MOMIJI: 日本語大規模インターリーブ視覚言語データセット

塩野 大輝 [1,2]　横井 慎吾 [1]　犬塚 眞太郎 [1]　高橋 翼 [1]　鈴木 潤 [2,3,4]　山口 祐 [1]
[1]Turing 株式会社　[2] 東北大学　[3] 理化学研究所　[4] 国立情報学研究所 LLMC
daiki.shiono.s1@dc.tohoku.ac.jp　yu.yamaguchi@turing-motors.com

概要

インターリーブ視覚言語データセットは，大規模視覚言語モデル（LVLM）の事前学習において中核的な役割を担っている．しかし，日本語の大規模なインターリーブ視覚言語データセットは少なく，また Interleave レイアウトの有効性に関する証拠も一貫していない．そこで我々は，Common Crawl から構築したオープンな日本語大規模インターリーブ視覚言語データセット MOMIJI を提案する．さらに，Interleave レイアウトと画像-テキスト間の意味的関連性を互いに独立に変化させる 10 通りの統制されたデータセット派生版を構築し，これらで LVLM を事前学習した際の性能を比較することで，Interleave レイアウトが有効となるデータ特性を調査した．

1　はじめに

インターリーブ視覚言語データとは，画像とテキストセグメントが共有されたコンテキスト内で交互に現れる形式を有するウェブ文書のようなデータである．このようなデータは大規模視覚言語モデル（LVLM）の事前学習の要となっており，基盤となる大規模言語モデル（LLM）の推論能力や文脈内学習能力を継承できるようにすることが期待される [1, 2, 3]．英語・中国語については大規模なインターリーブデータセットが豊富に存在する一方で [4, 5, 6, 7]，日本語に対しては高品質な資源が依然として限られている．さらに，先行研究ではインターリーブデータのクリーニングの重要性を強調してきたが [8]，具体的なフィルタリングパイプラインを開示したり，Interleave レイアウトと画像-テキスト間の意味的関連性がそれぞれ LVLM の性能にどのように個別に寄与するかを定量化している研究は少ない．その結果，これら 2 つの要因の相対的な重要性は依然として不明確である．

このギャップを埋めるため，私たちは **MOMIJI**（**M**odern **O**pen **M**ult**I**modal **J**apanese f**I**ltered Dataset）を提案する．MOMIJI は，Common Crawl から抽出したウェブ文書に厳格なフィルタリングパイプラインを適用することで構築された高品質な日本語インターリーブ視覚言語データセットである（図 1）．また，レイアウトの寄与と画像-テキスト間の意味的関連性の寄与が LVLM の事前学習にどのように影響するか調査するため，MOMIJI を基盤として，レイアウトと意味的関連性を変化させた 10 通りの制御されたデータセット派生版を作成する（図 2）．包括的な実験を通じて，意味的関連性が改善の主要な駆動要因であり，関連性が高い場合に Interleave レイアウトが追加の利得をもたらすことが確認された．

2　データセット構築

本節では，日本語のインターリーブ視覚言語データセットである **MOMIJI** の構築に用いたパイプラインを詳述する（図 1）．このワークフローは 3 つの主要な段階から構成される: (1) 画像 URL を含む日本語テキストの抽出，(2) テキスト品質に基づく文書フィルタリング，(3) 画像フィルタリング．最終的に，MOMIJI は約 56M 件の日本語ウェブ文書から構成され，1 文書あたりの平均文字数は 1,959 文字，平均画像枚数は 4.45 枚となった．

画像 URL を含む日本語テキストの抽出　本研究では，2024 年 2 月から 2025 年 1 月にかけて収集された Common Crawl の WARC アーカイブ[1)]を使用した．日本語の文書は Common Crawl 全体の約 5% に過ぎないこととテキスト抽出は言語判定よりも時間を要するという背景から，大規模日本語コーパス Swallow Corpus のパイプライン [9] を参考に，最初に粗いかつ迅速な日本語判定を行い，その後に高精度な日本語判定を適用する．粗いかつ迅速な日本語判定には，ひらがな，カタカナ，または漢字の文字集合を一切含まない文書を除外することを実施する．次に，`warcio` ライブラリ[2)]を用いて

