

# WAON: Large-Scale and High-Quality Japanese Image-Text Pair Dataset for Vision-Language Models

Issa Sugiura<sup>♣,◇</sup>, Shuhei Kurita<sup>‡,◇</sup>, Yusuke Oda<sup>◇</sup>,  
Daisuke Kawahara<sup>♡,◇</sup>, Yasuo Okabe<sup>♣</sup>, Naoaki Okazaki<sup>†,◇</sup>

♣Kyoto University, ◇NII LLMC, ‡NII, ♡Waseda University, †Institute of Science Tokyo

## Abstract

Large-scale and high-quality image-text pair datasets play an important role in developing high-performing Vision-Language Models (VLMs). In this work, we introduce **WAON**, a large-scale and high-quality Japanese image-text pair dataset containing approximately 155 million examples, collected from Common Crawl. Our dataset construction pipeline employs various techniques, including filtering and deduplication, which have been shown to be effective in previous studies. To evaluate its effectiveness, we also construct **WAON-Bench**, a manually curated benchmark for Japanese cultural image classification, consisting of 374 classes. To assess the effectiveness of our dataset, we conduct experiments using both WAON and the Japanese subset of ReLAION, one of the most widely used vision-language datasets. We fine-tune SigLIP2, a strong multilingual model, on both datasets. The results demonstrate that WAON enhances model performance on WAON-Bench more efficiently than ReLAION and achieves higher accuracy across all evaluated benchmarks. Furthermore, the model fine-tuned on WAON achieves state-of-the-art performance on several Japanese cultural benchmarks. We release our dataset, model, and code at <https://speed1313.github.io/WAON>.

**Keywords:** Vision-Language Models, Image-Text Pair Dataset, Japanese

## 1. Introduction

The scale and quality of image-text pair datasets, which consist of images and their corresponding textual descriptions, are critical factors in developing high-performing Vision-Language Models (VLMs) (Schuhmann et al., 2022; Cherti et al., 2023; Gadre et al., 2023). However, most existing dataset construction efforts have focused on English and Chinese, and large-scale, high-quality datasets for other languages remain scarce. In this study, we aim to construct a large-scale and high-quality dataset for Japanese language and cultural understanding.

Although several Japanese image-text pair datasets have been developed (Schuhmann et al., 2022; Sugiura et al., 2025; Sasagawa et al., 2025), they still face notable limitations. One of the most widely used open resources is the Japanese subset of ReLAION (Schuhmann et al., 2022), which contains around 120 million Japanese image-text pairs. However, this dataset has several limitations: its snapshot is outdated, with 20-30% of image URLs no longer accessible as of June 2025. Furthermore, its filtering relies on mCLIP (Chen et al., 2023), a model with considerably lower performance than recent state-of-the-art CLIP variants, leading to reduced data quality. Another line of work translates the captions from the English subset of ReLAION into Japanese using LLMs, but translation errors often introduce noise and mixed-language captions (Sugiura et al., 2025). Moreover, since the English subset of ReLAION primarily consists of Western cultural image-text pairs, such data con-

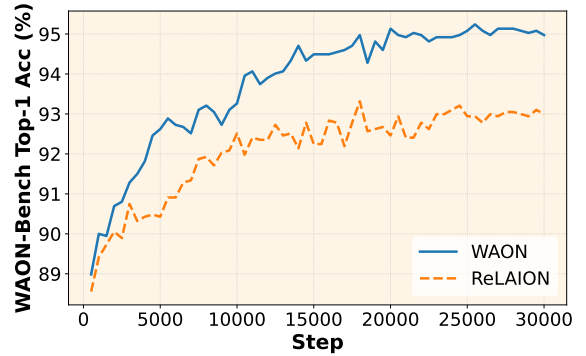


Figure 1: WAON-Bench Top-1 accuracy over training. Fine-tuning google/siglip2-base-patch16-256 on WAON improves performance on Japanese cultural benchmarks more efficiently than fine-tuning on the ReLAION (Japanese subset).

tribute little to improving performance on Japanese cultural tasks.

