 Vision 11B	[26]	Llama 3.2	11B	meta-llama/Llama-3.2-11B-Vision
Qwen2-VL 7B	[27]	Qwen2	7B	Qwen/Qwen2-VL-7B-Instruct
InternVL2 8B	[28]	InternLM2-Chat	8B	OpenGVLab/InternVL2-8B
LLaVA-1.5 7B	[4]	Llama2	7B	llava-hf/llava-1.5-7b-hf
LLaVA-1.6 7B	[29]	Mistral	7B	llava-hf/llava-v1.6-mistral-7b-hf
LLaVA-CALM2-SigLIP	[30]	CALM2	7B	cyberagent/llava-calm2-siglip
Japanese Stable VLM	[31]	Japanese Stable LM Instruct Gamma	7B	stabilityai/japanese-stable-vlm
Japanese InstructBLIP Alpha	[32]	Japanese StableLM Instruct Alpha	7B	stabilityai/japanese-instructblip-alpha
Llama-3-EvoVLM-JP-v2	[24]	Mantis-8B-SigLIP-Llama-3 Merged Llama-3-ELYZA-JP-8B Bunny-v1.1-Llama-3-8B-V	8B	SakanaAI/Llama-3-EvoVLM-JP-v2
GPT-4o	[33]	GPT-4	-	gpt-4o-2024-05-13

A ベースラインモデル

表 4 に、評価に用いたベースラインモデルの詳細を示す。

B ベンチマークデータセットの詳細

ここでは、評価で使用したデータセットの詳細情報を示す。

Heron Bench 日本に関する映画や建築物などの画像 21 枚に対して、 合計 103 個の Visual Question Answering (VQA) 問題からなるデータセット。評価は LLM-as-a-Judge によって行われる。

JA-VLM-Bench-In-the-Wild 日本に関連する 42 枚の画像に対して 50 個の VQA 問題からなるデータセット。質問と回答の構築には GPT-4V [33] が利用されている。データセットの品質を確保するために人間によるフィルタリングが行われている。

JA-VG-VQA500[14] Visual Genome [34] に基づいて作成された Japanese Visual Genome VQA データセット [22] のテストセットから 500 件のサンプルを抽出して構築されたデータセット。

C 評価の詳細

評価は主にベンチマーク提供者のデフォルトの評価設定に従った。LLM-as-a-Judge を用いる際、評価値の変動を小さくするため、Heron Bench および JA-VLM-Bench-In-the-Wild では 5 回の評価値を平均した. JA-VG-VQA-500 は事例数が多かったため、1 回の実行スコアを使用した。LLM-as-a-Judge における評価モデルは Azure OpenAI API の `gpt-4o-2024-05-13` を採用した。再現性を確保するために、評価者の temperature を 0、シード値を 0 に設定した。ただし、同じシードを使用しても出力が必ずしも決定的でない場合があることに注意する必要がある[15]。

Heron Bench Heron-Bench における LLM-as-a-Judge は、総合的なスコアが GPT-4 による参照文のスコアの比率によって定量化される。そのため、スコアが 100%を超える場合があるが、これはモデルが平均的に GPT-4 を上回ったことを示す。

14) https://huggingface.co/datasets/SakanaAI/JA-VG-VQA-500
15) 詳細については、https://platform.openai.com/docs/advanced-usage/reproducible-outputs を参照