Constructing Multimodal Datasets from Scratch for Rapid Development of a Japanese Visual Language Model

Keito Sasagawa[♣,‡], **Koki Maeda**[♢,‡*], **Issa Sugiura**[♠,‡*],
Shuhei Kurita[†,‡], **Naoaki Okazaki**[♢,†,‡], **Daisuke Kawahara**[♣,†,‡]
[♣]Waseda University, [♢]Institute of Science Tokyo, [♠]Kyoto University,
[†]National Institute of Informatics, [‡]NII LLMC

Abstract

To develop high-performing Visual Language Models (VLMs), it is essential to prepare multimodal resources, such as image-text pairs, interleaved data, and instruction data. While multimodal resources for English are abundant, there is a significant lack of corresponding resources for non-English languages, such as Japanese. To address this problem, we take Japanese as a non-English language and propose Japanese multimodal datasets for rapidly developing a Japanese multimodal model. We collect Japanese image-text pairs and interleaved data from web archives and generate Japanese instruction data using an existing large language model and a VLM. Our experimental results show that a VLM trained on these native datasets outperforms those relying on machine-translated content. The resulting VLM, dataset and code used for training is publicly available[1].

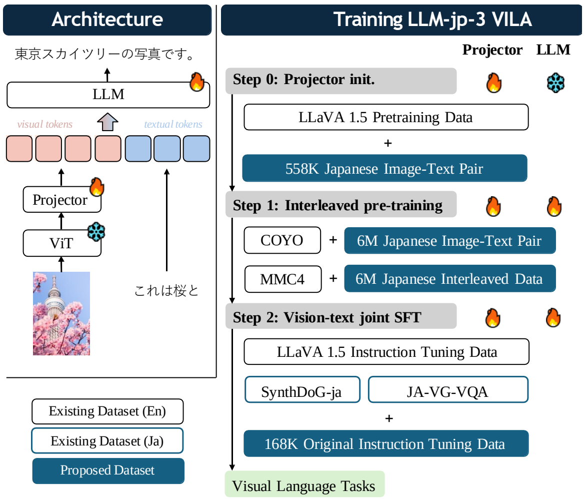

Architecture
東京スカイツリーの写真です。
LLM
visual tokens
textual tokens
Projector
ViT
これは桜と
Existing Dataset (En)
Existing Dataset (Ja)
Proposed Dataset
Training LLM-jp-3 VILA
Projector
LLM
Step 0: Projector init.
LLaVA 1.5 Pretraining Data
+
558K Japanese Image-Text Pair
Step 1: Interleaved pre-training
COYO
+
6M Japanese Image-Text Pair
MMC4
+
6M Japanese Interleaved Data
Step 2: Vision-text joint SFT
LLaVA 1.5 Instruction Tuning Data
SynthDoG-ja
JA-VG-VQA
+
168K Original Instruction Tuning Data
Visual Language Tasks

Figure 1: We propose **LLM-jp-3 VILA**, a novel Japanese visual language model. For each step of pretraining and instruction tuning, we construct tailored million-scale image-text dataset (■) from interleaved data.

1 Introduction

We develop a multimodal resource for high-performing Visual Language Models (VLMs) in Japanese. While English multimodal resources are relatively abundant in existing studies, there is a significant lack of corresponding datasets for non-English languages, such as Japanese. One potential solution is translating multimodal resources from English to Japanese. However, this approach often produces suboptimal results, as the translation does not account for the contextual relationship between the image and the text. Such a translation approach also cannot follow the cultural backgrounds of the image domains as they are collected in English websites.

To address this multilingual gap, we propose Japanese multimodal datasets for rapidly developing Japanese multimodal models. We build two types of datasets: *pretraining data* and *instruction data* in Japanese. For the pretraining data, we curate a large-scale dataset of Japanese image-text pairs and interleaved data by extracting Japanese data from web crawls. For the instruction tuning data, we follow the LLaVA (Liu et al., 2023) methodology to build Japanese instruction data from image captions and object bounding boxes. Also, we directly generate Japanese instruction data by giving Japanese images to an existing VLM via APIs, enhancing inference capabilities for images unique to Japan.

We demonstrate that models trained on our proposed datasets achieve higher accuracy than those relying on machine-translated datasets, and this performance gain is especially noticeable in the instruction data. For this purpose, we apply the VILA (Lin et al., 2023) architecture to our Japanese VLM with English publicly available data and Japanese newly collected data. We assume that this approach is adaptable to any languages and not

[*] Equally contributed.

[1] https://huggingface.co/llm-jp/llm-jp-3-vila-14b, https://github.com/llm-jp/llm-jp-VILA

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), pages 470–484
April 30, 2025 ©2025 Association for Computational Linguistics

Data	# Images	# Step	Full
English			
LLaVA-1.5 pretrain data	558K	0	✓
LLaVA-1.5 instruction data (subset)	358K	2	✓
COYO (Byeon et al., 2022) (subset)	6M	1	✓
mmc4-core (Zhu et al., 2023) (subset)	6M	1	✓
Japanese			
Translated data[2]	620K	2	✗
(Proposed) Image-text pairs	6.6M	0 & 1	✓
(Proposed) Interleaved	6M	1	✓
(Proposed) Instruction tuning	369K	2	✓

Table 1: Data size for LLM-jp-3 VILA. Full means the dataset is used in our full model.

limited to Japanese. Our contributions significantly enhance the resources available for Japanese VLM studies, enabling more effective regional localization and cultural understanding in VLMs.

2 Related Work

Resource for Visual Language Models Visual language models require a large amount of image-text pairs for both pretraining and instruction tuning. Visual instruction tuning of LLaVA (Liu et al., 2023, 2024a) relies on synthesized data by OpenAI GPT (Brown et al., 2020). This synthesized instruction data are used in the various VLM developments such as InstructBLIP (Dai et al., 2023) and VILA (Lin et al., 2023). Following these successes, we examine the use of the synthesized data in instruction tuning in Japanese VLM development.

Japanese Visual Language Resources The task of answering questions about documents that include both visual and textual contents is emerging as Document Visual Question Answering (Document VQA) (Kembhavi et al., 2016, 2017; Krishnamurthy et al., 2016; Kafle et al., 2018; Sampat et al., 2020; Tanaka et al., 2023). However, these datasets are mostly developed in English and do not reflect the domain-specific knowledge in other languages. Recently, JDocQA was proposed for a Japanese document visual question answering (Onami et al., 2024). Heron Bench (Inoue et al., 2024) was also proposed for evaluating Japanese VLM ability. However, Japanese resources are quite limited compared to English resources, and existing ones are often intended for fine-tuning models. In this paper, we construct a large-scale text-image corpora for Japanese VLM construction.

3 Dataset Construction

We build two types of datasets: pretraining data and instruction data in Japanese. As the pretraining data, we construct datasets that include data reflecting the cultural background of Japan by collecting images and texts from Japanese websites. We also create Japanese multimodal instruction data without using machine translation to improve the model's ability to generate Japanese text and follow user instructions when images and text are provided.

3.1 Pretraining Data

Pretraining data consists of two types: interleaved data and image-text pairs data.

Interleaved data Interleaved dataset is constructed in a way similar to MMC4 (Zhu et al., 2023) in Japanese. The texts in the dataset come from Japanese texts extracted from the 2020-2022 Common Crawl dumps in the llm-jp-corpus (LLM-jp, 2024). We use `bunkai` (Hayashibe and Mitsuzawa, 2020) to break down Japanese text into sentences. After this process, sentences consisting only of symbols with no numbers, English or Japanese characters were combined with the previous sentence. In addition, if the end of the parentheses come at the beginning of the next sentence, it is moved to the end of the previous sentence.