To address these issues, we present **WAON** (Web-scale image text Aligned Open Nihongo), a large-scale, high-quality dataset of 155M Japanese image-text pairs sourced from Common Crawl. Our dataset construction pipeline leverages recent advances in data curation and filtering, including deduplication and CLIP score based filtering (Schuhmann et al., 2022) using a strong multilingual model (Tschannen et al., 2025). WAON is designed to efficiently improve model performance in Japanese, particularly on Japanese cul-

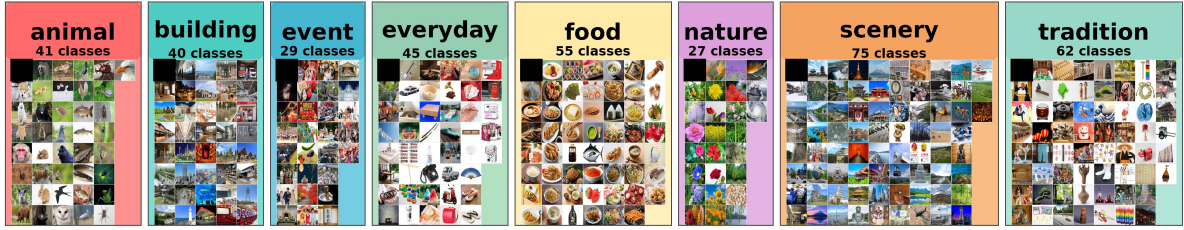


Figure 2: Overview of WAON-Bench. WAON-Bench consists of 8 categories with a total of 374 classes, each containing 5 images, resulting in a total of 1,870 images.

| Dataset                                                  | # Examples | Source                     |
|----------------------------------------------------------|------------|----------------------------|
| WIT (ja subset) (Srinivasan et al., 2021)                | 1M         | Wikipedia                  |
| ReLAION (ja subset) (Schuhmann et al., 2022)             | 120M       | Common Crawl               |
| ReLAION-ja (Sugiura et al., 2025)                        | 2.1B       | Common Crawl (translation) |
| llm-jp-japanese-image-text-pairs (Sasagawa et al., 2025) | 6.6M       | Common Crawl               |
| WAON (Ours)                                              | 155M       | Common Crawl               |

Table 1: Comparison of Japanese image-text pair datasets.

tural content. Since the proposed dataset construction pipeline uses Japanese only for the language identification step, it can be readily adapted to other languages.

To evaluate the effectiveness of WAON, we also introduce **WAON-Bench**, a manually curated benchmark for Japanese cultural image classification. WAON-Bench consists of 1,870 images across 374 classes related to Japanese culture (e.g., Shiba Inu, Jōmon pottery). We fine-tune SigLIP2 (Tschannen et al., 2025) on WAON and the Japanese subset of ReLAION, and demonstrate that WAON improves model performance more efficiently than ReLAION on WAON-Bench and consistently outperforms it across all tasks in the Japanese cultural benchmark. The fine-tuned model on WAON achieves the best performance on Japanese cultural benchmarks compared to existing models. To foster the development of VLMs, we release our dataset, model, and code.

## 2. Related Work

Image-text pair datasets have played a central role in advancing Vision-Language Models (VLMs). 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, including OpenCLIP (Cherti et al., 2023), diffusion models (Rombach et al., 2022),

and Visual Language Models (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. Data Filtering Network (DFN) (Fang et al., 2024) developed a data filtering model and provides a dataset curated with this model that enables training high-performance CLIP models. MetaCLIP (Xu et al., 2024) reverse-engineered OpenAI’s CLIP data curation strategy and proposed a metadata-based approach to curate and balance datasets, demonstrating that this approach yields higher-quality data than the original CLIP dataset. No-Filter (Pouget et al., 2024) experimentally demonstrated that pre-training with global data before fine-tuning on English content could improve cultural understanding without sacrificing performance on Western-centric benchmarks.

