WAON: A Large-Scale Japanese Image-Text Dataset for Cultural Adaptation in Contrastive Vision-Language Models

Anonymous ACL submission

Abstract

Contrastive vision-language models have achieved remarkable progress through large-scale pretraining. Recent work has shown that removing English-only caption filters and pretraining on global data is effective for improving multicultural performance. We study whether such global pretraining is sufficient for culture-specific understanding, or whether further adaptation with natively sourced data can boost performance beyond what global pretraining alone achieves. To enable this investigation, we present WAON, the largest publicly available native Japanese image-text dataset constructed from native Japanese web content in Common Crawl, containing approximately 155 million examples. We also introduce WAON-Bench, a manually curated Japanese cultural benchmark spanning 374 classes. Through comparative fine-tuning experiments on multiple Japanese image-text datasets, we observe that models fine-tuned on WAON consistently achieve stronger performance on Japanese cultural benchmarks than those fine-tuned on English-to-Japanese translated data. We release our dataset and code.

1 Introduction

Contrastive pre-training on large-scale image-text pair datasets has achieved remarkable progress in learning aligned representations of images and text (Dosovitskiy et al., 2021; Radford et al., 2021; Zhai et al., 2023). Early work typically filtered training data to English-only captions (Changpinyo et al., 2021; Radford et al., 2021; Fang et al., 2024), resulting in models heavily biased toward English and Western content. Recent work has shown that removing this filter retains images from diverse cultures and that pretraining on such globa data is effective for improving multicultural performance (Pouget et al., 2024; Nguyen et al., 2024), leading to the release of strong open-weight multilingual models such as SigLIP2 and MetaCLIP2 (Tschannen et al., 2025; Chuang et al., 2025).

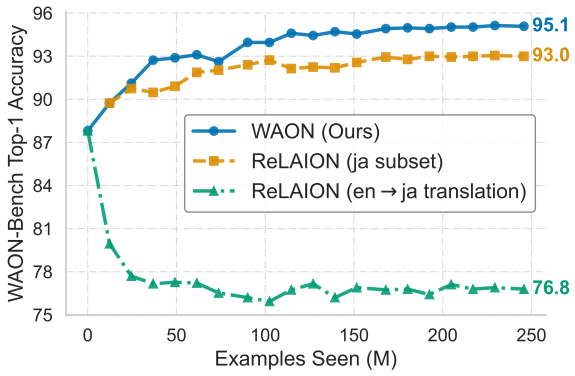


Figure 1: Zero-shot top-1 accuracy on WAON-Bench (a Japanese cultural benchmark) over training steps. Fine-tuning on ReLAION (en→ja translation) degrades performance from 87.8% to 76.8%, while WAON achieves the highest accuracy of 95.1%.

We study whether such global pretraining is sufficient for culture-specific understanding, or whether further adaptation with natively sourced data can boost performance beyond what global pretraining alone achieves, taking Japanese as a case study. To enable this investigation, we present **WAON**[1] (**W**eb-scale image text **A**ligned **O**pen **N**ihongo), the largest publicly available native Japanese image-text dataset constructed from native Japanese web content in Common Crawl, containing 155M examples, filtered using SigLIP2. We also introduce **WAON-Bench**, a manually curated benchmark for Japanese cultural image classification with 1,870 images across 374 classes spanning 8 diverse categories, covering more than twice as many cultural concepts as the previous benchmark (Honda and Arai, 2024).

Our experiments reveal that fine-tuning on data translated from English into Japanese consistently

[1] *Waon* (和音) means "harmony" in Japanese.

Dataset	Examples	Source	Caption Source	Quality Filter
WIT (ja subset) (Srinivasan et al., 2021)	1M	Wikipedia	Native	-
llm-jp-japanese-image-text-pairs (Sasagawa et al., 2025)	6.6M	Common Crawl	Native	OpenCLIP + Ja CLIP
ReLAION (ja subset) (Schuhmann et al., 2022)	120M	Common Crawl	Native	mCLIP
ReLAION (en→ja translation) (Sugiura et al., 2025)	1,452M	Common Crawl	Translation	mCLIP
WAON (Ours)	155M	Common Crawl	Native	SigLIP2

Table 1: Comparison of publicly available Japanese image-text pair datasets. "Native" indicates that captions are sourced directly from Japanese web content; "Translation" indicates captions obtained by translating those from English datasets. WAON is the largest dataset with natively sourced Japanese captions.

degrades Japanese cultural understanding, and WAON achieves the best performance on Japanese cultural benchmarks among publicly available models with comparable architectures. These results suggest that translation alone cannot compensate for the absence of natively sourced data.

In summary, our contributions are: (1) we release WAON, the largest publicly available natively sourced Japanese image-text dataset containing 155M examples; (2) we introduce WAON-Bench, a culturally comprehensive Japanese benchmark that more than doubles the class coverage of the prior benchmark; and (3) we provide empirical observations suggesting that data origin is a critical factor for cultural adaptation in contrastive vision-language models.

