Asagi: 合成データセットを活用した大規模日本語 VLM

上原康平 [1,2]　黒瀬優介 [1,2]　安道健一郎 [2,1]　Chen Jiali[1]　Gao Fan[1]　金澤爽太郎 [1]
坂本拓彌 [1]　竹田悠哉 [1]　Yang Boming[1]　Zhao Xinjie[1]　村尾晃平 [3]　吉田浩 [3]
田村孝之 [4]　合田憲人 [4,3]　喜連川優 [4,1]　原田達也 [1,2,3]
[1] 東京大学　[2] 理化学研究所　[3] 国立情報学研究所　[4] 情報・システム研究機構
{uehara, kurose, ando, harada}@mi.t.u-tokyo.ac.jp
{chenssr, fangao0802, kanazawa-sotaro317, takuya50719,
takeda-yuya190, boming, xinjie-zhao}@g.ecc.u-tokyo.ac.jp
{k-murao, h-yoshida, aida}@nii.ac.jp
tamura.takayuki@rois.ac.jp, kitsure@tkl.iis.u-tokyo.ac.jp

概要

大規模言語モデル（LLM）の発展として，画像など他のモダリティも扱う大規模マルチモーダルモデルの研究が進んでいる．本研究で注目する Vision & Language モデル（VLM）は，画像とテキストを同時に入力可能なモデルである．しかしながら，データが潤沢な英語モデルに比べ，データの不足する日本語モデルの開発は十分とはいえない．本研究では，日本語能力に特化した VLM の開発のため，大規模な日本語テキスト・画像ペアの合成データセットを新たに構築した．なお，データセットの構築時に，ライセンスによってデータ利用が制限される LLM は用いていない．構築したデータセットを用いて日本語 VLM を訓練し，その性能を評価した．

1　はじめに

大規模言語モデル（LLM）の研究開発は世界的に急速な進歩を遂げ，テキストのみならず画像や音声など複数モダリティを扱う研究が活発化している．特に，画像とテキストを同時に入力可能な大規模 Vision & Language モデル（VLM）は，複数の商用サービスにも導入され，その有用性が広く認識されつつある．しかしながら，これらの商用 VLM の多くは API を通じてのみ利用可能であり，モデルのパラメータや学習過程などの情報は一般に非公開である．学術コミュニティでは，よりオープンな形で研究が進められているが，大半は英語圏のデータに特化して学習された英語用モデルであり，日本語を主対象とした VLM はほとんど存在しない．

日本語 VLM を開発する上での最大の課題は，学習データの不足である．VLM の訓練には，画像とテキストのペアデータセットが必要になるが，日本語のテキスト・画像ペアは既存の公開データセットを総合しても数百万件規模にとどまる．英語圏の VLM が数千万件規模のデータで訓練されていることを鑑みると，これは十分とはいいがたい．

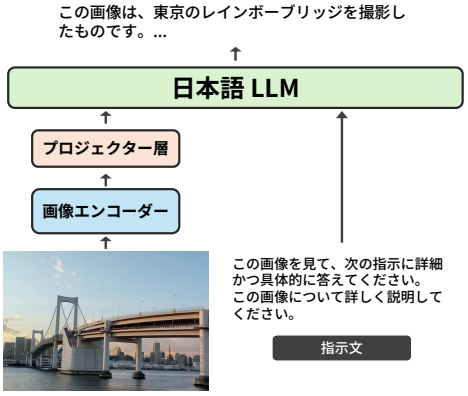

図 1　構築したモデルの概要図．画像エンコーダー，日本語 LLM，両者をつなげるプロジェクター層からなる．

本研究では，Web からクロールした画像などをもとに，英語 VLM や日本語 LLM を用いて画像・日本語テキストデータセットを合成し，日本語 VLM を学習した．本研究の主な貢献は，以下の通りである．

- 日本語テキスト・画像のペアデータセットを合成した．この際，生成物の利用が制限されるサービス（OpenAI GPT など）は用いていない．
- 大規模 VLM の効率的な学習を可能にする実装を行い，学習を実行した．学習した VLM は，利用制限のあるデータを用いていない既存モデルと比較して高い性能を示すことが確認された．