For Japanese, several datasets exist as shown in Table 1. WIT (Srinivasan et al., 2021) is constructed from Wikipedia articles and their corresponding image captions, but its Japanese subset contains only 1M image-text pairs, which is too small to train VLMs. The Japanese subset of ReLAION (Schuhmann et al., 2022) is large-scale, containing about 120M examples. However, it has several limitations, such as insufficient deduplication and quality filtering performed using relatively low-performing multilingual CLIP models trained on smaller datasets compared to the latest state-of-the-art models (Chen et al., 2023). ReLAION-ja (Sugiura et al., 2025) is a large-scale dataset by translating ReLAION English captions to Japanese with Gemma 2 (Gemma Team, 2024), but the dataset introduced translation errors, and cultural deviations derived from the English data. As shown in the

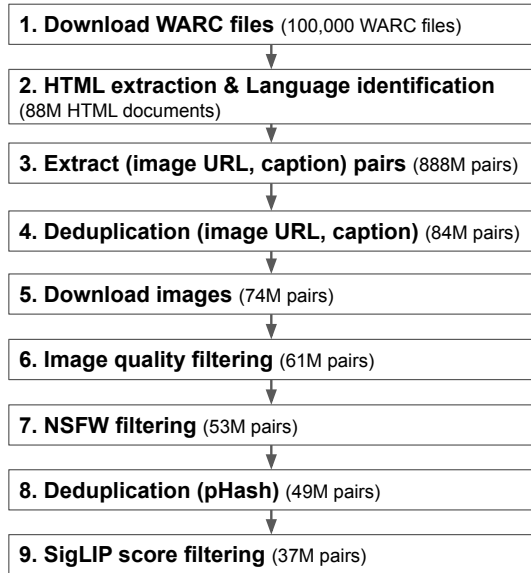


Figure 3: Construction pipeline of WAON. The numbers in parentheses indicate the remaining data count after each processing step (based on the 2025-18 snapshot).

evaluation of Sugiura et al. (2025), CLIP models pre-trained on ReLAION-ja exhibit limited understanding of Japanese culture.

### 3. Building WAON

We construct **WAON**, a large-scale and high-quality Japanese image-text pair dataset based on Common Crawl<sup>1</sup>. We extract (image, caption) pairs from the HTML pages contained in these snapshots and perform filtering and deduplication to obtain 155M Japanese (image, text) pairs. We describe the construction pipeline of WAON-Bench below.

#### 3.1. Dataset construction pipeline

Figure 3 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 deduplication that does not require image data, followed by image downloading. Subsequently, we apply filtering and deduplication processes that rely on image content. We apply this pipeline to the latest six Common Crawl snapshots (2024-26, 2024-33, 2024-42, 2024-51, 2025-08, and 2025-18). Following established practices in large-scale web dataset construction (Schuhmann et al., 2022; Lee et al., 2022; Zhu et al., 2023), we

incorporate several filtering and deduplication techniques while iteratively refining each step through manual inspection of sampled subsets. Below, we describe the details of each step.

**Download WARC files.** Common Crawl provides archives in three formats: WARC, WAT, and WET. In this work, we use WARC files, which contain complete HTML documents and corresponding request metadata. We first download crawl snapshots from Common Crawl. Each snapshot contains approximately 90,000 to 100,000 WARC files.

**HTML extraction & language identification.** Each WARC file contains multiple raw HTML documents. To construct Japanese image-text pairs, we first filter and retain only Japanese HTML documents. Following Okazaki et al. (2024), we initially determine language based on the lang attribute in the HTML header, as applying model-based language identification to every HTML body is computationally inefficient. We then apply Trafilatara (Barbresi, 2021) to extract underlying texts, and Lingua<sup>2</sup> to detect language of the document. We also discard documents with an empty <title> tag, assuming that high-quality image-text pairs are more likely to appear in HTML pages with well-written titles.

**Extracting (image URL, caption) pairs.** From the extracted HTML, we obtain (image URL, caption) pairs, where captions for each image are collected from either the image’s alt attribute (a short text description embedded in the HTML tag) or the corresponding <figcaption> element (the visible caption shown below an image). We remove entries with invalid image URLs and those whose captions contain no Japanese characters, as determined by Unicode code points.

**Deduplication (image URL, caption).** Since duplicates on the web often comprise low-diversity content such as advertising images, logos, and simple graphics, deduplication is widely used to improve dataset quality by eliminating redundancy and enhancing diversity (Lee et al., 2022; Zhu et al., 2023). In this step, we perform deduplication on both image URLs and captions. For captions, only the first occurrence of each (image URL, caption) pair is retained, and all subsequent duplicates are removed. Both image URLs and captions are converted into hash values and deduplicated independently. To improve memory efficiency during deduplication, we use a Bloom filter, a memory-efficient

<sup>1</sup><https://commoncrawl.org/>

<sup>2</sup><https://github.com/pemistahl/lingua-py>

probabilistic data structure that may introduce false positives but guarantees no false negatives.

