DEJIMA: A Novel Large-scale Japanese Dataset for Image Captioning and Visual Question Answering

Toshiki Katsube[1], **Taiga Fukuhara**[1], **Kenichiro Ando**[2,1], **Yusuke Mukuta**[1,2], **Kohei Uehara**[1], **Tatsuya Harada**[1,2]
[1]The University of Tokyo, [2]RIKEN
{katsube, fukuhara, ando, mukuta, uehara, harada}@mi.t.u-tokyo.ac.jp

Abstract

Vision-and-Language (V&L) models depend on large-scale, high-quality datasets, yet most resources are English-centric, and existing Japanese V&L datasets face a fundamental trade-off: manually annotated corpora offer quality but limited scale, translated datasets introduce unnatural phrasing and cultural bias, and web-crawled collections achieve scale but suffer from noise and poor grounding. To resolve this trade-off, we propose DEJIMA, a novel pipeline whose key idea is *detection-guided LLM refinement*: object detection first extracts visually verifiable evidence (labels and bounding boxes), then an LLM generates or refines Japanese text conditioned on this evidence, ensuring both factual grounding and linguistic naturalness without costly human annotation. Using this pipeline, we build two resources: an image–caption dataset (DEJIMA-Cap) and a VQA dataset (DEJIMA-VQA), each containing approximately 3.88M image–text pairs—over 20× larger than existing Japanese V&L datasets. Human evaluations demonstrate that DEJIMA achieves substantially higher Japaneseness and linguistic naturalness than translation- or annotation-based baselines, while maintaining factual correctness comparable to human-annotated corpora. Models trained on DEJIMA show consistent improvements across multiple Japanese multimodal benchmarks, confirming that culturally grounded, large-scale resources play a key role in enhancing model performance. All pipeline components are commercially licensed, and we publicly release the dataset and metadata to support further research and applications. Our project page is available at `https://mil-tokyo.github.io/DEJIMA-dataset/`.

Keywords: vision-and-language resources, Japanese multimodality, dataset creation, licensing and ethics

1. Introduction

Recent years have seen rapid progress in Vision-and-Language (V&L) models, which jointly process visual and textual information for tasks such as image captioning, Visual Question Answering (VQA), and image retrieval. The performance of these models critically depends on the availability of high-quality, large-scale datasets. However, most existing V&L datasets have been constructed primarily in English (Nguyen et al., 2024), and there remains a severe shortage of large, culturally appropriate resources for Japanese language.

Existing Japanese V&L datasets face clear trade-offs among quality, cultural adequacy, and scalability. Manually annotated datasets such as STAIR Captions (Yos, 2017) (164K images) and Japanese Visual Genome (Shimizu et al., 2018) (99K images) provide high-quality but small-scale data. Machine-translated datasets inherit biases and linguistic unnaturalness from English sources, while automatically crawled datasets like Japanese Image Text Pairs (Sasagawa et al., 2024) achieve large scale but often include noise, poor grounding, or alt-texts that are not natural Japanese sentences. Moreover, existing web-crawled and automatically constructed datasets are limited to caption-style or interleaved image–text formats.

To address these limitations, we propose DEtection-based Japanese Image-text dataset for Multi-modal Analysis (DEJIMA), a scalable pipeline for constructing culturally grounded, high-quality Japanese V&L datasets. Our approach integrates: (1) **web collection, filtering, and deduplication** of images and corresponding alt-texts from Japanese web pages in Common Crawl[1]; (2) **object detection** for extracting visually verifiable evidence; and (3) **LLM-based refinement** that enforces fluency and factual grounding under explicit constraints.

Using this pipeline, we created two resources: DEJIMA-Cap (image–caption pairs) and DEJIMA-VQA (image–question–answer triples), each containing approximately 3.88 million entries. This corresponds to 23.7× the scale of STAIR Captions and 39.1× that of Japanese Visual Genome. All models used in our pipeline are under commercially permissible licenses, enabling the resulting dataset to be used safely for research and commercial purposes.

Human evaluations demonstrate that DEJIMA achieves markedly higher scores in Japaneseness and linguistic naturalness compared with both manual and translation-based baselines, while maintaining factual correctness at a level comparable to human-annotated data. These results indicate that DEJIMA successfully balances scale, cultural richness, and quality. Furthermore, V&L models trained on DEJIMA outperform prior Japanese

[1]`https://commoncrawl.org/`

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9478–9489
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

datasets on benchmarks such as JA-VLM-Bench-In-the-Wild (aki, 2024) and Heron-bench (Inoue et al., 2024), highlighting the effectiveness of our detection-guided and LLM-refined approach.

In summary, our contributions are threefold:

- We introduce a detection-based Large Language Model (LLM) refinement pipeline that enhances fluency and factual grounding through evidence-first generation.
- We construct a large-scale Japanese Vision-and-Language (V&L) dataset (3.88M pairs) that surpasses existing resources by over 20× in scale while preserving cultural and linguistic naturalness.
- We demonstrate that models trained with DEJIMA achieve strong performance and natural, culturally coherent Japanese text generation, approaching human annotation quality.

2. Related Work

2.1. Dataset Construction Methods

Four common paradigms exist for constructing V&L datasets.

Human Annotation. Manual annotation offers strong image–text alignment and high textual quality but is prohibitively costly: building MS COCO (Lin et al., 2014) alone required over 70,000 crowd-worker hours.

Web Crawling. Large-scale efforts collect images and their associated alt-text from the web, as in CC3M (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), and LAION-400M (Schuhmann et al., 2021). Despite the large scale, alt-text can be noisy (partial descriptions, mismatched semantics, or unnatural sentences).

