WAON: 視覚言語モデルのための大規模かつ高品質な日本語画像・テキスト対データセット

杉浦一瑳 ♣,◇ 栗田修平 ‡,◇ 小田悠介 ◇ 河原大輔 ♡,◇ 岡部寿男 ♣ 岡崎直観 †,◇
♣ 京都大学 ◇NII LLMC ‡ 国立情報学研究所 ♡ 早稲田大学 † 東京科学大学
sugiura.issa.q29@kyoto-u.jp

概要

視覚言語モデルの性能は事前学習に用いられる画像・テキスト対データセットの規模と品質に大きく依存する．しかし，英語や中国語に比べて，大規模かつ高品質な日本語画像・テキスト対データセットは不足している．この課題を解決するために，本研究では近年のデータキュレーションに関する知見を取り入れたデータセット構築パイプラインを整備し，Common Crawl を基に約 1 億 5,500 万件の事例の日本語画像・テキスト対データセット **WAON** を開発した．評価実験の結果，WAON は既存のデータセットと比較して日本文化タスクにおける性能を効率的に向上させることを示した．データセット，モデル，およびコードは公開する．[1)]．

1 はじめに

高性能な視覚言語モデル（VLM）を開発する上で画像・テキスト対データセットの規模と質は重要な要素である [1, 2, 3]．しかし，既存のデータセット構築の取り組みの多くは英語と中国語に焦点を当てており，大規模かつ高品質なデータセットをその他の言語向けに構築した事例は少ない．本研究では，日本語の言語および文化理解のための大規模かつ高品質なデータセットの構築を目指す．

表 1 に示すように，日本語の画像・テキスト対データセットはすでに複数提案されているが [1, 4, 5]，いくつかの問題がある．約 1 億 2,000 万件の日本語画像・テキスト対を含む ReLAION [1] の日本語サブセットは数年前に構築されたデータセットであり，2025 年 6 月時点で画像 URL の 30%弱がアクセス不能となっている．さらに，フィルタリングに用いられている mCLIP [6] は最新モデルと比べて性能が劣り，データ品質が低い [7]．また，ReLAION の英語サブセットのキャプションを LLM により日本語へ翻訳するアプローチも提案されているが，この方法では翻訳ミスや言語の混在が生じやすい．加えて，主に西洋文化圏の画像・テキスト対で構成されているため，日本文化に関連するタスクの性能向上に寄与しないという問題が報告されている [4]．

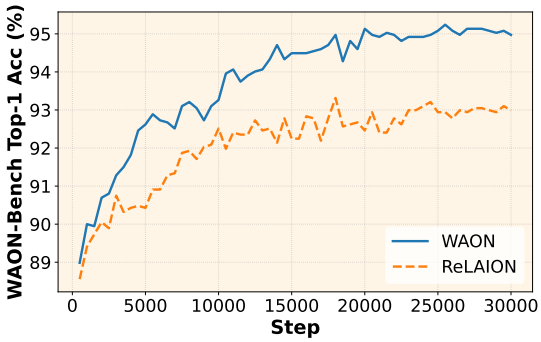

図 1 学習中の WAON-Bench の Top-1 accuracy.

表 1 日本語画像・テキスト対データセットの比較.

データセット	事例数	ソース
WIT (ja subset) [8]	1M	Wikipedia
ReLAION (ja subset) [1]	120M	Common Crawl
ReLAION-ja [4]	1.5B	Common Crawl
llm-jp-japanese-image-text [5]	6.6M	Common Crawl
WAON (Ours)	155M	Common Crawl

本研究ではこれらの問題を解決するために，Common Crawl から構築された 1 億 5,500 万件の大規模かつ高品質な日本語画像・テキスト対データセットである WAON (**W**eb-scale image text **A**ligned **O**pen **N**ihongo) を提案する．WAON の構築では，重複削除や強力な多言語モデル [9] を用いたフィルタリング [1] など，データキュレーションにおける近年の進展を活用している．提案する構築パイプラインは，言語識別のステップにおいてのみ日本語を使用しているため，他言語への適用も容易である．WAON と ReLAION の日本語サブセットを用いて SigLIP2 [9] をファインチューニングした結果，