We download images from URLs that are extracted from web texts. To avoid overloading on specific servers, we downsample URLs from frequent domains. Next, we remove duplicate images within the same document. We use `ImageHash`[3] to calculate the phash value of the image, and for images with a Hamming distance of 5 or less, we keep the one with the highest resolution. We also remove duplicate images across multiple documents. For data from each year, we remove images that have more than 10 duplicates in the 60K images sampled. This operation is repeated until the total number of sampled images is the same as the original number of images. This removes application icons, advertisements, and images that are inserted when links are broken. Then, we apply the NSFW image classifier from `dataset2metadata`[4] to remove potential NSFW images (Gadre et al., 2023) that have higher than 0.1 classifier scores.

For images that have passed through these filters, we calculate the similarity of all pairs of images and sentences in the document using the OpenCLIP (Ilharco et al., 2021) trained on LAION5B

[2] https://huggingface.co/datasets/turing-motors/LLaVA-v1.5-Instruct-620K-JA

[3] https://github.com/JohannesBuchner/imagehash

[4] https://github.com/mlfoundations/dataset2metadata

dataset[5] (Schuhmann et al., 2022). We remove images that do not have a CLIP similarity of at least 0.20 with any of the sentences in the document. In the same way as the construction method for MMC4, we map images to text by solving an assignment problem using `lapjv`[6]. Finally, we use the harmful text filtering applied in llm-jp-corpus. Furthermore, we only kept samples with the number of images between 2 and 5, the number of sentences between 10 and 100, the token length of the sample within the max length of the Large Language Model (LLM), and the similarity of all assigned image and text pairs above 0.20. As a result, the number of images in the dataset is 9.9M. For training, we use a subset of this dataset to balance the Japanese image-text pair dataset.

Image-text pairs We collected alt attributes for images after NSFW image filtering in interleaved data. We performed text filtering, based on the filtering method used in constructing the COYO dataset (Byeon et al., 2022). First, we use a regular expression to remove all text that does not contain any hiragana, katakana, or common kanji characters. We also filter too short alt texts and specific file names of images that are typically set for screenshots. Next, we filter NSFW content using the `DiscardAdultContentJa` filter of the `Hojichar`[7]. For text that pass through these filters, the first and last consecutive whitespace characters are removed, and if there are two or more consecutive whitespace characters, they are replaced with a half-width space. Then we deduplicate the data. Alt text that appeared more than 10 times was removed, and duplicates were removed for (image phash value, alt text) pairs.

Finally, the similarity of each image and alt text pair is calculated using OpenCLIP trained on LAION5B dataset and Japanese CLIP[8]. We also filter lower 30 percentile of CLIP alignment score data. The resulting dataset contains 6.6M images.

3.2 Instruction Data

There is already a Japanese multimodal instruction dataset, the LLaVA-Instruct dataset translated into Japanese using DeepL, but it contains unnatural Japanese due to translation errors.

Module	# Params
Vision Encoder	428M
Projector	32M
LLM	13B

Table 2: Model parameters for LLM-jp-3 VILA.

So we construct a Japanese instruction dataset, named **llava-instruct-ja**, from COCO images (Lin et al., 2014) in the same way as the LLaVA-Instruct dataset. Specifically, we first create few-shot examples to be input when generating data. These few-shot examples are designed to generate Japanese instruction data from the English captions attached to the images. We create these examples for three types of the instruction data: conversation type, detail description type, and complex reasoning type. For the detail description and complex reasoning types, we also input bounding boxes that represent the positional information of objects within the image. We used GPT-4o mini to generate data through via Azure OpenAI API. As a result, we obtained a dataset with 156K samples.

In addition, we use GPT-4o to generate multi-turn conversation data from the Japan Diverse Images Dataset[9] and develop **Japanese-photos-conv** dataset. Japan Diverse Images Dataset consists of images taken in Japan. For each image in this dataset, we generate a multi-turn question answer via `gpt-4o-2024-05-13`. In generating, we adopted zero-shot manner with the image as input. The system prompt for QA generation is shown in Table 6. Except several images filtered by Azure OpenAI, we collected dataset with 12K samples. The meaning of this dataset is that while the llava-instruct-ja dataset mostly made up of images taken in English-speaking countries, this dataset consists of images taken in Japan. By training on such a dataset, the model is expected to be able to make better inferences about images unique to Japan, such as landmarks in Japan, Japanese culture, and Japanese text.

We summarize the dataset used in LLM-jp-3 VILA in Table 1 and details in Appendix A.3.

4 Experiments

4.1 Model Training

Our model architecture integrates the vision encoder and LLM through the projector, similarly to

[5] https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k

[6] https://pypi.org/project/lapjv/

[7] https://github.com/HojiChar/HojiChar

[8] https://huggingface.co/line-corporation/clip-japanese-base

[9] https://huggingface.co/datasets/ThePioneer/japanese-photos

Hyperparameter	Step 0	Step 1	Step 2
Batch Size	256	1024	128
Learning Rate (lr)	1e-3	5e-5	1e-5
lr Scheduler Type	cosine		
lr Warmup Ratio	0.03		
Weight Decay	0		
Epoch	1		
Optimizer	AdamW		
DeepSpeed Stage	2	3	3

Table 3: Hyperparameters for each training step

VILA. We use SigLIP[10] (Zhai et al., 2023) for vision encoder, llm-jp-3-13b-instruct[11] for LLM, and two-layer MLP for projector.

We train our model in three training stages, inspired by VILA as shown in Figure 1. The first stage is the projector initialization stage, where only the projector parameters are tuned on English and Japanese image-text pair datasets. We use 558K samples of English and Japanese image-text pairs, where the English dataset is sourced from LLaVA-Pretrain dataset[12] and the Japanese dataset is from the subset of the dataset we constructed. The second stage is a multimodal continual pretraining stage, in which the parameters of the projector and LLM are tuned on image-text pair datasets and interleaved datasets. For the English dataset, we use a subset of 6M images from mmc4-core and a subset of 6M images from COYO. For the Japanese datasets, we use our pair dataset and our interleaved dataset, each with 6M images.

The third stage is the multimodal instruction tuning stage, where the projector and LLM are tuned so that the model can follow instructions. For the Japanese datasets, we use the llava-instruct-ja dataset and the japanese-photos-conv dataset described in Section 3. In addition, we use the JA-VG-VQA (Shimizu et al., 2018) dataset, a high-quality, manually created Japanese multimodal QA dataset. We concatenate the multiple QA pairs attached to each image in this dataset and convert it into a multi-turn conversation. Since the answers in this dataset are phrases or short sentences, we specify the response format to be that way. Specifically, we add the prompt "語句または短い文で答えてください。 (*Please answer in phrases or short sentence.*)" to the questions in the first turn. We exclude the samples from JA-VG-VQA-500[13] benchmark dataset, resulting in a dataset of 99K samples. Furthermore, to enhance the Japanese OCR capabilities, we use the synthdog-ja (Kim et al., 2022) dataset, which is composed of synthetic OCR data. We use a subset of 102K samples to match the data volume of the English OCR task dataset. We prepare several templates and convert the dataset into a QA format. For the English dataset, a subset of the instruction data from LLaVA-1.5 is used for training to match the amount of data in the Japanese dataset. We use 158K samples from the LLaVA-Instruct dataset as a synthetic dataset with GPT-4. As VQA datasets, we use a 53K sample subset of VQAv2 (Goyal et al., 2017) and a 46K sample subset of GQA (Hudson and Manning, 2019), which are multi-turn datasets similar to JA-VG-VQA. We also use the OCRVQA (Mishra et al., 2019) dataset with 80K samples and the TextCaps (Sidorov et al., 2020) dataset with 22K samples as datasets for the English OCR task. In total, the Japanese instruction dataset has 369K samples, and the English instruction dataset has 358K samples.