Translation. English resources are often translated to other languages (e.g., LLaVA-Instruct-150K-JA[2] from LLaVA Visual Instruct 150K (Liu et al., 2023)). While scalable, translations can introduce unnatural phrasing and attenuate culture-specific knowledge.

Generation/Augmentation. LLM-based augmentation can rewrite captions to enrich linguistic variety (e.g., LaCLIP (Fan et al., 2023)), whereas fully synthetic pipelines (e.g., SynthVLM (Liu et al., 2024b)) can generate image–text pairs from scratch. Risks include hallucinations when text is detached from image evidence and cultural bias if generators were not trained on diverse, multilingual sources.

2.2. Existing Japanese V&L Datasets

Existing Japanese V&L datasets span several of the paradigms above but each faces limitations. Human-annotated resources such as STAIR Captions (Yos, 2017) (164K images) and Japanese Visual Genome (Shimizu et al., 2018) (99K images) provide high quality but limited scale. Flickr30k (Young et al., 2014), Visual Genome (Krishna et al., 2017), GQA (Hudson and Manning, 2019), and MS COCO Captions (Lin et al., 2014) are major English benchmarks whose Japanese translations inherit the biases noted above. The Japanese Image Text Pairs dataset (Sasagawa et al., 2024) achieves large scale via web crawling but suffers from noise and poor grounding typical of alt-text. DEJIMA bridges this gap by combining web-scale collection with detection-guided LLM refinement, achieving both scale and quality.

3. Dataset

Figure 1 outlines our three-stage pipeline: (1) collecting URLs and alt-text from the web with rigorous filtering and deduplication, (2) object detection and tag calibration, and (3) LLM-based caption/VQA generation with grounding constraints.

3.1. Creation Pipeline

Stage 1: Web Collection, Filtering, and Deduplication. We use Common Crawl (pages crawled Aug 7, 2022–Jan 26, 2025; ∼50B pages) as the source and retain only Japanese pages and safe content through a multi-step filtering pipeline. Language and adult filtering are conducted using `HojiChar`[3], applying `AcceptJapanese()` to keep only Japanese text and `DiscardAdultContentJa()` to exclude adult material. Text normalization is performed (e.g., converting full-width spaces to half-width, trimming, and compressing consecutive spaces), and only entries whose alt-text passes all filters are kept together with their image URLs.

High-frequency alt-texts are then removed to promote diversity by discarding entries with frequency ≥ 10. For near-duplicate removal, we compute perceptual hashes (`imagehash.phash`) and deduplicate by the pair $(\texttt{pHash}, \texttt{alt})$, retaining only the first occurrence to eliminate visually identical images with the same caption.

To further ensure semantic alignment, we calculate the cosine similarity between image and

[2] https://huggingface.co/datasets/turing-motors/LLaVA-Instruct-150K-JA

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

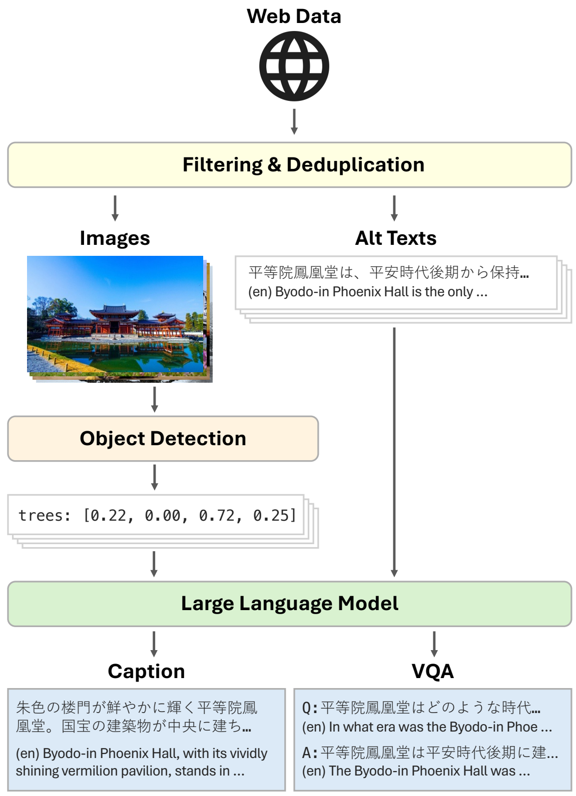

Figure 1: Pipeline: (1) web collection with strict filtering & deduplication, (2) detection-driven evidence extraction, (3) LLM refinement with grounding/safety constraints.

alt-text embeddings using the Japanese CLIP model `line-corporation/clip-japanese-base`[4]. Only entries above the 30th percentile of the global similarity distribution are kept, removing the noisiest bottom $\sim 30\%$. Each retained pair is stored with its alignment score. Finally, we retain only images with supported extensions (`jpeg`, `jpg`, `png`), filter out extreme aspect ratios ($>3{:}1$), and download the remaining images via `img2dataset`[5].

Stage 2: Object Detection and Tag Calibration. Visual evidence is extracted using Recognize Anything (RAM) (Zhang et al., 2024) with the RAM++(14M) checkpoint[6], which provides comprehensive object tags. Bounding boxes for each detected object are then obtained using Grounding DINO (Liu et al., 2024a), ensuring spatial grounding of visual entities. This combination yields both rich object-level coverage and accurate localization information for downstream text generation.

Stage 3: LLM-guided Generation with Grounding Constraints. We refine or generate texts using `qwen2.5-bakeneko-32b`[7], conditioning on both alt-text and detection outputs (labels and bounding boxes) to provide explicit visual grounding. A few-shot prompting setup is employed, in which curated exemplars guide the model toward culturally natural and semantically faithful text generation.

For the caption dataset, the LLM refines noisy alt-texts into fluent and descriptive Japanese captions consistent with detected objects. For the VQA dataset, both questions and answers are generated by the LLM under the same grounding signals, ensuring coherence between the visual evidence and linguistic reasoning.