1) https://speed1313.github.io/WAON

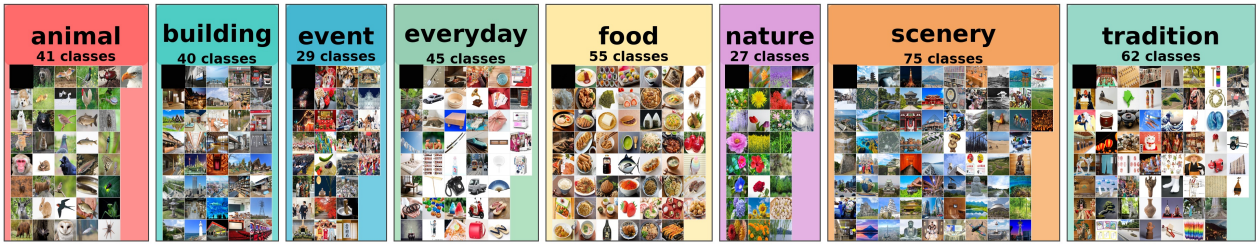

図 2　WAON-Bench の概要. 8 つのカテゴリ，374 のクラス，各クラス 5 枚の画像で構成された合計 1,870 事例の日本文化画像分類評価データセット.

図 3　WAON の構築パイプライン. 括弧内の数字は 2025-18 スナップショットにおける各プロセス実行後に残った事例数を示す. 6 スナップショット合わせて約 1 億 5,500 万対得られた.

WAON は ReLAION より効率的日本文化タスクでのモデル性能を向上させ，その他の日本語タスクにおいても一貫して ReLAION を上回ることが示された.

2　WAON の構築

図 3 に今回用いたデータセット構築パイプラインを示す. 大規模 Web データセット構築における既存研究 [1, 10, 11] に従い，複数のフィルタリングおよび重複削除を組み込みつつ，サンプルの人手による確認を通して各工程を反復的に改善した. 本パイプラインを最新の 6 つの Common Crawl スナップショット（2024-26, 2024-33, 2024-42, 2024-51, 2025-08, 2025-18）に適用した結果，合計 1 億 5,500 万件の対データセットが得られた. 以下に，各ステップの詳細を述べる.

WARC ファイルダウンロード　本研究では，HTML を含む WARC ファイルを利用する. まず，Common Crawl からクロールスナップショットをダウンロードする. 各スナップショットには，およそ 90,000 ~ 100,000 個の WARC ファイルが含まれている.

HTML 抽出 & 言語識別　各 WARC ファイルには複数の HTML 文書が含まれている. そこで，まず日本語 HTML 文書のみを抽出する. Swallow Corpus [12] に従い，まず HTML ヘッダ内の lang 属性に基づいて言語を高速に判定する. その後，Trafilatura [13] を用いて文書本文を抽出し，Lingua[2] により言語を正確に検出する. <title> タグが空の文書は，高品質な画像・テキスト対が含まれる可能性が低いと判断し除外する.

(画像 URL，キャプション) 対の抽出　抽出した HTML から (画像 URL，キャプション) 対を取得する. キャプションは，<img>要素に埋め込まれた短い説明文である alt 属性，または画像の下に表示される<figcaption>要素から収集する. 無効な画像 URL や，Unicode のコードポイントにおける日本語文字を含まないキャプションは除外する.