Parameters of LLM-jp-3 VILA Parameter counts for each module in LLM-jp-3 VILA is shown in the Table 2.

Computational Budget Training for step 0 takes about 14-15 hours on 1 node with 8xA100 (40GB). Step 1 takes about 130 hours to training on 8 nodes with 8xA100 (40GB). Step 2 takes about 11 hours to training on 4 nodes with 8xA100 (40GB).

Hyperparameters Table 3 shows the hyperparameters for each training step.

4.2 Benchmark Datasets

To verify the comprehensive capabilities of LLM-jp-3 VILA, we employed three benchmarks: Heron Bench (Inoue et al., 2024), JA-VLM-Bench-In-the-Wild (Akiba et al., 2024), and JA-VG-VQA500.

Heron Bench evaluates the Japanese language capabilities of VLMs using a dataset of 21 images and 103 image-question-answer triplets specifically designed within the cultural and linguistic context of Japan.

JA-VLM-Bench-In-the-Wild comprises 42 images paired with 50 curated questions focusing on a diverse range of culturally specific elements and objects commonly found in Japan. In developing the benchmark, authors leveraged GPT-4V (OpenAI,

[10] https://huggingface.co/google/siglip-so400m-patch14-384

[11] https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

[12] https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain

[13] https://huggingface.co/datasets/SakanaAI/JA-VG-VQA-500

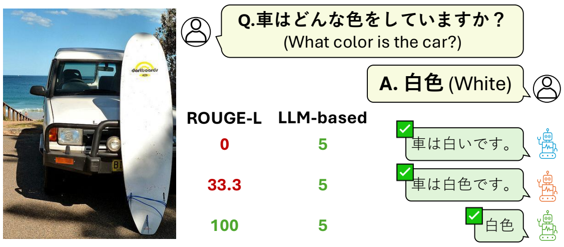


Figure 2: Example of how ROUGE-L scores can be misleading when evaluating Japanese VQA responses. Three different LLM-generated answers are shown alongside their corresponding ROUGE-L scores.

2024) and conducted human-in-the-loop filtering process to ensure the dataset quality.

JA-VG-VQA500[14] is a subset of the Japanese Visual Genome VQA dataset (Shimizu et al., 2018), which is based on Visual Genome (Krishna et al., 2017), extracted 500 samples from the test set.

4.3 Evaluation Settings

We compare our model's overall Japanese language performance against several competitive Japanese and multilingual VLMs.[15] We largely follow the original evaluation settings; however, we have introduced certain modifications to the evaluation methods to better reflect the objectives of our study. We averaged scores of 5 runs in Heron Bench and JA-VLM-Bench-In-the-Wild and employed a single run score of JA-VG-VQA-500. We followed LLM-as-a-judge approach via Azure OpenAI API and employed `gpt-4o-2024-05-13`.

In **Heron Bench**, the model's performance is quantified by the ratio of its average score of answers evaluated in LLM-as-a-judge process to that obtained by GPT-4. Consequently, scores can exceed 100%, which indicates that the model outperformed GPT-4 on average. For reproducibility, the evaluator's temperature was set to 0 and the seed to 0. This setting is also the case for other LLM-as-a-Judge-based evaluations. Note that even the same seed is used, the output may not be deterministic[16].

In **JA-VLM-Bench-In-the-Wild** and **JA-VG-VQA500**, ROUGE-L (Lin, 2004) is commonly used as an evaluation metric. However, ROUGE scores vary greatly depending on the answer style in Japanese question answering. Figure 2 presents a typical example where ROUGE scores vary significantly due to differences in sentence structure and wording. VLMs are asked to identify the color of a car in an image, and the reference answer is "白色 (White)". Three different answers are shown, each with varying levels of detail and grammatical structure:

1. "車は白いです。" (The car is white.)
2. "車は白色です。" (The car is white.)
3. "白色" (White)

While all three answers are factually correct, their ROUGE-L scores differ significantly. The simplest answer received the highest ROUGE-L score, even though it lacks the grammatical completeness. In contrast, other answers received a score of 0, despite conveying the same information. This example highlights the limitations of ROUGE-L in capturing the semantic nuances and stylistic variations of Japanese language. It suggests that relying solely on ROUGE-L might lead to an inaccurate assessment of LLM performance in VQA tasks, particularly in languages where word order is less rigid and contextual understanding is crucial.

To prevent such underestimation, we employed the LLM-as-a-judge framework (Zheng et al., 2023) with GPT-4o. We evaluated the generated responses on a five-point Likert scale, with some modifications to a publicly available standard VQA prompt template[17].

4.4 Results

Table 4 presents the performance on three benchmarks[18]. Compared to current VLMs of similar size, LLM-jp-3 VILA consistently achieved state-of-the-art performance, as measured by the LLM-as-a-Judge score. Additionally, on the JA-VG-VQA-500 benchmark, our model surpassed even the performance of GPT-4o. While ROUGE-L is commonly used for evaluation, it tends to deteriorate when the generated output deviates significantly from the provided reference. This is evidenced by GPT-4o's low ROUGE-L score despite its demonstrably high capabilities, suggesting that ROUGE-L may no longer be a suitable metric for evaluating performance in this context.

[14] https://huggingface.co/datasets/SakanaAI/JA-VG-VQA-500

[15] For details regarding the baseline models, please refer to the Appendix B.1.

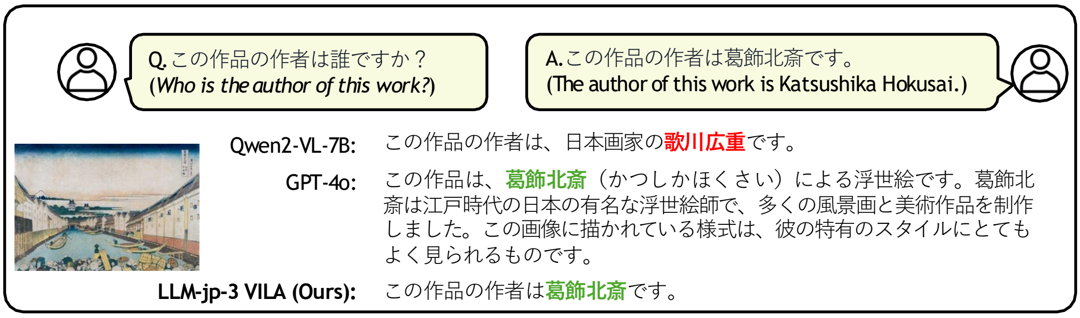
[16] https://platform.openai.com/docs/advanced-usage/reproducible-outputs

[17] The actual prompt is provided in Appendix B.2. You can also refer to https://cloud.google.com/vertex-ai/generative-ai/docs/models/metrics-templates.

[18] The category-wise scores of Heron Bench are provided in the Appendix B.3.