1）https://data.commoncrawl.org/
2）https://github.com/webrecorder/warcio

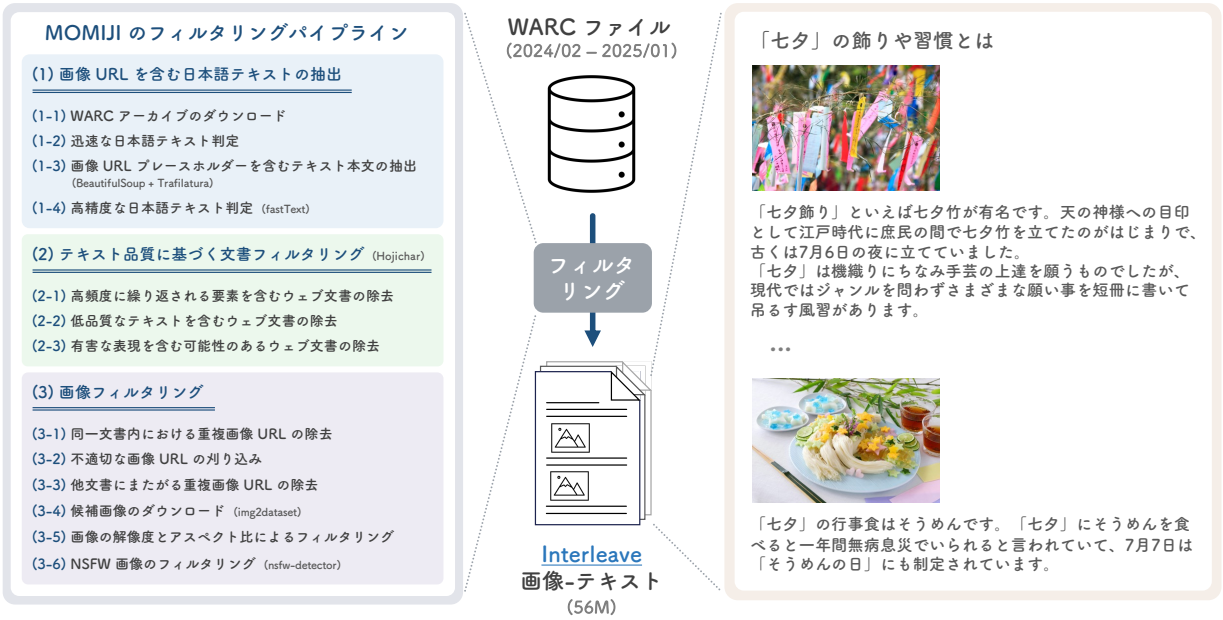

図 1　MOMIJI のフィルタリングパイプライン． Common Crawl WARC アーカイブに 3 段階のフィルタリングパイプラインを施すことで，高品質な日本語画像-テキスト混在（インターリーブ）データを抽出する．これらの手順により，計 56M 件のインターリーブデータが得られる．右: MOMIJI に含まれるデータの一例．

WARC ファイルから HTML コンテンツを抽出し，`bs4` ライブラリ[3]で解析する．そして，以下の条件の少なくとも 1 つを満たすウェブ文書のみを保持する: (i) 文書が日本語を明示的に宣言している場合，例えば `<html lang="ja">`; または (ii) `<title>` 要素が，高精度言語判定器である `fastText` ライブラリ[4] [10] により日本語として分類される場合．次に，`bs4` を `Trafilatura` ライブラリ[5] [11] と併用することで，画像 URL のプレースホルダーを保持したままテキスト本文を抽出する．これらのプレースホルダーを一時的に除去した後，テキスト本文を対象に `fastText` による高精度な言語判定を行い，日本語と判定されたウェブ文書のみを保持する．

テキスト品質に基づく文書フィルタリング　Swallow Corpus [9]，LLM-jp Corpus [12]，および OBELICS [5] の構築で用いられた手法を参考に，画像とテキストの自然な対応関係を保持しつつ，テキスト品質に基づく文書フィルタリングを実施する．このフィルタリングパイプラインは `Hojichar` ライブラリ[6]を用いて実装されている．各文書には，以下の 3 つの連続するフィルタを順に適用する: (1) 高頻度に繰り返される要素を含むウェブ文書の除去，(2) 低品質なテキストを含むウェブ文書の除去，そして (3) 有害な表現を含む可能性のあるウェブ文書の除去．このフィルタリング処理により，低品質なテキストを含むウェブ文書を除去し，事前に定義された品質基準を満たすもののみが保持される．文書フィルタリング処理の詳細は，参考情報 A.1.1 を参照されたい．