Efficiency Comparison. While manually annotated datasets such as MS COCO (123K images) required over 70,000 crowd-worker hours to complete (Lin et al., 2014), our automated DEJIMA pipeline generates approximately 1 million captions or VQA pairs in about 40 GPU hours using an NVIDIA A100 GPU. This represents an efficiency gain of several orders of magnitude.

Dataset Variants. We prepare multiple dataset variants to isolate the contributions of each component in our pipeline. For the captioning task, we construct four datasets: **DEJIMA-Cap-Simple** consists of raw image–alt-text pairs after filtering; **DEJIMA-Cap-Refined** uses only LLM-refined alt-texts; **DEJIMA-Cap-Detection** contains captions generated by the LLM from only detection tags; and **DEJIMA-Cap-All** integrates both alt-texts and detection tags as inputs to the LLM.

Similarly, for the VQA task, we prepare three corresponding datasets: **DEJIMA-VQA-Refined**, **DEJIMA-VQA-Detection**, and **DEJIMA-VQA-All**.

Dataset Example. Figure 2 shows one example each from DEJIMA-Cap and DEJIMA-VQA.

3.2. Dataset Analysis

Statistical Comparison. Table 1 provides an overview of the quantitative properties of existing Japanese Vision–Language datasets and our DEJIMA resources. For reference, we also include MS COCO Translation and GQA Translation, which are machine-translated versions of the original English datasets (MS COCO Captions and GQA) produced using `qwen2.5-bakeneko-32b`.

[4] https://huggingface.co/line-corporation/clip-japanese-base
[5] https://github.com/rom1504/img2dataset
[6] https://huggingface.co/xinyu1205/recognize-anything-plus-model/blob/main/ram_plus_swin_large_14m.pth
[7] https://huggingface.co/rinna/qwen2.5-bakeneko-32b

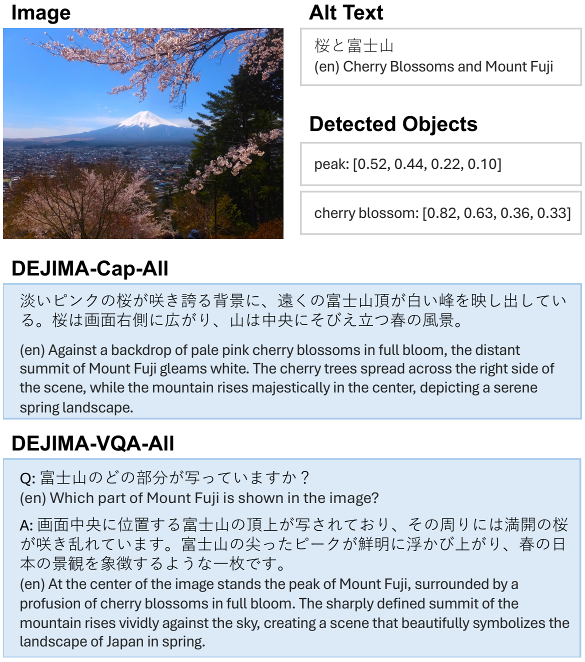

Figure 2: Example instances from DEJIMA-Cap and DEJIMA-VQA. The figure illustrates the pipeline flow: (1) the input image is processed by the object-detection model to obtain *detected objects* (labels and bounding boxes); (2) the *alt text* and detected objects are provided as inputs to the LLM, which generates a refined *caption* and grounded *VQA* pairs in Japanese. English translations are also included in the figure.

From a scalability perspective, DEJIMA-Cap and DEJIMA-VQA each comprise roughly 3.88 million image–text pairs—23.7 times larger than STAIR Captions (164K images) and 39.1 times larger than Japanese Visual Genome (99K images). This vast scale allows DEJIMA to cover a substantially broader range of visual scenes and linguistic phenomena than prior Japanese datasets. Note that the exact counts differ slightly between DEJIMA-Cap variants (e.g., 3,884,632) and DEJIMA-VQA variants (e.g., 3,882,892): this is because Stage 3 VQA generation occasionally produces outputs that fail format-validation checks (e.g., malformed question–answer structure), causing a small number of entries to be discarded.

In terms of linguistic diversity, captions in DEJIMA-Cap-All are longer and more informative, with an average length of 79.6 characters, compared to around 21 characters in STAIR Captions and MS COCO Translation. The vocabulary also expands significantly, reaching approximately 287K unique types, indicating rich lexical variety and descriptive expressiveness.

From DEJIMA-Cap-Simple to DEJIMA-Cap-Refined, the average caption length nearly doubles (18 → 38 characters), showing that LLM refinement enhances fluency and expressiveness. Incorporating object-detection information (DEJIMA-Cap-Detection) further adds visually grounded terms, and combining both detection results and alt-text information (DEJIMA-Cap-All) leads to larger vocabulary expansion (from about 30K to 280K types). Interestingly, the vocabulary size of DEJIMA-Cap-All is slightly smaller than that of DEJIMA-Cap-Simple, which can be attributed to the LLM refinement process filtering out overly specific or uncommon proper nouns that appeared in raw web-derived alt-texts. This indicates that the inclusion of web-derived alt-text contributes a rich variety of linguistic expressions—such as specific object names, proper nouns, and contextual descriptions—while the LLM-based refinement selectively retains more general and natural ones. A similar pattern is observed for the VQA data. These statistics collectively demonstrate that our pipeline not only scales up dataset size but also improves linguistic diversity and descriptive depth by effectively combining visual grounding and naturally written web text.