2 Related Work

Cultural understanding in contrastive vision-language models. Early work on contrastive vision-language models, spanning both model development and dataset construction, commonly filtered training data to English-only image-text pairs under the assumption that this would improve data quality (Radford et al., 2021; Schuhmann et al., 2022; Gadre et al., 2023; Fang et al., 2024). However, recent work has shown that incorporating non-English and geographically diverse data improves cultural understanding in vision-language models (Pouget et al., 2024; Nguyen et al., 2024). Nguyen et al. (2024) show that non-English data contains unique visual information and broader cultural concepts not captured by English-only corpora, and that training on translated versions of such data improves both English-centric (Deng et al., 2009) and geographically diverse tasks (Ramaswamy et al., 2023). No Filter (Pouget et al., 2024) highlights the importance of culturally diverse data for training contrastive models. Meta CLIP 2 (Chuang et al., 2025) trains contrastive vision-language models on global data at scale. These works highlight the importance of incorporating non-English data during training. Building on these findings, our work further demonstrates that, in the setting of Japanese cultural adaptation via fine-tuning vision-language models pre-trained on global data, native Japanese data remains crucial: fine-tuning with English-to-Japanese translated data degrades Japanese cultural understanding, whereas natively collected Japanese data more effectively improves performance on Japanese cultural tasks.

Image-text pair datasets. image-text pair datasets have played a central role in advancing contrastive vision-language models (Radford et al., 2021). Early efforts relied on small-scale datasets annotated by human experts (Deng et al., 2009; Lin et al., 2015). With the introduction of CLIP (Radford et al., 2021), which leveraged web-scale data, the construction of large-scale datasets from sources such as Common Crawl became a major research focus. Among these, LAION-5B (Schuhmann et al., 2022) has been the most influential, serving as the first public billion-scale dataset that has powered a wide range of models (Cherti et al., 2023; Rombach et al., 2022; Lin et al., 2024). Alongside scaling datasets, recent research has prioritized data quality. DataComp (Gadre et al., 2023) introduced a systematic benchmark for evaluating dataset curation methods. DFN (Fang et al., 2024) developed a data filtering model that enables training high-performance CLIP models. Meta CLIP (Xu et al., 2024) proposed a metadata-based curation approach that yields higher-quality data than the original CLIP dataset.

Japanese vision-language resources. For Japanese, several image-text pair datasets exist as shown in Table 1, yet large-scale datasets free from English bias remain scarce. WIT (Srinivasan et al., 2021) is constructed by collecting image-text pairs

from Wikipedia, but contains only 1M Japanese examples, insufficient for extensive fine-tuning. ReLAION (ja subset) (Schuhmann et al., 2022) is the Japanese subset of ReLAION-5B (Schuhmann et al., 2022), a large-scale multilingual image-text dataset, containing 120M examples. However, it relies on mCLIP (Chen et al., 2023a) for quality filtering. mCLIP is trained on a machine-translated extension of CC12M (Changpinyo et al., 2021), which was constructed by filtering web-crawled data to English captions only, and is substantially smaller in scale than recent multilingual models (Tschannen et al., 2025; Chuang et al., 2025). These limitations introduce English bias (Pouget et al., 2024) and limited performance, potentially causing mCLIP to inadequately score Japanese culture-specific image-text pairs during filtering. ReLAION (en→ja translation) (Sugiura et al., 2025) is constructed by translating the English captions of the ReLAION English subset into Japanese, covering 1,452M image-text pairs, but introduces translation errors and inherits cultural biases from the source data. Evaluation resources are equally limited: the only existing Japanese cultural benchmark, Recruit (Honda and Arai, 2024), comprises 161 classes heavily concentrated in food categories, offering limited coverage of Japanese culture as a whole. WAON and WAON-Bench address both gaps simultaneously.

3 Building WAON

We construct **WAON**, a large-scale Japanese image-text pair dataset based on Common Crawl (Common Crawl, 2025). We first extract (image URL, caption) pairs from Japanese HTML pages in Common Crawl snapshots, and then apply filtering and deduplication to obtain 155M Japanese image-text pairs. We describe the construction pipeline below.

3.1 Dataset Construction Pipeline

Figure 2 illustrates our dataset construction pipeline. We first extract Japanese HTML documents from WARC files as efficiently as possible to reduce the number of samples processed in later, computationally intensive steps such as model-based filtering. After extracting image-text pairs from the HTML, we perform text-based deduplication before downloading images, followed by image-based filtering and deduplication. We apply this pipeline to six recent Common Crawl snapshots (2024-26, 2024-33, 2024-42, 2024-51, 2025-08, and 2025-18), iteratively refining each step through manual inspection of sampled subsets following established practices (Schuhmann et al., 2022; Lee et al., 2022; Zhu et al., 2023).

Figure 2: Construction pipeline of WAON. Beginning with six Common Crawl snapshots, we extract Japanese image-text pairs and apply cascaded filtering and deduplication, yielding 155M high-quality pairs.

Japanese HTML extraction. Common Crawl provides archives in three formats: WARC, WAT, and WET. We use WARC files, which contain complete HTML documents and corresponding request metadata. Each snapshot contains approximately 90,000 to 100,000 WARC files, each containing multiple raw HTML documents.

