Developing Japanese CLIP Models Leveraging an Open-weight LLM for Large-scale Dataset Translation

Issa Sugiura[♣,‡], **Shuhei Kurita**[♢,‡], **Yusuke Oda**[‡],
Daisuke Kawahara[♠,‡], **Naoaki Okazaki**[♡,‡]
[♣]Kyoto University, [‡]NII LLMC, [♢]National Institute of Informatics,
[♠]Waseda University, [♡]Institute of Science Tokyo
sugiura.issa.q29@kyoto-u.jp, {skurita, odashi}@nii.ac.jp
dkw@waseda.jp, okazaki@c.titech.ac.jp

Abstract

CLIP is a foundational model that bridges images and text, widely adopted as a key component in numerous vision-language models. However, the lack of large-scale open Japanese image-text pairs poses a significant barrier to the development of Japanese vision-language models. In this study, we constructed a Japanese image-text pair dataset with 1.5 billion examples using machine translation with open-weight LLMs and pre-trained Japanese CLIP models on the dataset. The performance of the pre-trained models was evaluated across seven benchmark datasets, achieving competitive average scores compared to models of similar size without the need for extensive data curation. However, the results also revealed relatively low performance on tasks specific to Japanese culture, highlighting the limitations of translation-based approaches in capturing cultural nuances. Our dataset[1], models[2], and code[3] are publicly available.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) has emerged as a powerful framework for aligning images and text within a shared embedding space. By leveraging contrastive learning, CLIP has demonstrated remarkable capability in bridging visual and textual modalities, thereby being adopted in numerous multimodal models such as visual-language models and diffusion models (Liu et al., 2023; Lin et al., 2024; Ramesh et al., 2022).

While the size and quality of the pre-training dataset is critical for CLIP's performance (Cherti et al., 2023; Xu et al., 2024), the availability of large-scale, high-quality Japanese image-text pairs remains limited, posing challenges for advancing research of Japanese vision-language models. As of this writing, the largest publicly available Japanese dataset is the Japanese subset of ReLAION-5B (Schuhmann et al., 2022), comprising approximately 120 million image-text pairs. This size is smaller than the 2.1 billion image-text pairs available in the English subset of ReLAION-5B, highlighting a gap in data size. Moreover, while the English subset is filtered using OpenAI's CLIP, which has high performance, the Japanese subset is filtered using mCLIP (Chen et al., 2023a), where the filtering quality may be suboptimal due to mCLIP's lower performance on Japanese.

To construct large-scale Japanese image-text pair datasets, there are two primary approaches: web crawling using resources such as Common Crawl (Schuhmann et al., 2022) and translating existing English datasets. However, web crawling presents challenges due to the relatively small proportion of Japanese web pages in Common Crawl, which account for only about 5% compared to the about 43% occupied by English pages[4], indicating a nearly ninefold disparity. Consequently, machine translation emerges as a viable alternative.

In this paper, we constructed a dataset of 1.5 billion Japanese image-text pairs by leveraging open-weight LLMs for translation. We also pre-trained Japanese CLIP models using the constructed dataset to assess its effectiveness. Our experimental evaluations demonstrate that our models achieve competitive performance across various benchmark datasets, compared to other models of similar size. However, the performance on tasks related to Japanese culture was relatively low, highlighting the limitations of translation-based approaches in effectively enhancing understanding of Japanese culture.

[1] https://huggingface.co/llm-jp/relaion2B-en-research-safe-japanese-translation

[2] https://huggingface.co/llm-jp/llm-jp-clip-vit-base-patch16, https://huggingface.co/llm-jp/llm-jp-clip-vit-large-patch14

[3] https://github.com/llm-jp/clip-eval

[4] https://commoncrawl.github.io/cc-crawl-statistics

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 162–170

English Caption	Japanese Caption
Iron Man Movie Poster	アイアンマン映画ポスター
Unique 14k Gold Yellow and Blue Diamond Engagement Ring 2.64ct.	ユニークな14金イエローゴールドとブルーダイヤの婚約指輪 2.64ct.
Hot Chocolate With Marshmallows, Warm Happiness To Soon Follow	マシュマロ入りホットチョコレート、まもなく幸せが訪れる。
Herd of cows on alpine pasture among mountains in Alps, northern Italy. Stock Photo	アルプス北部、イタリアのアルプス山脈の山々の中にある高地草地に群れでいる牛の写真

Table 1: Examples of original English captions of ReLAION-5B and their Japanese translations by gemma.