**Download images.** To construct a high-quality image-text pair dataset, we need to perform filtering based on the images, including both the quality of the images themselves and the alignment between images and text. Therefore, at this stage, we download images from the extracted URLs. To efficiently download images at scale from the image URL list, we use `img2dataset` (Beaumont, 2021), a tool designed for parallel downloading of web images.

**Image quality filtering.** Web images often contain low-quality content such as advertisements, which can be removed to some extent using simple heuristics. For example, images with abnormal aspect ratios (e.g., excessively wide banners) are indications of advertisements. Following mmC4 (Zhu et al., 2023), we filter out images with width or height below 150 pixels and aspect ratios outside the range of 0.5 to 2.0. In addition, we apply our own color diversity filter, requiring each image to contain more than 32 unique colors.

**NSFW filtering.** As many web images contain adult or other unsafe content, it is necessary to remove such images. We use `dataset2metadata`<sup>3</sup>, an NSFW classification model based on OpenCLIP (Cherti et al., 2023), to detect unsafe images. Images with an unsafe score exceeding 0.1 are filtered out. The threshold of 0.1 follows prior work (Zhu et al., 2023; Sasagawa et al., 2025), and we confirmed its suitability through analysis of the score distribution and manual inspection of randomly sampled data. This process effectively removed most photographic and comic adult content.

**Deduplication (pHash).** Even after the previous curation steps, there exist many near-duplicate images that visually look similar despite having different URLs and captions. To remove such near-duplicate images, we compute perceptual hashes (pHash) using ImageHash<sup>4</sup> and perform deduplication using a Bloom filter. pHash generates a hash value that differ little from visually similar images, enabling detection of near-duplicates. We opt for exact matching rather than Hamming distance-based deduplication, as the latter would require computing pairwise distances across all images, which is computationally prohibitive at our scale.

<sup>3</sup><https://github.com/mlfoundations/dataset2metadata>

<sup>4</sup><https://pypi.org/project/ImageHash/>

| Snapshot | # Examples  |
|----------|-------------|
| 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 |

Table 2: Number of examples per snapshot after deduplication.

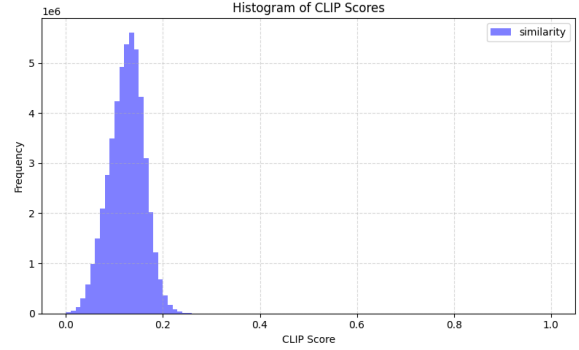


Figure 4: SigLIP similarity score distribution of image-text pairs from the 2025-18 snapshot. We exclude image-text pairs with similarity scores below 0.1.

**SigLIP score filtering.** Since captions associated with images are created either manually or automatically during HTML authoring, many image-text pairs exhibit poor semantic alignment between visual and textual content. To remove misaligned pairs, we apply CLIP score-based filtering (Schuhmann et al., 2022), which measures the cosine similarity between image and text embeddings. Specifically, we use `google/siglip2-base-patch16-256` (Tschannen et al., 2025) to compute embeddings. Figure 4 shows the distribution of SigLIP scores on the 2025-18 snapshot before filtering. Based on the distribution and manual inspection of randomly sampled examples, we set the threshold to 0.1 and discard pairs with lower scores.

