Evaluating Multimodal Large Language Models on Vertically Written Japanese Text

Keito Sasagawa[*‡], **Shuhei Kurita**[†‡], **Daisuke Kawahara**[*‡]
[*]Waseda University, [†]NII, [‡]NII LLMC
Tokyo, Japan
kate@fuji.waseda.jp, skurita@nii.ac.jp, dkw@waseda.jp

Abstract

Multimodal Large Language Models (MLLMs) have seen rapid advances in recent years and are now being applied to visual document understanding tasks. They are expected to process a wide range of document images across languages, including Japanese. Understanding documents from images requires models to read what are written in them. Since some Japanese documents are written vertically, support for vertical writing is essential. However, research specifically focused on vertically written Japanese text remains limited. In this study, we evaluate the reading capability of existing MLLMs on vertically written Japanese text. First, we generate a synthetic Japanese OCR dataset by rendering Japanese texts into images, and use it for both model fine-tuning and evaluation. This dataset includes Japanese text in both horizontal and vertical writing. We also create an evaluation dataset sourced from the real-world document images containing vertically written Japanese text. Using these datasets, we demonstrate that the existing MLLMs perform worse on vertically written Japanese text than on horizontally written Japanese text. Furthermore, we show that training MLLMs on our synthesized Japanese OCR dataset results in improving the performance of models that previously could not handle vertical writing. The datasets and code are publicly available (https://github.com/llm-jp/eval_vertical_ja).

Keywords: Multimodal Datasets, Multimodal LLM, Japanese OCR

1. Introduction

Research on multimodal large language models (MLLMs) has advanced rapidly (Bai et al., 2025; Zhu et al., 2025; Team et al., 2025; OpenAI, 2025b), and these models are now applied to a wide range of multimodal tasks. Document image question answering is one such task, where a model answers questions about given document images, and various datasets have been proposed (Mathew et al., 2021, 2022; Tanaka et al., 2021, 2023; Tito et al., 2023; Onami et al., 2024). To solve such tasks, MLLMs are required to read the text within the input document image. A variety of multimodal datasets have been developed for English, and current MLLMs demonstrate high accuracy in reading English text on English-language benchmarks.

Japanese documents often include vertically written text, necessitating explicit support for vertical writing. First, we explain the difference in reading order between horizontally and vertically written Japanese text. Figure 1 shows the reading order for horizontally and vertically written Japanese text. In horizontal Japanese writing, as in English, characters are read from left to right and lines progress from top to bottom. In vertical writing, characters are read from top to bottom and lines progress from right to left. It is essential to examine whether models can read vertically written text as well as horizontally written text. However, most Optical Character Recognition (OCR) benchmarks for MLLMs are primarily concerned with horizontally written text, often neglecting the evaluation of vertically written Japanese text.

To address this gap, we evaluate the OCR capability of MLLMs for reading vertically written Japanese text. We first build a dataset of synthetic images rendered with Japanese text for training and evaluation. This dataset comprises images containing text in both horizontal and vertical writing and exhibiting multi-column layouts (1-4 columns). We also construct an OCR evaluation dataset from real-world PDF pages that contain vertically written Japanese text.

In our experiments, we evaluate multiple open and closed MLLMs, as well as variants fine-tuned on the synthetic image dataset, using our constructed test dataset. We show that existing MLLMs read vertically written Japanese text less accurately than horizontally written text. We further demonstrate that fine-tuning on the synthetic dataset improves models that initially struggle with vertically written Japanese text.

The contributions of this study are as follows:

1. We evaluate the Japanese OCR capabilities of MLLMs and quantitatively demonstrate that current MLLMs perform worse with vertical writing than with horizontal writing.

2. We release scripts for synthesizing images of multi-column text in both horizontal and vertical writing styles, along with a dataset of the

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9065–9081
11–16 May 2026.

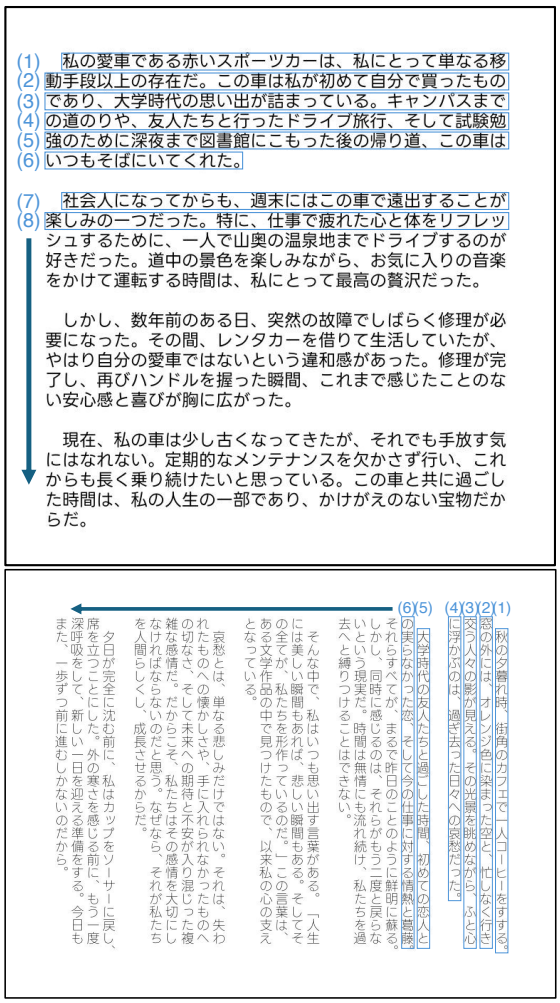


Figure 1: Examples of reading order for Japanese documents (Top: Horizontal writing, Bottom: Vertical writing). The blue numbers attached to each line indicate the order in which the text on that line should be read. In horizontal writing, similar to English documents, characters in each line are read from left to right, and lines are read from top to bottom. In vertical writing, characters in each line are read from top to bottom, and lines are read from right to left.

synthesized images (**JSSODa**). This dataset enables training MLLMs on vertically written text and evaluation.

3. We also release an OCR dataset containing images of vertically written Japanese text from the real-world PDF pages (**VJRODa**). This dataset enables the evaluation of MLLM's OCR capabilities on real-world document images containing vertically written Japanese text.

The datasets and code are publicly available (https://github.com/llm-jp/eval_vertical_ja).

2. Related Work

2.1. Multimodal Large Language Models

Recent MLLM (Liu et al., 2023, 2024; Li et al., 2025) architectures employ a structure that connects the vision encoder and LLM via a projection layer. It converts the input image into features using a vision encoder, feeds them into a projection layer to transform them into image tokens that the LLM can handle, and then inputs them into the LLM together with text tokens. MLLMs are trained through multimodal instruction tuning, enabling them to handle a wide range of tasks such as document image question answering.

Extensive research has been conducted on multimodal models capable of understanding document images (Xu et al., 2020, 2021; Huang et al., 2022; Ye et al., 2023a,b; Hu et al., 2024, 2025; Dong et al., 2024; Luo et al., 2024). Among these, notable MLLMs that excel at understanding visual Japanese texts and Japanese document images include Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and Gemma 3 (Team et al., 2025). Qwen2.5-VL introduces dynamic resolution processing, enabling native handling of images at various resolutions. InternVL3 improves performance by employing techniques such as a pixel unshuffle operation, a dynamic resolution strategy that divides images into multiple tiles, and the variable visual position encoding. Gemma 3 also handles images of various resolutions by dividing the input image into multiple non-overlapping crops of the same size when necessary and feeding them into the vision encoder.

2.2. Japanese OCR Datasets and Vertical Text Datasets

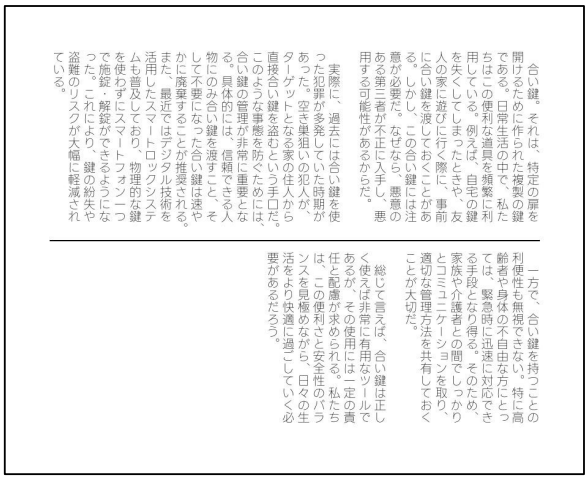
NDLOCR (NDL Lab, 2021) is an OCR program developed by the National Diet Library, Japan (NDL), using materials held in the NDL Digital Collections. For the evaluation of this program, they used in-house evaluation dataset, which is not publicly available. Kindai-OCR (Le et al., 2019) is an OCR system for modern Japanese magazines based on an attention-based encoder-decoder architecture. The document images targeted by this system include vertically written Japanese text, but they are somewhat old and not recent documents.

SVTD and VTD142 (Choi et al., 2019) are datasets for scene text recognition focusing on vertical text. SVTD is a synthetic dataset of vertical text images generated following the procedure of Cheng et al. (2018), whereas VTD142 is a dataset comprising real-world vertical text collected from web pages. In Orihashi et al. (2022), a synthetic dataset of horizontal and vertical text was constructed based on the synthesis method de-

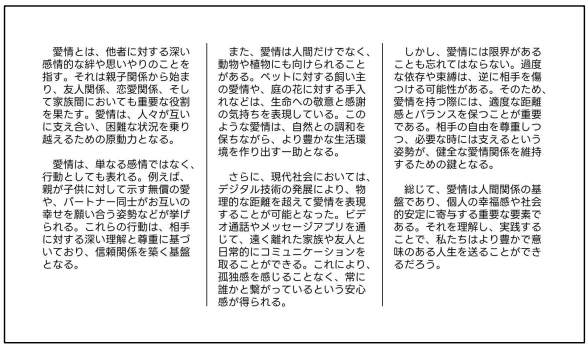
scribed in Gupta et al. (2016) and used for model training. They also employed the Japanese horizontal and vertical text subsets from the ICDAR 2019 Multi-lingual Scene Text Detection and Recognition competition dataset (Nayef et al., 2019) in their experiments. As the images in these datasets primarily comprise word-level text, sentence-level evaluation is not feasible.