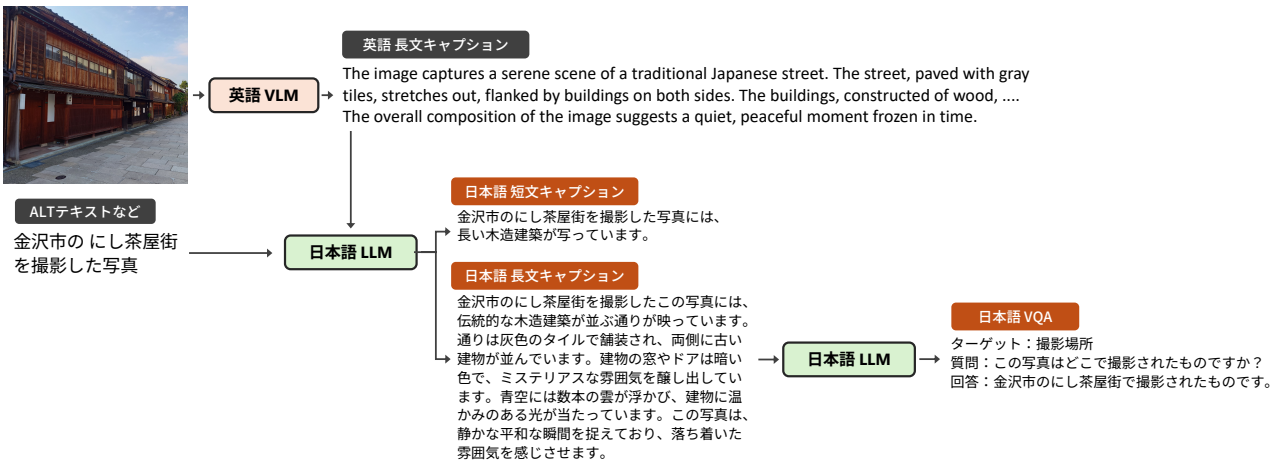

図 2　データセット構築のパイプライン．

2　データセット構築

大規模 VLM の構築にあたっては，大量の画像とテキストをペアとしたデータセットを用いることが不可欠である．たとえば英語圏においては，Conceptual Captions [1] や LAION [2] など，数千万から数十億規模の画像・テキストペアデータが活用されている．一方，日本語圏では，画像と日本語テキストのペアデータはごく限られており，数十万から数百万件規模にとどまる [3, 4, 5, 6, 7]．

2.1　合成データセットの構築

本研究では，大規模な画像・日本語テキストのペアデータセットを作成するにあたり，LLM を活用したデータ合成パイプラインを構築した．合成データの作成は，画像 I と参照テキスト T_{ref} のペアを入力とし，合成キャプション S を出力する処理として定式化できる．ここで，参照テキストは，ALT テキストや，画像に関連する記事の文章など，画像に関連したテキスト情報を指す．

データのソースには，（A）あらかじめ T_{ref} が付与されているデータ（Wikipedia の記事・画像ペアなど）と，（B）T_{ref} が付与されていないデータ（Web からクロールした画像など）の 2 種類が存在する．

以下，データ合成の各処理について説明する．

Web データからの参照テキスト取得（B のみ）
合成データ作成のため，日本語 Web サイトを対象にクローリングを実施し，HTML と画像をダウンロードした．データの取得は，2024 年 8 月から 11 月にかけて実施された．クロールされた画像や HTML から，VLM の訓練に適さない品質や内容のデータを除外するため，ルールベースのフィルタリングを適用した（詳細は補足 **A** に記載）．

画像を含む HTML から，画像に最も適したテキスト（これを参照テキスト T_{ref} とする）を取得するため，(1) `Trafilatura` [8] を用いて HTML データからテキストを抽出，(2) 画像とテキスト各文ごとに日本語版 CLIP [9] を用いて類似度を計算，(3) 画像との類似度が最も高いテキストを参照テキスト（T_{ref}）として採用，という手順で処理を行った．