We retain only Japanese HTML documents, as they are more likely to contain Japanese captions and culturally relevant images (Nguyen et al., 2024). Following Okazaki et al. (2024), we first determine language based on the `lang` attribute in the HTML header, and apply model-based language identification only to documents that pass this initial filter, as running it on all HTML bodies would be computationally prohibitive at our scale. We then apply Trafilatura (Barbaresi, 2021) to extract the underlying text, and Lingua (Stahl, 2022) to confirm the document language. We also discard documents with an empty `<title>` tag, as high-quality image-text pairs are more likely to appear in well-structured HTML pages.

Image-text pair extraction. From the extracted HTML, we obtain (image URL, caption) pairs.

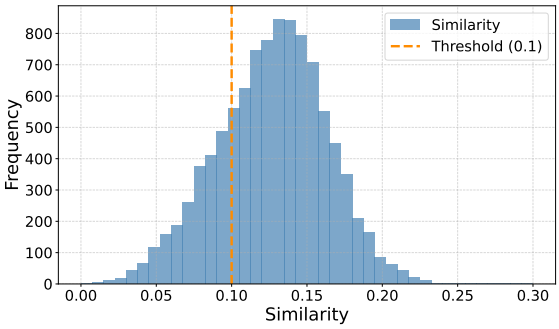


Figure 3: Distribution of SigLIP2-base cosine similarity scores for 10,000 randomly sampled image-text pairs prior to alignment filtering. Pairs below the threshold of 0.1 are removed.

Snapshot	Examples
CC-MAIN-2025-18	37,445,634
CC-MAIN-2025-08	36,043,758
CC-MAIN-2024-51	28,178,004
CC-MAIN-2024-42	20,221,965
CC-MAIN-2024-33	17,910,213
CC-MAIN-2024-26	15,433,133
Total	155,232,707

Table 2: Number of image-text pairs per Common Crawl snapshot in WAON after cross-snapshot deduplication of image URLs, captions, and pHash values. Earlier snapshots contain a higher proportion of previously seen content, resulting in fewer remaining pairs after deduplication.

Captions are collected from either the `alt` attribute of the image tag or the corresponding `<figcaption>` element. We remove entries with invalid image URLs and those whose captions contain no Japanese characters (i.e., no Unicode code points in the Hiragana, Katakana, or CJK Unified Ideographs ranges).

Pair deduplication. Duplicates on the web often comprise low-diversity content such as advertising images, logos, and simple graphics. Following prior work (Lee et al., 2022; Zhu et al., 2023), we deduplicate on both image URLs and captions, retaining only the first occurrence of each (image URL, caption) pair. To improve memory efficiency, we use a Bloom filter, a probabilistic data structure that may introduce false positives but guarantees no false negatives.

Image quality filtering. We download images from the deduplicated URL list using `img2dataset` (Beaumont, 2021), a tool designed for parallel downloading of web images. We then apply heuristic filtering to remove low-quality content: following mmC4 (Zhu et al., 2023), we discard images whose width or height is below 150 pixels or whose aspect ratio falls outside $[0.5, 2.0]$, as abnormal aspect ratios are indicative of advertisements or banners.

To remove unsafe content, we apply `dataset2metadata` (Gadre et al., 2023), an NSFW classifier based on OpenCLIP (Cherti et al., 2023). We set the filtering threshold to 0.1, determined by analyzing the score distribution and manually inspecting randomly sampled images across score ranges, following prior work (Zhu et al., 2023; Sasagawa et al., 2025). This process effectively removes most photographic and comic adult content.

pHash deduplication. Even after the previous steps, many near-duplicate images remain that are visually similar despite having different URLs and captions. We compute perceptual hashes (pHash) using ImageHash (Buchner, 2021) and deduplicate using a Bloom filter. We opt for exact matching rather than Hamming distance-based deduplication, as the latter requires computing pairwise distances across all images, which is computationally prohibitive at our scale.

Image-text alignment filtering. Many image-text pairs on the web exhibit poor semantic alignment between visual and textual content, as captions are created during HTML authoring without strict consistency requirements. Following LAION (Schuhmann et al., 2022), we filter misaligned pairs using cosine similarity between image and text embeddings computed by SigLIP2 (`siglip2-base-patch16-256`) (Tschannen et al., 2025). We use SigLIP2 instead of mCLIP (Chen et al., 2023a), which was used in prior work (Schuhmann et al., 2022), because SigLIP2 is trained on the multilingual WebLI dataset (Chen et al., 2023b) containing 12B image-text pairs across 109 languages, whereas mCLIP relies on translation-based multilingual transfer from a substantially smaller dataset. In our preliminary evaluations, SigLIP2 exhibited strong performance on Japanese cultural understanding tasks, making it better suited for evaluating semantic alignment in Japanese image-text pairs. Figure 3 shows the score distribution for 10,000 randomly sampled pairs from the CC-MAIN-2025-18 snapshot prior to filtering. Based on manual inspection of 100 randomly sampled pairs across score ranges, we set the threshold to

Figure 4: Randomly sampled images from WAON and ReLAION (en→ja translation) (Sugiura et al., 2025). WAON contains many images with Japanese text, scenes captured in Japan, and culturally Japanese content, whereas ReLAION (en→ja) more frequently contains images associated with English text and Western cultural contexts.

0.1 and discard pairs below this value, as pairs above this threshold were generally well aligned semantically.