SynthDoG (Kim et al., 2022) is a dataset that comprises images synthesized by compositing document textures onto background images and rendering text. Although the images in this dataset may partially contain vertical Japanese texts, the proportion is small; moreover, they are unrealistic document images that would not appear in real-world settings, making them unsuitable for an evaluation dataset. CC-OCR (Yang et al., 2024) is a dataset for evaluating MLLM's OCR capabilities and includes Japanese OCR data. This dataset includes vertical Japanese text, but it is text within natural images and only exists at the word-level, making sentence-level evaluation impossible. MangaOCR (Aizawa et al., 2020; Matsui et al., 2017; Baek et al., 2025) is an OCR dataset focused on textual elements in manga, such as dialogue and sound effects. It includes vertical Japanese text; however, its images are restricted to manga pages.

In this work, we construct OCR datasets containing contemporary Japanese text in vertical writing at the sentence-level and evaluate MLLMs.



(a) vertical, 2-columns



(b) horizontal, 3-columns

Figure 2: Example images from JSSODa.

3. Dataset Construction

To evaluate the OCR capabilities of MLLMs on vertically written Japanese text, we construct two types of datasets. One is a synthetic dataset that can be generated at scale and can also be used for training MLLMs. The other is derived from real-world documents, enabling evaluation under realistic conditions. These datasets consist of pairs of images and texts written within them. The first one is **JSSODa** (**J**apanese **S**imple **S**ynthetic **O**CR **Da**taset), which is constructed by rendering Japanese text generated by an LLM into images. The other one is **VJRODa** (**V**ertical **J**apanese **R**eal-world **O**CR **Da**taset), which consists of images containing vertical Japanese texts sourced from the real-world PDF pages.

3.1. Construction of JSSODa

To generate large-scale evaluation data, we synthesize images from Japanese texts. During pretraining, MLLMs are trained using vast amounts of text data, including web text. If we use web text for our dataset, it may have been used for model training, potentially preventing proper evaluation of the models. We instead use Japanese text generated by an LLM.

Text Generation Specifically, we input Japanese nouns into an LLM and generated sentences about them. We used nouns from the JUMAN dictionary[1] for Japanese nouns, and llm-jp-3.1-instruct4 (LLM-jp, 2024) for the LLM. We removed generated sentences with fewer than 100 characters or more than 3,000 characters.

Image Synthesis Next, we synthesized images based on the generated Japanese texts. We generated a total of 8 layout types of images for both vertical and horizontal writing, each with 1 to 4 columns layout. Horizontal writing is the same as English text: characters are drawn from left to right, and lines are drawn from top to bottom. In multi-column layouts, each column is rendered in order from left to right. In vertical writing, characters are drawn from top to bottom, and lines are drawn from right to left. In multi-column layouts, each column is rendered in order from top to bottom. Based on the above character drawing order, we synthesized the images by drawing each character of the gener-

[1] https://github.com/ku-nlp/JumanDIC

ated sentences into images using the Pillow library[2]. We collected Japanese fonts from Google Fonts[3] and free-fonts.jp[4], and used them for synthesizing images. The number of font files collected is approximately 200. For the text used, the top 25% in length was set to four columns, the next 25%-50% to three columns, the next 50%-75% to two columns, and the top 75%-100% to one column. The total number of images reached 22,493. Figure 2 shows examples of the generated images.

We split the constructed dataset into train:val:test at a ratio of 8:1:1 while maintaining the same proportion of the 8 layout types.

3.2. Construction of VJRODa

We extract pages containing vertically written Japanese text from Japanese PDF documents and annotate each page image with a transcription of its content in Japanese reading order.

First, we collected Japanese PDFs from NDL WARP project[5], and converted each page to an image. Using these images, we created the dataset following the procedure below. (1) Filter images that do not contain vertically written Japanese text. (2) Transcribe the text in the images.

(1) Filter images that do not contain vertically written Japanese text We filtered out images that do not contain vertically written Japanese text by using a fast projection profile method (Akiyama and Hagita, 1988), followed by a slower, more detailed vertically written text detection based on character bounding boxes from Tesseract OCR (Smith, 2007). These processes enable the rapid, automatic filtering of images that do not contain vertically written text.

First, each image was converted to grayscale, inverted to obtain black text on a white background, and binarized at a fixed threshold of 128. We then calculated the total number of binarized 1-valued pixels per row and per column. For each of them, we calculated the coefficient of variation (CV, standard deviation divided by mean) to examine the degree of variation. If the column-wise CV is larger, it is treated as a candidate for containing vertically written text.

For images considered to contain vertically written text, we used Tesseract OCR to extract the character-level bounding boxes within them. We merged the character-level bounding boxes in both the horizontal and vertical directions. If the number of vertical merges exceeds the number of horizontal merges, we determine that the image contains vertically written text.

Figure 3: An example image from VJRODa. The blue number above each line indicates the order in which the text on that line should be read. Characters in each line are read from top to bottom, and lines are read from right to left. Each column is read from top to bottom. (https://warp.ndl.go.jp/info:ndljp/pid/11712522/www.vill.kariwa.niigata.jp/open/info/000000001_0000000609.pdf, page 8, Personal information has been masked.)

(2) Transcribe the text in the images We extracted texts from the collected images using PyMuPDF[6] for pages with embedded text and PaddleOCR (Cui et al., 2025b) for those without, such as scanned images. Then, we manually corrected the extracted texts to Japanese reading order. We also corrected OCR errors at the same time.

Figure 3 shows an example image in this dataset.

[2]https://github.com/python-pillow/Pillow
[3]https://fonts.google.com/
[4]https://free-fonts.jp/
[5]https://warp.ndl.go.jp/
[6]https://github.com/pymupdf/PyMuPDF

	JSSODa (train)	JSSODa (test)	VJRODa
# Img	17,991	2,256	100
avg. Char	706.19	705.63	1143.91
min. Char	119	192	85
max. Char	1704	1399	3386

Table 1: Statistics of our datasets.

3.3. Statistics of Datasets

Table 1 shows the statistics of our datasets.

4. Experiments

4.1. Experimental Setup

4.1.1. Fine-tuning Models

We trained open MLLMs using the JSSODa train set. As models for training, we used three models capable of reading Japanese text: "Qwen2.5-VL-7B-Instruct", "InternVL3-8B-hf", and "Gemma 3 12b IT". These models consist of a vision encoder, a multimodal projector, and an LLM. In this study, we tuned the parameters of all modules. The number of images used for training was 18k, and we trained the models for one epoch. We set the batch size to 32. We used AdamW (Loshchilov and Hutter, 2019) as the optimizer and set the learning rate to 2e-5.

4.1.2. Evaluation

We evaluated multiple open MLLMs and closed models on the JSSODa test set and VJRODa. For open MLLMs, we used Qwen2.5-VL models with 7B and 32B parameters, InternVL3 models with 8B and 38B parameters, and Gemma 3 models with 12B and 27B parameters. We also evaluated three models fine-tuned on the JSSODa train set, as described in Section 4.1.1. During inference, texts were generated using greedy decoding. As closed models, we used GPT-4.1 (OpenAI, 2025a) and GPT-5 (OpenAI, 2025b). For GPT-4.1, we set the temperature to 0. For GPT-5, we set "reasoning_effort" to "minimal". For all other parameters, we used the default settings. We set max new tokens to 1024 for JSSODa test set, and 3072 for VJRODa. Additionally, we used the same user prompt that was used during the model fine-tuning.

4.1.3. Evaluation Metric

For evaluation metrics, we used Character Error Rate (CER) and BLEU (Papineni et al., 2002). Before calculating the scores, we applied Unicode NFKC normalization and whitespace removal to both the reference and predicted texts.

CER CER is defined by the following formula:

$$\text{CER} = \frac{\text{EditDistance}(pred, ref)}{|ref|} \times 100,$$

where $pred$ is the model's predicted text, and ref is the correct text. $\text{EditDistance}(p, r)$ is the edit distance between strings p and r. Lower CER values are generally considered indicative of better performance.

BLEU We used SacreBLEU (Post, 2018) for calculating BLEU. The reference texts and predicted texts were split into character units.

MLLMs sometimes generate the same tokens repeatedly, up to max new tokens. When this behavior happens, the scores tend to become extremely low and may not be reliable as a reference for evaluation. To cope with this problem, we also report the scores when repetitive sections are removed from the predicted texts. We use a regular expression to remove the last ten or more consecutive occurrences of a string from the entire string, leaving only the first occurrence.

4.2. Results

4.2.1. Result on JSSODa

Tables 2 and 3 show the evaluation results on the JSSODa test set. Each shows the results for horizontal and vertical writing. The part below "Raw Output" shows the results when MLLM's output was used directly for evaluation, while the part below "Remove Repetition" shows the results after removing repeated sections. "(+FT)" denotes the model fine-tuned on the JSSODa train set.