合成キャプションの生成（A・B 共通）（図 2）
参照テキスト T_{ref} と画像 I から，合成キャプション S を生成する方法を説明する．ここで，参照テキスト T_{ref} は，画像についての背景情報を与えてはいるが，画像に写っているものについて視覚的に説明するものとは限らない．そこで，各画像について，英語 VLM（Phi-3.5-vision [10]）を用いて英語の長文キャプションを生成した．この英語キャプションと参照テキストを組み合わせて，視覚的な情報と，背景情報の両方を含む日本語キャプションを生成する．

テキストの組み合わせには，日本語 LLM である CALM3-22B-Chat [11] を用い，両テキストを組み合わせて要約するように指示した．データの多様性を確保するため，日本語 LLM への指示を工夫し，1 つの画像に対して長短 2 種類の合成キャプション（S_{long}, S_{short}）を生成した．合成キャプション生成のために用いたプロンプトは，補足 **C** に示す．

VQA データの合成（A・B 共通） 合成キャプションの作成に加えて，多様な指示文に適切に回答できるようにするため，VQA（Visual Question Answering）データ（S_{VQA}）の合成も行った．ここでは，先ほど作成した合成長文キャプション（S_{long}）

をもとに，CALM3-22B-Chat へ質問文と回答を生成するように指示した．さらに，多様な指示への対応能力を高めるため，多肢選択式 VQA データ（S_{MCVQA}）の合成も行った．ここでは，S_{VQA} に対して，CALM3-22B-Chat を用いて誤りの選択肢候補を複数生成し，元の正解と組み合わせた．

以上により，日本語キャプション（S_{long}, S_{short}）が約 800 万件，VQA（S_{VQA}）が約 750 万件，多肢選択式 VQA（S_{MCVQA}）が約 100 万件生成された．なお，同一画像に対して，S_{long}, S_{short}, S_{VQA}, S_{MCVQA} の最大 4 種類のテキストが生成される場合がある．同じ画像を過度に学習に用いるとモデルが過学習を起こす可能性があるため，1 画像につき 1 テキストのみ採用するようにサンプリングを行った．

2.2 既存データセットの活用

既存の日本語データと，CALM3-22B-Chat を用いて日本語訳した英語データも活用した．データの内訳は，補足 **B** に示す．なお，Visual Genome [12] 由来のデータについて，ベンチマークである JA-VG-VQA-500 [13] に含まれる画像は除外した．

3 実験

3.1 モデルと実装

`Asagi` モデルは，LLaVA [15, 16] をベースとしたエンコーダ・デコーダ型の VLM である．LLaVA は，画像特徴を抽出するエンコーダ，テキストを生成するデコーダ（LLM），そして両者を結合するプロジェクター（2 層の MLP）によって構成されている．

`Asagi` モデルでは，画像エンコーダに SigLIP [17] を，テキストデコーダに LLM-jp が提供する日本語 LLM（`llm-jp-3-1.8b-instruct`, `llm-jp-3-3.7b-instruct`, `llm-jp-3-13b-instruct`） [18] を用いた（それぞれ `Asagi-2B`, `Asagi-4B`, `Asagi-14B`）．モデルの詳細は補足 **D** に示す．

実装　モデルの実装には，分散学習フレームワークである Megatron-LM [19] を用いた．Megatron-LM は，LLM を効率的に訓練するためのモデル並列化手法を備えている．本研究の実装では，エンコーダ・デコーダ構造およびマルチモーダル入力に対応できるよう，Megatron-LM を拡張した．

訓練　LLaVA の訓練方法に従い，本研究でも 2 段階学習を採用した．**Stage1**: 画像エンコーダとテキストデコーダをフリーズし，プロジェクター層のみを学習．**Stage2**: 画像エンコーダのみフリーズし，プロジェクター層とテキストデコーダを学習．

Stage1 の訓練には，準備したデータから，キャプション約 1800 万件を用い，Stage2 の訓練には，VQA データも含めて約 2100 万件のデータを用いた．