	Heron-Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
Models	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
Japanese InstructBLIP Alpha	14.0	20.8	2.42	–	–
Japanese Stable VLM	24.2	23.3	2.47	–	–
Llama-3-EvoVLM-JP-v2	39.3	41.4	2.92	**23.5**	2.96
LLaVA-CALM2-SigLIP	43.3	47.2	3.15	17.4	3.21
LLaVA-1.6 7B	25.8	28.6	2.40	11.7	2.67
LLaVA-1.5 7B	34.8	40.6	2.48	13.9	2.66
Llama 3.2 11B Vision	36.5	27.4	2.77	13.8	2.95
InternVL2 8B	45.2	33.7	2.98	11.6	3.13
Qwen2-VL 7B Instruct	54.8	45.3	3.53	16.2	3.48
VILA1.5 13B	34.3	41.7	2.62	12.9	2.80
LLM-jp-3 VILA (Ours)	**57.2**	**52.3**	**3.69**	16.2	**3.62**
GPT-4o	87.6	37.6	3.85	12.1	3.58

Table 4: Comparison on Japanese benchmarks between current VLMs and **LLM-jp-3 VILA**. "–" indicates that the score cannot be calculated as the benchmark dataset is included for training in such models. **Bold** indicates the best score except GPT-4o. "LLM" is an abbreviation for LLM-as-a-Judge. Detailed information for baseline models is in Appendix B.1 and a breakdown of Heron Bench scores by category in Appendix B.3.



Q. この作品の作者は誰ですか？
(*Who is the author of this work?*)

A. この作品の作者は葛飾北斎です。
(The author of this work is Katsushika Hokusai.)

Qwen2-VL-7B: この作品の作者は、日本画家の**歌川広重**です。

GPT-4o: この作品は、**葛飾北斎**（かつしかほくさい）による浮世絵です。葛飾北斎は江戸時代の日本の有名な浮世絵師で、多くの風景画と美術作品を制作しました。この画像に描かれている様式は、彼の特有のスタイルにとてもよく見られるものです。

LLM-jp-3 VILA (Ours): この作品の作者は**葛飾北斎**です。

Figure 3: Examples of text generated by each model in response to a question from the Japanese-Heron-Bench. **Green** indicates the correct word and **red** indicates the wrong word.

Qualitative Evaluation Figure 3 shows the qualitative comparisons between LLM-jp-3 VILA and existing models. While Qwen2-VL-7B misidentifies the answer, confusing it with another famous Japanese artist 歌川広重 *(Utagawa Hiroshige)*, our model precisely answers the question, showing capability for Japanese culture-specific knowledge. GPT-4o also provides a good answer with rich information. We provide additional examples in Appendix B.4.

Ablation of the training dataset To validate the effectiveness of our constructed dataset, we did ablation study of the constructed dataset. We ablated our instruction data, step 1 training, and our interleaved data. For the ablation of instruction data, We replaced our proposed dataset with LLaVA-v1.5-Instruct-620K-JA[19], where LLaVA-v1.5-Instruct-620K-JA is the Japanese machine translation version of the instruction data from LLaVA-1.5, excluding text-only data. Table 5 highlights that the use of translated instruction data degrades performance significantly compared to our proposed dataset. This is because the quality of the dataset is reduced due to translation errors that occur during machine translation. In addition, while there is some improvement with step 1 training, the performance improvement by interleaved data is still limited for several evaluation metrics. One reason for this is that the number of images in our dataset is approximately half of those used in the step 1 training of the VILA's dataset, which limits the effectiveness of the proposed dataset. Increasing the amount of data is the future work of this study.

5 Conclusion

In this paper, we constructed Japanese multimodal datasets to develop a high-performance Japanese visual language model. We collected images and texts from Japanese websites to construct datasets that include data reflecting the cultural background of Japan. We also constructed Japanese multimodal instruction data using existing LLMs. By using pro-

[19] https://huggingface.co/datasets/turing-motors/LLaVA-v1.5-Instruct-620K-JA

			Heron-Bench	JA-VLM-Bench-In-the-Wild		JA-VG-VQA-500	
Step-0	Step-1	Step2	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
✓	✓	translated	47.2	45.6	3.19	15.7	3.33
✓	✗	✓	56.5	**57.3**	3.47	16.1	3.54
✓	w/o interleaved	✓	**58.6**	52.2	3.50	**16.7**	3.61
✓	✓	✓	57.2	52.3	**3.69**	16.2	**3.62**

Table 5: **Ablation study of LLM-jp-3 VILA. Bold** indicates the best score. "LLM" is an abbreviation for LLM-as-a-Judge.

posed datasets, we developed LLM-jp-3 VILA, a Japanese visual language model designed to integrate multiple images with natural language understanding. Our experiments demonstrate the model's effectiveness across a range of multimodal tasks in Japanese. For future work, we plan to extend our model to wide branch of Japanese visual datasets.

Acknowledgments

This research project has benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program for the use of GPT-4o through which leading foundation models hosted by Microsoft Azure along with access to Azure credits were provided to conduct the research.

Limitation

Since we generate training dataset with proprietary models, the quality relies on such models. We need to find ways to synthesize better quality, less hallucinated datasets. Also, visual knowledge of Japan may depend on the original vision encoder. We may need to build another vision encoder from scratch or conduct additional tuning.

Ethical Consideration

We are implementing NSFW filtering for images and text, so it is thought that there are almost no such images. In addition, when generating data using GPT-4o, NSFW images are rejected by Azure OpenAI, so it is likely that there will be almost no such images in the generated instruction data.

References

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2024. Evolutionary optimization of model merging recipes. *Preprint*, arXiv:2403.13187.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

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. Curran Associates, Inc.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yuta Hayashibe and Kensuke Mitsuzawa. 2020. Sentence boundary detection on line breaks in Japanese.

In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 71–75, Online. Association for Computational Linguistics.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

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.

Aozora Inagaki. 2024. llava-calm2-siglip.

Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. 2024. Heron-Bench: A Benchmark for Evaluating Vision Language Models in Japanese. *Preprint*, arXiv:2404.07824.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision – ECCV 2016*, pages 235–251, Cham. Springer International Publishing.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123:32–73.

Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. 2016. Semantic parsing to probabilistic programs for situated question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 160–170, Austin, Texas. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *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, Cham. Springer International Publishing.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

LLM-jp. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *CoRR*, abs/2407.03963.

Meta. 2024. Llama-3.2-11b-vision.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. JDocQA: Japanese document question answering dataset for generative language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9503–9514, Torino, Italy. ELRA and ICCL.

Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2024. JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-linguistic question answering (VLQA) challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616, Online. Association for Computational Linguistics.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa 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 *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928. Association for Computational Linguistics.

Makoto Shing and Takuya Akiba. 2023. Japanese instructblip alpha.

Makoto Shing and Takuya Akiba. 2024. Japanese stable vlm.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *AAAI*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

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 *Advances in Neural Information Processing Systems*, volume 36, pages 8958–8974. Curran Associates, Inc.

A Dataset Construction Details

A.1 Image Downloading

When we download images from URLs, we limit the extensions of the image file to “.jpg”, “.jpeg”, and “.png”, and if the words “logo”, “button”, “icon”, “plugin”, or “widget” is included, we skip them. If the size of the image is less than 150px in either the height or width, it will be removed. Also, if the aspect ratio of the image is less than 0.5 or greater than 2, it will be removed.

A.2 Filtering for Image-text Pair Dataset

Some data has text that is set when the alt attribute is not set. We remove such data, where the text begin with “画像にalt属性が指定されていません。” or “この画像にはalt属性が指定されておらず、”.