The results highlight a clear disparity: while all models handled horizontally written text reasonably well, they struggled significantly with vertically written text. In particular, Qwen2.5-VL sometimes read vertically written text in horizontal reading order, as shown in Section 4.3. InternVL3, Gemma 3, GPT-4.1, and GPT-5 seemed to have a certain understanding of character reading order, but compared to horizontally written text, they made more errors in character recognition. For both InternVL3 and Gemma 3, we observe that performance on vertically written text tends to improve as the model's parameter count increases. GPT-5 sometimes produced no text output at all. The reasoning trace alone may have reached the max new tokens limit. Additionally, removing repetitions improves the score, indicating that MLLMs generate tokens repeatedly. When the model is trained

Horizontal Writing								
Columns	1		2		3		4	
Models	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
Raw Output								
Qwen2.5-VL-7B	7.75	91.7	19.4	87.2	21.6	87.7	23.5	82.6
Qwen2.5-VL-32B	**0.101**	**99.8**	17.0	99.1	10.6	**99.2**	6.79	**98.9**
InternVL3-8B-hf	0.462	99.3	25.3	95.4	13.6	96.3	8.20	96.9
InternVL3-38B-hf	0.894	98.9	10.4	**99.3**	5.15	99.0	3.78	98.0
Gemma 3 12B IT	3.06	95.6	16.8	92.6	15.3	92.3	17.7	84.1
Gemma 3 27B IT	2.13	97.3	9.81	97.2	14.7	96.6	14.8	90.7
GPT-4.1	1.88	97.8	**1.06**	98.5	**1.03**	98.7	**1.10**	98.4
GPT-5	2.09	97.5	1.26	98.3	1.28	98.6	1.96	97.4
Qwen2.5-VL-7B (+FT)	**0.0637**	**99.9**	**0.0251**	**99.9**	**0.0266**	**99.9**	**0.0383**	**99.9**
InternVL3-8B-hf (+FT)	0.391	99.5	0.0418	**99.9**	0.0624	**99.9**	0.363	99.6
Gemma 3 12B IT (+FT)	0.162	99.7	0.267	99.5	0.364	99.4	1.31	98.0
Remove Repetition								
Qwen2.5-VL-7B	7.63	91.6	17.2	86.9	20.4	87.5	22.0	82.2
Qwen2.5-VL-32B	**0.101**	**99.8**	17.0	99.1	10.6	**99.2**	6.79	**98.9**
InternVL3-8B-hf	0.462	99.3	25.3	95.4	13.6	96.3	7.61	96.8
InternVL3-38B-hf	0.894	98.9	10.4	**99.3**	5.15	99.0	3.78	98.0
Gemma 3 12B IT	2.75	95.6	16.1	92.6	14.5	92.3	16.8	83.9
Gemma 3 27B IT	1.66	97.4	9.81	97.2	14.7	96.6	14.8	90.7
GPT-4.1	1.88	97.8	**1.06**	98.5	**1.03**	98.7	**1.10**	98.4
GPT-5	2.09	97.5	1.26	98.3	1.28	98.6	1.96	97.4
Qwen2.5-VL-7B (+FT)	**0.0637**	**99.9**	**0.0251**	**99.9**	**0.0266**	**99.9**	**0.0383**	**99.9**
InternVL3-8B-hf (+FT)	0.391	99.5	0.0418	**99.9**	0.0624	**99.9**	0.363	99.6
Gemma 3 12B IT (+FT)	0.162	99.7	0.267	99.5	0.364	99.4	1.31	98.0

Table 2: The result on JSSODa test set (horizontal)

on the train set, the score for vertically written text improves substantially.

4.2.2. Result on VJRODa

Table 4 shows the evaluation result on VJRODa. None of the models performed well, indicating that they struggle with text in real-world vertical-writing document images. When fine-tuned on the JSSODa train set, Qwen2.5-VL-7B exhibited substantial performance gains under both the "Raw Output" and "Remove Repetition" evaluation settings, whereas the other two models showed no appreciable improvement. A plausible explanation is that Qwen2.5-VL-7B initially lacked a robust understanding of the reading order in vertically written Japanese text, while the other models had already internalized this to some extent. These findings suggest that the JSSODa train set may be effective for enhancing vertical-text reading capabilities in models that do not yet capture Japanese vertical reading order. By contrast, improving performance on vertically written text in real-world documents likely requires training with real-world OCR datasets.

4.3. Case Study

Figure 4 shows example outputs for images from the JSSODa test set (vertical). The original Qwen2.5-VL-7B incorrectly read vertically written text from top-left to right. In the fine-tuned model, vertically written text could now be read correctly, in order from the top right downward.

Figure 5 shows example outputs from real-world data (VJRODa). Qwen2.5-VL-7B was only able to output a portion of the text. On the other hand, the fine-tuned model output text mostly following Japanese reading order, though with some errors. In GPT-4.1, the output appears to adhere to the reading order of vertically written Japanese text; however, some portions of the output do not correspond to the actual text present in the image.

5. Conclusion

In this paper, we evaluated the OCR capability of MLLMs for vertically written Japanese text. To this end, we constructed a synthetic image dataset of horizontally and vertically written Japanese text, as well as a dataset of document images containing

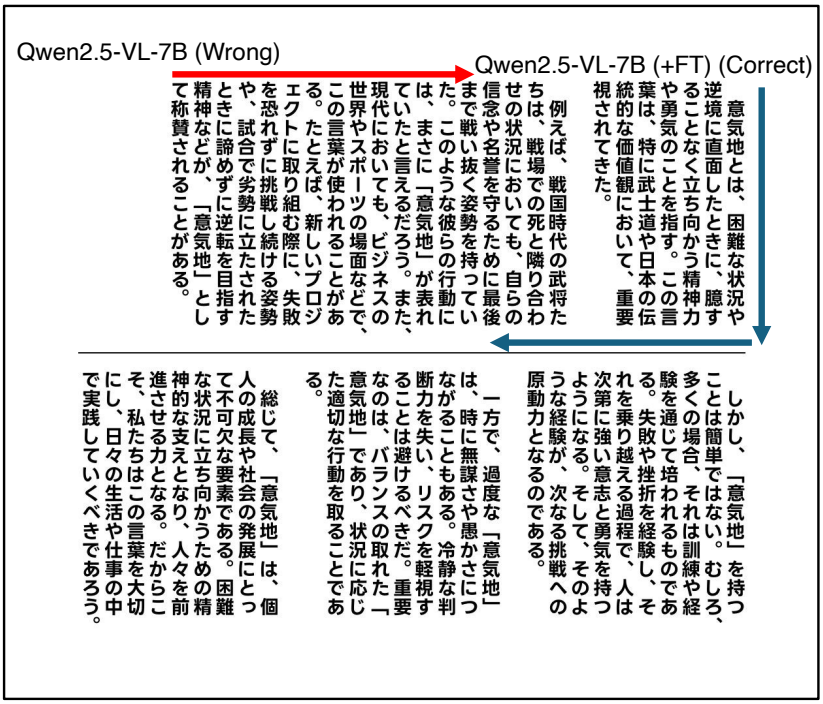
Vertical Writing								
Columns	1		2		3		4	
Models	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
Raw Output								
Qwen2.5-VL-7B	112	26.8	100	21.1	104	19.6	102	18.7
Qwen2.5-VL-32B	128	29.7	107	18.6	104	18.7	97.9	16.6
InternVL3-8B-hf	81.0	50.0	64.8	61.1	64.8	61.6	61.9	59.8
InternVL3-38B-hf	22.1	81.5	**43.7**	**75.4**	40.1	76.5	42.7	**72.0**
Gemma 3 12B IT	20.3	79.2	63.3	52.8	42.8	66.4	61.1	53.0
Gemma 3 27B IT	**7.62**	**91.4**	47.0	63.9	**28.3**	**81.3**	**37.7**	70.6
GPT-4.1	18.2	82.8	56.0	65.8	45.9	65.2	40.7	69.4
GPT-5	21.3	83.1	66.0	61.7	59.4	61.6	48.7	64.7
Qwen2.5-VL-7B (+FT)	**0.104**	**99.8**	**0.202**	**99.9**	**0.113**	**99.8**	**0.284**	**99.6**
InternVL3-8B-hf (+FT)	0.619	99.4	0.315	99.8	0.330	99.6	1.10	98.8
Gemma 3 12B IT (+FT)	0.502	99.1	1.04	98.5	1.47	98.1	4.09	94.8
Remove Repetition								
Qwen2.5-VL-7B	73.2	31.2	83.8	19.3	88.1	17.5	87.0	15.5
Qwen2.5-VL-32B	84.2	45.7	90.6	21.0	88.6	19.3	89.6	15.9
InternVL3-8B-hf	39.4	66.6	47.6	68.1	50.2	70.5	49.0	67.9
InternVL3-38B-hf	12.3	89.5	39.3	**74.8**	34.1	80.6	35.7	**79.3**
Gemma 3 12B IT	14.3	83.3	48.3	57.2	35.0	69.3	52.6	54.4
Gemma 3 27B IT	**7.62**	**91.4**	**37.3**	70.0	**26.2**	**82.6**	**35.6**	70.8
GPT-4.1	17.4	82.7	53.8	65.5	45.5	65.1	38.6	69.1
GPT-5	21.3	83.1	66.0	61.7	59.4	61.6	48.7	64.7
Qwen2.5-VL-7B (+FT)	**0.104**	**99.8**	**0.202**	**99.9**	**0.113**	**99.8**	**0.284**	**99.6**
InternVL3-8B-hf (+FT)	0.619	99.4	0.315	99.8	0.330	99.6	1.10	98.8
Gemma 3 12B IT (+FT)	0.502	99.1	1.04	98.5	1.47	98.1	4.09	94.8

Table 3: The result on JSSODa test set (vertical)

	Raw Output		Remove Repetition	
Models	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
Qwen2.5-VL-7B	154	20.1	88.5	22.0
Qwen2.5-VL-32B	128	**42.6**	63.0	**58.5**
InternVL3-8B-hf	121	26.0	66.5	40.8
InternVL3-38B-hf	138	27.7	64.0	45.1
Gemma 3 12B IT	125	17.5	67.9	23.3
Gemma 3 27B IT	**67.9**	35.0	**56.7**	34.2
GPT-4.1	101	29.2	61.7	34.1
GPT-5	70.1	40.9	69.4	41.0
Qwen2.5-VL-7B (+FT)	**65.1**	**51.5**	**40.5**	**61.1**
InternVL3-8B-hf (+FT)	251	26.1	73.5	54.9
Gemma 3 12B IT (+FT)	77.6	27.9	67.4	27.2

Table 4: The result on VJRODa

vertically written Japanese text collected from real-world PDFs, and conducted evaluations. Evaluation results indicate that the current MLLMs struggle with reading vertically written Japanese text. We also found that training on our dataset of synthetic images can improve the performance of a model that does not handle vertically written Japanese text well. In the future, we would like to explore methods for building models that can handle a variety of Japanese document images.