`Asagi-14B` の学習には H100 GPU が 8 枚搭載された計算ノードを 4 ノード使用し，stage1 の学習には約 3 日，stage2 の学習には約 9 日を要した．

3.2 実験と考察

3.2.1 評価ベンチマーク

Heron-Bench [20]，JA-VLM-Bench-In-The-Wild [13]，JA-VG-VQA-500 [13] の 3 つのベンチマークを用いてモデルの性能を評価した．評価計算には，VILA-jp の論文 [14] で用いられた `eval-mm` ライブラリを用いた．JA-VG-VQA-500 は簡潔な回答を求める形式であるため，「この画像を見て、次の質問に簡潔に答えてください。」という指示文を，他のベンチマークでは「この画像を見て、次の質問に詳細かつ具体的に答えてください。」という指示文を与えた．評価を行う際は，モデルの Temperature を 0.0 に設定した．また，Heron-Bench および LLM-as-a-Judge での評価には，`gpt-4o-2024-05-13` を用いた．LLM による評価はそれぞれ 5 回行い，その平均値を最終スコアとした（JA-VG-VQA-500 は 1 回のみ）．

3.2.2 評価結果

結果を，表 1 に示す．また，モデルの出力例を図 3 に示している．まず，本研究が提案する手法は，現状の最高性能モデルである VILA-jp と遜色ないスコアを示した．特に，制限付きライセンスデータで学習したモデルを除外した比較においては，Japanese InstructBLIP Alpha [21] や Japanese Stable VLM [22]，LLaVA-CALM2-SigLIP [23] などの既存モデルと比べると，Asagi-14B は大幅に性能が向上している（表の下線部分）．

3.2.3 データセット・パラメータサイズの影響

本研究で新たに構築した合成データがモデル性能に与える影響を検証した．partial s1 モデル，partial s2 モデルは，それぞれ stage1，stage2 の学習において，今回作成した合成データを用いていないモデルである．結果，partial s2 モデルにおいては，評価スコアの顕著な低下が見られた．stage2 では指示追従

表 1　各モデルの性能比較．Asagi モデル以外のスコアは，VILA-jp の論文 [14] の結果を引用した．GPT などによって生成された制限付きデータを訓練に用いていないモデルは，† を付与している．太字は GPT-4o 以外のモデルの中で最高のスコアを示し，下線は GPT によって生成されたデータを用いないモデルの中で最高のスコアを示す．

		Heron-Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
	LM size	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
Japanese InstructBLIP Alpha†	7B	14.0	20.8	2.42	-	-
Japanese Stable VLM†	7B	24.2	23.3	2.47	-	-
LLaVA-CALM2-SigLIP†	7B	43.3	47.2	3.15	17.4	3.21
Llama-3-EvoVLM-JP-v2	8B	39.3	41.4	2.92	23.5	2.96
VILA-jp	13B	**57.2**	52.3	**3.69**	16.2	3.62
Asagi-2B†	1.8B	44.7	48.8	3.26	53.7	3.69
Asagi-4B†	3.7B	49.3	49.6	3.38	55.6	3.78
Asagi-14B $_{\text{full}}$†	13B	55.8	50.8	3.44	**56.8**	**3.84**
Asagi-14B $_{\text{partial s1}}$†	13B	56.4	**52.5**	3.58	56.7	3.83
Asagi-14B $_{\text{partial s2}}$†	13B	20.6	42.7	2.43	56.5	3.77
GPT-4o	-	87.6	37.6	3.85	12.1	3.58

能力を強化するような訓練が行われるため，ここで十分なデータを用いない場合，最終モデルの指示理解や応答品質が大きく損なわれたと考えられる．一方，partial s1 においては，評価スコアはむしろ向上する傾向がみられた．これは，stage2 の学習に用いた合成データが非常に大規模であったことから，stage1 の学習における合成データの有無が最終性能に与える影響が相対的に小さかった可能性がある．stage1 における最適なデータ量・設定については，今後さらなる検証が必要と考えられる．