2 Constructing a Japanese Image-Text Pair Dataset

To construct a Japanese image-text pair dataset, we translated the captions of the English subset of ReLAION-5B[5] into Japanese using gemma-2-9b-it[6], a high-performance open-weight LLM. ReLAION-5B is a refined version of LAION-5B (Schuhmann et al., 2022), with Child Sexual Abuse Material (CSAM) removed. It is a large-scale dataset of image-text pairs, where images and their corresponding IMG-alt text are collected from Common Crawl and filtered using existing CLIP models. The dataset is divided into three subsets: English, multilingual, and no-language.

To enable the rapid translation of large datasets, we developed `text2dataset`[7], a translation tool for LLMs. This tool utilizes `vLLM` (Kwon et al., 2023), a fast LLM inference library, to efficiently translate large-scale English datasets into Japanese.

Prompt To translate text using LLMs, it is crucial to provide both the text for translation and a clear instruction prompt (Zhu et al., 2024). In this study, we used the following prompt:

```
You are an excellent English-Japanese translator. Please translate the following sentence into Japanese.\n You must output only the translation.\n Sentence :{passage}\n Translation:
```

The `{passage}` is replaced with the source text for translation. The LLM is then expected to generate the translated text based on this prompt.

[5] https://laion.ai/blog/relaion-5b
[6] https://huggingface.co/google/gemma-2-9b-it
[7] https://github.com/llm-jp/text2dataset

Translation Results We translated the entire captions of the English subset of ReLAION-5B, consisting of 2,097,693,557 examples. This process was completed in about 9 days using 32 NVIDIA A100 40GB GPUs.

Table 1 shows translated examples. It is evident that the English captions were successfully translated into Japanese. However, a manual check of the first 10,000 examples revealed some translation issues. Despite explicitly specifying the target language in the prompt, there were examples where the translation was incorrectly performed into Chinese or Korean, which accounted for about 1% of the cases. Additionally, a phenomenon specific to instruction-tuned LLMs was observed: for example, an expression like "`Please let me know if you have any questions.`" was added at the end of the translated text, which accounts for about 0.1% of the examples. These issues could be improved by utilizing higher-performance translation LLMs or applying post-processing to the translation results. We leave them as future work.

We used `img2dataset` (Beaumont, 2021) to download images. Due to issues such as broken URL links or preprocessing failures, the success rate of downloading was approximately 70%, resulting in a final dataset of 1,451,957,221 Japanese image-text pairs.

3 Training CLIP

We describe the training settings of llm-jp-clip-ViT-B/16 as our default model in this section.

We pre-trained CLIP models using the constructed dataset. In this study, we used ViT-B/16 (Dosovitskiy et al., 2021) as the image encoder and $\text{RoBERTa}_{\text{BASE}}$ (Liu et al., 2019) as the text encoder. The output dimension of each encoder was set to 512, and both were trained from

English Template	Japanese Template
a photo of the {}	{}の写真
a sketch of a {}	{}のスケッチ
a photo of the cool {}	かっこいい{}の写真

Table 2: Examples of prompt template.

Dataset	Examples	Classes	Language
Image Classification			
ImageNet	50,000	1,000	En
Recruit	7,654	161	Ja
CIFAR10	10,000	10	En
CIFAR100	10,000	100	En
Food101	25,250	101	En
Caltech101	8,677	101	En
Image-Text Retrieval			
XM3600	3,600	–	En, Ja, etc

Table 3: Details of evaluation datasets.

scratch. We used the `llm-jp-tokenizer`[8] as the base tokenizer and applied custom modifications tailored for CLIP. The text encoder's maximum context length was set to 76 tokens. The image resolution was set to 224×224.