Qwen2.5-VL-7B (Wrong reading order, CER: 93.3)

て精とやをエるこ世現てはたま信せち称神き、恐エク。の世界現代い、で念のは例さ的は勇こ境意賛なに試れトた言やにたまこの戦や状、えれな、気とに気さど諦合ずにと葉スおとさのい名況戦ばて価特のな直地れがめでに取えばがポい言によ抜誉に場、き値にこく面とはる、ず劣挑りば使Iてえるようくをおで戦た観武 …

Qwen2.5-VL-7B (+FT) (Correct reading order, CER: 0)

意気地とは、困難な状況や逆境に直面したときに、臆することなく立ち向かう精神力や勇気のことを指す。この言葉は、特に武士道や日本の伝統的な価値観において、重要視されてきた。¥n\n例えば、戦国時代の武将たちは、戦場での死と隣り合わせの状況においても、自らの信念や名誉を守るために最後まで戦い抜く…

Figure 4: The example outputs on JSSODa (vertical) generated by Qwen2.5-VL-7B and its fine-tuned (+FT) variant. The red arrow (→) represents the reading order of Qwen2.5-VL-7B, and the blue arrow (→) represents the reading order of the fine-tuned model.

6. Acknowledgement

This work was supported by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models” project of the Ministry of Education, Culture, Sports, Science and Technology. In this work, we used the “mdx: a platform for building data-empowered society” (Suzumura et al., 2022).

7. Limitations

The Need for a More Visually Diverse Synthetic Japanese OCR Dataset for Model Training Our JSSODa dataset comprises clean, minimalist document images consisting of black text on a white background; no synthetic noise or other degradations are introduced, and no embedded images, tables, or figures are included. To better generalize to real-world document images, it is necessary to develop a dataset that encompasses greater visual diversity.

Evaluation on Documents with Multiple Plausible Reading Orders In real-world document images, there can be multiple plausible reading orders. For instance, the text within an image caption may have several possible reading sequences. In our VJRODa dataset, only a single reading order is annotated for each document. Consequently, under metrics such as CER and BLEU, even a correct reading order can potentially receive a low score if it differs from the single ground truth. Ideally, one would annotate all possible ground-truth reading orders and select the one that yields the highest score; however, this approach is very expensive. Therefore, developing an efficient and effective method for evaluating the OCR capabilities of MLLMs is an important direction for future work.

Qwen2.5-VL-7B (CER: 93.3, all outputs are shown)

国民年金\nねんきん特別便について\nQ1\nねんきん特別便に関するよくある質問\nA1\n年金特別便の一年金記録のお知らせ\nQ2\nA2\nお問い合わせ先\nTEL [REDACTED]\n2008.10.10 広報かりわ

Qwen2.5-VL-7B (+FT) (CER: 44.0)

民年金\n\nねんきん特別便について\n\n社会保険庁ではみなさまにご自分の年金加入記録を確認していただくため平成十九年十二月から平成二十一年十月にかけて、年金加入記録をお送りしています。\n\nご自分の年金加入記録を確認し、間違いの有る・無しを問わず必ず回答してくださいますようお願いします。\n\nねんきん特別便に関するよくある質問\n\nQ1\n\n先日、ねんきん特別便が届きました。「年金記録のお知らせ」の内容を確認したところ、結婚前に旧姓で加入していた厚生年金の記録が漏れています。どのようにしたらいいですか？…

GPT-4.1(CER: 70.7)

国民年金\nねんきん特別便について\n\n社会保険庁では、みなさまにご自分の年金加入期間の記録を確認していただくことが重要と考え、平成十九年十月にかけて「年金加入記録のお知らせ」を送付し、ご自分の記録を確認して間違いの有無を問わずご回答してくださるようお願いしています。\n\nねんきん特別便に関するよくある質問\n\nQ1\n先日、「ねんきん特別便」が届きました。「年金記録のお知らせ」内容を確認したところ、結婚前に国民年金に加入していた時期に旧姓で加入していたように記載されています。今の姓で加入していた記録がありませんが、…

Figure 5: The example outputs on VJRODa generated by Qwen2.5-VL-7B, its fine-tuned (+FT) variant, and GPT-4.1. (https://warp.ndl.go.jp/info:ndljp/pid/11712522/www.vill.kariwa.niigata.jp/open/info/000000001_0000000609.pdf, page 8, Personal information has been masked.)

Image Diversity in the VJRODa Dataset The images in the VJRODa dataset are sourced primarily from PDFs obtained from public-sector websites. As a result, the domain of document images is relatively narrow and may lack visual diversity.

8. Ethical Considerations

Since our JSSODa dataset is constructed from LLM-generated texts, it does not infringe third-party copyrights. Regarding the data sources of our VJRODa dataset, their use for information analysis is permitted under the copyright law of the country in which the research was conducted.

9. Bibliographical References

Teruo Akiyama and Norihiro Hagita. 1988. Automatic reading system for printed documents. In *Proceedings of IAPR Workshop on COMPUTER VISION*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-vl technical report.

Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. 2018. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng Cui, Ting Sun, Suyin Liang, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Xueqing Wang, Changda Zhou, Hongen Liu, Manhui Lin, Yue Zhang, Yubo Zhang, Handong Zheng, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025a. Paddleocr-vl: Boosting multilingual document parsing via a 0.9b ultra-compact vision-language model.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025b. Paddleocr 3.0 technical report.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei

Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2024. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. In *Advances in Neural Information Processing Systems*, volume 37, pages 42566–42592. Curran Associates, Inc.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, Miami, Florida, USA. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2025. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5817–5834, Vienna, Austria. Association for Computational Linguistics.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4083–4091, New York, NY, USA. Association for Computing Machinery.

Anh Duc Le, Daichi Mochihashi, Katsuya Masuda, Hideki Mima, and Nam Tuan Ly. 2019. Recognition of japanese historical text lines by an attention-based encoder-decoder and text line generation. In *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, HIP '19, page 37–41, New York, NY, USA. Association for Computing Machinery.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. 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, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15630–15640.

NDL Lab. 2021. Development of japanese ocr software in fy2021. https://lab.ndl.go.jp/data_set/ocr_en/r3_software/. Accessed: 2025-09-24.

OpenAI. 2025a. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-09-24.

OpenAI. 2025b. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-09-24.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto,

Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. 2022. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 1–7.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-docowl: Mod-

ularized multimodal large language model for document understanding.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023b. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models.

10. Language Resource References

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. 2020. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18.

Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. 2025. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding.

Chankyu Choi, Youngmin Yoon, Junsu Lee, and Junseok Kim. 2019. Simultaneous recognition of horizontal and vertical text in natural images. In *Computer Vision – ACCV 2018 Workshops*, pages 202–212, Cham. Springer International Publishing.

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 (ECCV)*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2017. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838.

Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, and Jean-Marc Ogier. 2019. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition — rrc-mlt-2019. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1582–1587.

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, Italia. ELRA and ICCL.

Shota Orihashi, Yoshihiro Yamazaki, Mihiro Uchida, Akihiko Takashima, and Ryo Masumura. 2022. Shared modeling of horizontal and vertical writing using character counting for japanese scene text recognition. In *Proceedings of the 21st Forum on Information Technology (FIT2022)*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13636–13645.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13878–13888.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy.

11. Appendix

11.1. Evaluation of OCR-Specialized MLLMs

We evaluated two recent OCR-specialized MLLMs, DeepSeek-OCR (Wei et al., 2025) and PaddleOCR-VL (Cui et al., 2025a), on our proposed dataset. For DeepSeek-OCR, we employed the Gundam (Dynamic Resolution) mode with the prompt "`<image>\nFree OCR.` ". The max new tokens parameter was set to 1024 for JSSODa and 3072 for VJRODa. For PaddleOCR-VL, we adopted a pipeline in which document layout analysis was first performed and the resulting segments were fed to the MLLM; the final answer was obtained by concatenating the text according to the predicted reading order. No max new tokens constraint was specified for PaddleOCR-VL. Note that these settings, including the prompt and max new tokens configurations, differ from those used for other models.

Table 5 presents the results on the JSSODa test set. DeepSeek-OCR exhibited strong text recognition performance for horizontally written text but performed poorly on vertically written text. In contrast, PaddleOCR-VL showed much less degradation on vertically written text than DeepSeek-OCR. Table 6 reports the results on the VJRODa dataset. DeepSeek-OCR obtained low scores, indicating weak performance on vertically written Japanese text. Across all models, including those listed in Table 4, PaddleOCR-VL achieved the highest scores.