Cross-Snapshot Deduplication and Statistics. We run the pipeline described above independently for each snapshot, then perform cross-snapshot deduplication of image URLs, captions, and pHash values in reverse chronological order from CC-MAIN-2025-18 to CC-MAIN-2024-26. As earlier snapshots contain a higher proportion of previously seen content, more pairs are removed during deduplication, resulting in fewer remaining examples (Table 2). In total, WAON contains approximately 155M image-text pairs.

3.2 Qualitative Analysis

To qualitatively examine the characteristics of WAON, we randomly sample image-text pairs from WAON and compare them with samples from ReLAION (en→ja translation) (Sugiura et al., 2025). Figure 4 shows 15 randomly sampled images from each dataset. WAON contains more images with Japanese text, scenes photographed in Japan, and content closely related to Japanese culture. In contrast, ReLAION (en→ja) more frequently includes images containing English text and content associated with Western cultural contexts. These observations suggest that the two datasets follow notably different distributions, and that WAON contains a larger proportion of Japanese-associated visual content. We also present additional representative examples from WAON in Appendix C.

4 Building WAON-Bench

The Recruit dataset (Honda and Arai, 2024) has been widely used to evaluate Japanese cultural understanding in contrastive vision-language models (Yokoo et al., 2024), but its limited cultural coverage of only 161 classes heavily skewed toward food leaves many aspects of Japanese culture underrepresented.

To address this limitation, we construct WAON-Bench, a manually curated Japanese cultural image classification benchmark designed to address the limited cultural coverage of the Recruit dataset (Honda and Arai, 2024), which contains only 161 classes heavily skewed toward food. As illustrated in Figure 5, WAON-Bench comprises 374 classes spanning 8 categories (animal, building, event, everyday, food, nature, scenery, and tradition), with 5 images per class for a total of 1,870 images, providing more than twice the cultural coverage of the previous benchmark with a more balanced distribution across diverse cultural domains.

4.1 Dataset Construction Pipeline

Category definition. We define eight categories to broadly cover diverse concepts in Japanese culture: animal, building, event, everyday, food, nature, scenery, and tradition.

Class definition. We collect class names related to Japanese culture from diverse sources, including existing datasets (Honda and Arai, 2024), Japanese Wikipedia articles, Google searches, conversations with LLMs, and direct observation during visits to Japanese cities such as Kyoto, resulting in 374 class

Figure 5: Overview of WAON-Bench. WAON-Bench comprises 374 classes spanning 8 categories (animal, building, event, everyday, food, nature, scenery, and tradition), with one randomly sampled image shown for each class.

names (e.g., Shiba Inu, Tokyo Tower).[2] We iteratively add classes while maintaining approximate balance across categories. Although the correspondence between categories and classes is inherently ambiguous, we assign each class to a single category for organizational convenience.

Image collection. We collect images for each class by using the class name as a search query in Google Image Search and manually selecting five images from the results. We limit collection to five images per class because, as observed in the Recruit dataset (Honda and Arai, 2024), retrieving more images frequently introduces noisy examples that do not match the class label, and for many culturally specific classes, only a limited number of clearly relevant images could be reliably identified. While this limits within-class visual diversity, expanding image counts via in-the-wild photography collection remains an important direction for future work. During selection, we prioritize diversity in composition, perspective, and setting, and exclude images containing objects corresponding to other classes to avoid label ambiguity.

Dataset verification. To ensure data quality, we recruit two native Japanese-speaking undergraduate student annotators and ask them to verify the correspondence between each image and its class label. We ask the annotators to check not only whether the image matches the assigned label but also whether it contains elements corresponding to other class labels.[3] Each annotator reviews half of the samples, identifying a total of 16 labeling errors, which we subsequently correct.

[2] See Appendix D for the complete class list.
[3] See Appendix E for the annotation interface.

4.2 Visual Diversity of WAON-Bench

To quantitatively evaluate the visual diversity of WAON-Bench, we compare it with the Recruit dataset. We randomly sample 1,000 images from each dataset, compute image embeddings using the vision encoder of `siglip2-base-patch16-256`, and measure the average pairwise cosine distance after L2 normalization. WAON-Bench achieves 0.495, higher than Recruit (0.425), indicating that WAON-Bench images are more widely dispersed in the embedding space and thus more visually diverse.

To further examine the semantic distribution of WAON-Bench, we apply t-SNE to the same embeddings (Figure 6). The animal, nature, and food categories form relatively distinct clusters, while the remaining categories are more intermixed, reflecting the inherent visual overlap between cultural concepts that span multiple categories. Since category labels are not used in our experiments, we do not further refine the category-class mapping and leave this as future work.

5 Experiments

To evaluate the quality of WAON and existing Japanese image-text datasets, we fine-tune `siglip2-base-patch16-256` (Tschannen et al., 2025), a strong multilingual model pretrained on global data, on WAON and two existing Japanese datasets. We conduct an experiment in which only the fine-tuning dataset is varied, while the model

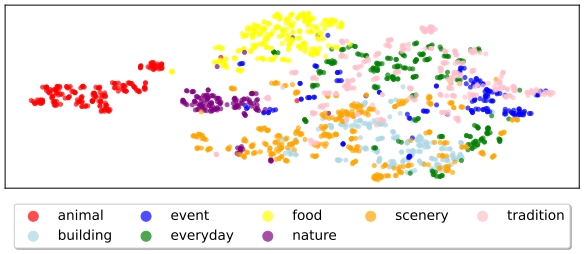


Figure 6: t-SNE visualization of image embeddings for WAON-Bench images, colored by category. The animal, food, and nature categories form relatively distinct clusters, while the remaining categories are more intermixed, reflecting the inherent visual overlap between cultural concepts that span multiple domains.

architecture, compute budget, and hyperparameter configuration are kept fixed, ensuring that the training data itself is the primary factor affecting model performance.