(画像 URL，キャプション) 対の重複削除　Web 上には，広告画像やロゴ，単純な図など，重複した画像が多数含まれている. このような重複を削除することは，先行研究において有効であることが示されており，広く用いられている [10, 11]. 本ステップでは，画像 URL とキャプションの双方に対して重複削除を行う. 画像 URL とキャプションはそれぞれハッシュ値に変換し，独立に重複削除を行う. 重複がある事例については最初に見つかった事例のみを残し，他の事例を削除する. また，重複削除時のメモリ効率を高めるために Bloom フィルタ [14] を使用する. Bloom フィルタはメモリ効率の高い確率的データ構造であり，偽陽性は生じ得るが偽陰性は生じないという特性を持つ.

2)　https://github.com/pemistahl/lingua-py

画像ダウンロード　ここでは抽出した URL から画像をダウンロードする. 大規模な画像 URL リストから効率的に画像を取得するため，Web 画像を並列にダウンロードできる img2dataset [15] を用いた.

画像品質フィルタリング　Web 上の画像には広告などの低品質なコンテンツが多く含まれているが，単純なヒューリスティクスによって一定程度除去できる. たとえば極端に横長のバナーなど異常なアスペクト比を持つ画像は広告である可能性が高い. ここでは幅または高さが 150 ピクセル未満の画像，アスペクト比が 0.5 ~ 2.0 の範囲外にある画像を除外した [11]. さらに独自の色多様性フィルタを適用し，画像内のユニークな色数が 32 色以下の画像は除去した.

NSFW フィルタリング　多くの Web 画像にはアダルト画像などの不適切な内容 (NSFW) が含まれるためこれらを除去する必要がある. 本研究では不適切画像検出のために，OpenCLIP [2] を基盤とした NSFW 分類モデルの dataset2metadata[3] を用いる. unsafe スコアが 0.1 を超える画像を除外する. 閾値 0.1 は既存研究 [11, 5] に従ったものであり，スコア分布の分析およびランダムサンプルの手動確認により，この設定が適切であることを確認した. この処理により，NSFW コンテンツの大部分を効果的に除去できた.

pHash ベース類似画像重複削除　前述のキュレーション工程を経ても，URL やキャプションが異なるものの，見た目がほぼ同一の類似画像が多数残っている. これらを除去するため，ImageHash[4] を用いて pHash を計算し，Bloom フィルタによる重複削除を行う. pHash は視覚的に類似した画像同士でハッシュ値がほとんど変わらないように設計されており，類似画像を検出できる. 本研究ではハミング距離に基づく近似重複削除ではなく，ハッシュ値の完全一致による重複削除を採用した.

SigLIP スコアベースフィルタリング　画像に付随するキャプションは HTML 作成時に手動または自動で生成されるため，画像とテキストの対応が不適切な対が多く含まれる. これらの事例を除去するため，画像とテキストの埋め込み間のコサイン類似度を用いたフィルタリング [1] を適用した. ここでは siglip2-base-patch16-256 [9] を用いて類似度を計算した. 分布に基づく分析とランダムサンプルの手動確認の上，類似度の閾値は 0.1 に設定し，それ未満の対を除外した.

3) https://github.com/mlfoundations/dataset2metadata
4) https://pypi.org/project/ImageHash

3　WAON-Bench の構築

日本文化的概念の理解を適切に評価できる画像分類評価データセットとして利用されるものとして Recruit データセット [16] がある. しかし Recruit データセットにはいくつかの問題が存在している. 第一に，161 クラス中，101 クラスが食カテゴリに属しており偏りがある. 第二に，多数のラベル誤りデータが存在している. 例えば,「明治神宮」とラベル付けされた画像が実際には芝生だけを写している例が存在する. そこで，我々は日本文化画像分類ベンチマーク WAON-Bench を新たに構築した. 図 2 に WAON-Bench の概要を示す. WAON-Bench は 8 つのカテゴリ（animal, building, event, everyday, food, nature, scenery, tradition）にまたがる 374 クラスから構成される. 各クラスには 5 枚の画像が含まれ，合計 1,870 枚の画像を含む.

3.1　ベンチマーク構築パイプライン

WAON-Bench は以下の手順で構築した.