11.2. Prompts

Table 7 shows the prompt used to instruct the LLM to generate Japanese sentences from Japanese nouns during the construction of the JSSODa dataset. Table 8 shows the user prompt used for fine-tuning and evaluation.

11.3. Training Budget

When trained on our JSSODa train set, Qwen2.5-VL-7B-Instruct requires 2.5 hours on 1 node with 8 $\times$ NVIDIA A100 (40 GB) GPUs; InternVL3-8B-hf requires 2 hours on 2 nodes, each with 8 $\times$ NVIDIA A100 (40 GB) GPUs; and Gemma 3 12B IT requires 1 hour on 2 nodes, each with 8 $\times$ NVIDIA A100 (40 GB) GPUs.

11.4. Regular Expression for Removing Repetition

The regular expressions used to remove repeated segments from LLM-generated text are presented in Table 9.

11.5. API Versions

The API versions used were `gpt-4.1-2025-04-14` for GPT-4.1 and `gpt-5-2025-08-07` for GPT-5.

11.6. Additional Examples of Images from Our Datasets

We show additional image examples from our JSSODa dataset in Figures 6 and 7, and from our VJRODa dataset in Figure 8.

Columns	1		2		3		4	
Models	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
horizontal Writing								
Raw Output								
DeepSeek-OCR	0.190	99.6	6.09	99.3	3.08	99.0	3.10	96.7
PaddleOCR-VL	0.496	99.1	3.58	96.7	12.6	93.7	18.5	90.4
Remove Repetition								
DeepSeek-OCR	0.190	99.6	6.09	99.3	3.08	99.0	3.10	96.7
PaddleOCR-VL	0.496	99.1	3.58	96.7	12.6	93.7	18.5	90.4
Vertical Writing								
Raw Output								
DeepSeek-OCR	108	35.7	160	11.7	153	12.4	130	9.80
PaddleOCR-VL	27.3	97.2	11.0	90.4	7.75	93.0	6.67	94.1
Remove Repetition								
DeepSeek-OCR	108	35.8	158	11.8	152	12.6	129	9.87
PaddleOCR-VL	27.2	97.2	3.75	96.4	4.61	96.1	4.00	96.5

Table 5: The result on JSSODa test set

	Raw Output		Remove Repetition	
Models	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
DeepSeek-OCR	182	10.8	181	10.9
PaddleOCR-VL	20.1	91.4	19.1	91.3

Table 6: The result on VJRODa

以下の単語について、その単語をテーマにした日本語の文章を出力してください。文章は500文字以上にしてください。与えられた単語が必ずしも出力に含まれている必要はありません。文章の一部に英単語や数字が含まれていてもよいです。文章の文体はどのようなものでもよく、教科書風の文章、ニュース記事、小説、エッセイ、プレスリリース、官公庁の文章、SNSなど、日本語として破綻していなければ何でもよいです。出力は文章のみとし、余計なものは出力しないでください。

単語: {word}

(*For the following word, output a Japanese text themed around that word. Make the text at least 500 characters long. The given word does not necessarily need to be included in the output. It is acceptable for parts of the text to contain English words or numbers. Any writing style is fine—textbook-like writing, a news article, a novel, an essay, a press release, an official government-style document, social media, etc.—as long as it is coherent in Japanese. Output text only, and do not include anything unnecessary.*

Word: {word})

Table 7: Prompt for generating Japanese sentences from Japanese nouns

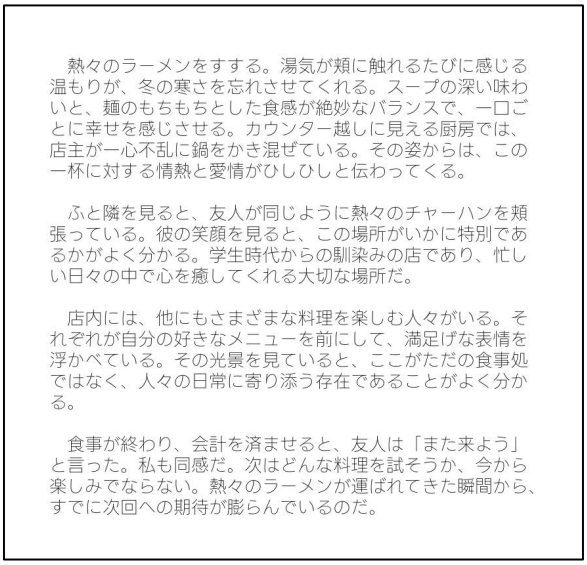
この画像内のテキストを日本語の読み順に従って全て出力してください。出力は画像内のテキストのみとしてください。

(*Please output all the text in this image following the Japanese reading order. The output should contain only the text in the image.*)

Table 8: Prompt for fine-tuining and evaluation

```
repetition_pattern = r'((\S .*?)\2{9,})(?!.*(\S .*?)\3{9,})'
pred_text = re.sub(repetition_pattern, r'\2', pred_text, flags=re.DOTALL)
```

Table 9: Regular expression for removing repetition



熱々のラーメンをすする。湯気が頬に触れるたびに感じる温もりが、冬の寒さを忘れさせてくれる。スープの深い味わいと、麺のもちもちとした食感が絶妙なバランスで、一口ごとに幸せを感じさせる。カウンター越しに見える厨房では、店主が一心不乱に鍋をかき混ぜている。その姿からは、この一杯に対する情熱と愛情がひしひしと伝わってくる。

ふと隣を見ると、友人が同じように熱々のチャーハンを頬張っている。彼の笑顔を見ると、この場所がいかに特別であるかがよく分かる。学生時代からの馴染みの店であり、忙しい日々の中で心を癒してくれる大切な場所だ。

店内には、他にもさまざまな料理を楽しむ人々がいる。それぞれが自分の好きなメニューを前にして、満足げな表情を浮かべている。その光景を見ていると、ここがただの食事処ではなく、人々の日常に寄り添う存在であることがよく分かる。

食事が終わり、会計を済ませると、友人は「また来よう」と言った。私も同感だ。次はどんな料理を試そうか、今から楽しみでならない。熱々のラーメンが運ばれてきた瞬間から、すでに次回への期待が膨らんでいるのだ。

(a) horizontal, 1-column

ある晴れた日の午後、小さな町の一角にある幼稚園では、子供たちが元気いっぱいに遊んでいた。その中でも特に目を引いたのは、一人の小さな男の子だった。彼の名前は健太といい、その無邪気な笑顔と愛らしい仕草で、周囲の人々を魅了していた。

健太の母親である美咲さんは、毎朝早く起きてお弁当を作り、彼を送り出すのが日課だった。彼女はいつも「健太、今日も一日楽しく過ごしてね」と優しく声をかけ、彼を抱きしめる。その瞬間、健太の瞳には純粋な愛情が映し出され、美咲さんもまた、息子への深い愛を感じるのだった。

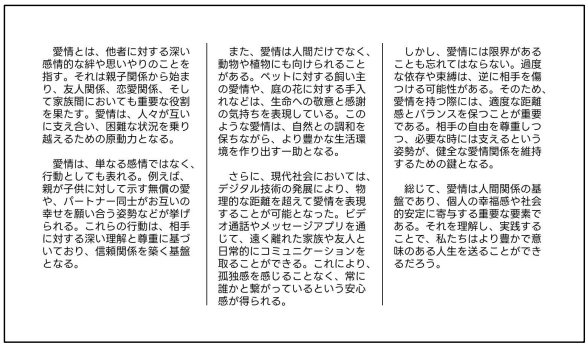
しかし、そんな幸せな日々にも影が忍び寄る。数ヶ月前、美咲さんの夫が突然の病でこの世を去ったのだ。愛する人を失った悲しみは計り知れず、美咲さんはどうやってこの子を育てていけばいいのか途方に暮れた。それでも、彼女は心の中で夫との約束を思い出し、健太のために強く生きようと決意した。

ある日、美咲さんは近所の公園で子供たちと一緒に遊ぶ健太の姿を見て、ふと微笑んだ。そこには、亡き夫がいつも言っていた「子供たちの笑顔が一番の宝物だ」という言葉が重なって見えたからだ。美咲さんは心の中で夫に感謝し、これからも健太と共に歩んでいくことを誓った。

それからというもの、美咲さんはますます愛情深く健太に接し、彼の成長を見守ることに喜びを感じていた。そして、健太もまた、母親の深い愛に包まれながら、毎日を楽しく過ごしていた。

このようにして、愛児・健太とその母・美咲さんの物語は、小さな町の温かい人々によって支えられながら続いていくのであった。

(b) horizontal, 2-columns



愛情とは、他者に対する深い感情的な絆や思いやりのことを指す。それは親子関係から始まり、友人関係、恋愛関係、そして家族間においても重要な役割を果たす。愛情は、人々が互いに支え合い、困難な状況を乗り越えるための原動力となる。

愛情は、単なる感情ではなく、行動としても表れる。例えば、親が子供に対して示す無償の愛や、パートナー同士がお互いの幸せを願い合う姿勢などが挙げられる。これらの行動は、相手に対する深い理解と尊重に基づいており、信頼関係を築く基盤となる。

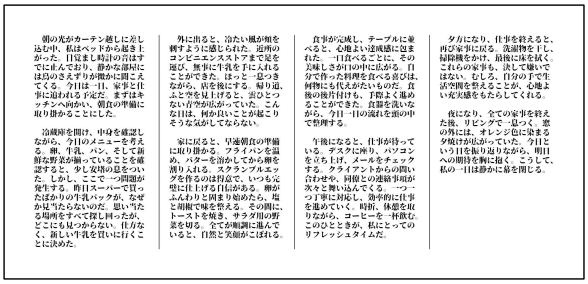
また、愛情は人間だけでなく、動物や植物にも向けられることがある。ペットに対する飼い主の愛情や、庭の花に対する手入れなどは、生命への敬意と感謝の気持ちを表現している。このような愛情は、自然との調和を保ちながら、より豊かな生活環境を作り出す一助となる。