In addition, some data have the image file name set automatically when a screenshot is taken, etc., as alt text. For example, “写真 2015-01-20 18 12 33”. Specifically, if the data does not contain Japanese after “写真,” “キャプチャ,” “画像,” “スクリーンショット,” “全画面キャプチャ,” “ファイル,” “コメント,” or “コピー,” the data will be removed.

A.3 Japanese Photos Conv Dataset

The system prompt for QA generation of Japanese-photos-conv dataset is shown in Table 6.

Table 6: The prompt used in generating japanese-photos-conv.

Prompt

~~~~
あなたは、マルチモーダルinstruction tuningデータの優秀なアノテーターです。

これから、1枚の画像が与えられるので、その画像に関するinstruction tuningのための高品質なデータセットを生成する必要があります。このデータは、モデルが異なるモーダリティ間での関連性や相互作用を学習できるよう設計されなければなりません。

データにはオブジェクトの種類、オブジェクトの数、オブジェクトの動作、オブジェクトの位置、オブジェクト間の相対位置など、画像の内容を尋ねる質問を含めてください。また、明確な答えがある質問のみを含め、自信を持って答えられない質問はしないようにしてください。

また、画像の内容に関連した複雑な質問、例えば、画像に写っているオブジェクトの背景知識を尋ねる質問、画像の中で起こっている出来事について議論するよう求める質問なども含めてください。この場合も、不確かな詳細については質問しないようにしてください。
複雑な質問に回答する際は、詳細な回答にしてください。例えば、詳細な例や推論の手順を示すことで、内容に説得力を持たせ、よく整理された回答にすることができます。

出力形式は次のようにしてください。
```
Q:
{質問}
A:
{回答}
===
Q:
{質問}
A:
{回答}
===
・・・(省略)
===
Q:
{質問}
A:
{回答}
```
~~~~

Table 7: Comparison of Vision Language Models (VLMs) highlighting their base language models and its sizes (in billions of parameters), and corresponding Hugging Face repositories or API.

VLM	Reference	Base LM	LM Size	Hugging Face / API
Llama 3.2 Vision 11B	(Meta, 2024)	Llama 3.2	11B	meta-llama/Llama-3.2-11B-Vision
Qwen2-VL 7B	(Wang et al., 2024)	Qwen2	7B	Qwen/Qwen2-VL-7B-Instruct
InternVL2 8B	(Chen et al., 2024)	InternLM2-Chat	8B	OpenGVLab/InternVL2-8B
LLaVA-1.5 7B	(Liu et al., 2024a)	Llama2	7B	llava-hf/llava-1.5-7b-hf
LLaVA-1.6 7B	(Liu et al., 2024b)	Mistral	7B	llava-hf/llava-v1.6-mistral-7b-hf
VILA-1.5 13B	(Lin et al., 2023)	Vicuna 1.5	13B	Efficient-Large-Model/VILA1.5-13b
LLaVA-CALM2-SigLIP	(Inagaki, 2024)	CALM2	7B	cyberagent/llava-calm2-siglip
Japanese Stable VLM	(Shing and Akiba, 2024)	Japanese Stable LM Instruct Gamma	7B	stabilityai/japanese-stable-vlm
Japanese InstructBLIP Alpha	(Shing and Akiba, 2023)	Japanese StableLM Instruct Alpha	7B	stabilityai/japanese-instructblip-alpha
Llama-3-EvoVLM-JP-v2	(Akiba et al., 2024)	Mantis-8B-SigLIP-Llama-3 Merged Llama-3-ELYZA-JP-8B Bunny-v1.1-Llama-3-8B-V	8B	SakanaAI/Llama-3-EvoVLM-JP-v2
GPT-4o	(OpenAI, 2024)	GPT-4	-	gpt-4o-2024-05-13

B Evaluation Details

B.1 Baseline Models

We present a overview of the vision language models and their corresponding base language models, sizes, and Hugging Face repositories or APIs in Table 7.

B.2 Prompts for LLM-as-a-Judge

We also provide the actual prompt used in the evaluation in Table 8.

B.3 Detailed Result of Heron Bench

Table 9 provides a breakdown of Heron-Bench scores by category for each model. Our model demonstrates state-of-the-art performance across all categories among open models.

B.4 Additional Qualitative Examples

Here we provide additional examples in Figure 4,5,6, and 7.

B.5 About JMMMU

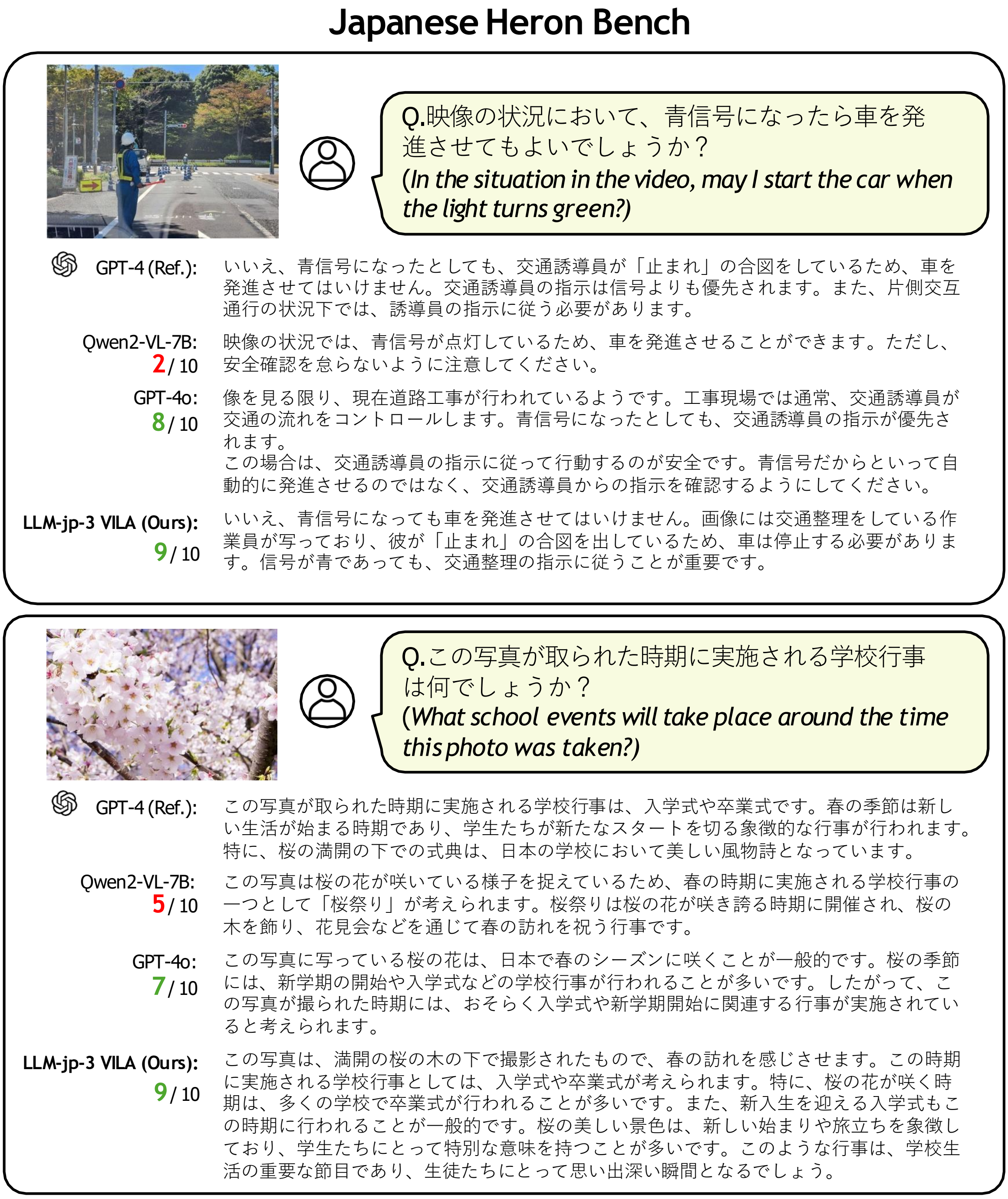
We are aware of the JMMMU (Onohara et al., 2024), a valuable resource for evaluating Japanese vision and language models. However, our instruction tuning process for LLM-jp-3 VILA focused on generating free-form answers rather than selecting from a predefined set of options. As JMMMU primarily consists of multiple-choice questions, it was deemed unsuitable for assessing the performance of our model in its current stage of training. We plan to explore fine-tuning strategies specifically for multiple-choice QA in future work.