Representational Coverage. To examine the representational coverage of DEJIMA relative to existing datasets, we analyzed the 2D feature distributions obtained by applying PCA to CLIP (Ilharco et al., 2021) image embeddings. All datasets were jointly projected into a shared two-dimensional space, and a common 60×60 grid (with 2% padding) was used to discretize the plane into probability maps $p_d(i, j)$ for each dataset d.

For each dataset, we computed two complementary measures: (1) the asymmetric *coverage rate* $\mathrm{Coverage}(P|Q) = \sum_{b \in \mathrm{occ}_Q} p_P(b)$, which quantifies how much probability mass of dataset P lies within the bins occupied by Q, and (2) the bidirectional Kullback–Leibler divergences $\mathrm{KL}(P||Q)$ and $\mathrm{KL}(Q||P)$, computed with numerical stabilization ($\varepsilon = 10^{-12}$). These metrics respectively capture the spatial overlap and the distributional divergence between datasets in the shared embedding space.

Using the domestic dataset *recruit-jp*[8] as the reference target, DEJIMA achieved the highest coverage rate $\mathrm{Coverage}(\text{target}|\text{DEJIMA}) = 0.785$, substantially exceeding Japanese Visual Genome (0.435), STAIR Captions (0.430), MS COCO (0.406), and GQA (0.342). This indicates that DEJIMA spans approximately 79% of the visual domain occupied by real Japanese imagery. Conversely, when measuring $\mathrm{Coverage}(\text{dataset}|\text{target})$, Japanese Visual Genome exhibited the highest value (0.534), followed by MS COCO (0.502) and STAIR Captions (0.492), while DEJIMA scored

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

Dataset	Type	# Images	# Texts	Avg. # Chars	Vocabulary Size
Caption					
STAIR Captions	Human-annotated	123,287	616,435	23.80	30,195
MS COCO Translation	Machine-translated	123,287	616,767	22.41	32,960
DEJIMA-Cap-Simple (Ours)	Alt	3,884,632	3,884,632	18.21	336,924
DEJIMA-Cap-Refined (Ours)	Alt + LLM	3,884,629	3,884,629	38.03	314,900
DEJIMA-Cap-Detection (Ours)	Detection + LLM	3,884,632	3,884,632	49.55	30,674
DEJIMA-Cap-All (Ours)	Alt + Detection + LLM	3,884,632	3,884,632	79.62	287,434
VQA					
Japanese Visual Genome	Human-annotated	99,208	793,664	19.50	20,797
GQA Translation	Machine-translated	71,067	3,999,765	22.58	11,856
DEJIMA-VQA-Refined (Ours)	Alt + LLM	3,875,343	3,875,343	56.62	321,720
DEJIMA-VQA-Detection (Ours)	Detection + LLM	3,883,943	3,883,943	77.00	31,929
DEJIMA-VQA-All (Ours)	Alt + Detection + LLM	3,882,892	3,882,892	108.86	278,860

Table 1: Statistical comparison of Japanese V&L Datasets.

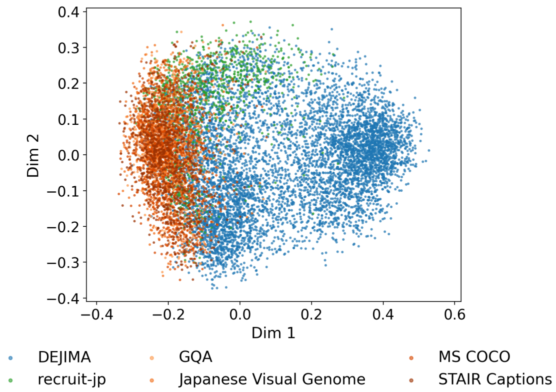

Figure 3: PCA projection of CLIP image embeddings. DEJIMA covers the entire region of recruit-jp while extending into broader global contexts.

lower (0.192), suggesting broader support beyond the domestic domain.

The KL divergences show a consistent pattern: $\mathrm{KL}(\text{recruit-jp}||\text{DEJIMA}) = 6.03$ (lowest among all pairs), followed by Japanese Visual Genome (≈12.2), STAIR (≈12.3), MS COCO (≈12.8), and GQA (≈14.2). In the reverse direction, $\mathrm{KL}(\text{DEJIMA}||\text{recruit-jp}) = 16.4$, indicating that DEJIMA includes additional regions not present in recruit-jp.

Figure 3 visualizes these relationships. DEJIMA exhibits the widest and most continuous distribution among the compared datasets, covering the full domestic (Japanese) domain while extending smoothly toward broader, globally oriented image clusters. These results collectively demonstrate that DEJIMA provides both dense coverage of Japanese visual concepts and expanded representational diversity beyond existing Japanese datasets.

Human Evaluation. We also conducted human evaluations to assess the quality of the caption and VQA datasets. Pairwise comparisons were performed on 150 random samples per dataset. Annotations were collected via crowdsourcing (80 workers total), comparing two datasets side-by-side with randomized order. To ensure quality, we inserted a control item for each worker and excluded inconsistent annotations (22.5% of workers); final sample sizes n are reported in each table.

Before the evaluation, all workers were presented with detailed descriptions of the evaluation metrics. For caption evaluation, the metrics included *Japaneseness of image*, which measures how strongly the image reflects Japanese culture or scenery; *Japaneseness of text*, which captures the presence of Japan-specific expressions or contexts in the caption; *Naturalness of text*, which assesses the fluency and naturalness of the Japanese language; *Image–text consistency*, which evaluates factual alignment between the caption and visible evidence such as color, number, or position; *Coverage*, which measures how well the caption mentions major objects or regions in the image; and *Expressiveness*, which evaluates the richness and vividness of description beyond simple enumeration. For VQA evaluation, the metrics included *Japaneseness of image*; *Japaneseness of text* and *Naturalness of text* for both question and answer; *Q–A relevance*, which measures whether the answer appropriately addresses the question's intent; *Q–image consistency*, which assesses whether the question is grounded in the image; and *Answer correctness*, which checks whether the answer is factually correct given the image content.