さらに、現代社会においては、デジタル技術の発展により、物理的な距離を超えて愛情を表現することが可能となった。ビデオ通話やメッセージアプリを通じて、遠く離れた家族や友人と日常的にコミュニケーションを取ることができる。これにより、孤独感を感じることなく、常に誰かと繋がっているという安心感が得られる。

しかし、愛情には限界があることも忘れてはならない。過度な依存や束縛は、逆に相手を傷つける可能性がある。そのため、愛情を持つ際には、適度な距離感とバランスを保つことが重要である。相手の自由を尊重しつつ、必要な時には支えるという姿勢が、健全な愛情関係を維持するための鍵となる。

総じて、愛情は人間関係の基盤であり、個人の幸福感や社会的安定に寄与する重要な要素である。それを理解し、実践することで、私たちはより豊かで意味のある人生を送ることができるだろう。

(c) horizontal, 3-columns



朝の光がカーテン越しに差し込む中、私はベッドから起き上がった。目覚まし時計の音はすでに止んでおり、静かな部屋には鳥のさえずりが聞かに聞こえてくる。今日は一日、家事と仕事に追われる予定だ。まずはキッチンへ向かい、朝食の準備に取り掛かることにした。

冷蔵庫を開け、中身を確認しながら、今日のメニューを考える。卵、牛乳、パン、そして新鮮な野菜が揃っていることを確認すると、少し安堵の息をついた。しかし、ここで一つ問題が発生する。昨日スーパーで買ったばかりの牛乳パックが、なぜか見当たらないのだ。思い当たる場所をすべて探し回ったが、どこにも見つからない。仕方なく、新しい牛乳を買いに行くことに決めた。

外に出ると、冷たい風が頬を刺すように感じられた。近所のコンビニエンスストアまで足を運び、無事に牛乳を手に入れることができた。ほっと一息つきながら、店を後にする。帰り道、ふと空を見上げると、雲ひとつない青空が広がっていた。こんな日は、何か良いことが起こりそうな気がしてならない。

家に戻ると、早速朝食の準備に取り掛かる。フライパンを温め、バターを溶かしてから卵を流し入れる。スクランブルエッグを作るのは得意で、いつも完璧に仕上げる自信がある。卵がふんわりと固まり始めたら、塩と胡椒で味を整える。その間に、トーストを焼き、サラダ用の野菜を切る。全てが順調に進んでいると、自然と笑顔がこぼれる。

食事が完成し、テーブルに並べると、心地よい達成感に包まれた。一口食べることに、その美味しさが口の中に広がる。自分で作った料理を食べる喜びは、何物にも代えがたいものだ。食後の後片付けも、手際よく進めることができた。食器を洗いながら、今日一日の流れを頭の中で整理する。

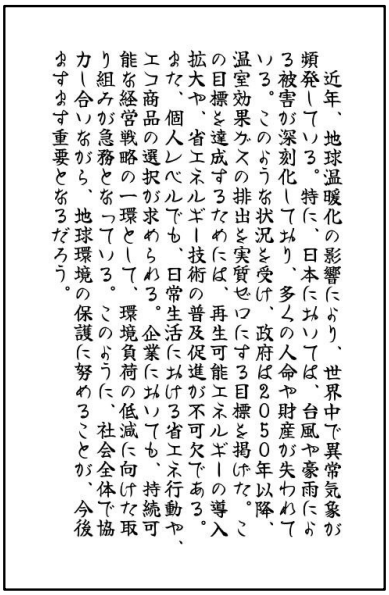
午後になると、仕事が待っている。デスクに座り、パソコンを立ち上げ、メールをチェックする。クライアントからの問い合わせや、同僚との連絡事項が次々と舞い込んでくる。一つ一つ丁寧に対応し、効率的に仕事を進めていく。時折、休憩を取りながら、コーヒーを一杯飲む。このひとときが、私にとってのリフレッシュタイムだ。

夕方になり、仕事を終えると、再び家事に戻る。洗濯物を干し、掃除機をかけ、晩ご飯の準備に取りかかる。これらの家事も、決して嫌いではない。むしろ、自分の手で生活空間を整えることが、心地よい充実感をもたらしてくれる。

夜になり、全ての家事を終えた後、リビングで一息つく。窓の外には、オレンジ色に染まる夕焼けが広がっていた。今日という日を振り返りながら、明日への期待を胸に抱く。こうして、私の一日は静かに幕を閉じる。

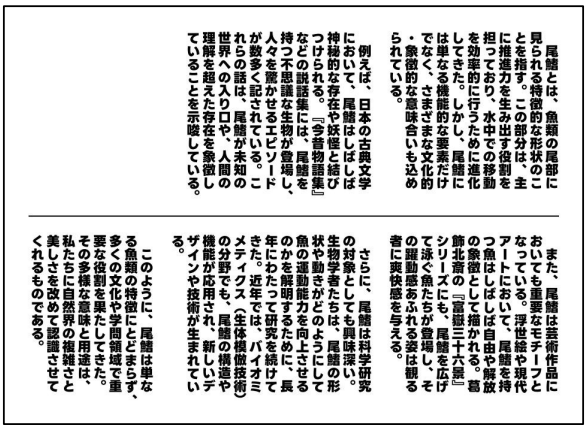
(d) horizontal, 4-columns

Figure 6: Additional example images from JSSODa (horizontal writing)



近年、地球温暖化の影響により、世界中で異常気象が頻発している。特に、日本においては、台風や豪雨による被害が深刻化しており、多くの人命や財産が失われている。このような状況を受け、政府は2050年以降、温室効果ガスの排出を実質ゼロにする目標を掲げた。この目標を達成するためには、再生可能エネルギーの導入拡大や、省エネルギー技術の普及促進が不可欠である。また、個人レベルでも、日常生活における省エネ行動や、エコ商品の選択が求められる。企業においても、持続可能な経営戦略の一環として、環境負荷の低減に向けた取り組みが急務となっている。このように、社会全体で協力し合いながら、地球環境の保護に努めることが、今後ますます重要となるだろう。

(a) vertical, 1-column



尾髄とは、魚類の尾部に見られる特徴的な形状のことを指す。この部分は、主に推進力を生み出す役割を担っており、水中の移動を効率的に行うために進化してきた。しかし、尾髄には単なる機能的な要素だけでなく、さまざまな文化的・象徴的な意味合いも込められている。

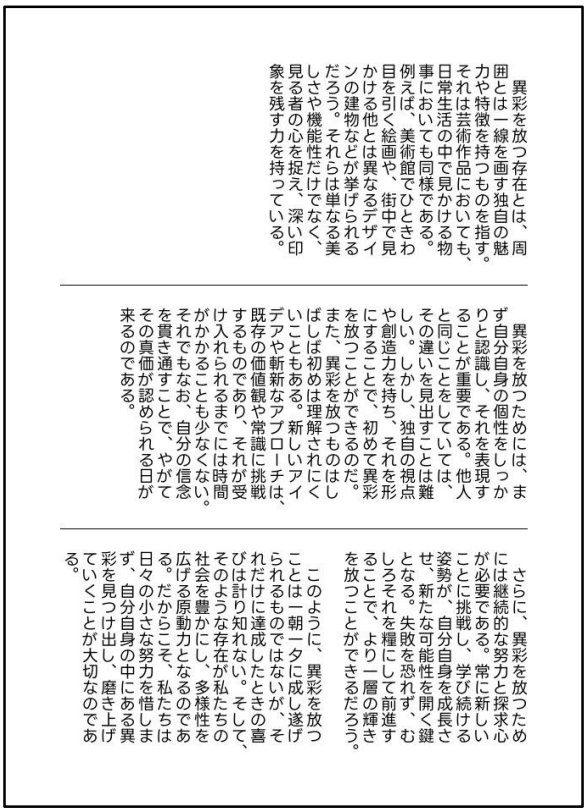
例えば、日本の古典文学において、尾髄はしばしば神秘的な存在や妖怪と結びつけられる。『今昔物語集』などの説話集には、尾髄を持つ不思議な生物が登場し、人々を驚かせるエピソードが数多く記されている。これらの話は、尾髄が未知の世界への入り口や、人間の理解を超えた存在を象徴していることを示唆している。

また、尾髄は芸術作品においても重要なモチーフとなっている。浮世絵や現代アートにおいて、尾髄を持つ魚は、しばしば自由や解放の象徴として描かれる。葛飾北斎の『富嶽三十六景』シリーズにも、尾髄を広げて泳ぐ魚たちが登場し、その躍動感あふれる姿は観る者に爽快感を与える。

さらに、尾髄は科学研究の対象としても興味深い。生物学者たちは、尾髄の形状や動きがどのようにして魚の運動能力を向上させるのかを解明するため、長年にわたって研究を続けてきた。近年では、バイオミメティクス（生体模倣技術）の分野でも、尾髄の構造や機能が応用され、新しいデザインや技術が生まれている。

このように、尾髄は単なる魚類の特徴にとどまらず、多くの文化や学問領域で重要な役割を果たしてきた。その多様な意味と用途は、私たちに自然界の複雑さと美しさを改めて認識させてくれるものである。