画像フィルタリング　有用な画像だけが残るようにするため，私たちは 6 段階の画像フィルタリングパイプラインを適用する: (1) 同一文書内における重複画像 URL の除去，(2) 不適切な画像 URL の刈り込み，(3) 他文書にまたがる重複画像 URL の除去，(4) 候補画像のダウンロード，(5) 画像の解像度とアスペクト比によるフィルタリング，そして (6) NSFW 画像のフィルタリング．このパイプラインにより，ロゴ，バナー広告，NSFW 素材，その他の望ましくない画像を破棄する．また，最終的に有効な画像を 1 枚も含まないウェブ文書は，データセットから除外される．画像フィルタリング処理の詳細は，参考情報 A.1.2 を参照されたい．

3　実験設定

図 2 に示すように，MOMIJI に対して複数の後処理戦略を適用することで，10 通りの制御されたデータセット派生版を作成する．この制御設計により，レイアウトの寄与と意味的関連性の寄与を分離でき，LVLM の事前学習に有効となる視覚言語データ

3) https://github.com/wention/BeautifulSoup4
4) https://github.com/facebookresearch/fastText
5) https://trafilatura.readthedocs.io/
6) https://github.com/HojiChar/HojiChar

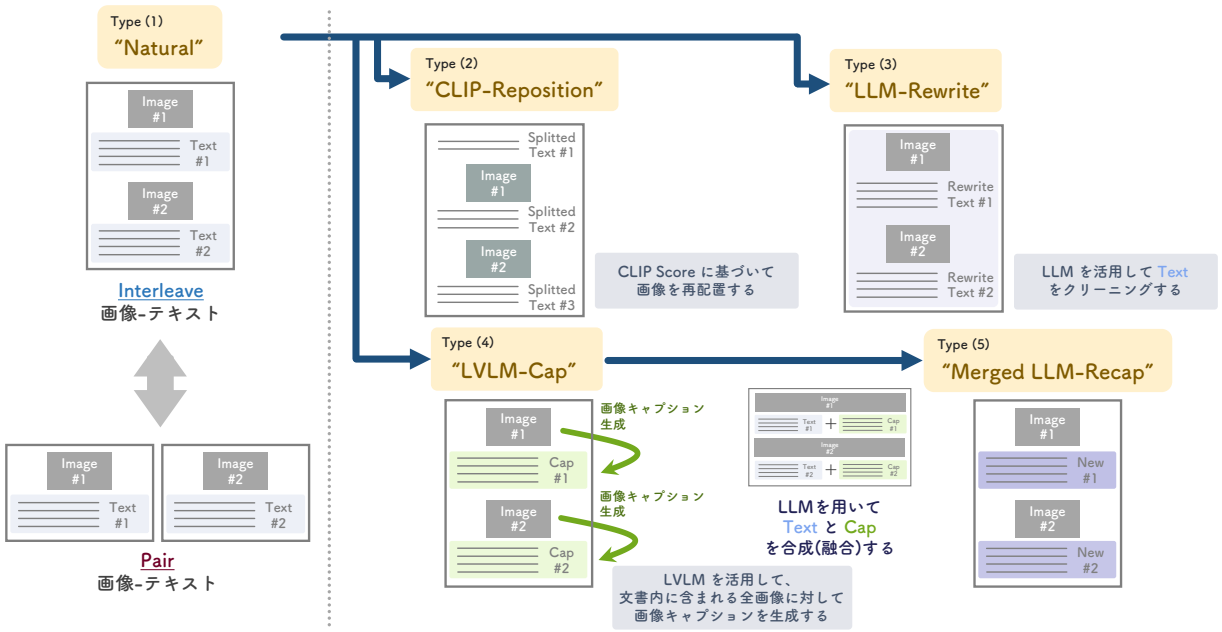

図 2　実験のためのデータセット後処理手法. *Natural* のインターリーブ画像-テキストデータ（左上）を，*CLIP-Reposition*，*LLM-Rewrite*，*LVLM-Cap*，*Merged LLM-Recap* の 4 つの派生版へと変換する．さらに，文書を連続した画像-テキストペアに分割することで，各派生版について Pair レイアウトも導出する（左下）．