Table 8: The prompt used in LLM-as-a-judge process. {input_text}, {answer}, and {pred} indicate the place to insert the question, answer and VLM's prediction, respectively.

Prompt
Instruction You are an expert evaluator. Your task is to evaluate the quality of the responses generated by AI models. We will provide you with the user prompt and an AI-generated responses. You should first read the user prompt carefully for analyzing the task, and then evaluate the quality of the responses based on and rules provided in the Evaluation section below. # Evaluation ## Metric Definition You will be assessing question answering quality, which measures the overall quality of the answer to the question in the user prompt. Pay special attention to length constraints, such as in X words or in Y sentences. The instruction for performing a question-answering task is provided in the user prompt. The response should not contain information that is not present in the context (if it is provided). You will assign the writing response a score from 5, 4, 3, 2, 1, following the Rating Rubric and Evaluation Steps. Give step-by-step explanations for your scoring, and only choose scores from 5, 4, 3, 2, 1. ## Criteria Definition Instruction following: The response demonstrates a clear understanding of the question answering task instructions, satisfying all of the instruction's requirements. Groundedness: The response contains information included only in the context if the context is present in the user prompt. The response does not reference any outside information. Completeness: The response completely answers the question with sufficient detail. Fluent: The response is well-organized and easy to read. ## Rating Rubric 5: (Very good). The answer follows instructions, is grounded, complete, and fluent. 4: (Good). The answer follows instructions, is grounded, complete, but is not very fluent. 3: (Ok). The answer mostly follows instructions, is grounded, answers the question partially and is not very fluent. 2: (Bad). The answer does not follow the instructions very well, is incomplete or not fully grounded. 1: (Very bad). The answer does not follow the instructions, is wrong and not grounded. ## Evaluation Steps STEP 1: Assess the response in aspects of instruction following, groundedness,completeness, and fluency according to the criteria. STEP 2: Provide overall score based on the rubric in the format of ‘Score: X‘ where X is the score you assign to the response. # Question, Reference Answer, and AI-generated Response ## Question {input_text} ## Reference Answer {answer} ## AI-generated Response {pred}

				Japanese Heron-Bench			
				Detail	**Conv**	**Complex**	**Average**
Japanese InstructBLIP Alpha				12.4	13.9	15.7	14.0
Japanese Stable VLM				18.9	30.7	23.0	24.2
Llama-3-EvoVLM-JP-v2				43.1	37.9	36.9	39.3
LLaVA-CALM2-SigLIP				45.4	45.8	38.8	43.3
LLaVA-1.6 7B				21.3	27.5	28.7	25.8
LLaVA-1.5 7B				34.7	33.8	35.7	34.8
Llama 3.2 11B Vision				34.4	40.0	35.1	36.5
InternVL2 8B				48.9	41.1	45.5	45.2
Qwen2-VL 7B Inst				57.2	54.2	53.1	54.8
	Step-0	Step-1	Step2				
LLM-jp-3 VILA (Ours)	✓	✓	translated	46.3	43.4	52.0	47.2
	✓	✗	✓	**60.5**	50.3	58.8	56.5
	✓	w/o interleaved	✓	57.7	**54.9**	**63.3**	**58.6**
	✓	✓	✓	57.2	54.4	60.0	57.2
GPT-4o				94.3	80.4	88.2	87.6

Table 9: Performance of VLMs on Japanese Heron Bench dataset.



Japanese Heron Bench

Q.映像の状況において、青信号になったら車を発進させてもよいでしょうか？
(In the situation in the video, may I start the car when the light turns green?)

GPT-4 (Ref.): いいえ、青信号になったとしても、交通誘導員が「止まれ」の合図をしているため、車を発進させてはいけません。交通誘導員の指示は信号よりも優先されます。また、片側交互通行の状況下では、誘導員の指示に従う必要があります。

Qwen2-VL-7B: 2 / 10 映像の状況では、青信号が点灯しているため、車を発進させることができます。ただし、安全確認を怠らないように注意してください。

GPT-4o: 8 / 10 像を見る限り、現在道路工事が行われているようです。工事現場では通常、交通誘導員が交通の流れをコントロールします。青信号になったとしても、交通誘導員の指示が優先されます。
この場合は、交通誘導員の指示に従って行動するのが安全です。青信号だからといって自動的に発進させるのではなく、交通誘導員からの指示を確認するようにしてください。

LLM-jp-3 VILA (Ours): 9 / 10 いいえ、青信号になっても車を発進させてはいけません。画像には交通整理をしている作業員が写っており、彼が「止まれ」の合図を出しているため、車は停止する必要があります。信号が青であっても、交通整理の指示に従うことが重要です。

Q.この写真が取られた時期に実施される学校行事は何でしょうか？
(What school events will take place around the time this photo was taken?)