2B，4B，14B モデルについては，LLM のパラメータ数が増加するにつれ，VLM の最終的な性能も向上しており，VLM の性能向上に LLM のパラメータ数が寄与していることが示唆された．

4　結論と今後の展望

本研究では，大規模な日本語画像・テキストの合成データセットを構築し，日本語 VLM の学習を行った．構築されたモデルは，出力の利用が制限されている LLM によって生成されたデータを用いないモデルとしては，最高のスコアを達成した．一方，データセットの合成戦略や，各訓練ステージにおけるデータ利用量など，モデル性能に影響を与える要因については今後の検証が必要である．

今回の学習で開発した訓練コードは，100B を超える大規模なモデルにも対応可能であるため，今後は更なるモデルのスケールアップを検討している．現状，本実験と同様の計算ノード 24 台を用いて，

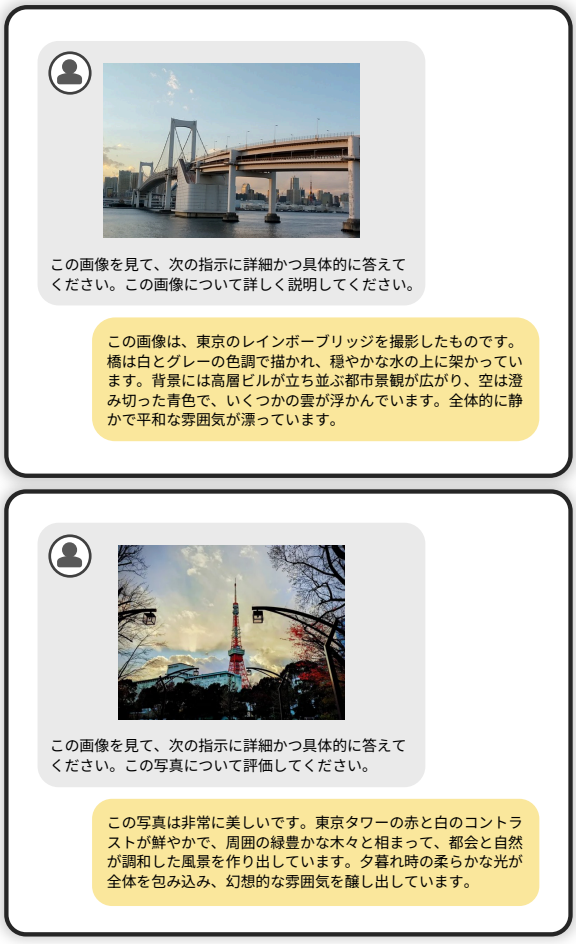

図 3　Asagi-14B の出力結果の例．

LLM-jp がリリースしている 172B 日本語 LLM をベースとした Asagi-173B の学習に取り組んでいる．

謝辞

本研究は，戦略的イノベーション創造プログラム (SIP)「統合型ヘルスケアシステムの構築」JPJ012425，JST ムーンショット型研究開発事業 JPMJMS2011，CREST 課題番号 JP-MJCR2015，JSPS 科研費 JP23K16990，JP23K19971，及び東京大学 Beyond AI 研究推進機構の基礎研究費 (AI 自体の進化) の支援を受けたものです.

参考文献

[1] Piyush Sharma, Nan Ding, Sebastian Goodman, Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In **ACL**, 2018.

[2] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. **NeurIPS Data-centric AI Workshop**, 2021.

[3] Yuya Yoshikawa, Yutaro Shigeto, Akikazu Takeuchi. Stair captions: Constructing a large-scale japanese image caption dataset. In **ACL**, 2017.

[4] Takashi Miyazaki, Nobuyuki Shimizu. Cross-lingual image caption generation. In **ACL**, 2016.

[5] Nobuyuki Shimizu, Na Rong, Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In **COLING**, 2018.

[6] Ruka Funaki, Hideki Nakayama. Image-mediated learning for zero-shot cross-lingual document retrieval. In **EMNLP**, 2015.

[7] Hideki Nakayama, Akihiro Tamura, Takashi Ninomiya. A visually-grounded parallel corpus with phrase-to-region linking. In **LREC**, 2020.