1. クラス名の定義: 日本文化に関連する 374 個のクラス名（例: 柴犬，東京タワー）を定義した. クラス名は，Wikipedia，Google 検索，ChatGPT や，京都など実際の街歩きでの観察をもとに用意した.
2. カテゴリの分類: 8 つのカテゴリを定義し，各クラスをいずれかに割り当てた. ImageNet [20] のように階層的クラス関係を持つ WordNet を用いた厳密な対応関係が理想ではあるが，日本語版 WordNet はカバレッジが不十分で現代的な用語が多く欠けていたため，手作業でカテゴリへの割り当てを行った.
3. 画像収集と選定: 各クラスについてクラス名を Google 画像検索のクエリとし，検索結果から手動で 5 枚の画像を選んだ. 選定にあたっては構図，視点，背景などの多様性を重視し，クラス内で幅広い視覚的バリエーションが得られるよう配慮した. また他クラスの要素を含む画像を避け，ミスラベルの発生に注意した.

これらの作業は選定基準の一貫性と品質を保つため，クラウドソーシングを用いず著者が行った.

表 2　各タスクの評価結果.

モデル	パラメータ数	検索	画像分類			平均
		$\text{XM3600}_{\text{ja}}$	$\text{ImageNet}_{\text{ja}}$	Recruit	WAON-Bench	
siglip2-base-patch16-256 (fine-tuned on WAON)	375M	73.75	49.61	**83.14**	**94.97**	**75.37**
siglip2-base-patch16-256 (fine-tuned on ReLAION)	375M	72.39	47.38	81.65	92.99	73.60
siglip2-base-patch16-256 [9]	375M	38.28	48.12	76.98	87.81	62.80
clip-japanese-base [17]	196M	**78.00**	48.90	81.65	90.05	74.65
siglip-base-patch16-256-mult [18]	371M	43.22	53.26	75.10	89.25	65.21
Japanese Stable CLIP ViT-L-16 [19]	414M	66.03	**55.97**	71.29	82.03	68.83
LAION-CLIP-ViT-H-14 [2]	1193M	72.64	47.67	70.62	85.88	69.20

4　評価

WAON の品質を評価するために，siglip2-base-patch16-256 [9] を WAON でファインチューニングし，いくつかのベンチマークにおいてベースラインモデルと性能を比較する.

4.1　学習設定

学習ステップ数は 30,000，バッチサイズは 8,192 とした. 学習率は最大 1e-5 とし，コサイン減衰スケジューリングを採用，ウォームアップは 1,500 ステップ，最小学習率は 1e-7 とした.

4.2　学習データセット

学習データは WAON と ReLAION の日本語サブセットを用いた場合で比較した. ReLAION の日本語サブセットは合計 1 億 2,000 万例を含むが，2025 年 6 月時点でダウンロード可能だった 8,500 万枚の事例を使用した. WAON と ReLAION 日本語サブセットはデータ規模が異なるものの，学習ステップ数とハイパーパラメータを揃えて公平な比較を行った.

4.3　評価データセット

我々はゼロショット画像分類タスクとゼロショット画像・テキスト検索タスクで評価した. 画像分類には WAON-Bench，ImageNet [20]，Recruit [16] を使用し，画像・テキスト検索には XM3600 [21] を使用した.

4.4　ベースラインモデル

ベースラインモデルとして，日本語モデルの clip-japanese-base [17]，Japanese Stable CLIP ViT-L-16 [19]，多言語モデルの LAION-CLIP-ViT-H-14 [2]，siglip-base-patch16-256-mult [18]，siglip2-base-patch16-256 [9] を用いた.