### 3.2. Dataset statistics

We first run the pipeline for each snapshot, then deduplicate image URLs, captions, and pHashes across snapshots from 2025-18 down to 2024-26. Table 2 reports the number of examples per snapshot after deduplication. As earlier snapshots contain a higher proportion of previously seen examples, more data are removed during deduplication, resulting in fewer remaining examples. In total, we

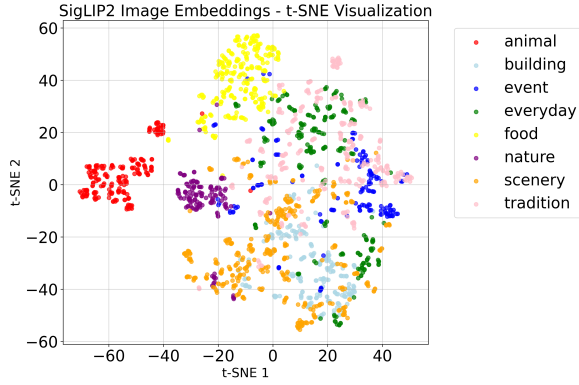


Figure 5: t-SNE map of image embeddings of WAON-Bench.

obtain a Japanese image-text pair dataset containing approximately 155M examples.

## 4. Building WAON-Bench

Evaluating the quality of large-scale image-text pair datasets is challenging, as their scale makes exhaustive manual inspection infeasible. A common approach is to pre-train or fine-tune a CLIP-style (Radford et al., 2021) model on the dataset and then evaluate it on downstream benchmarks such as image classification or image-text retrieval.

Since WAON focuses on Japanese culture, it requires a benchmark that can specifically evaluate models’ understanding of Japanese cultural concepts. There exists Recruit dataset (Honda and Arai, 2024), a Japanese cultural image classification dataset, which is commonly used for this purpose. However, we identified two major limitations with the existing benchmark. First, the class labels are heavily biased toward food-related categories—among the 161 classes, 101 belong to the food category. Second, the images associated with ground-truth classes are often ambiguous. For example, the dataset contains an image labeled as *Meiji Shrine* that actually depicts only a lawn. We suspect this issue arises because the dataset was automatically constructed using the Flickr API<sup>5</sup>, where images are collected based on class names as search queries. While this approach enables large-scale data collection, it often results in misaligned labels, duplicate samples, and limited diversity.

To address these limitations, we construct **WAON-Bench**, a new Japanese cultural image classification benchmark that is manually curated to ensure accurate class-image alignment and greater diversity. Figure 2 illustrates an overview of WAON-Bench. The dataset consists of 374 classes across

8 categories: animals, buildings, events, everyday life, food, nature, scenery, and traditions. Each class contains 5 images, resulting in a total of 1,870 images.

### 4.1. Dataset construction pipeline

WAON-Bench is constructed as follows:

1. We first define 374 class names related to Japanese culture (e.g., *Shiba Inu*, *Tokyo Tower*). To collect class names, we referred to Japanese Wikipedia, Google searches, and ChatGPT conversations, as well as observations from walking around cities such as Kyoto.
2. We then define eight categories: *animal*, *building*, *event*, *everyday*, *food*, *nature*, *scenery*, and *tradition*, and assign each class to one of these categories. While using WordNet, which defines hierarchical relationships between classes as in ImageNet (Deng et al., 2009), would be ideal, the Japanese version of WordNet has insufficient coverage and lacks many modern terms. Therefore, we manually assign each class to one of the eight categories.
3. For each class, we use the class name as a search query in Google Image Search and manually select five images from the results. During selection, we prioritize diversity in composition, perspective, and setting to capture a wide range of visual variations within the class. We also take care to avoid images that might include elements of other classes, minimizing the risk of mislabeling.

This curation process was performed entirely by the first author without crowdsourcing to ensure consistent selection criteria and to minimize mislabeling.

### 4.2. Analysis of Semantic Distribution in WAON-Bench Images

To examine the semantic distribution and potential biases of images in WAON-Bench, we applied t-SNE to the image embeddings obtained from the image encoder of `google/siglip2-base-patch16-256`.

Figure 5 shows the t-SNE map of image embeddings from WAON-Bench. The *animal*, *nature*, and *food* categories form distinct clusters, while the other categories are more intermixed. This suggests that these three categories are semantically more distinct from the others, whereas the remaining categories share more overlapping visual or conceptual characteristics. Therefore, we recommend treating the category information as auxiliary metadata rather than as a strict hierarchical label.