5.1 Training Settings

We fine-tune each model for 30,000 steps with a batch size of 8,192, totaling approximately 245M example exposures. We use AdamW (Loshchilov and Hutter, 2019) with an epsilon of 1e-8, a maximum learning rate of 1e-5 with cosine decay scheduling, a warmup of 1,500 steps, and a minimum learning rate of 1e-7. We use the SigLIP loss (Zhai et al., 2023) as our training objective. All experiments are conducted on eight NVIDIA H200 GPUs.

5.2 Training Datasets

To evaluate the effectiveness of WAON and existing large-scale Japanese image-text pair datasets as fine-tuning data, we use three Japanese large-scale datasets: WAON, ReLAION (ja subset) (Schuhmann et al., 2022), and ReLAION (en→ja translation) (Sugiura et al., 2025).

WAON is our newly constructed dataset consisting of 155M image-text pairs collected from Common Crawl. ReLAION (ja subset) (Schuhmann et al., 2022) is a Japanese subset extracted from the multilingual ReLAION-5B dataset (Schuhmann et al., 2022) using CLD3 language identification (Ooms, 2020). It contains 120M examples, of which 85M pairs were successfully downloaded as of June 2025 due to broken image URLs. This dataset's quality filtering relies on mCLIP (Chen et al., 2023a). ReLAION (en→ja translation) (Sugiura et al., 2025) is a Japanese-translated version derived from the English subset of ReLAION-5B. It was constructed by translating the original English captions into Japanese using Gemma 2 (Gemma Team, 2024), resulting in 1,452M image-text pairs.

For all datasets, we remove evaluation images via pHash deduplication to prevent data contamination. We exclude smaller datasets such as WiT (Srinivasan et al., 2021) and llm-jp-japanese-image-text-pairs (Sasagawa et al., 2025), as their substantially smaller scale would require more training epochs and lead to different training dynamics, making fair comparison difficult.

5.3 Evaluation Datasets

To evaluate Japanese, Western, and broader global cultural understanding, we use four datasets: WAON-Bench, Recruit (Honda and Arai, 2024), ImageNet (Deng et al., 2009), and XM3600 (Thapliyal et al., 2022).

WAON-Bench is our newly constructed Japanese cultural image classification benchmark, consisting of 374 classes and 1,870 examples. Recruit dataset (Honda and Arai, 2024) is also a Japanese cultural image classification benchmark, containing 161 classes and 7,654 examples. ImageNet is a large-scale image classification benchmark that contains many images associated with Western culture, comprising 1,000 classes and 50,000 examples. We use the Japanese-translated class labels provided by Sawada et al. (2024). XM3600 (Thapliyal et al., 2022) is a global image-caption retrieval benchmark consisting of 3,600 images collected from regions associated with 36 languages. Each image is paired with captions in all 36 languages. We use the Japanese captions in our experiments.

We report top-1 accuracy for classification tasks and Recall@1 for retrieval tasks.

5.4 Baseline Models

To provide broader comparison beyond our fine-tuned models, we compare against publicly released models that use ViT-B/16 (Dosovitskiy et al., 2021) as the vision encoder. Specifically, we include `clip-japanese-base` (Yokoo et al., 2024) as a Japanese-specific CLIP model, and `siglip-base-patch16-256-mult` (Zhai et al., 2023) and `siglip2-base-patch16-256` (Tschannen et al., 2025) as strong multilingual baselines.

6 Results

Table 3 reports the performance of both baseline and fine-tuned models across all benchmarks. Fig-

Model		Zero-Shot Classification			Retrieval	Avg.
		Japanese Culture		Western Culture	Global	
		WAON-Bench	Recruit	ImageNet	XM3600	
Baseline Models						
clip-japanese-base (Yokoo et al., 2024)		90.1	81.7	48.9	78.0	74.7
siglip-base-patch16-256-mult (Zhai et al., 2023)		89.3	75.1	53.3	43.2	65.2
siglip2-base-patch16-256 (Tschannen et al., 2025)		87.8	77.0	48.1	38.3	62.8
Fine-tuned Models	Fine-Tuning Data					
siglip2-base-patch16-256	**WAON (Ours)**	**95.1**	**83.3**	49.6	73.5	**75.4**
siglip2-base-patch16-256	ReLAION (ja subset)	93.0	81.7	47.4	72.4	73.6
siglip2-base-patch16-256	ReLAION (en→ja translation)	76.8	67.4	**51.7**	**73.8**	67.4

Table 3: Comparison of baseline and fine-tuned models on zero-shot classification and retrieval tasks. Fine-tuning on WAON achieves the best performance on Japanese cultural benchmarks and the best overall average, surpassing both baseline models and those fine-tuned on ReLAION variants. Notably, fine-tuning on ReLAION (en→ja) consistently degrades performance on Japanese cultural benchmarks relative to the base model.

ure 1 shows the accuracy curves on WAON-Bench during fine-tuning with each training dataset.