4.5　評価結果

WAON は効率的にモデル性能を向上させる. 図 1 に学習中の WAON-Bench の Top-1 Accuracy の推移を示す. ReLAION で学習したモデルは約 93%で飽和する一方，WAON で学習したモデルは継続的に向上し，約 95%に達する. さらに WAON は各学習ステップにおいて ReLAION を上回っており，より効率的に性能を向上させることがわかる. また，各モデルのベンチマーク横断での性能を示した表 2 では，WAON は評価したすべてのタスクで ReLAION の日本語サブセットを一貫して上回り幅広いベンチマークに対する有効性を示している.

WAON は日本文化タスクで最高性能を示す. 表 2 に示すように，WAON でファインチューニングしたモデルは日本文化タスクの WAON-Bench と Recruit の両方で最先端の性能を達成した. WAON を用いたファインチューニングによりすべてのタスクで性能向上が見られ，特に XM3600，Recruit，WAON-Bench で大きな向上が見られた. ImageNet に対しては WAON でのファインチューニングによる改善は約 1.5 ポイントと小さい. これは，ImageNet のクラス名は日本語に翻訳されているものの，画像自体は英語圏文化に関連したカテゴリが中心であり，WAON に含まれる画像と分布が異なるためと考えられる.

5　おわりに

本研究では，大規模かつ高品質な日本語画像・テキスト対データセット WAON を構築した. さらに，信頼性の高い日本文化タスクの評価のために，日本文化画像分類データセット WAON-Bench を新たに構築した. 実験では，WAON は既存データセットより効率的にモデルの性能を向上させることを示した.

謝辞

本研究成果は，データ活用社会創成プラットフォーム mdx を利用して得られたものです．また，総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用しました．

参考文献

[1] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R 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 **NeurIPS (Datasets and Benchmarks Track)**, 2022.

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In **CVPR**, 2023.

[3] 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 **NeurIPS (Datasets and Benchmarks Track)**, 2023.

[4] Issa Sugiura, Shuhei Kurita, Yusuke Oda, Daisuke Kawahara, and Naoaki Okazaki. Developing Japanese CLIP models leveraging an open-weight LLM for large-scale dataset translation. In **NAACL (Student Research Workshop)**, 2025.

[5] 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)**, 2025.

[6] Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. mCLIP: Multilingual CLIP via cross-lingual transfer. In **ACL**, 2023.

[7] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, Xinlei Chen, Zhuang Liu, Saining Xie, Wen tau Yih, Shang-Wen Li, and Hu Xu. Meta CLIP 2: A worldwide scaling recipe. arXiv preprint arXiv:2507.22062, 2025.

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

[9] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

[10] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **ACL**, 2022.

[11] 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 **NeurIPS (Datasets and Benchmarks Track)**, 2023.

[12] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **COLM**, 2024.

[13] Adrien Barbaresi. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In Heng Ji, Jong C. Park, and Rui Xia, editors, **ACL (System Demonstrations)**, 2021.

[14] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. **ACM**, 1970.

[15] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset, 2021.

[16] Shion Honda and Hidehisa Arai. japanese-image-classification-evaluation-dataset. https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset, 2024.

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

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

[19] Makoto Shing and Takuya Akiba. Japanese Stable CLIP ViT-L/16. https://huggingface.co/stabilityai/japanese-stable-clip-vit-l-16, 2023.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In **CVPR**, 2009.

[21] Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In **EMNLP**, 2022.

[22] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In **LREC-COLING**, 2024.

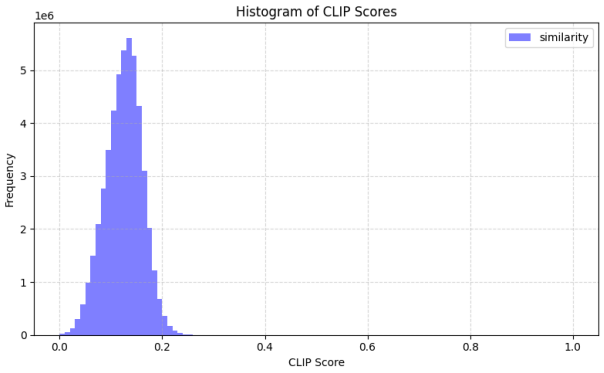

図 4　2025-18 snapshot の画像・テキスト対の SigLIP 類似度スコア分布. SigLIP スコアフィルタリングのステップでは，0.1 を下回る対を削除した.