Two-sided binomial tests were conducted to determine whether observed preference rates significantly deviated from chance (50%). Statistical significance was evaluated at the 5% level, and the Holm–Bonferroni correction was applied across metrics within each comparison. In the following tables, * indicates significance at the 5% level.

Tables 2 and 3 show the results of pairwise human evaluations comparing our datasets with existing ones. Across most metrics, DEJIMA-Cap-All and DEJIMA-VQA-All were consistently preferred, with statistically significant differences in nearly all linguistic and cultural dimensions.

For captions, compared with existing datasets such as MS COCO Translation and STAIR Captions, DEJIMA-Cap-All achieved clearly higher preference rates for *Japaneseness of image* and *Japaneseness of text*. This improvement stems from the fact that existing datasets are either translations of English captions or annotations on English-centric image resources, while DEJIMA collects images from Japanese websites, naturally reflecting local culture and scenery. DEJIMA also outperformed these baselines in *Expressiveness*, showing that combining alt-text and detection enables more multi-perspective and vivid descriptions. However, *image–text consistency* was lower than in human-annotated datasets, reflecting the continued advantage of manual alignment in grounding accuracy. In *Naturalness of text*, DEJIMA surpassed MS COCO Translation, likely because Japanese LLM generation yields more fluent expressions than machine translation or crowd-sourced captions. Ablation comparisons confirm that excluding either alt-text or detection reduces overall quality—the two inputs complement each other, balancing contextual richness and grounding precision.

For VQA, DEJIMA-VQA-All also outperformed translation-based datasets (GQA Translation) and human-annotated ones (Japanese Visual Genome) in *Japaneseness of image*, *Japaneseness of text*, and *Naturalness of text*. Again, this advantage arises from DEJIMA's use of Japanese web imagery, which better captures domestic cultural elements. Compared with Japanese Visual Genome, DEJIMA-VQA-All performed lower in *Q–A relevance* and *Answer correctness*, showing that human annotation remains more precise for factual grounding. In contrast, against GQA Translation, DEJIMA showed no significant difference in *Q–A relevance* or *Answer correctness*, and no gap in *Q–image consistency*, demonstrating that high-quality Japanese VQA data can be constructed automatically when grounded by detection. Ablation results again show that alt-text and detection each compensate for the other's weaknesses.

4. Experiments

To examine how much the constructed dataset contributes to improving VLM performance, we trained Vision-and-Language Models (VLMs) using our dataset.

4.1. Setup

Following the two-stage training scheme of LLaVA (Liu et al., 2023), we use a `siglip-so400m-patch14-384`[9] vision encoder and `llm-jp/llm-jp-3-7.2b`[10] as the language model. In the first stage, we train the model to align images and captions using the caption datasets, and in the second stage, we fine-tune it on VQA data to enhance multimodal reasoning and grounding abilities. We evaluate the resulting models on two representative Japanese multimodal benchmarks: JA-VLM-Bench-In-the-Wild (aki, 2024) and Heron-bench (Inoue et al., 2024).

Benchmarks. **JA-VLM-Bench-In-the-Wild** is a benchmark developed by Sakana AI to evaluate Japanese VLMs in culturally grounded, open-domain scenarios. It measures the model's ability to understand and describe Japanese cultural and social contexts through diverse real-world images. The dataset is constructed through a semi-automatic process: GPT-4V generates questions and answers for 42 curated Japanese images, and human annotators filter and refine them. Thus, it primarily tests linguistic fluency, cultural knowledge, and contextual relevance in free-form Japanese.

Heron-bench was created by Turing Corporation and focuses on more fine-grained multimodal reasoning. It contains 21 carefully selected Japanese images across seven domains (e.g., anime, art, food, culture, landscape, landmark, transportation), with 102 manually written questions by researchers. It emphasizes precise perception, factual correctness, and reasoning grounded in Japanese context, making it a stricter test for factual alignment and multimodal understanding.

4.2. Benchmark Results

Table 4 reports model performance on both benchmarks (LLM-as-a-Judge scores; higher is better).

The divergent results on the two benchmarks highlight a critical relationship between the granularity of the training data and the specific demands of the evaluation task.

On JA-VLM-Bench-In-the-Wild, the top performance of the DEJIMA-Cap-Simple & DEJIMA-VQA-Refined model is instructive. This benchmark primarily assesses a model's ability to handle general-purpose queries about widely recognizable Japanese subjects, such as traditional clothing (e.g., kimono) or major landmarks (e.g., Tokyo

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

[10]https://huggingface.co/llm-jp/llm-jp-3-7.2b

Compared Dataset	n	Japaneseness of image	Japaneseness of text	Naturalness of text	Image-text consistency	Coverage	Expressiveness
MS COCO Translation	105	82.86*	87.62*	86.67*	20.00*	39.05*	92.38*
STAIR Captions	135	74.07*	77.78*	62.22*	20.74*	43.70	68.89*
DEJIMA-Cap-Refined	105	–	86.67*	64.76*	52.38	61.90*	91.43*
DEJIMA-Cap-Detection	135	–	76.30*	65.93*	62.96*	70.37*	81.48*

Table 2: Caption: Pairwise preference of **DEJIMA-Cap-All** vs. baselines. * indicates significance at 5%.