For optimization, we used AdamW with hyperparameters of $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-6}$. Learning rate scheduling consisted of 2,000 steps of linear warmup followed by cosine decay, with a peak learning rate of 5.0×10^{-4} and a minimum learning rate of 0.0. We trained the model for 9 epochs, processing a total of 13 billion examples.

We employed the contrastive loss function proposed by Radford et al. (2021). The batch size was set to 8,192, with gradient accumulation over four steps. Notably, the accumulated loss differs from the contrastive loss computed directly with a batch size of 32,768.

We used `OpenCLIP` (Ilharco et al., 2021) as the training framework and trained the model on 16 NVIDIA H100 80GB GPUs, requiring two weeks for training.

4 Evaluation

We evaluated the performance of our models by comparing it with Japanese and multilingual baseline CLIP models on zero-shot image classification and image-text retrieval tasks.

[8] https://github.com/llm-jp/llm-jp-tokenizer

4.1 Evaluation Settings

Zero-shot Image Classification We followed the evaluation methodology proposed by Radford et al. (2021) for zero-shot image classification. First, we convert class labels corresponding to the target images into natural language sentences using prompt templates. For example, a label will be inserted into the placeholder `{label}` in a template "`a photo of a {label}`" to convert the label into a natural sentence. Next, we compute the similarity scores between images and texts, and the label with the highest similarity is selected as the predicted class for the image. In this study, we used Japanese prompt templates provided by `japanese-clip` (Shing et al., 2022). Table 2 shows examples of the Japanese templates used in this experiment. For evaluation, we used accuracy@1 as the metric.

Zero-shot Image-Text Retrieval Image-text retrieval involves two main tasks: text-to-image retrieval and image-to-text retrieval. In text-to-image retrieval, the goal is to find the most relevant images based on a textual query by computing the similarity between the text embedding and the embeddings of all candidate images, then ranking the images accordingly. In contrast, image-to-text retrieval aims to retrieve the most relevant textual descriptions for a given image query. For evaluation, we used recall@1 as the metric.

Evaluation Datasets Table 3 provides details of the evaluation datasets used in our experiments.

In zero-shot image classification task, we used ImageNet-1K (Deng et al., 2009), Recruit[9], CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009), Food101 (Bossard et al., 2014), and Caltech101 (Li et al., 2022). For ImageNet, we used Japanese class labels from `japanese-clip`. Recruit consists of four image classification tasks related to concepts and objects unique to Japan: jafood101, jaflower30, jafacility20, and jalandmark10, with 7,586 images successfully retrieved from 7,654. For CIFAR10, CIFAR100, Food101, and Caltech101, class labels were translated into Japanese using DeepL.

In zero-shot image-text retrieval task, we used CrossModal-3600 (XM3600) (Thapliyal et al., 2022). XM3600 is a dataset containing multilingual annotations for 3,600 images. In this exper-

[9] https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset

Model	# Params (M)	Image Classification						Retrieval		Avg.
		ImageNet	Recruit	CIFAR10	CIFAR100	Food101	Caltech101	XM3600 I → T	XM3600 T → I	
Japanese CLIP										
Rinna ViT-B/16	196	50.6	39.9	90.7	64.0	53.2	84.6	53.8	54.0	61.4
Rinna ViT-B/16 cloob	196	54.6	41.6	88.2	60.3	57.2	80.2	53.4	53.4	61.1
LY ViT-B/16	196	52.0	**83.8**	96.3	76.7	73.9	**88.4**	**76.9**	**78.0**	**78.3**
llm-jp-clip-ViT-B/16	248	54.2	59.4	91.8	69.2	82.2	85.6	73.6	72.7	73.6
StabilityAI ViT-L/16	414	**62.4**	70.5	97.6	**84.1**	74.0	86.7	67.3	66.0	76.1
llm-jp-clip-ViT-L/14	467	59.5	62.9	96.4	77.0	**88.2**	87.8	74.1	74.1	77.5
Multilingual CLIP										
SigLIP B/16-256 multi	370	51.9	71.2	92.4	65.8	78.6	85.6	45.9	43.0	66.8
jina-clip-v2	865	35.8	48.1	95.1	58.3	52.0	69.4	67.3	66.4	61.6
LAION ViT-H/14 multi	1193	53.0	74.5	**97.9**	78.4	74.3	85.1	75.0	72.0	76.3

Table 4: Performance of each model in zero-shot image classification and image-text retrieval tasks. **Bold** indicates first place, and underline indicates second place.