の望ましい特性を調査できる．

3.1　派生データセットの準備

MOMIJI から 1M 件のデータを無作為抽出し，5 つの後処理手法を適用した（図 2）：

1. ***Natural*** オリジナルの MOMIJI データ．画像とテキストの元の順序を保持している．
2. ***CLIP-Reposition*** MMC4 [4] に従い，ウェブ文書内に含まれる各画像と各文との CLIP 類似度[7] [13, 14] に基づいて画像を再配置する．具体的には，日本語文境界判定器 bunkai[8] [15] でウェブ文書内のテキストを文に分割し，すべての文と画像間の CLIP スコアを計算し，スコアが最も高い文の直前に画像を移動する．目的: 意味的に近い位置に画像を再配置することがモデル性能を改善するかどうか調査する．
3. ***LLM-Rewrite*** レイアウトは *Natural* と同一であるが，すべてのテキストセグメントが LLM（Sarashina2.2-3B-Instruct[9]）で洗練される．例えば，文法的でない文や ". ", "!", "?" で終わらないものは，統語的一貫性を確保するために書き換えられる．目的: テキスト部分だけを洗練した際の効果を調査する．
4. ***LVLM-Cap*** LVLM として Qwen2.5-VL[10] [8] を用い，元のテキストを破棄して各画像から日本語キャプションを生成する．目的: 画像内容に強く焦点を当てた記述による利得を調査する．
5. ***Merged LLM-Recap*** *Natural* のテキストセグメントと *LVLM-Cap* の画像キャプションを，LLM（CALM3-22B-Chat[11] [16]）で合成（融合）する．目的: 元のコンテキストと画像中心のキャプションを組み合わせることで相乗的な利得が得られるかどうかを調査する．

また，レイアウトそのものの寄与を切り分けるため，各派生版について**画像-テキストペア**のレイアウトも導出する．具体的には，各文書を上から下へ走査し，各画像を，その次の画像に先行する連続したテキストセグメントとペアにする（図 2 左）．

この手順により 10 通りの異なる派生版（{Interleave, Pair} × {Natural, CLIP-Reposition, LLM-Rewrite, LVLM-Cap, Merged LLM-Recap}）が得られ，各データで LVLM を事前学習し，性能を比較する．

3.2　モデルアーキテクチャと学習手順

すべての実験は，NVILA [17] から導入された自己回帰型 LVLM である **Heron-NVILA** を用いて実施す

7）https://huggingface.co/line-corporation/clip-japanese-base
8）https://github.com/megagonlabs/bunkai
9）https://huggingface.co/sbintuitions/sarashina2.2-3b-instruct-v0.1
10）https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct
11）https://huggingface.co/cyberagent/calm3-22b-chat

る．このモデルは 3 つのモジュール，すなわち視覚エンコーダ[12] [18, 19]，プロジェクタ[13]，LLM[14] [20] から構成される．Heron-NVILA は，2 つの連続する段階で学習し，各段階につき 1 エポック学習する．

Stage 1: プロジェクタアライメント． この段階では，視覚エンコーダと LLM は凍結し，プロジェクタのみを学習可能とする．また，学習データセットには，すべての設定で常に，LLaVA-Pretrain (558K)[15] [21] と，llm-jp-japanese-image-text-pairs (6M サブセット)[16] [22] を用いる．

Stage 2: 事前学習． この段階では，視覚エンコーダは凍結し，プロジェクタと LLM は学習可能とする．学習データセットにはそれぞれ 3.1 節で導入した MOMIJI 由来の 10 個の派生版でモデルを事前学習する．この設計により，各データ条件がモデルの性能に与える影響を切り分けられる．

3.3 評価ベンチマーク

本研究では，日本語の視覚言語評価ベンチマークとして JMMMU [23]（多肢選択の推論および知識評価），Heron-Bench [24] および JA-VLM-Bench-In-the-Wild [25]（自由記述回答形式の推論および世界知識評価），JA-Multi-Image-VQA [26]（複数画像推論），JA-VG-VQA500 [27, 28]（一般的な視覚的質問応答），および JDocQA [29]（文書理解）を用いている．表 1 で報告されている平均スコアは上記のベンチマークスコアを正規化したスコアの平均値である．

4 結果

LVLM の事前学習におけるインターリーブ視覚言語データの有効性を調査するため，複数の視覚言語ベンチマークにわたって包括的な評価を実施した．

4.1 レイアウト比較（Interleave vs. Pair）

ほぼすべての設定において，事前学習した Heron-NVILA は Pair レイアウトよりも Interleave レイアウトを好む結果となった．Interleave レイアウトを用いることで *LVLM-Cap* は +6.8 ポイント（行 (d) と (i) を参照），*Merged LLM-Recap* は +4.1 ポイント（行 (e) と (j) を参照）を獲得した．