Translated data degrades Japanese cultural understanding. As shown in Figure 1, fine-tuning on ReLAION (en→ja translation) consistently degrades Japanese cultural understanding, with WAON-Bench performance steadily declining throughout fine-tuning from 87.8% to 76.8%. Furthermore, Table 3 shows that the fine-tuned model underperforms the base model on both WAON-Bench and the Recruit dataset.

In contrast, ReLAION (en→ja translation) achieves the largest improvement among the three training datasets on ImageNet (ja), with a gain of +3.6 points, suggesting that translated data remains effective for improving performance on benchmarks more closely aligned with Western cultural distributions. These results indicate that translated datasets inherit cultural biases from English source data and are not well suited for Japanese culture-specific fine-tuning despite their large scale.

Natively collected data yields superior cultural understanding. Fine-tuning on WAON yields consistent improvements on Japanese cultural benchmarks, reaching 95.1% on WAON-Bench and 83.3% on Recruit, outperforming all baseline models. ReLAION (ja subset), also natively sourced, shows gains on Japanese cultural benchmarks, yet lags behind WAON on both WAON-Bench and Recruit. One possible reason is that its filtering pipeline relies on mCLIP (Chen et al., 2023a), which is trained on small-scale English data (Changpinyo et al., 2021) and may be less effective at preserving some Japanese-specific image-text pairs during filtering. This suggests that stronger multilingual models such as SigLIP2, trained on large-scale global data, could yield more culturally robust filtering and further improve data quality.

Notably, although `clip-japanese-base` is trained from scratch with heavy pretraining on a large-scale in-house Japanese dataset comprising approximately 540M pairs (Yokoo et al., 2024), our model achieves superior performance through fine-tuning on top of a globally pretrained backbone (Tschannen et al., 2025), supporting prior findings (Pouget et al., 2024) that global pretraining followed by culture-specific fine-tuning is an effective strategy.

WAON maintains competitive performance on global benchmarks. Fine-tuning on WAON also improves performance beyond Japanese cultural tasks. On XM3600 (ja), WAON yields a substantial gain of 35.2 points over the base model (38.3→73.5), approaching the performance of ReLAION (en →ja) (73.8). Increasing the number of training steps may allow WAON to achieve performance comparable to clip-japanese-base (78.0).

7 Conclusion

We presented WAON, a large-scale Japanese image-text dataset containing 155M examples collected from native Japanese web content, and WAON-Bench, a manually curated Japanese cultural benchmark. Through comparative fine-tuning experiments across multiple Japanese image-text datasets, we observed that while models fine-tuned on English-to-Japanese translated data show degraded performance on Japanese cultural benchmarks, those fine-tuned on WAON consistently achieve the best performance across all baselines.

Limitations

Japanese as a single case study. We take Japanese as a case study, and whether our findings generalize to other languages and cultures remains an open question. Extending this investigation to other low-resource or culturally distinct languages is an important direction for future work.

Single model architecture. We fine-tune `siglip2-base-patch16-256` only, and do not assess whether our findings consistently hold across different model architectures or scales. Evaluating the generalizability of our findings to larger models or different architectural designs remains an important direction for future work.

Dataset construction pipeline. Our construction pipeline includes multiple filtering and deduplication steps, but we do not provide a quantitative analysis of each step's contribution to overall dataset quality. A more systematic analysis of each pipeline component would be valuable for guiding future large-scale vision-language dataset construction.

Ethical Considerations

Web data and copyright. WAON is constructed from publicly available web data collected via Common Crawl. We do not redistribute any image files; instead, we release only image URLs and associated metadata. All images remain hosted by their original sources, and their use is subject to the respective website terms and applicable copyright laws. Users of WAON are responsible for ensuring compliance with the terms of service of the original sources.

Content filtering and risk mitigation. To mitigate potential risks, we apply NSFW filtering to remove unsafe content during dataset construction. We acknowledge that despite this filtering, some harmful, biased, or privacy-sensitive content may remain, as is inherent in large-scale web-crawled datasets. We provide clear instructions for ethical use and will maintain a post-release mechanism for reporting and removing problematic samples, including those involving copyright infringement, personal privacy, sensitive content, or cultural bias. We encourage users to exercise caution and report any concerns to the dataset maintainers.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. *Preprint*, arXiv:1906.02569.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *ACL (System Demonstrations)*.

Romain Beaumont. 2021. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset. Accessed: 2026-05-19.

Johannes Buchner. 2021. Imagehash. https://github.com/JohannesBuchner/imagehash. Accessed: 2026-05-19.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568.

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

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, and 10 others. 2023b. PaLI: A jointly-scaled multilingual language-image model. In *ICLR*.

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

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. 2025. Meta CLIP 2: A worldwide scaling recipe. *Preprint*, arXiv:2507.22062.

Common Crawl. 2025. Common Crawl. https://commoncrawl.org. Accessed: 2026-05-19.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. 2024. Data filtering networks. In *ICLR*.