(b) vertical, 2-columns



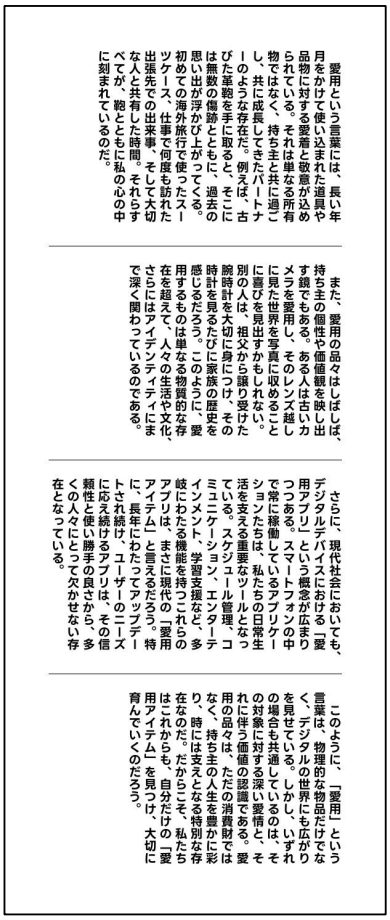
異彩を放つ存在とは、周囲とは一線を画す独自の魅力や特徴を持つものを指す。それは芸術作品においても、日常生活の中で見かける物事においても同様である。例えば、美術館でひときわ目を引く絵画や、街中で見かける他とは異なるデザインの建物などが挙げられるだろう。それらは単なる美しさや機能性だけでなく、見る者の心を捉え、深い印象を残す力を持っている。

異彩を放つためには、まず自分自身の個性をしっかりと認識し、それを表現することが重要である。他人と同じことをしていては、その違いを見出すことは難しい。しかし、独自の視点や創造力を持ち、それを形にすることで、初めて異彩を放つことができるのだ。また、異彩を放つものはしばしば初めは理解されにくいこともある。新しいアイデアや斬新なアプローチは、既存の価値観や常識に挑戦するものであり、それが受け入れられるまでには時間がかかることも少なくない。それでもなお、自分の信念を貫き通すことで、やがてその真価が認められる日が来るのである。

さらに、異彩を放つためには継続的な努力と探求心が必要である。常に新しいことに挑戦し、学び続ける姿勢が、自分自身を成長させ、新たな可能性を開く鍵となる。失敗を恐れず、むしろそれを糧にして前進することで、より一層の輝きを放つことができるだろう。

このように、異彩を放つことは一朝一夕に成し遂げられるものではないが、それだけに達成したときの喜びは計り知れない。そして、そのような存在が私たちの社会を豊かにし、多様性を広げる原動力となるのである。だからこそ、私たちは日々の小さな努力を惜しまず、自分自身の中にある異彩を見つけ出し、磨き上げていくことが大切なのである。

(c) vertical, 3-columns



愛用という言葉には、長い年月をかけて使い込まれた道具や品物に対する愛着と敬意が込められている。それは単なる所有物ではなく、持ち主と共に過ごし、共に成長してきたパートナーのような存在だ。例えば、古びた革鞄を手に取ると、そこには無数の傷跡とともに、過去の思い出が浮かび上がってくる。初めての海外旅行で使ったスーツケース、仕事で何度も訪れた出張先での出来事、そして大切な人と共有した時間。それらすべてが、鞄とともに私の心の中に刻まれているのだ。

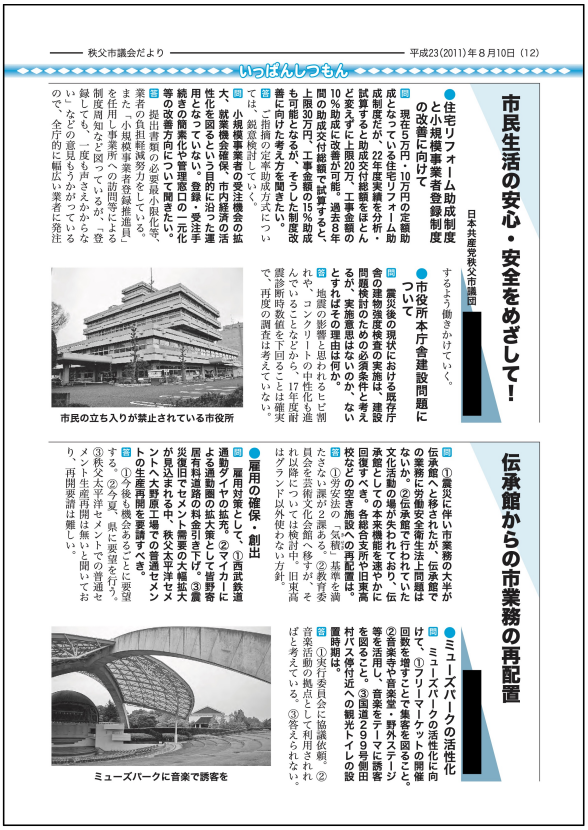
また、愛用の品々はしばしば、持ち主の個性や価値観を映し出す鏡でもある。ある人は古いカメラを愛用し、そのレンズ越しに見た世界を写真に収めることに喜びを見出すかもしれない。別の人は、祖父から譲り受けた腕時計を大切に身につけ、その時計を見るたびに家族の歴史を感じるだろう。このように、愛用するものは単なる物質的な存在を超えて、人々の生活や文化、さらにはアイデンティティにまで深く関わっているのである。

さらに、現代社会においても、デジタルデバイスにおける「愛用アプリ」という概念が広まりつつある。スマートフォンの中で常に稼働しているアプリケーションたちは、私たちの日常生活を支える重要なツールとなっている。スケジュール管理、コミュニケーション、エンターテインメント、学習支援など、多岐にわたる機能を持つこれらのアプリは、まさに現代の「愛用アイテム」と言えるだろう。特に、長年にわたってアップデートされ続け、ユーザーのニーズに応え続けるアプリは、その信頼性と使い勝手の良さから、多くの人々にとって欠かせない存在となっている。

このように、「愛用」という言葉は、物理的な物品だけでなく、デジタルの世界にも広がりを見せている。しかし、いずれの場合も共通しているのは、それぞれの対象に対する深い愛情と、それに伴う価値の認識である。愛用の品々は、ただの消費財ではなく、持ち主の人生を豊かに彩り、時には支えとなる特別な存在なのだ。だからこそ、私たちはこれからも、自分だけの「愛用アイテム」を見つけ、大切に育んでいくのだろう。

(d) vertical, 4-columns

Figure 7: Additional example images from JSSODa (vertical writing)



秋父市議会だより　平成23(2011)年8月10日 (12)

いっぱんしつもん

市民生活の安心・安全をめざして！

日本共産党秋父市議団

●住宅リフォーム助成制度と小規模事業者登録制度の改善に向けて

問　現在5万円・10万円の定額助成となっている住宅リフォーム助成制度だが、22年度実績を分析・試算すると助成交付総額をほとんど変えずに上限20万、工事金額の10%助成に改善が可能。過去8年間の助成交付総額で試算すると、上限30万円、工事金額の15%助成も可能となるが、そうした制度改善に向けた考え方を聞きたい。

答　ご指摘の定率助成方式については、鋭意検討していく。

問　小規模事業者の受注機会の拡大、就業機会確保、市内経済の活性化を図るという目的に沿った運用となっていない。登録・受注手続きの簡素化や管理窓口の一元化等の改善方向について聞きたい。

答　提出書類の必要最小限化等、業者の負担軽減努力をしている。また、「小規模事業者登録推進員」を任用し事業所への訪問等による制度周知など図っているが、「登録しても、一度も声さえかからない」などの意見もうかがっているので、全庁的に幅広い業者に発注するよう働きかけていく。

●市役所本庁舎建設問題について

問　震災後の現状における既存庁舎の建物強度検査の実施は、建設問題検討の必須条件と考えるが、実施意思はないのか、ないとすればその理由は何か。

答　地震の影響と思われるヒビ割れや、コンクリートの中性化も進んでいることなどから、17年度耐震診断時数値を下回ることは確実で、再度の調査は考えていない。

市民の立ち入りが禁止されている市役所

伝承館からの市業務の再配置

●ミューズパークの活性化

問　①震災に伴い市業務の大半が伝承館へと移されたが、伝承館での業務に労働安全衛生法上の問題はないか。②伝承館で行われていた文化活動の場が失われており、伝承館としての本来機能を速やかに回復すべき。各級公共施設への再配置は。

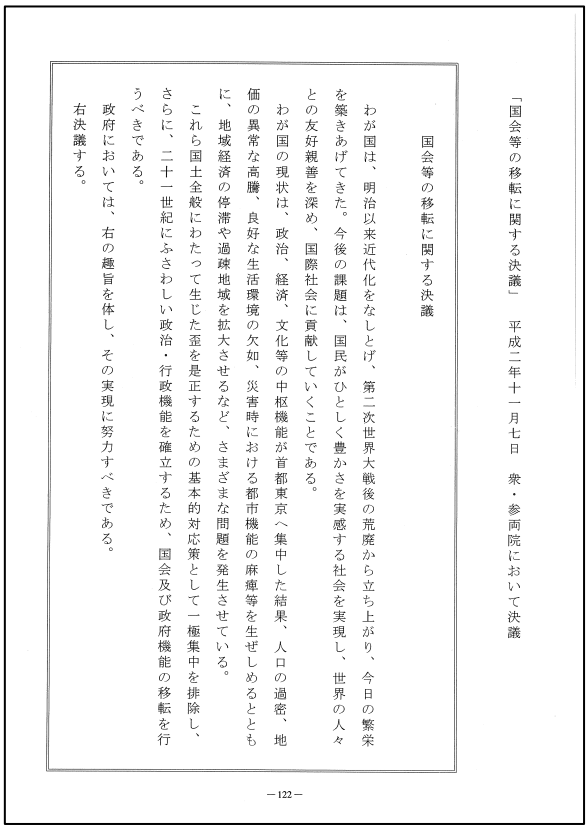
答　①労安法の「気積」基準を満たさない課が2課ある。②教育委員会を芸術文化会館へ移すが、それ以降については検討中。旧東高はグランド以外使えない。