<sup>5</sup><https://www.flickr.com/services/api/>

| Task                 | Dataset                                         | # Classes | # Examples |
|----------------------|-------------------------------------------------|-----------|------------|
| Image Classification | ImageNet (Deng et al., 2009) (ja translation)   | 1,000     | 50,000     |
|                      | Recruit (Honda and Arai, 2024)                  | 161       | 7,654      |
|                      | WAON-Bench (Ours)                               | 374       | 1,870      |
| Image-Text Retrieval | XM3600 (Thapliyal et al., 2022) (ja annotation) | –         | 3,600      |

Table 3: Comparison of Japanese evaluation datasets.

| Model                                                 | Params | Retrieval    | Image Classification |                        |              |              | Avg |
|-------------------------------------------------------|--------|--------------|----------------------|------------------------|--------------|--------------|-----|
|                                                       |        |              | XM3600 <sub>ja</sub> | ImageNet <sub>ja</sub> | Recruit      | WAON-Bench   |     |
| siglip2-base-patch16-256 (fine-tuned on WAON)         | 375M   | 73.75        | 49.61                | <b>83.14</b>           | <b>94.97</b> | <b>75.37</b> |     |
| siglip2-base-patch16-256 (fine-tuned on ReLAION)      | 375M   | 72.39        | 47.38                | 81.65                  | 92.99        | 73.60        |     |
| siglip2-base-patch16-256 (Tschannen et al., 2025)     | 375M   | 38.28        | 48.12                | 76.98                  | 87.81        | 62.80        |     |
| clip-japanese-base (Yokoo et al., 2024)               | 196M   | <b>78.00</b> | 48.90                | 81.65                  | 90.05        | 74.65        |     |
| siglip-base-patch16-256-mult (Zhai et al., 2023)      | 371M   | 43.22        | 53.26                | 75.10                  | 89.25        | 65.21        |     |
| Japanese Stable CLIP ViT-L-16 (Shing and Akiba, 2023) | 414M   | 66.03        | <b>55.97</b>         | 71.29                  | 82.03        | 68.83        |     |
| LAION-CLIP-ViT-H-14 (Cherti et al., 2023)             | 1193M  | 72.64        | 47.67                | 70.62                  | 85.88        | 69.20        |     |

Table 4: Performance of models on each dataset. Fine-tuning google/siglip2-base-patch16-256 on the WAON dataset achieves state-of-the-art performance on the Japanese cultural benchmarks, Recruit and WAON-Bench. Using WAON consistently yields higher performance across all benchmark datasets compared to fine-tuning on ReLAION (ja subset).

## 5. Evaluation

To evaluate the quality of WAON, we fine-tune google/siglip2-base-patch16-256 on WAON and compare the model’s performance against baseline models on several benchmarks.

### 5.1. Training settings

We fine-tune our models for 30,000 steps with a training batch size of 8,192. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer with an epsilon value of  $1e-8$ . We employ 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 objective function. All experiments are conducted on eight NVIDIA H200 GPUs.

### 5.2. Training dataset

We compare a model fine-tuned on WAON with one fine-tuned on the Japanese subset of ReLAION. Its Japanese subset, identified via language detection of captions using CLD3<sup>6</sup>, contains 120M examples, of which 85M images were available for download as of June 2025. These 85M examples were used for fine-tuning. Although WAON and the Japanese subset of ReLAION differ in size, we ensure a fair comparison by training both models for the same

number of steps and using identical training hyper-parameters.

### 5.3. Evaluation dataset

We evaluate our model on zero-shot image classification and zero-shot image-text retrieval tasks. For image classification, we use WAON-Bench, ImageNet (Deng et al., 2009), and Recruit (Honda and Arai, 2024). For image-text retrieval, we use CrossModal-3600 (XM3600) (Thapliyal et al., 2022).

Table 3 summarizes the statistics of these datasets. ImageNet is an image classification dataset consisting of 1,000 classes. While the original class names are in English, we use a Japanese-translated version (Sawada et al., 2024) to evaluate whether models can correctly associate Japanese text with corresponding images. Recruit (Honda and Arai, 2024) is an image classification dataset focused on Japanese culture, containing 161 classes across four categories: food, flower, facility, and japanese landmark. XM3600 (Thapliyal et al., 2022) is an image-text retrieval dataset that provides annotations in 36 languages for 3,600 images. For our experiments, we use the first Japanese annotation assigned to each image. We use top-1 accuracy as the metric for the image classification task and recall@1 as the metric for the image-text retrieval task.