Model	# Params (M)	Recruit				
		jafacility20	jafood101	jaflower30	jalandmark10	Overall
Japanese CLIP						
Rinna ViT-B/16	196	63.0	28.4	56.5	60.3	39.9
Rinna ViT-B/16 cloob	196	61.5	27.3	63.5	69.4	41.6
LY ViT-B/16	196	**82.0**	**83.8**	**90.5**	**91.8**	**83.8**
llm-jp-clip-ViT-B/16	248	72.4	52.7	67.0	82.2	59.4
StabilityAI ViT-L/16	414	70.8	65.1	89.0	78.6	70.5
llm-jp-clip-ViT-L/14	467	75.3	55.8	73.5	84.7	62.9
Multilingual CLIP						
SigLIP B/16-256 multi	370	64.9	70.7	88.5	68.0	71.2
jina-clip-v2	865	80.0	47.1	44.0	48.5	48.1
LAION ViT-H/14 multi	1193	80.5	69.1	85.4	89.1	74.5

Table 5: Performance of each model in zero-shot image classification across each subtask of Recruit.

iment, we used the first Japanese annotations assigned to each image.

Baseline Models To compare the performance of our models, we used Japanese CLIP and multilingual CLIP models. For Japanese CLIP models, we used Rinna ViT-B/16 (Sawada et al., 2024), Rinna ViT-B/16 cloob (Sawada et al., 2024), LY ViT-B/16 (Shuhei et al., 2024), and StabilityAI ViT-L/16 (Shing and Akiba, 2023). For multilingual CLIP models, we used SigLIP B/16-256 multi (Zhai et al., 2023), jina-clip-v2 (Koukounas et al., 2024), and LAION ViT-H/14 multi (Schuhmann et al., 2022). Details of the baseline models can be found in Appendix A.

4.2 Results

The performance of each model is shown in Table 4. Our llm-jp-clip-ViT-B/16 model achieves the second highest average score among Japanese CLIP models of similar size, following LY ViT-B/16. On ImageNet, a key benchmark dataset for CLIP, llm-jp-clip-ViT-B/16 achieved a high score of 54.2, second only to Rinna ViT-B/16 cloob's 54.6 among models of similar size. However, Rinna ViT-B/16 cloob, which was trained on the relatively small CC12M (Changpinyo et al., 2021) dataset, shows limited generalization performance outside ImageNet. We suspect that this is due to the limited diversity and scale of CC12M, which restricts the ability of the Rinna ViT-B/16 cloob to generalize

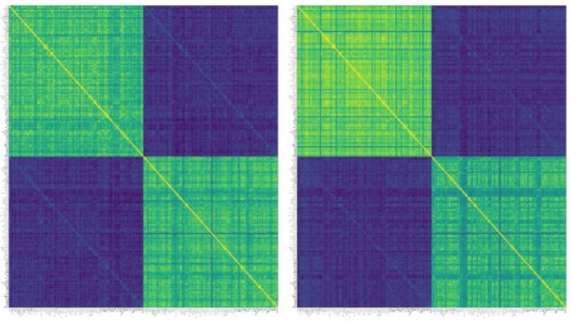

Figure 1: Cosine similarity matrices of text and image embeddings. Left: LY ViT-B/16. Right: llm-jp-clip-ViT-B/16. The top-left block represents similarities among text embeddings, the bottom-right block represents similarities among image embeddings, and the top-right/bottom-left blocks represent similarities between text and image embeddings. Brighter colors indicate higher similarity.