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, and 15 others. 2023. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS (Datasets and Benchmarks Track)*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Shion Honda and Hidehisa Arai. 2024. japanese-image-classification-evaluation-datase. https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset. Accessed: 2026-05-19.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *ACL*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *CVPR*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common objects in context. *Preprint*, arXiv:1405.0312.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. In *NeurIPS*.

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

Jeroen Ooms. 2020. Google's compact language detector 3. https://github.com/google/cld3. Accessed: 2026-05-19.

Angéline Pouget, Lucas Beyer, Emanuele Bugliarello, Xiao Wang, Andreas Peter Steiner, Xiaohua Zhai, and Ibrahim Alabdulmohsin. 2024. No filter: Cultural and socioeconomic diversity in contrastive vision-language models. In *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Vikram V. Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Bryan Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2023. GeoDE: a geographically diverse evaluation dataset for object recognition. In *NeurIPS (Datasets and Benchmarks Track)*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. 2025. Constructing multimodal datasets from scratch for rapid development of a Japanese visual language model. In *NAACL (System Demonstrations)*.

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

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. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS (Datasets and Benchmarks Track)*.

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

Peter Stahl. 2022. Lingua. https://github.com/pemistahl/lingua-py. Accessed: 2026-05-19.

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

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

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. 2025. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *Preprint*, arXiv:2502.14786.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2024. Demystifying CLIP data. In *ICLR*.

Shuhei Yokoo, Shuntaro Okada, Peifei Zhu, Shuhei Nishimura, and Naoki Takayama. 2024. CLIP Japanese base. https://huggingface.co/line-corporation/clip-japanese-base. Accessed: 2026-05-19.

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

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An open, billion-scale corpus of images interleaved with text. In *NeurIPS (Datasets and Benchmarks Track)*.

A Licenses for Our Resources

WAON is released under the Apache 2.0 License. Note that we distribute only image URLs rather than the images themselves. WAON-Bench, our fine-tuned SigLIP2 model, and our code are also released under the Apache 2.0 License.

B Use of AI Assistants

We used AI assistants to correct typographical errors, improve the clarity and naturalness of expressions, and generate scripts for plotting figures.

C Representative Examples in WAON

Table 4 shows representative examples from WAON, selected to highlight the cultural diversity of the dataset. The examples illustrate that WAON covers a wide variety of Japan-specific content, including event flyers, location maps, and food imagery, reflecting the breadth of Japanese web content captured in the dataset.

D WAON-Bench Class Names

Table 5 lists the class names for each category in WAON-Bench.

E WAON-Bench Verification Tool

Figure 7 shows the interface of the verification tool developed using Gradio (Abid et al., 2019), which is used to validate the correspondence between classes and images in WAON-Bench.

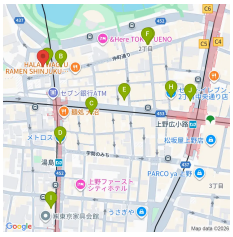
Image			
Caption	4年ぶりの開催となる「新宿街バルウィーク2023」 (“Shinjuku Machi Bar Week 2023,” held for the first time in four years)	パールフルートギャラリー東京周辺のコンビニエンスストア一覧地図 (Map of convenience stores around Pearl Flute Gallery Tokyo.)	こしあん、黒砂糖とメープルシロップの2種のあんペーストと、[...] (《An Toast》—a set featuring two types of sweet bean paste (koshian; and a blend of brown sugar and maple syrup),[...])
Similarity	0.154	0.165	0.182

Table 4: Representative examples of image-text pairs from WAON, covering event announcements, location maps, and food imagery, reflecting the cultural diversity of native Japanese web content.

Figure 7: A screenshot of the annotation interface used to validate the correspondence between classes and images in WAON-Bench.

Table 5: Class names for each category in WAON-Bench.