表 1 10 個の MOMIJI 由来のデータ派生版（{Interleave, Pair} × {Natural, CLIP-Reposition, LLM-Rewrite, LVLM-Cap, Merged LLM-Recap}）で事前学習した Heron-NVILA（7B）の性能．“平均スコア” 列は 3.3 節で紹介した 6 つの日本語視覚言語ベンチマークにわたる平均スコアを報告する．

ラベル	レイアウト	後処理タイプ	平均スコア (%)
(a)	Interleave	Natural	36.5
(b)	Interleave	CLIP-Reposition	37.9
(c)	Interleave	LLM-Rewrite	40.4
(d)	Interleave	LVLM-Cap	**45.2**
(e)	Interleave	Merged LLM-Recap	**42.6**
(f)	Pair	Natural	37.1
(g)	Pair	CLIP-Reposition	36.8
(h)	Pair	LLM-Rewrite	36.1
(i)	Pair	LVLM-Cap	38.4
(j)	Pair	Merged LLM-Recap	38.5

4.2 前処理タイプ別の比較

***Natural* vs. *LLM-Rewrite*.** 本実験設定の下では，LLM によるテキストの洗練だけでは有意な利点は得られなかった．*LLM-Rewrite* は，すべてのレイアウトにおいて *Natural* ベースラインと同等であるか，わずかに下回った（行 (a)，(c)，(f)，(h) を参照）．

***Natural* vs. *CLIP-Reposition*.** CLIP 類似度に基づく画像の再配置は，レイアウトにかかわらず，差分が僅少であった（行 (a)，(b)，(f)，(g) を参照）．

***Natural* vs. *LVLM-Cap* and *Merged LLM-Recap*.** *LVLM-Cap* と *Merged LLM-Recap* の両タイプがすべての前処理タイプの中で高いスコアを示しており，1M 規模のデータが利用可能な場合には，画像の内容と密に関係するテキストを注入することが重要な可能性が高い（行 (a)，(d)，(e)，(f)，(i)，(j) を参照）．

5 おわりに

本論文では，約 56M 件の画像とテキストの自然な位置関係を保持する高品質な日本語大規模インターリーブデータセット MOMIJI を構築した．また，インターリーブデータのレイアウトの寄与と画像-テキスト間の意味的関連性による寄与を分離した 10 通りの派生版を作成し，LVLM の事前学習に有効なデータ特性を調査したところ，意味的に強く関連した画像-テキストペアを含むインターリーブデータが LVLM の性能を大きく向上させる事が分かった．今後の研究では，画像-テキストペアの意味的関連性を強化させたインターリーブデータの規模をさらに拡大した場合の下流タスク性能を調査したい．

12） https://huggingface.co/Efficient-Large-Model/paligemma-siglip-so400m-patch14-448
13） 3×3 の spatial-to-channel（STC）を備えた 2 層 MLP
14） https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
15） https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain
16） https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-japanese-image-text-pairs

謝辞

本研究は，経済産業省および NEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）が実施する GENIAC 第 2 期 JPNP20017，JST ムーンショット型研究開発事業 JPMJMS2011-35 (fundamental research)，文部科学省の補助事業「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」，JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2421 の支援を受けたものです.

参考文献

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2022.

[2] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, et al. MM1: methods, analysis and insights from multimodal LLM pre-training. In **European Conference on Computer Vision (ECCV)**, pp. 304–323, 2024.

[3] Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In **The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)**, 2024.

[4] Wanrong Zhu, Jack Hessel, Anas Awadalla, et al. Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text. In **Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)**, 2023.

[5] Hugo Laurençon, Lucile Saulnier, Leo Tronchon, et al. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In **Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)**, 2023.

[6] Anas Awadalla, Le Xue, Oscar Lo, et al. MINT-1T: Scaling Open-Source Multimodal Data by 10x: A Multimodal Dataset with One Trillion Tokens. In **The Thirty-eight Conference on Neural Information Processing Systems (NeurIPS)**, 2024.

[7] Qingyun Li, Zhe Chen, Weiyun Wang, et al. OmniCorpus: A Unified Multimodal Corpus of 10 Billion-Level Images Interleaved with Text. In **The Thirteenth International Conference on Learning Representations (ICLR)**, 2025.