Compared Dataset	n	Japaneseness of image	Japaneseness of text	Naturalness of text	Q-A relevance	Q-Image consistency	Answer correctness
GQA Translation	105	91.43*	92.38*	89.52*	41.90	51.43	41.90
Japanese Visual Genome	135	92.59*	87.41*	78.52*	31.85*	38.52*	34.81*
DEJIMA-VQA-Refined	90	–	76.67*	74.44*	57.78	56.67	57.78
DEJIMA-VQA-Detection	120	–	79.17*	72.50*	61.67*	65.83*	63.33*

Table 3: VQA: Pairwise preference of **DEJIMA-VQA-All** vs. baselines. * indicates significance at 5%.

Tower). For this purpose, the concise, natural language of raw alt-texts in DEJIMA-Cap-Simple appears sufficient to build a strong foundational image-text alignment. The highly detailed and lengthy captions of DEJIMA-Cap-All, while factually rich, may have inadvertently biased the model toward object-level specifics, potentially hindering its ability to generate the holistic, general-scene descriptions that this benchmark rewards.

In sharp contrast, Heron-bench demands a much deeper and more specialized understanding of Japanese culture. Its questions probe niche topics requiring fine-grained reasoning, such as knowledge of specific film directors, the context of historical documents, or details about less-famous regions. The decisive victory of the DEJIMA-Cap-All & DEJIMA-VQA-All model here is a testament to our pipeline's effectiveness. We attribute this success to the synergistic combination of broad contextual cues from alt-text and precise, visually-grounded facts from object detection. This rich, multi-faceted data is essential for equipping the model with the detailed knowledge required to answer such expert-level questions, a task for which simpler descriptions are inadequate.

Ultimately, while strong performance on JA-VLM-Bench confirms robust general capabilities, the success on Heron-bench is particularly significant. It demonstrates that our All pipeline enables a model to progress beyond superficial recognition of cultural icons towards a genuinely deep and nuanced understanding of Japan. This capacity for grounded, in-depth reasoning on specialized cultural knowledge represents a more challenging and valuable achievement, showcasing the true potential of the DEJIMA dataset.

4.3. Human Evaluation of VLM Outputs

In addition to automatic evaluations, we conducted human evaluations of VLM outputs on Heron-bench to complement the benchmark results. All outputs for 102 questions were manually assessed following the same protocol as the dataset-level evaluation described in Section 3.2. Each output was evaluated by one Japanese-speaking crowd worker who compared two model outputs side by side (e.g., between models trained on different dataset pipelines), with randomized display order to avoid positional bias. A total of 408 samples (102 questions × 4 model comparisons) were evaluated by 28 workers in total. Each worker was assigned 15 samples, except for four workers who were assigned 13 samples, and one identical control item was inserted for quality checking. Annotations from workers whose responses to the control were inconsistent were excluded, accounting for approximately 21.4% of workers. The resulting effective sample size n for each comparison is reported in Table 5.

Before the evaluation, all workers were presented with the definitions of the evaluation criteria for pairwise model comparison. The four metrics were as follows: *Japaneseness of text*, which measures the presence of Japanese-specific expressions and cultural contexts; *Naturalness of text*, which evaluates the fluency and naturalness of the Japanese language; *Q–A relevance*, which measures whether the answer appropriately addresses the question's intent; and *Answer correctness*, which assesses factual accuracy given the image and question.

As shown in Table 5, the model trained on the **DEJIMA-Cap-All & DEJIMA-VQA-All** pipeline achieved the highest preference in nearly all metrics, including *Japaneseness of text*, *Naturalness of text*, and notably *Answer correctness*. This indicates that the linguistic and cultural richness of DEJIMA datasets translates directly into improved downstream VLM performance.

When compared to the MS COCO Translation & Japanese Visual Genome pipeline, the All-pipeline outputs were preferred overwhelmingly

Stage 1	Stage 2	JA-VLM-Bench-In-the-Wild (LLM-as-a-Judge ↑)	Heron-bench (LLM-as-a-Judge ↑)
STAIR Captions	Japanese Visual Genome	3.04	31.57
STAIR Captions	GQA Translation	1.58	20.54
MS COCO Translation	Japanese Visual Genome	2.88	33.94
DEJIMA-Cap-Simple	DEJIMA-VQA-Refined	**3.12**	44.82
DEJIMA-Cap-Refined	DEJIMA-VQA-Refined	1.96	15.89
DEJIMA-Cap-Detection	DEJIMA-VQA-Detection	1.36	21.95
DEJIMA-Cap-All	DEJIMA-VQA-All	2.48	**52.26**

Table 4: Evaluation of trained models on Japanese V&L benchmarks.

(97.5% for Japaneseness, 92.5% for Naturalness), confirming the strong impact of using Japanese-grounded data. Against ablated versions, alt-only or detection-only pipelines showed lower scores, while the combined pipeline consistently achieved the best balance of fluency and factuality. Notably, the Detection-based variant performed well in *Q–A relevance* and *Answer correctness*, while the Refined variant excelled in Naturalness, again supporting the complementary relationship between the two information sources.

4.4. VLM Output Example

Figure 4 presents a qualitative example comparing generations across training pipelines for the question "What is the name of the mountain visible in the background?" The correct answer is "Mount Yotei." We show outputs from five representative pipelines. As illustrated, baseline and ablated variants tend to either misidentify the mountain (e.g., producing a popular landmark) or remain generic despite object detections. In contrast, the All pipeline correctly names *Mount Yotei* and provides a visually grounded rationale that refers to salient scene attributes.

5. Conclusion

We introduced DEJIMA, a detection-guided and LLM-refined pipeline for constructing culturally grounded Japanese Vision-and-Language resources at scale. Our approach yields two datasets—DEJIMA-Cap and DEJIMA-VQA—each with $\sim$3.88M entries, surpassing existing Japanese resources by more than an order of magnitude while preserving linguistic naturalness and visual grounding. Human evaluations showed consistent gains in *Japaneseness* and *naturalness* over translation- and manual-annotation baselines, and downstream experiments demonstrated strong improvements on Japanese multimodal benchmarks (JA-VLM-Bench-In-the-Wild and Heron-bench).