Category	Classes	Class Names
Animal	41	'アユ', 'イタチ', 'イノシシ', 'ウグイス', 'ウサギ', 'ウナギ', 'エゾシカ', 'カタツムリ', 'カブトムシ', 'カマキリ', 'カラス', 'キジ', 'キツネ', 'クモ', 'コイ', 'コウモリ', 'スズメ', 'セミ', 'タヌキ', 'チョウ', 'ツキノワグマ', 'ツバメ', 'ツル', 'トキ', 'トンボ', 'ニジマス', 'ニホンカモシカ', 'ニホンザル', 'ニホンリス', 'ニワトリ', 'ハト', 'フクロウ', 'フナ', 'ホタル', 'メダカ', 'モモンガ', '日本スピッツ', '日本猫', '柴犬', '秋田犬', '鷲'
Building	40	'ガソリンスタンド', 'コンビニ', 'スーパーマーケット', 'バス停', 'パチンコ屋', '交番', '体育館', '倉庫', '公園', '劇場', '博物館', '合掌造り', '商店街', '図書館', '学校', '屋上庭園', '展望台', '工場', '旅館', '映画館', '橋', '水族館', '消防署', '灯台', '灯籠', '町家', '病院', '空港', '納屋', '縁側', '茶室', '遊園地', '郵便局', '鉄塔', '銭湯', '駅', '駐車場', '高層ビル', '高床倉庫', '鳥居'
Event	29	'お正月', 'お盆', 'きつねの嫁入り行列', 'ひな祭り', 'よさこい祭り', '七五三', '七夕', '修学旅行', '卒業式', '同窓会', '地域清掃', '地蔵盆', '展示会', '成人式', '書き初め', '月見', '盆踊り', '相撲大会', '節分', '紅葉', '結婚式', '花火大会', '花見', '葬式', '運動会', '遠足', '防災訓練', '雪まつり', '餅つき'
Everyday	45	'うちわ', 'おはじき', 'お年玉', 'かき氷', 'こたつ', 'そろばん', 'ちゃぶ台', 'カラオケ', 'ガチャガチャ', 'クレーンゲーム', 'シーサー', 'トミカ', 'プリクラ機', 'ランドセル', '三味線', '下駄', '原付バイク', '囲炉裏', '囲碁盤', '射的', '屋台', '市民プール', '扇子', '数珠', '敷布団', '新幹線', '日本刀', '洗濯物', '温泉', '瓶ラムネ', '竹とんぼ', '竹馬', '線香', '自動販売機', '花札', '草履', '蓄音機', '蚊取り線香', '路面電車', '軽トラ', '郵便ポスト', '金魚すくい', '風呂敷', '風呂桶', '黒電話'
Food	55	'いくら', 'いちご', 'いなり寿司', 'うどん', 'うなぎの蒲焼', 'おでん', 'おにぎり', 'お好み焼き', 'きんぴらごぼう', 'しゃぶしゃぶ', 'しらす', 'すき焼き', 'そば', 'たけのこ', 'たこ焼き', 'ちゃんぽん', 'ちらし寿司', 'とんかつ', 'どら焼き', 'ひつまぶし', 'りんご飴', 'わたあめ', 'オムライス', 'カレーライス', 'チャーハン', 'チョコバナナ', 'ベビーカステラ', 'マグロ', 'ラーメン', '味噌', '味噌汁', '和菓子', '唐揚げ', '大福', '天ぷら', '弁当', '抹茶', '日本酒', '春巻き', '松茸', '椎茸', '海苔', '漬物', '焼きそば', '焼き鳥', '焼肉', '白米', '精進料理', '納豆', '緑茶', '茄子', '茶碗蒸し', '豆腐', '醤油', '餃子'
Nature	27	'つばき', 'アサガオ', 'アジサイ', 'キク', 'キンモクセイ', 'タンポポ', 'チューリップ', 'ハイビスカス', 'バラ', 'ヒマワリ', 'マリーゴールド', 'ラベンダー', '台風', '噴火', '彼岸花', '杉', '松', '柳', '桜', '梅', '津波', '牡丹', '竹', '菜の花', '蓮', '藤', '銀杏'
Scenery	75	'ご当地マンホール', 'カプセルホテルの内部', 'ゴーヤー畑', 'サロベツ湿原', 'スーパーの試食コーナー', 'ビルの屋上給水タンク', 'プラモデル専門店', 'メイドカフェ', 'ユニバーサルスタジオジャパン', 'ランタンフェスティバル', 'ロープウェイ', 'ワカメ干し', '不動明王像', '五稜郭', '京都タワー', '伏見稲荷大社', '兼六園', '原爆', '原爆ドーム', '厳島神社', '名古屋城', '地下鉄改札口', '地元ゆるキャラ', '地蔵', '大阪城', '天橋立', '姫路城', '富士山', '小樽運河', '三内丸山遺跡', '山火事防止の看板', '幕張メッセ', '平等院鳳凰堂', '戦艦大和', '東京スカイツリー', '東京タワー', '東京ディズニーランド', '東大寺', '松島', '桜島', '棚田', '法隆寺', '流氷観光船', '浅草寺', '清水寺', '漁港', '火炎放射器', '牛丼チェーン店のカウンター席', '特急列車の車内販売', '痛車', '皇居', '知床', '石垣', '石垣島のサンゴ礁', '砂防ダム', '秋葉原の電気街', '競艇', '競輪', '競馬', '茶畑', '蜂の巣駆除', '通天閣', '道の駅', '金閣寺', '鎌倉大仏', '防波堤', '雪国の街並み', '零戦', '電車の吊り広告', '首都高のジャンクション', '首里城', '駅前ロータリー', '鳥取砂丘', '鴨川', '黒部ダム'
Tradition	62	'お遍路さん', 'かるた', 'けん玉', 'こけし', 'ししおどし', 'しめ縄', 'だるま落とし', 'ねぶた', 'のれん', '修行僧', '前方後円墳', '剣道', '勾玉', '千羽鶴', '印鑑', '埴輪', '太鼓', '妖怪', '将棋', '尺八', '弓道', '弥生土器', '御朱印', '御神木', '折り紙', '招き猫', '提灯', '枕草子', '柔道', '歌舞伎', '浮世絵', '漆器', '灯籠流し', '熊手', '牛車', '狂言', '狛犬', '獅子舞', '琴', '琵琶', '盆栽', '石庭', '破魔矢', '神輿', '空手', '節句人形', '絵馬', '縄文土器', '能舞台', '花魁', '華道', '観光人力車', '赤べこ', '足袋', '銅鐸', '鐘つき', '鐘楼門', '陶器', '雅楽', '風鈴', '鯉のぼり', '鯛車'