[8] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. **arXiv preprint**, cs.CL/2502.13923v1, 2025.

[9] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, et al. Building a Large Japanese Web Corpus for Large Language Models. In **First Conference on Language Modeling (COLM)**, 2024.

[10] Armand Joulin, Edouard Grave, Piotr Bojanowski, et al. Bag of Tricks for Efficient Text Classification. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, 2017.

[11] Adrien Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing: System Demonstrations**, 2021.

[12] Akiko Aizawa, Eiji Aramaki, Bowen Chen, et al. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **arXiv preprint**, cs.CL/2407.03963v2, 2024.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from natural language supervision. In **International conference on machine learning (ICML)**, pp. 8748–8763, 2021.

[14] Gabriel Ilharco, Mitchell Wortsman, et al. OpenCLIP, 2021.

[15] Yuta Hayashibe and Kensuke Mitsuzawa. Sentence Boundary Detection on Line Breaks in Japanese. In **Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)**, pp. 71–75, 2020.

[16] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.

[17] Zhijian Liu, Ligeng Zhu, Baifeng Shi, et al. NVILA: Efficient Frontier Visual Language Models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 4122–4134, June 2025.

[18] 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 (ICCV)**, 2023.

[19] Lucas Beyer, Andreas Steiner, André Susano Pinto, et al. PaliGemma: A versatile 3B VLM for transfer. **arXiv preprint**, cs.CL/2407.07726v2, 2024.

[20] An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 Technical Report. **arXiv preprint**, cs.CL/2412.15115v2, 2024.

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In **Advances in Neural Information Processing Systems (NeurIPS)**, pp. 34892–34916, 2023.

[22] Keito Sasagawa, Koki Maeda, Issa Sugiura, et al. Constructing Multimodal Datasets from Scratch for Rapid Development of a Japanese Visual Language Model. **arXiv preprint**, cs.CL/2410.22736v1, 2024.

[23] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, et al. JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark for Culture-aware Evaluation. In **Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)**, 2025.

[24] Yuichi Inoue, Kento Sasaki, Yuma Ochi, et al. Heron-Bench: A Benchmark for Evaluating Vision Language Models in Japanese. **arXiv preprint**, cs.CL/2404.07824v1, 2024.

[25] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary Optimization of Model Merging Recipes. **Nature Machine Intelligence**, pp. 1–10, 2025.

[26] Inoue Yuichi, Akiba Takuya, and Makoto Shing. Llama-3-EvoVLM-JP-v2, 2024.

[27] 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, 2018.

[28] Ranjay Krishna, Yuke Zhu, Oliver Groth, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. **International Journal of Computer Vision**, Vol. 123, pp. 32–73, 2017.

[29] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9503–9514, 2024.

[30] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. **Journal of machine Learning research**, Vol. 3, No. Jan, pp. 993–1022, 2003.

[31] Xiang Yue, Yuansheng Ni, Kai Zhang, et al. 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)**, 2024.

A　参考情報

A.1　MOMIJI フィルタリングの詳細

A.1.1　テキスト品質に基づく文書フィルタリング

ここでは，2 章の「テキスト品質に基づく文書フィルタリング」で導入した以下の 3 つの連続するフィルタについて詳細に説明する: (1) 高頻度に繰り返される要素を含むウェブ文書の除去，(2) 低品質なテキストを含むウェブ文書の除去，(3) 有害な表現を含む可能性のあるウェブ文書の除去.

(1) 高頻度に繰り返される要素を含むウェブ文書の除去

- 他の行と重複している行の総行数に対する比率（0.30 以上）
- 他の段落と重複している段落の総段落数に対する比率（0.30 以上）
- 重複している行に含まれる文字数の総文字数に対する比率（0.20 以上）
- 重複している段落に含まれる文字数の総文字数に対する比率（0.20 以上）
- 全 n-gram のうち最頻の n-gram の出現頻度（n=2: 0.20 以上; n=3: 0.18 以上; n=4: 0.16 以上）

(2) 低品質なテキストを含むウェブ文書の除去

- 文書内の総文字数（400 未満）
- 文書内におけるひらがなの文字割合（0.20 未満）
- 平均文長（文字数ベース）（20 未満または 90 超）
- 省略記号で終わる文の割合（0.20 以上）