[8] Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In **ACL**, 2021.

[9] Shuhei Yokoo, Shuntaro Okada, Peifei Zhu, Shuhei Nishimura, Naoki Takayama. Clip japanese base, 2024. https://huggingface.co/line-corporation/clip-japanese-base.

[10] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allison Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Young Jin Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xianmin Song, Olatunji Ruwase, Praneetha Vaddamanu, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Cheng-Yuan Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. **arXiv**, 2024.

[11] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024. https://huggingface.co/cyberagent/calm3-22b-chat.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **IJCV**, Vol. 123, No. 1, pp. 32–73, 2017.

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

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

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

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee. Improved baselines with visual instruction tuning. In **CVPR**, 2024.

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

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

[19] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. **arXiv**, 2019.

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

[21] Makoto Shing, Takuya Akiba. Japanese instructblip alpha. https://huggingface.co/stabilityai/japanese-instructblip-alpha.

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

[23] Aozora Inagaki. Llava-calm2-siglip. https://huggingface.co/cyberagent/llava-calm2-siglip.

[24] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, Marc Najork. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In **SIGIR**, 2021.

[25] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, Vittorio Ferrari. Connecting vision and language with localized narratives. In **ECCV**, 2020.

[26] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In **CVPR**, 2024.

[27] Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, Volodymyr Kuleshov. Commoncanvas: Open diffusion models trained on creative-commons images. In **CVPR**, 2024.

[28] toshi456. Llava-cc3m-pretrain-595k-ja, 2023. https://huggingface.co/datasets/toshi456/LLaVA-CC3M-Pretrain-595K-JA.

[29] Peiyuan Liao, Xiuyu Li, Xihui Liu, Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. **arXiv**, 2022.

[30] Drew A Hudson, Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In **CVPR**, 2019.

[31] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. **IJCV**, Vol. 127, pp. 398–414, 2016.

[32] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In **ECCV**, 2022.

[33] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In **CVPR**, 2019.

[34] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. **arXiv**, 2024.

[35] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, C.V. Jawahar. Infographicvqa. In **WACV**, January 2022.

[36] Kushal Kafle, Brian Price, Scott Cohen, Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In **CVPR**, 2018.

[37] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi. A diagram is worth a dozen images. In **ECCV**, 2016.

表 2　利用したデータセットの内訳.

データセット		Stage1	Stage2	件数
ROIS	合成（B）	✓	✓	8,408,364
Japanese image text pairs [14]	合成（A）	✓	✓	4,371,417
Wikipedia [24]	合成（A）	✓	✓	2,518,084
Open Images [25]	翻訳	✓	✓	676,384
DCI [26]	翻訳	✓	✓	7,269
CommonCatalog CC-BY [27]	翻訳	✓		3,515,193
LLaVA-Pretrain-JA [15, 28]		✓		551,603
STAIR Captions [3]		✓	✓	409,542
Flickr-JP [7]		✓	✓	158,666
YJ Captions [4]		✓	✓	129,724
Japanese Pascal [6]		✓	✓	4,983
ArtBench [29]	合成（A）		✓	104,180
GQA [30]	翻訳		✓	1,881,682
VQA v2 [31]	翻訳		✓	881,423
A-OKVQA [32]	翻訳		✓	34,081
OK-VQA [33]	翻訳		✓	17,915
Japanese Visual Genome [5]			✓	1,577,049
PangeaInstruct [34]			✓	92,885

A Web クロールデータのフィルタリング

以下の条件に該当する画像・テキストは除外した.

- 単色の画像
- 一辺が 100px 以下もしくは 2048px 以上の画像
- アスペクト比が 0.3 : 1 もしくは 3 : 1 を超える画像
- アルファベット・数字・空白しか含まないテキスト
- HojiChar [1] の有害キーワードを含むテキスト
- 3 文字以下もしくは 1000 文字以上のテキスト

クロールされた画像には，広告バナーや文書画像など，訓練には適さない画像も多く含まれる．そうしたデータを除くため，ドキュメント画像データセット [35, 36, 37] の画像特徴量と各画像の特徴量のコサイン類似度を計算し，ドキュメント画像との類似度が高い画像を除外した.