表 3　重複削除後の各スナップショットの事例数.

スナップショット	事例数
2025-18	37,445,634
2025-08	36,043,758
2024-51	28,178,004
2024-42	20,221,965
2024-33	17,910,213
2024-26	15,433,133
Total	155,232,707

A　WAON の構築

図 4 には，フィルタリング前の 2025-18 スナップショットにおける SigLIP スコア分布を示す.

A.1　WAON の統計

まず各スナップショットに対してパイプラインを実行し，その後，2025-18 から 2024-26 に向かって順に，画像 URL・キャプション・pHash を跨いで重複削除を行う. 早期のスナップショットほど既出データの割合が高いため，重複削除によって除去されるデータが多く，最終的に残る例数は少なくなる. 最終的に，約 1.55 億件の日本語画像—テキスト対データセットを得た. 表 3 に重複削除後のスナップショットごとの事例数を示す.

B　WAON-Bench の分析

WAON-Bench に含まれる画像の意味的分布や潜在的なバイアスを調べるため，siglip2-base-patch16-256 の画像エンコーダで得た画像埋め込みに t-SNE を適用した.

図 5 に WAON-Bench の画像埋め込みに対する t-SNE マップを示す. animal，nature，food の 3 つのカテゴリは明確なクラスタを形成している一方，それ以外のカテゴリは互いに混ざり合っている. これは，この 3 カテゴリが他と比べて意味的により独立しているのに対し，残りのカテゴリは視覚的または概念的特徴がより重なり合っていることを示唆している. したがって，カテゴリ情報は厳密な階層ラベルとして扱うのではなく，補助的なメタデータとして利用することを推奨する.

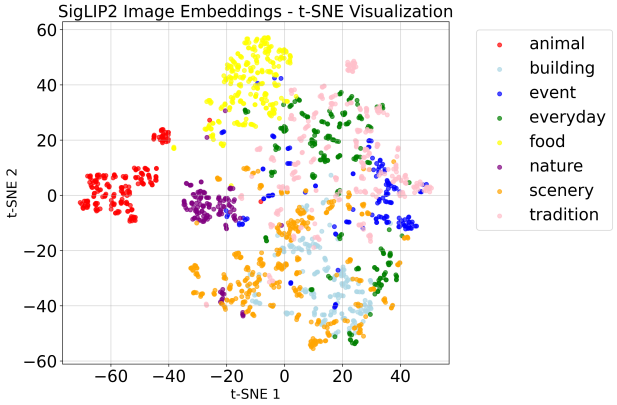

図 5　WAON-Bench に含まれる画像の埋め込みの t-SNE マップ.

表 4　日本語評価データセットの比較

タスク	データセット	クラス数	事例数
画像分類	ImageNet [20] (ja translation)	1,000	50,000
	Recruit [16]	161	7,654
	WAON-Bench (Ours)	374	1,870
検索	XM3600 [21] (ja annotation)	–	3,600

C　実験に用いた評価データセット

表 4 に各データセットの統計情報をまとめる.

ImageNet は 1,000 クラスからなる画像分類データセットであり，元のクラス名は英語だが，日本語で正しく画像と対応付けられるかを評価するため，日本語翻訳版 [22] を用いる. Recruit [16] は日本文化に特化した 161 クラスの画像分類データセットで，food・flower・facility・japanese landmark の 4 カテゴリから構成される. XM3600 [21] は 3,600 枚の画像に対して 36 言語のテキストが付与された画像・テキスト検索データセットである. 本研究では各画像に割り当てられた最初の日本語アノテーションを使用した. 評価指標として画像分類には top-1 accuracy，画像・テキスト検索には recall@1 を用いた.