Image Encoder	ImageNet	XM3600	
		I → T	T → I
Full Scratch	**54.2**	**73.6**	**72.7**
Continued	52.9	71.6	71.7
LiT	52.7	71.7	70.9

Table 6: Effect of training settings of image encoders.

beyond ImageNet.

On Recruit, which contains images specific to Japanese culture, its score was more than 30 points lower compared to LY ViT-B/16. The performance of each model in zero-shot image classification across each subtask of Recruit is shown in Table 5. We can observe that llm-jp-clip-ViT-B/16 significantly underperforms compared to LY ViT-B/16 on jafood101.

To investigate the cause of this performance gap, we visualized and analyzed the embeddings of LY ViT-B/16 and llm-jp-clip-ViT-B/16. We calculated the cosine similarities between all combinations of text and image embeddings for each class within jafood101. The similarity matrices for both models are shown in Figure 1. We can observe that LY ViT-B/16 separates positive and negative text embeddings more clearly than llm-jp-clip-ViT-B/16. This performance gap may be due to the lack of examples specific to Japanese culture in the translation data, leading to poor results on Recruit, which contains images specific to Japanese such as "交番" (police station), "おでん" (oden, a Japanese fishcake stew), and "鎌倉大仏" (the Great Buddha of Kamakura).

4.3 Ablation Study on Image Encoder

We performed several ablation studies to determine the optimal configuration of the image encoder.

Effect of Training Settings We experimented with the following three training settings for the image encoder: (1) Training from scratch, (2) Continued pre-training, and (3) Pre-training only the text encoder with a frozen pre-trained image encoder (Locked-image Tuning; LiT (Zhai et al., 2022)). For the continued pre-training and LiT settings, we initialized the weights of the image encoder model using the LAION's CLIP[10]. For all settings, the text encoder was trained from scratch. To prevent loss spikes in both the continued pre-training and LiT settings, the peak learning rate was reduced to 1.0×10^{-4}. Figure 2 shows the accuracy curve of ImageNet for each setting, and Table 6 reports the final performance. Similarly to previous research (Zhai et al., 2022), LiT exhibited a significant performance improvement in the early stages of training, but subsequent improvements were gradual. Although the initial performance of ImageNet was low when training from scratch, substantial performance improvements were observed as training progressed, surpassing both continued training and LiT settings in the end.

[10] https://huggingface.co/laion/CLIP-ViT-B-16-laion2B-s34B-b88K/tree/main

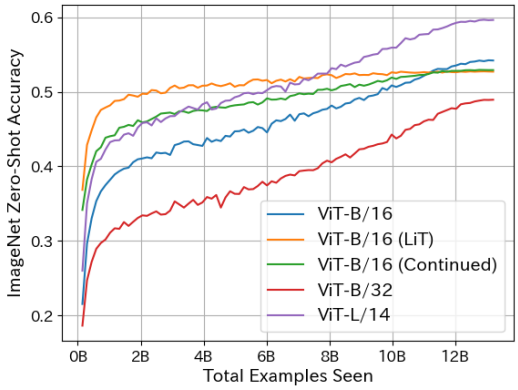

Figure 2: Accuracy curve of ImageNet zero-shot image classification.

Effect of Model Size We compared the performance of ViT-B/16 and ViT-L/14. All settings other than the image encoder were kept the same. The results are shown in Table 4. In all tasks, ViT-L/14 outperformed ViT-B/16. This reconfirmed that increasing the model size leads to better performance, as observed in the previous study (Cherti et al., 2023).

Effect of Patch Size We examined the performance differences on ImageNet caused by different patch size settings in the image encoder. In this study, we evaluated ViT-B/32 and ViT-B/16. For ViT-B/32, the batch size was set to 16,384, with gradient accumulation over two steps, and the peak learning rate set to 1.0×10^{-3}. Other settings were kept the same as those for ViT-B/16. The results of accuracy curve on ImageNet are shown in Figure 2. ViT-B/16 consistently outperformed ViT-B/32, aligning with previous findings (Radford et al., 2021), where smaller patch sizes yielded better performance.