B データセットの詳細

表 2 に，本研究で利用したデータセットの内訳を示す．ここで，「合成（A）」は，あらかじめ参照テキストが付与されているデータをもとに合成処理を行ったデータを指し，「合成（B）」は参照テキストが明示的に付与されていないデータをもとに合成処理を行ったデータを指す（2.1 節参照）．「翻訳」は，英語データセットを日本語に翻訳したデータを指す．特に指定がない場合は，日本語データセットをそのまま利用したものである.

Wikipedia は，WIT データセット [24] において，画像に対応して付与されている記事情報を参照テキストとして利用した．ArtBench は，絵画の作者に関する Wikipedia 記事の概要を参照テキストとして利用した.

C 合成キャプション生成のプロンプト

図 4 に，合成キャプション生成，英語キャプション翻訳，VQA データ作成，多肢選択式 VQA の選択肢作成のためのプロンプト例を示す．プロンプトは，共通の指示文の後に，各タスクごとの指示文が続く形式で構成されている．これに加えて，3 つ程度の例文を提示した.

1）　https://github.com/HojiChar/HojiChar

共通プロンプト

以下はタスクを説明する指示と、追加の背景情報を提供する入力の組み合わせです。要求を適切に満たす回答を書いてください。

指示
{ 各タスクごとの指示文 }

合成キャプション

ある画像についての 2 つの文章を簡潔に要約して，画像を説明するような 1 つの短い日本語文章にしてください．
1 つ目は，英語で画像の内容を説明する文章です．2 つ目は，その画像に関する日本語のテキストです．
英語テキストは，画像の大まかな内容については正確ですが，固有名詞などについては誤っていることがあります．日本語テキストは，画像の内容を表していないこともありますが，固有名詞などは正確であることがあります．
よって，最終的な回答は，英語テキストから大まかな内容を，日本語テキストから固有名詞などを取り入れてください．なお，画像中の文字についての言及は，最終的な日本語テキストから除いてください．

英語翻訳

ある画像についての英語の文章を翻訳して，画像を説明するような日本語の文章にしてください．

VQA 作成

ある画像についての文章から，質問のターゲットと，それをもとにした画像についての質問文と，その質問に対する回答を書いてください．
必ず日本語テキストから得られる情報を元にして，質問と回答を作成してください．質問のターゲットは，日本語テキストに含まれる情報を元にしてください．なお，画像中の文字についての言及は，最終的な日本語テキストから除いてください．

多肢選択式 VQA 作成

ある画像について，与えられた質問と回答をもとに，ダミーとなる回答選択肢を 3 つ生成してください．
A. ○○ B. ○○ C. ○○ といった形式にしてください．
回答は，なるべく正解とは明確に異なるものにしてください．生成した選択肢以外のテキストは生成しないでください．

図 4　合成データ生成のプロンプト例.

D モデル学習時のハイパーパラメータ

表 3, 4 に，本研究で用いたモデル学習時の主なハイパーパラメータや，モデルの詳細を示す.

表 3　モデル学習時の主なハイパーパラメータ.

	Stage1	Stage2
バッチサイズ	2048	144
訓練イテレーション	9,121	146,995
学習率関連		
— 最大学習率	1e-4	2e-5
— 最小学習率	1e-8	1e-8
— スケジューラー	コサイン	コサイン
— Warmup ratio	0.03	0.03
Weight decay	0.1	0.1
Optimizer	AdamW	AdamW

表 4　モデル設定の詳細.

	2B	4B	14B	173B
パラメータ数	2.3B	4.2B	14.2B	173.2B
画像エンコーダー	428M	428M	428M	428M
プロジェクター	13M	26M	64M	330M
LLM	1.8B	3.7B	13B	172B
Stage1				
テンソルパラレル	1	1	8	8
パイプラインパラレル	1	1	1	1
Stage2				
テンソルパラレル	1	2	4	8
パイプラインパラレル	1	1	1	8