6. Ethics Statement

This work involves the large-scale collection and automatic refinement of Japanese image–text data from the public web and the use of paid human evaluations. We describe the ethical considerations across data sourcing, licensing and redistribution, privacy, annotator welfare, cultural sensitivity, safety, transparency, and intended use.

Data sourcing and licensing. We collect candidate image–alt-text pairs from Common Crawl and retain only Japanese pages after language and adult filtering. Web content may be copyrighted or subject to site-specific terms. To respect rights of content owners, we do not redistribute images; we release only non-image artifacts (e.g., URLs, text, derived metadata, and model-generated text). Users of the resource must download images directly from origin servers and abide by the original licenses and terms of service. We provide a take-down channel on the project page; we will promptly remove or update entries upon substantiated requests from rights holders.

Privacy and personal data. Public web images can contain people or private locations. Our pipeline does not attempt to identify individuals, infer sensitive attributes, or link content to specific persons. We apply content filters to reduce exposure to adult content and other unsafe material and exclude pages that are likely to contain personally identifying information in the text layer. We do not release any additional annotations about identities or sensitive attributes. We ask downstream users to avoid tasks such as face recognition, re-identification, or surveillance, and to comply with applicable privacy laws in their jurisdictions.

Safety filtering and hallucination control. To mitigate propagation of harmful or unverifiable claims during LLM refinement, generation is conditioned on detected visual evidence and guided by constraints that discourage unsupported statements. In addition, we performed text-based safety filtering at the alt-text stage to remove entries con-

Compared Pipeline	n	Japaneseness of text	Naturalness of text	Q-A relevance	Answer correctness
MS COCO Translation & Japanese Visual Genome	75	97.33^{*}	92.00^{*}	61.33	77.33^{*}
DEJIMA-Cap-Simple & DEJIMA-VQA-Refined	88	79.55^{*}	79.55^{*}	53.41	50.00
DEJIMA-Cap-Detection & DEJIMA-VQA-Detection	73	89.04^{*}	87.67^{*}	84.93^{*}	90.41^{*}
DEJIMA-Cap-Refined & DEJIMA-VQA-Refined	88	63.64^{*}	97.73^{*}	77.27^{*}	70.45^{*}

Table 5: VLM output: pairwise preference vs. **DEJIMA-Cap-All & DEJIMA-VQA-All** pipeline. * indicates significance at 5%.

Figure 4: VLM Output Example for the question “What is the name of the mountain visible in the background?” We compare outputs from five pipelines against the correct answer (Mount Yotei). The All pipeline (alt-text + detection) yields correct identification with natural phrasing and explicit visual grounding, while translation-based and ablated variants either hallucinate specific locations or stay generic despite object detections. Extremely long or repetitive model outputs have been truncated for clarity.

taining adult, violent, or otherwise harmful content before downstream processing.

Cultural sensitivity and bias. The dataset emphasizes Japanese content. While this focus is central to our goal of culturally grounded Japanese V&L resources, it can also amplify culture-specific biases or stereotypes. We explicitly caution against using the resource to make normative claims about Japanese culture and encourage researchers to conduct bias and fairness analyses for their particular use cases.

Annotator welfare. Human evaluations were conducted with consenting, Japanese-speaking crowd workers. Instructions detailed the task, quality expectations, and the right to withdraw. We compensated workers at or above local minimum-wage equivalents for the expected time per assignment, paid for approved tasks, and minimized exposure to potentially sensitive content through pre-filtering. No personally identifying information about workers is collected or released.

Environmental considerations. Our pipeline is designed to be compute-efficient by leveraging detection-guided prompts and batching. We report approximate compute used for dataset construction and model training to support assessments of environmental impact and reproducibility. We encourage downstream users to consider energy-efficient settings and to reuse our released artifacts to avoid redundant computation.

Intended use and misuse. The resource is intended for research and development of multimodal systems that generate or understand Japanese text and imagery. It should not be used for surveillance, identity inference, discriminatory decision-making, medical or legal advice, or any context where errors could cause harm without appropriate human oversight. We release documentation (datasheet/model card), licensing guidance, and known limitations to support responsible deployment.

Compliance. This study uses publicly available web data and non-identifiable human annotations. No institutional review board approval was required under our local guidelines; we nonetheless adhered to principles of informed consent, data minimization, and respect for rights holders and annotators. Downstream users are responsible for ensuring compliance with local laws and the original content licenses when using the released artifacts.

7. Limitations

While DEJIMA presents a scalable and culturally grounded approach to constructing large-scale Japanese Vision-and-Language datasets, several limitations remain.

First, although we applied extensive filtering and grounding mechanisms, web-derived alt-text may still contain noise, inaccuracies, or cultural bias. Image–text consistency, in particular, is not always guaranteed and can be weaker than human annotation in some cases. Detection-guided grounding reduces unverifiable or hallucinated statements, but challenges persist for long-tail cultural references and abstract descriptions.

Second, our concept of “Japaneseness” was intentionally left open-ended and was evaluated based on annotators’ subjective judgment rather than a predefined operational definition. A more systematic or multidimensional framework for measuring cultural grounding would enhance reproducibility and interpretability.

Third, the current work focuses exclusively on Japanese data. Some pipeline components, notably the `HojiChar` filtering library used for language identification and adult-content filtering, currently support only Japanese (and English). However, the overall pipeline architecture is language-agnostic and could be applied to other languages by substituting these language-specific modules with appropriate alternatives. Extending the approach to multilingual settings remains an important direction for future work.