5 Conclusion

In this study, we constructed a large-scale Japanese image-text dataset using translation with open-weight LLMs and pre-trained Japanese CLIP models on the dataset. The results demonstrated competitive performance in the average score across the benchmark datasets compared to models of similar size. However, the performance on tasks related to Japanese culture was relatively low, highlighting the limitations of translation-based approaches in capturing cultural nuances. Future work includes building more diverse and high-quality Japanese image-text datasets and further improving the performance of Japanese CLIP models.

Limitations

In this study, we used open-weight LLMs for translation. While these models require GPUs, making large-scale processing costly, recent advancements have enabled access to smaller, high-performing LLMs that offer a more cost-effective alternative. For instance, assuming an average caption length of 50 characters, translating 2.1 billion examples with DeepL would cost approximately 260M JPY. In contrast, using an open-weight LLM reduced the cost to just 500K–1M JPY.

Acknowledgments

In this research work, we used the "mdx: a platform for building data-empowered society".

References

Romain Beaumont. 2021. img2dataset: Easily turn large sets of image urls to an image dataset.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *Computer Vision – ECCV 2014*, pages 446–461.

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 *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043. Association for Computational Linguistics.

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 Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023b. Pali: A jointly-scaled multilingual language-image model. *Preprint*, arXiv:2209.06794.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

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 *International Conference on Learning Representations*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *Preprint*, arXiv:2309.17425.

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 Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei W Koh, Olga Saukh, Alexander J Ratner, Shuran Song, Hannaneh 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 *Advances in Neural Information Processing Systems*, volume 36, pages 27092–27112.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. OpenCLIP.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott

Martens, Nan Wang, and Han Xiao. 2024. jina-clip-v2: Multilingual multimodal embeddings for text and images. *Preprint*, arXiv:2412.08802.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. 2022. Caltech 101.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26679–26689.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

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 *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905.

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 *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Makoto Shing and Takuya Akiba. 2023. Japanese Stable CLIP ViT-L/16.

Makoto Shing, Tianyu Zhao, and Kei Sawada. 2022. japanese-clip.

Yokoo Shuhei, Okada Shuntaro, Zhu Peifei, Nishimura Shuhei, and Takayama Naoki. 2024. CLIP Japanese Base.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM*, 59(2):64–73.

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 *The Twelfth International Conference on Learning Representations*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. LiT: Zero-shot transfer with locked-image text tuning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781.

A Details of Baseline Models

Table 7 shows the details of the baseline models: Rinna ViT-B/16[11], Rinna ViT-B/16 cloob[12], LY ViT-B/16[13], StabilityAI ViT-L/16[14], SigLIP B/16-256 multi[15], jina-clip-v2[16], and LAION ViT-H/14 multi[17].

[11] https://huggingface.co/rinna/japanese-clip-vit-b-16
[12] https://huggingface.co/rinna/japanese-cloob-vit-b-16
[13] https://huggingface.co/line-corporation/clip-japanese-base
[14] https://huggingface.co/stabilityai/japanese-stable-clip-vit-l-16
[15] https://huggingface.co/google/siglip-base-patch16-256-multilingual
[16] https://huggingface.co/jinaai/jina-clip-v2
[17] https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k

Model	# Params (M)	Training Dataset
Japanese CLIP		
Rinna ViT-B/16	196	CC12M (Changpinyo et al., 2021)
Rinna ViT-B/16 cloob	196	CC12M
LY ViT-B/16	196	CC12M, YFCC100M (Thomee et al., 2016), Common Crawl†
StabilityAI ViT-L/16	414	CC12M, MS-COCO (Lin et al., 2014)
Multilingual CLIP		
SigLIP B/16-256 multi	370	WebLI† (Chen et al., 2023b)
jina-clip-v2	865	DFN (Fang et al., 2023), CommonPool (Gadre et al., 2023)
LAION ViT-H/14 multi	1193	LAION-5B (Schuhmann et al., 2022)

Table 7: Details of the baseline models used in the experiment. Datasets marked with † are not publicly available. We report only the primary dataset used by the developers.