(3) 有害な表現を含む可能性のあるウェブ文書の除去

- NG 表現の割合（文字数ベース）が 0.05 以上の場合，その文書を除去する．NG 表現に該当する語は以下の通り:
  - 日本語の不適切語
  - 日本語の差別語
  - 日本語の暴力的および脅迫的な語
- 句読点，記号，空白，絵文字，その他の特殊文字がテキストの少なくとも 0.40 を占める文書を除去する．
- いずれかの単一文字が 200 回以上繰り返されている文書を除去する．

A.1.2　画像フィルタリング

本節では，2 章の「画像フィルタリング」で導入した以下の 6 つの連続するフィルタについて詳細に説明する: (1) 同一文書内における重複画像 URL の除去，(2) 不適切な画像 URL の刈り込み，(3) 他文書にまたがる重複画像 URL の除去，(4) 候補画像のダウンロード，(5) 画像の解像度とアスペクト比によるフィルタリング，(6) NSFW 画像のフィルタリング.

(1) 同一文書内における重複画像 URL の除去

各ウェブ文書について，同一の画像 URL が複数回出現する場合，最初の出現に対応する画像プレースホルダー，すなわち文書中で最も上位に現れる画像プレースホルダーのみを保持する.

(2) 不適切な画像 URL の刈り込み

画像 URL は，(i) ファイル拡張子が `jpeg`，`jpg`，`png`，`webp` のいずれでもない，(ii) URL が事前定義したブラックリスト中のいずれかの語（例: `logo`，`button`，`icon`，`plugin`，`widget`）を含む，または (iii) URL が不適切表現リストに含まれる語を含む場合に破棄される.

(3) 他文書にまたがる重複画像 URL の除去

各処理バッチ内において，収集したウェブ文書全体で少なくとも 10 回以上出現する画像 URL はすべて除外する.

(4) 候補画像のダウンロード

`img2dataset` ライブラリ[17]を用いて，この段階で残っているすべての画像 URL から画像を取得する．ダウンロード中に，URL が無効になっている画像，幅または高さが 150 px 未満の画像は除外する．また，画像をダウンロードする際，URL，高さ，幅，SHA-256 ハッシュなどのメタデータも同時に取得する.

(5) 画像の解像度とアスペクト比によるフィルタリング

辺長が 150 px 未満または 20,000 px を超える画像，あるいはアスペクト比が 2:1 を超える，または 1:2 を下回る画像を破棄する.

(6) NSFW 画像のフィルタリング

最後に，この段階で残存するすべての画像にオープンソースの NSFW 画像分類器[18]を適用し，NSFW と判定されたものを除去する.

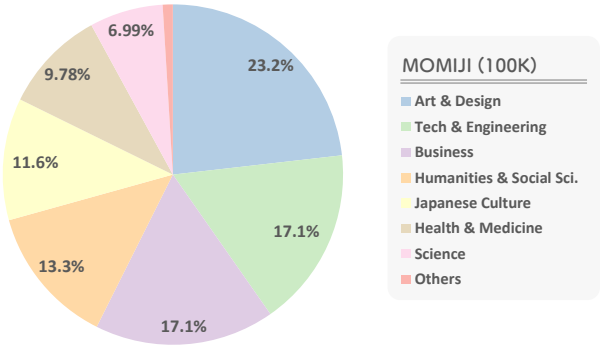

図 3　MOMIJI のウェブ文書（無作為抽出された 100K 件）におけるドメイン分布．GPT-4o は抽出された名詞を分析し，各文書を適切なドメインに割り当てる.

A.2　MOMIJI のデータ分析

MOMIJI に含まれるデータのドメインのカバレッジと潜在的なバイアスを調査するため，Latent Dirichlet Allocation（LDA）を適用した [30]．具体的には，MOMIJI から 100K 件のウェブ文書を無作為抽出し，LDA モデルを学習して 20 個のトピックを抽出した．各トピックに対する候補ドメインラベル集合は，MMMU [31] および JMMMU [23] のドメイン体系に着想を得て準備した．その後，GPT-4o を用いて，トピックの名詞集合と最も整合するラベルを選択した．得られた分布（図 3）は，MOMIJI が，単一のドメインに強く偏ることのない，広範でバランスの取れたドメイン分布を有していることを示している.

A.3　MOMIJI データセットの公開情報

本研究で構築した MOMIJI データセットは次の URL から入手可能である: https://huggingface.co/datasets/turing-motors/MOMIJI

17）https://github.com/rom1504/img2dataset
18）https://github.com/GantMan/nsfw_model