Finally, while we constructed datasets of approximately 3.88 million pairs, the pipeline is inherently scalable. Running additional iterations with expanded web sources could further enlarge the dataset. Future work should also investigate efficient quality control at even larger scales and assess environmental and computational impacts.

8. Data and Code Availability

To promote transparency and reproducibility, we have released the full dataset construction pipeline, metadata, and evaluation scripts on our project page.[11]

Specifically, we provide: (1) source code for the data collection, filtering, deduplication, detection, and LLM-based refinement stages; (2) JSON-formatted metadata containing image URLs, alt-texts and LLM-generated texts; and (3) configuration files and prompts used for generation and evaluation.

Following the licensing terms of the original web data, we do not redistribute image files themselves. Instead, we release URL lists and derived text annotations so that users can reproduce the dataset by downloading images directly from their original sources, respecting each website’s terms of service.

All released components (code, metadata, and annotations) are available under a permissive open license allowing both research and commercial use, except where restricted by upstream content licenses. We also provide detailed documentation (datasheet) describing the data sources, filtering criteria, model configurations, and ethical considerations to ensure responsible use and facilitate future extensions.

Acknowledgements

This work was partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

9. Bibliographical References

2017. Stair captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

2024. Evolutionary optimization of model merging recipes.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training

[11] https://mil-tokyo.github.io/DEJIMA-dataset/

with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575.

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. If you use this software, please cite it as below.

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. *arXiv preprint arXiv:2404.07824*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

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: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024b. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *CoRR*.

Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang Wei W Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459.

Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. 2024. Constructing multimodal datasets from scratch for rapid development of a japanese visual language model. *arXiv preprint arXiv:2410.22736*.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732.

A. Dataset Examples

This appendix presents four examples from the DEJIMA dataset. Each example shows the source image along with DEJIMA-Cap-All (caption) and DEJIMA-VQA-All (question–answer pair), with English translations.

(a) Byodo-in Phoenix Hall

Caption (DEJIMA-Cap-All):

朱色の楼門が鮮やかに輝く平等院鳳凰堂。国宝の建築物が中央に立ち、屋根の金箔が陽光を反射してきらめいている。平安時代の風格を今に伝える荘厳な姿が、緑豊かな庭園に映える。

Byodo-in Phoenix Hall with its vermilion gate glowing vividly. The National Treasure stands at center, its gilded roof glittering in the sunlight. Its majestic Heian-era form is set against a lush garden.

VQA (DEJIMA-VQA-All):

Q: 平等院鳳凰堂はどのような時代に建造されましたか？

Q: In what era was Byodo-in Phoenix Hall constructed?

A: 平等院鳳凰堂は平安時代後期（11世紀前半）に建立され、当時の建築様式や装飾が現在まで保存されています。

A: It was built in the late Heian period (early 11th century), and its architectural style and decorations have been preserved to the present day.

(b) Ukiyo-e by Hiroshige

Caption (DEJIMA-Cap-All):

歌川広重の「袋井 出茶屋ノ図」で、茶屋前で休む男性や枝ぶりの立派な老松が描かれる。画面右下の絵葉書風枠内に作品が収められ、人物たちの和服姿が江戸時代の雰囲気を漂わせる。

Hiroshige's 'Fukuroi: Dejima Teahouse' depicts a man resting in front of a teahouse and a magnificent old pine. The work is framed in a postcard-style border, with figures in kimono evoking the Edo period.

VQA (DEJIMA-VQA-All):

Q: 絵の中の人物は何人いますか？

Q: How many figures are in the painting?

A: 浮世絵『袋井 出茶屋ノ図』には、男性1人と別の人物2人が描かれています。男性は左下に立ち、他の2人は中央より右側に配置されています。

A: In the ukiyo-e, one man and two other figures are depicted. The man stands at the lower left, and the other two are positioned to the right of center.

(c) Shirakawa-go in Winter

Caption (DEJIMA-Cap-All):

白川郷の冬景色で、茅葺き屋根の民家が雪に覆われた森の中に佇んでいる。白銀の世界に黒い樹木がアクセントとなり、遠景には積雪した山並みが広がる。

A winter scene of Shirakawa-go, where a thatched-roof farmhouse stands in a snow-covered forest. Dark trees contrast against the silvery-white world, with snow-capped mountains in the background.

VQA (DEJIMA-VQA-All):

Q: 白川郷の雪景色の中で、最も目立つ建造物はどのような特徴がありますか？

Q: What are the features of the most prominent structure in this snowy scene?

A: 画面中央に位置する白川郷の民家は、伝統的な合掌造りの外観を持ち、屋根の雪に覆われた姿が特徴的です。

A: The farmhouse at center has a traditional gassho-zukuri appearance, characterized by its snow-covered roof.

(d) Shinkansen and Mount Fuji

Caption (DEJIMA-Cap-All):

新幹線が駅の跨線橋の下を通過し、背景には白銀の富士山がそびえ立つ。跨線橋は右半分に横たわり、新幹線は右斜め前方へ高速で走行中。

The Shinkansen passes beneath a station overpass, with snow-white Mount Fuji towering in the background. The overpass extends across the right half, while the Shinkansen speeds diagonally to the right.

VQA (DEJIMA-VQA-All):

Q: 写真に写っている交通機関は何ですか？

Q: What mode of transportation is shown in the photo?

A: 画面中央に新幹線が写っており、その速度感あふれるデザインが特徴的です。背景には人道橋が前景に配置されており、鉄道と富士山を連想させる構図となっています。

A: A Shinkansen is shown at center with its dynamic design. A pedestrian bridge in the foreground creates a composition evoking railways and Mount Fuji.

Figure 5: Four examples from the DEJIMA dataset. Each cell shows the source image, the DEJIMA-Cap-All caption, and a DEJIMA-VQA-All question–answer pair. English translations are shown.