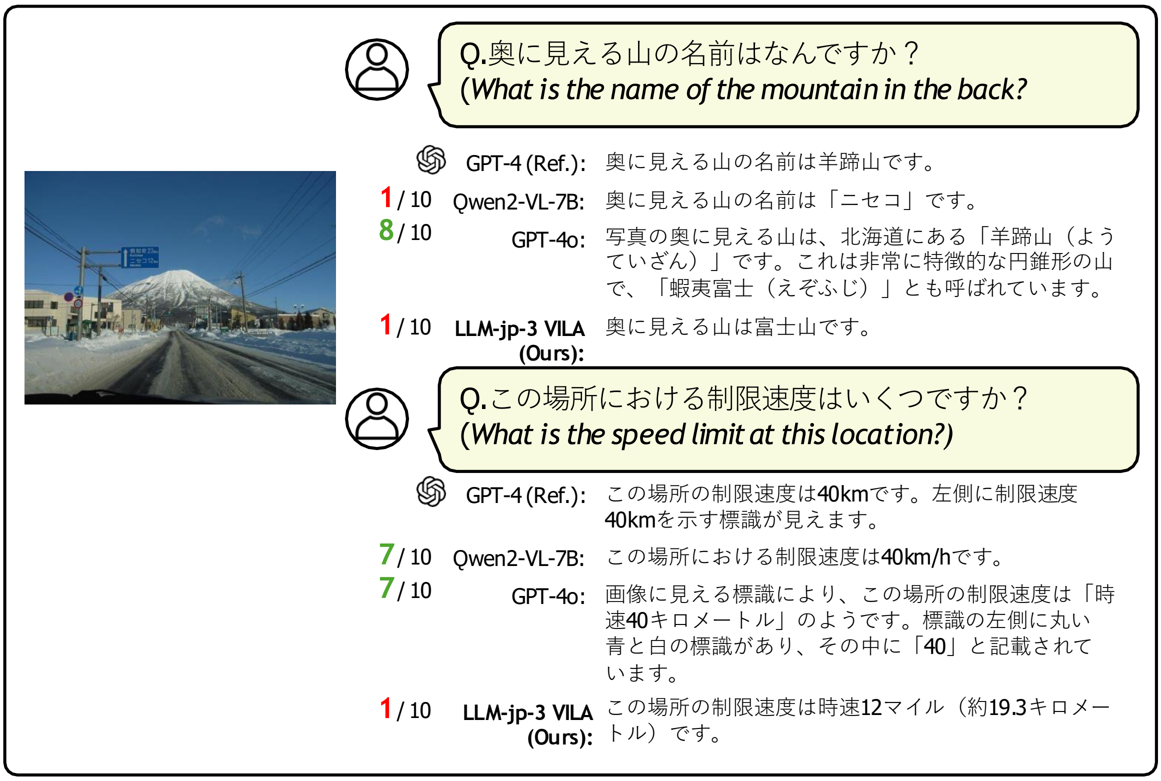
GPT-4 (Ref.): この写真が取られた時期に実施される学校行事は、入学式や卒業式です。春の季節は新しい生活が始まる時期であり、学生たちが新たなスタートを切る象徴的な行事が行われます。特に、桜の満開の下での式典は、日本の学校において美しい風物詩となっています。

Qwen2-VL-7B: 5 / 10 この写真は桜の花が咲いている様子を捉えているため、春の時期に実施される学校行事の一つとして「桜祭り」が考えられます。桜祭りは桜の花が咲き誇る時期に開催され、桜の木を飾り、花見会などを通じて春の訪れを祝う行事です。

GPT-4o: 7 / 10 この写真に写っている桜の花は、日本で春のシーズンに咲くことが一般的です。桜の季節には、新学期の開始や入学式などの学校行事が行われることが多いです。したがって、この写真が撮られた時期には、おそらく入学式や新学期開始に関連する行事が実施されていると考えられます。

LLM-jp-3 VILA (Ours): 9 / 10 この写真は、満開の桜の木の下で撮影されたもので、春の訪れを感じさせます。この時期に実施される学校行事としては、入学式や卒業式が考えられます。特に、桜の花が咲く時期は、多くの学校で卒業式が行われることが多いです。また、新入生を迎える入学式もこの時期に行われることが一般的です。桜の美しい景色は、新しい始まりや旅立ちを象徴しており、学生たちにとって特別な意味を持つことが多いです。このような行事は、学校生活の重要な節目であり、生徒たちにとって思い出深い瞬間となるでしょう。

Figure 4: Additional qualitative examples in the Heron Bench.



Q.奥に見える山の名前はなんですか？
(*What is the name of the mountain in the back?*

GPT-4 (Ref.): 奥に見える山の名前は羊蹄山です。

1 / 10 Qwen2-VL-7B: 奥に見える山の名前は「ニセコ」です。

8 / 10 GPT-4o: 写真の奥に見える山は、北海道にある「羊蹄山（ようていざん）」です。これは非常に特徴的な円錐形の山で、「蝦夷富士（えぞふじ）」とも呼ばれています。

1 / 10 **LLM-jp-3 VILA (Ours):** 奥に見える山は富士山です。

Q.この場所における制限速度はいくつですか？
(*What is the speed limit at this location?*)

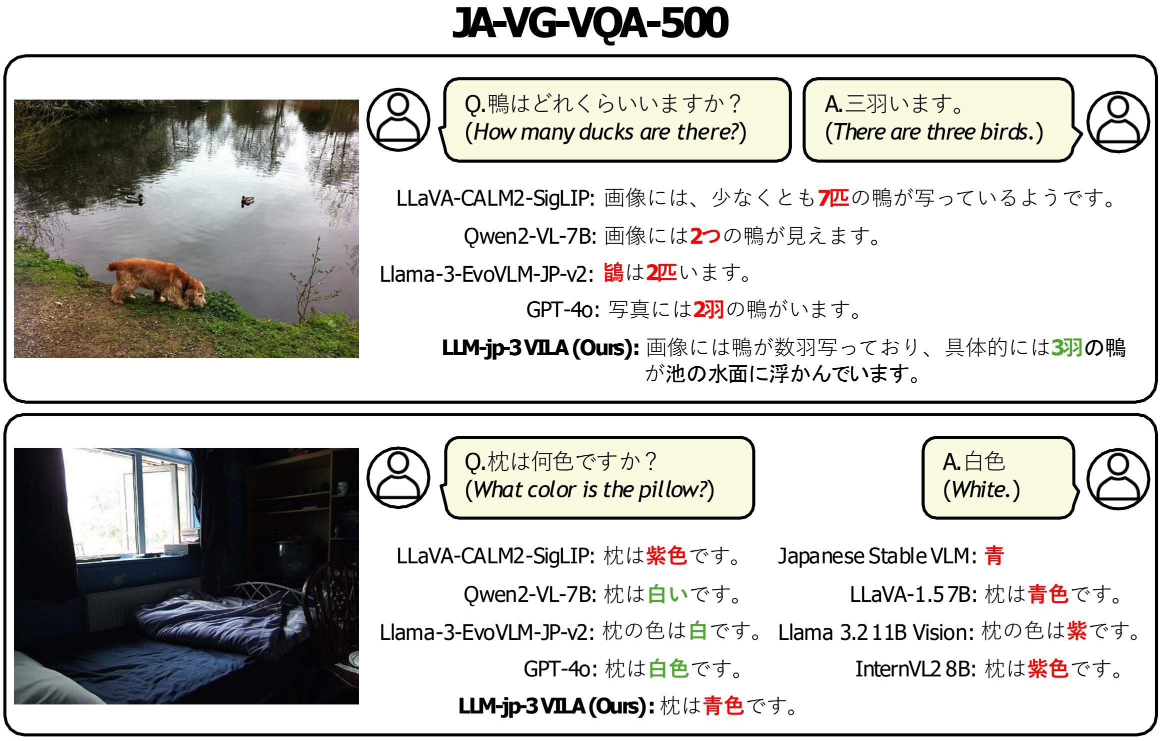
GPT-4 (Ref.): この場所の制限速度は40kmです。左側に制限速度40kmを示す標識が見えます。

7 / 10 Qwen2-VL-7B: この場所における制限速度は40km/hです。

7 / 10 GPT-4o: 画像に見える標識により、この場所の制限速度は「時速40キロメートル」のようです。標識の左側に丸い青と白の標識があり、その中に「40」と記載されています。

1 / 10 **LLM-jp-3 VILA (Ours):** この場所の制限速度は時速12マイル（約19.3キロメートル）です。

Figure 5: Continuous qualitative examples in the Heron Bench.