●雇用の確保・創出

問　雇用対策として、①西武鉄道通勤ダイヤの拡充、②マイカーによる通勤圏の拡大策として提案寄届有料道路の料金引き下げ、③震災復旧でセメント需要の大幅拡大が見込まれる中、秩父太平洋セメントへ大野原工場での普通セメントの生産再開を要請すべき。

答　①今後も機会あるごとに要望する。②今夏、県に要望を行う。③秩父太平洋セメントでの普通セメント生産再開は無いと聞いており、再開要請は難しい。

ミューズパークに音楽で誘客を



「国会等の移転に関する決議」　平成二年十一月七日　衆・参両院において決議

国会等の移転に関する決議

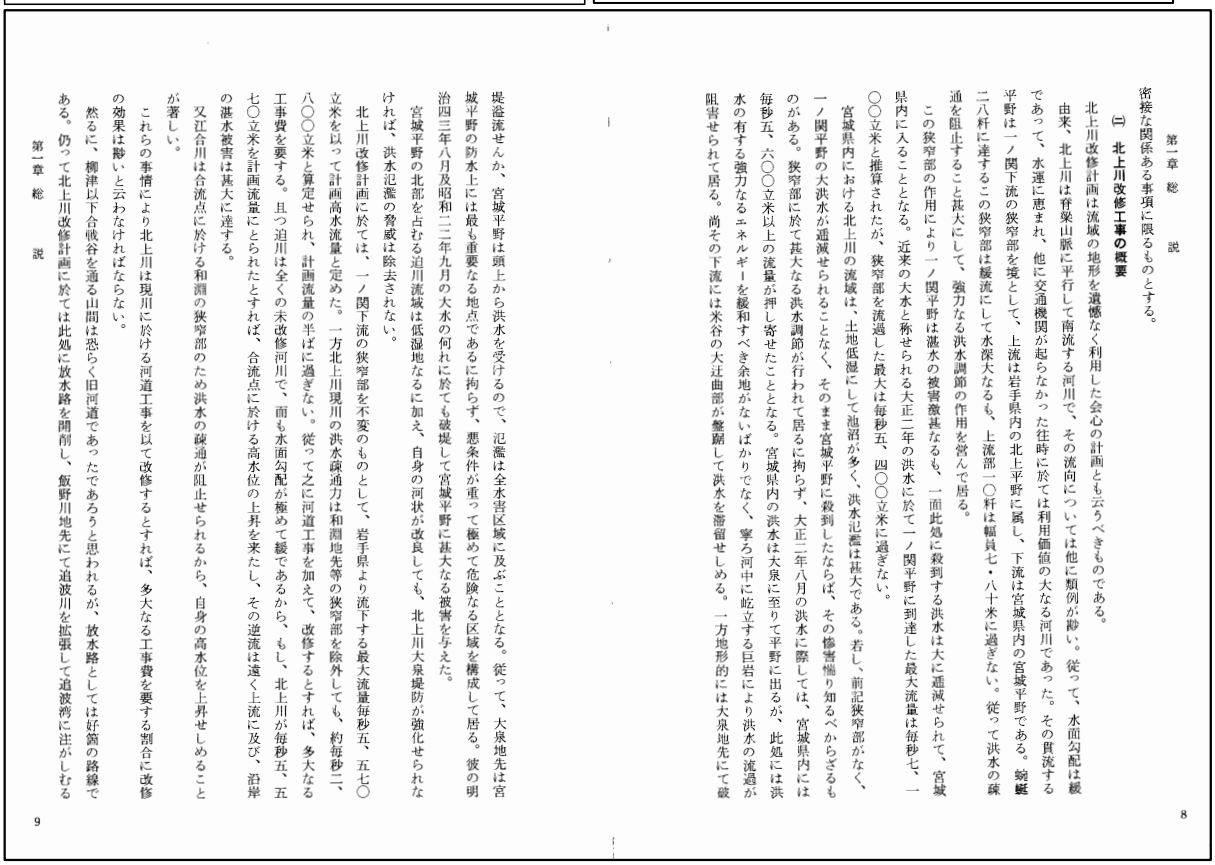
わが国は、明治以来近代化をなしとげ、第二次世界大戦後の荒廃から立ち上がり、今日の繁栄を築きあげてきた。今後の課題は、国民がひとしく豊かさを実感する社会を実現し、世界の人々との友好親善を深め、国際社会に貢献していくことである。

わが国の現状は、政治、経済、文化等の中枢機能が首都東京へ集中した結果、人口の過密、地価の異常な高騰、良好な生活環境の欠如、災害時における都市機能の麻痺等を生ぜしめるとともに、地域経済の停滞や過疎地域を拡大させるなど、さまざまな問題を発生させている。

これら国土全般にわたって生じた歪を是正するための基本的対応策として一極集中を排除し、さらに、二十一世紀にふさわしい政治・行政機能を確立するため、国会及び政府機能の移転を行うべきである。

政府においては、右の趣旨を体し、その実現に努力すべきである。

右決議する。



密接な関係ある事項に限るものとする。

(二)　北上川改修工事の概要

北上川改修計画は流域の地形を遺憾なく利用した会心の計画とも云うべきものである。

由来、北上川は脊梁山脈に平行して南流する河川で、その流向については他に類例が尠い。従って、水面勾配は緩であって、水運に恵まれ、他に交通機関が起らなかった往時に於ては利用価値の大なる河川であった。その貫流する平野は一ノ関下流の狭窄部を境として、上流は岩手県内の北上平野に属し、下流は宮城県内の宮城平野である。蛟蜒二八粁に達するこの狭窄部は緩流にして水深大なるも、上流部一〇粁は幅員七・八十米に過ぎない。従って洪水の疎通を阻止すること甚大にして、強力なる洪水調節の作用を営んで居る。

この狭窄部の作用により一ノ関平野は湛水の被害極甚なるも、一面此処に殺到する洪水は大に遮減せられて、宮城県内に入ることとなる。近来の大水と称せられる大正二年の洪水に於て一ノ関平野に到達した最大流量は毎秒七、一〇〇立米と推算されたが、狭窄部を流過した最大は毎秒五、四〇〇立米に過ぎない。

宮城県内における北上川の流域は、土地低湿にして池沼が多く、洪水氾濫は甚大である。若し、前記狭窄部がなく、一ノ関平野の大洪水が遮減せられることなく、そのまま宮城平野に殺到したならば、その惨害慮り知るべからざるものがある。狭窄部に於て甚大なる洪水調節が行われて居るに拘らず、大正二年八月の洪水に際しては、宮城県内には毎秒五、六〇〇立米以上の流量が押し寄せたこととなる。宮城県内の洪水は大泉に至りて平野に出るが、此処には洪水の有する強力なるエネルギーを緩和すべき余地がないばかりでなく、寧ろ河中に屹立する巨岩により洪水の流過が阻害せられて居る。尚その下流には米谷の大迂曲部が蟠踞して洪水を滞留せしめる。一方地形的には大泉地先にて破堤流せんか、宮城平野は頭上から洪水を受けるので、氾濫は全水害区域に及ぶこととなる。従って、大泉地先は宮城平野の防水上には最も重要なる地点であるに拘らず、悪条件が重って極めて危険なる区域を構成して居る。彼の明治四三年八月及昭和二二年九月の大水の何れに於ても破堤して宮城平野に甚大なる被害を与えた。

宮城平野の北部を占むる迫川流域は低湿地なるに加え、自身の河状が改良しても、北上川大泉堤防が強化せられなければ、洪水氾濫の脅威は除去されない。

北上川改修計画に於ては、一ノ関下流の狭窄部を不変のものとして、岩手県より流下する最大流量毎秒五、五七〇立米を以って計画高水流量と定めた。一方北上川現川の洪水疎通力は和淵地先等の狭窄部を除外しても、約毎秒二、八〇〇立米と算定せられ、計画流量の半ばに過ぎない。従って之に河道工事を加えて、改修するとすれば、多大なる工事費を要する。且つ迫川は全くの未改修河川で、而も水面勾配が極めて緩であるから、もし、北上川が毎秒五、五七〇立米を計画流量にとられたとすれば、合流点に於ける高水位の上昇を来たし、その逆流は遠く上流に及び、沿岸の湛水被害は甚大に達する。

又江合川は合流点に於ける和淵の狭窄部のため洪水の疎通が阻止せられるから、自身の高水位を上昇せしめることが著しい。

これらの事情により北上川は現川に於ける河道工事を以て改修するとすれば、多大なる工事費を要する割合に改修の効果は尠いと云わなければならない。

然るに、柳津以下合戦谷を通る山間は恐らく旧河道であったであろうと思われるが、放水路としては好個の路線である。仍って北上川改修計画に於ては此処に放水路を開削し、飯野川地先にて追波川を拡張して追波湾に注がしむる

Figure 8: Additional example images from VJRODa (top left: https://www.city.chichibu.lg.jp/secure/5707/%E5%B8%82%E8%AD%B0%E4%BC%9A%E3%81%A0%E3%82%88%E3%82%8A25%E5%8F%B712-15.pdf, page 1, top right: https://www.mlit.go.jp/kokudokeikaku/iten/information/council/pdf/toushin_sankou.pdf, page 120, bottom: https://www.thr.mlit.go.jp/Bumon/J73101/homepage/iport/kitakami/asobu_manabu/kitakamigawa/image/kitakami_3.pdf, page 14, Personal information has been masked.)