<sup>6</sup><https://github.com/google/cld3>

## 5.4. Baseline models

As baseline models, we use `line-corporation/clip-japanese-base` (Yokoo et al., 2024), Japanese Stable CLIP ViT-L-16 (Shing and Akiba, 2023) as Japanese CLIP models. We use `LAION-CLIP-ViT-H-14` (Cherti et al., 2023), `google/siglip-base-patch16-256-mult` (Zhai et al., 2023), and `google/siglip2-base-patch16-256` (Tschannen et al., 2025) as multilingual CLIP models.

## 5.5. Results

**WAON boosts performance more efficiently than ReLAION.** Figure 1 shows the Top-1 accuracy curve on WAON-Bench during training. While the model trained on ReLAION saturates around 93%, the model trained on WAON continues to improve, reaching approximately 95%. Moreover, WAON consistently outperforms ReLAION at every training step, demonstrating that WAON can more efficiently enhance the model’s performance on Japanese cultural understanding.

Table 4 shows the performance of each model across the benchmarks. WAON consistently outperforms the Japanese subset of ReLAION across all evaluated tasks, including those beyond WAON-Bench, demonstrating its effectiveness across a wide range of benchmarks.

**WAON achieves state-of-the-art performance on Japanese cultural benchmarks.** As shown in Table 4, `siglip2-base-patch16-256` (fine-tuned on WAON) achieves state-of-the-art performance on both the WAON-Bench and Recruit, which are benchmarks for Japanese cultural understanding. Starting from the base model `siglip2-base-patch16-256`, our fine-tuned model improves performance across all benchmark datasets, with particularly large gains on XM3600, Recruit, and WAON-Bench. These results demonstrate that fine-tuning on WAON significantly enhances the model’s Japanese cultural understanding. In contrast, fine-tuning on WAON yields a modest improvement of approximately 1.5 points on ImageNet. This is likely because, although the ImageNet class labels have been translated into Japanese, the images predominantly represent categories from English-speaking cultures, which are out of scope of WAON.

## 6. Conclusion

In this paper, we introduced WAON, a large-scale, high-quality Japanese image-text pair dataset consisting of approximately 155M examples. To evaluate its effectiveness, we constructed WAON-Bench,

a manually curated Japanese cultural image classification dataset. We show that fine-tuning on WAON boosts performance more efficiently than the Japanese subset of ReLAION and achieves state-of-the-art results on Japanese cultural benchmarks, including Recruit and WAON-Bench, compared to existing models.

## 7. Limitations and Ethical Considerations

Although WAON is large-scale with 155M image-text pairs, there remains about a tenfold gap compared to the 2B pairs in ReLAION’s English dataset. This limitation mainly arises from the language imbalance in Common Crawl, where English webpages are roughly nine times larger than Japanese ones in number<sup>7</sup>. While our dataset is based on six Common Crawl snapshots, expanding the number of snapshots could further increase the data scale.

Regarding WAON-Bench, current models already achieve accuracy exceeding 90%, suggesting that the benchmark may be approaching saturation. This may be due to the class design avoiding overly similar categories and the careful image selection process, which prioritized clarity and minimized label noise. Developing a more challenging benchmark for fine-grained Japanese cultural understanding remains an important direction for future work, particularly to support the research and development of vision-language models.

## 8. Acknowledgements

In this research work, we used the “mdx: a platform for building data-empowered society”. We used ABCI 3.0 provided by AIST and AIST Solutions with support from “ABCI 3.0 Development Acceleration Use”.

## 9. Bibliographical References

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>.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and

---

<sup>7</sup><https://commoncrawl.github.io/cc-crawl-statistics>

- Wenping Wang. 2023. [mCLIP: Multilingual CLIP via cross-lingual transfer](#). In *ACL*.
- 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*.
- 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*.
- 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, Dhruva 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, Hananeh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 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](#). arXiv preprint arXiv:2408.00118.
- Shion Honda and Hidehisa Arai. 2024. [japanese-image-classification-evaluation-dataset](#). <https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset>.
- 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](#). arXiv preprint arXiv:1405.0312.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *ICLR*.
- 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*.
- 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*.
- 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*.
- Makoto Shing and Takuya Akiba. 2023. [Japanese stable clip vit-l/16](#). <https://huggingface.co/stabilityai/japanese-stable-clip-vit-l-16>.
- 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*.

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](#). arXiv 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>.

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*.