JA-VG-VQA-500

Q.鴨はどれくらいいますか？
(*How many ducks are there?*)

A.三羽います。
(*There are three birds.*)

LLaVA-CALM2-SigLIP: 画像には、少なくとも7匹の鴨が写っているようです。

Qwen2-VL-7B: 画像には2つの鴨が見えます。

Llama-3-EvoVLM-JP-v2: 鵠は2匹います。

GPT-4o: 写真には2羽の鴨がいます。

LLM-jp-3 VILA (Ours): 画像には鴨が数羽写っており、具体的には3羽の鴨が池の水面に浮かんでいます。

Q.枕は何色ですか？
(*What color is the pillow?*)

A.白色
(*White.*)

LLaVA-CALM2-SigLIP: 枕は紫色です。

Qwen2-VL-7B: 枕は白いです。

Llama-3-EvoVLM-JP-v2: 枕の色は白です。

GPT-4o: 枕は白色です。

LLM-jp-3 VILA (Ours): 枕は青色です。

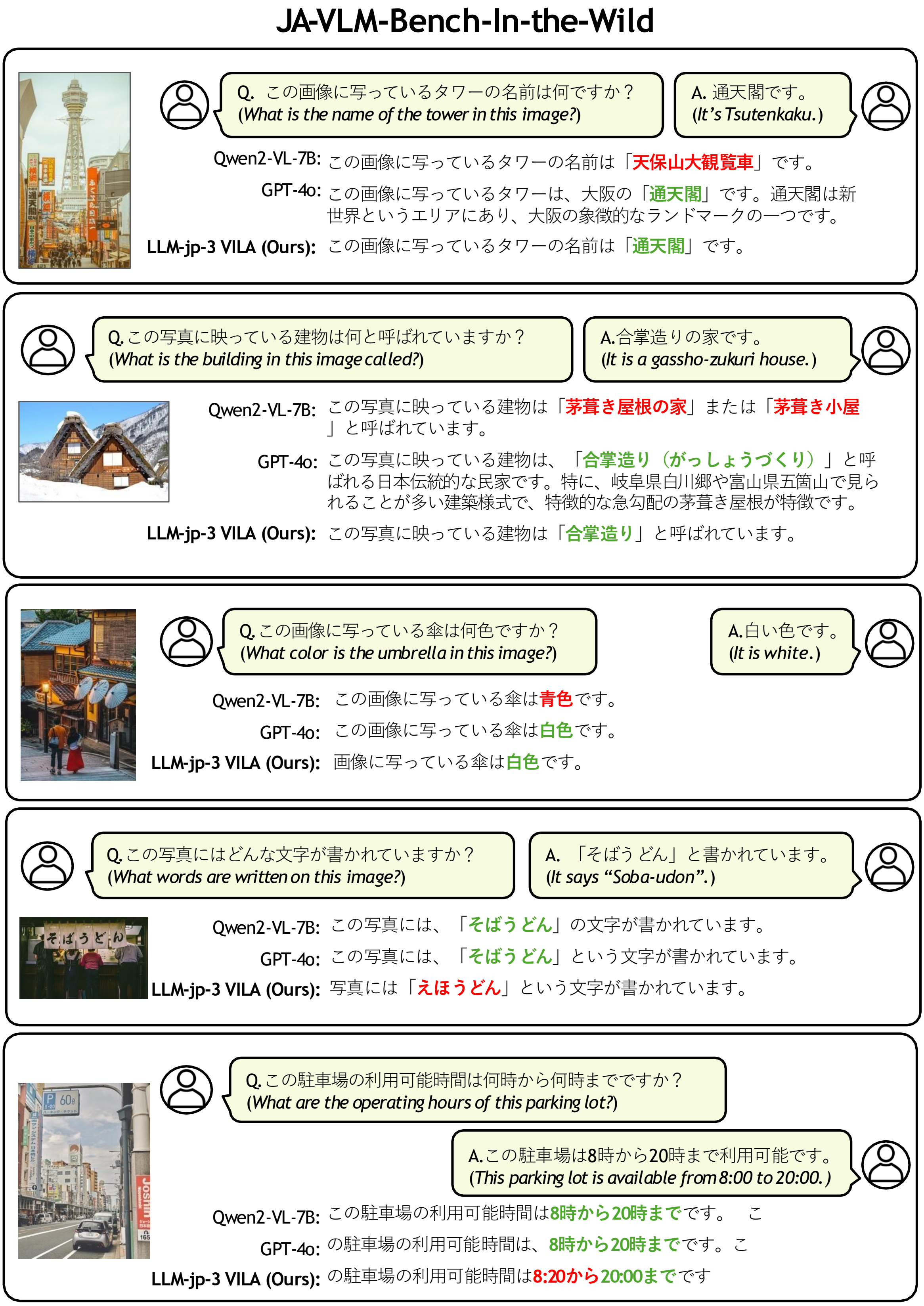
Japanese Stable VLM: 青

LLaVA-1.5 7B: 枕は青色です。

Llama 3.2 11B Vision: 枕の色は紫です。

InternVL2 8B: 枕は紫色です。

Figure 6: Additional qualitative examples in the JA-VG-VQA-500.



JA-VLM-Bench-In-the-Wild

Q. この画像に写っているタワーの名前は何ですか？
(*What is the name of the tower in this image?*)

A. 通天閣です。
(*It's Tsutenkaku.*)

Qwen2-VL-7B: この画像に写っているタワーの名前は「**天保山大観覧車**」です。

GPT-4o: この画像に写っているタワーは、大阪の「**通天閣**」です。通天閣は新世界というエリアにあり、大阪の象徴的なランドマークの一つです。

LLM-jp-3 VILA (Ours): この画像に写っているタワーの名前は「**通天閣**」です。

Q. この写真に映っている建物は何と呼ばれていますか？
(*What is the building in this image called?*)

A. 合掌造りの家です。
(*It is a gassho-zukuri house.*)

Qwen2-VL-7B: この写真に映っている建物は「**茅葺き屋根の家**」または「**茅葺き小屋**」と呼ばれています。

GPT-4o: この写真に映っている建物は、「**合掌造り（がっしょうづくり）**」と呼ばれる日本伝統的な民家です。特に、岐阜県白川郷や富山県五箇山で見られることが多い建築様式で、特徴的な急勾配の茅葺き屋根が特徴です。

LLM-jp-3 VILA (Ours): この写真に映っている建物は「**合掌造り**」と呼ばれています。

Q. この画像に写っている傘は何色ですか？
(*What color is the umbrella in this image?*)

A. 白い色です。
(*It is white.*)

Qwen2-VL-7B: この画像に写っている傘は**青色**です。

GPT-4o: この画像に写っている傘は**白色**です。

LLM-jp-3 VILA (Ours): 画像に写っている傘は**白色**です。

Q. この写真にはどんな文字が書かれていますか？
(*What words are written on this image?*)

A. 「そばう どん」と書かれています。
(*It says "Soba-udon".*)

Qwen2-VL-7B: この写真には、「**そばう どん**」の文字が書かれています。

GPT-4o: この写真には、「**そばう どん**」という文字が書かれています。

LLM-jp-3 VILA (Ours): 写真には「**えほうどん**」という文字が書かれています。

Q. この駐車場の利用可能時間は何時から何時までですか？
(*What are the operating hours of this parking lot?*)

A. この駐車場は8時から20時まで利用可能です。
(*This parking lot is available from 8:00 to 20:00.*)

Qwen2-VL-7B: この駐車場の利用可能時間は**8時から20時まで**です。　こ

GPT-4o: の駐車場の利用可能時間は、**8時から20時まで**です。こ

LLM-jp-3 VILA (Ours): の駐車場の利用可能時間は**8:20から20:00まで**です

Figure 7: Additional qualitative examples in the JA-VLM-Bench-In-the-Wild.