

JMMMU-Pro: Text-embedded Japanese Multi-discipline Multimodal Understanding Benchmark via Vibe Benchmark Construction

Anonymous ACL submission

Abstract

This paper introduces JMMMU-Pro, a text-embedded Japanese Multi-discipline Multimodal Understanding Benchmark, and Vibe Benchmark Construction, a scalable construction method. Following the evolution from MMMU to MMMU-Pro, JMMMU-Pro extends JMMMU by composing the question image and question text into a single image, thereby creating a benchmark that requires integrated visual-textual understanding through visual perception. To build JMMMU-Pro, we propose Vibe Benchmark Construction, a methodology in which an image generative model (*e.g.*, Nano Banana Pro) produces candidate visual questions, and humans verify the outputs and, when necessary, regenerate with adjusted prompts to ensure quality. By leveraging Nano Banana Pro’s highly realistic image generation capabilities and its ability to embed clean Japanese text, we construct a high-quality benchmark at low cost, covering a wide range of background and layout designs. Experimental results show that all open-source LMMs struggle substantially with JMMMU-Pro, underscoring JMMMU-Pro as an important benchmark for guiding future efforts in the open-source community. We believe that JMMMU-Pro provides a more rigorous evaluation tool for assessing the Japanese capabilities of LMMs and that our Vibe Benchmark Construction also offers an efficient guideline for future development of text-embedded VQA benchmarks.

1 Introduction

With the recent success of large multimodal models (LMMs) in English (OpenAI, 2024a; Liu et al., 2024a,b), there has been a growing interest in developing multilingual LMMs (Wang et al., 2024a; Dash et al., 2025; Yue et al., 2025a; Bai et al., 2025a) and LMMs specialized for non-English languages (SB Intuitions, 2025b; Team, 2025). Although LMM development in the Japanese domain

has emerged (SB Intuitions, 2025b; Baek et al., 2025a; Sasagawa et al., 2025), progress has been slower than in the English domain, in part due to the limited evaluation benchmarks. Given the large and rapidly growing population of Japanese LMM users, there is an increasing need to establish more Japanese benchmarks that can facilitate the development of LMMs capable of handling the Japanese language and culture seamlessly.

A major limitation of existing Japanese benchmarks is that the question image and the question text are provided to the model as separate modalities. This setup differs substantially from the core human cognitive skill: *Seamlessly integrating visual and textual information and interpreting them through visual perception*. Equipping LMMs with this cognitive ability in Japanese is a crucial step toward developing embodied agents and robotic systems (Zitkovich et al., 2023; Ahn et al., 2022; Huang et al., 2022; Li et al., 2024c) that can autonomously operate and explore real-world environments in Japan through visual perception. Furthermore, from the perspective of current LMMs’ use cases, users commonly provide LMMs with screenshots that include both Japanese text and images. Therefore, to foster human cognitive skills and support a wide range of real-world use cases, it is essential to evaluate LMMs on sufficiently complex tasks where both the question image and the question text are presented through a visual modality. Among the English benchmarks, MMMU-Pro (Yue et al., 2025b) extends MMMU (Yue et al., 2024) by constructing a benchmark in which both the question text and the question image are embedded within a single image, thereby enabling the evaluation of this dimension. However, there is no benchmark in Japanese that supports such evaluation. Therefore, developing a Japanese benchmark that enables the evaluation of this dimension is essential.

In this paper, we propose **JMMMU-Pro**, a text-

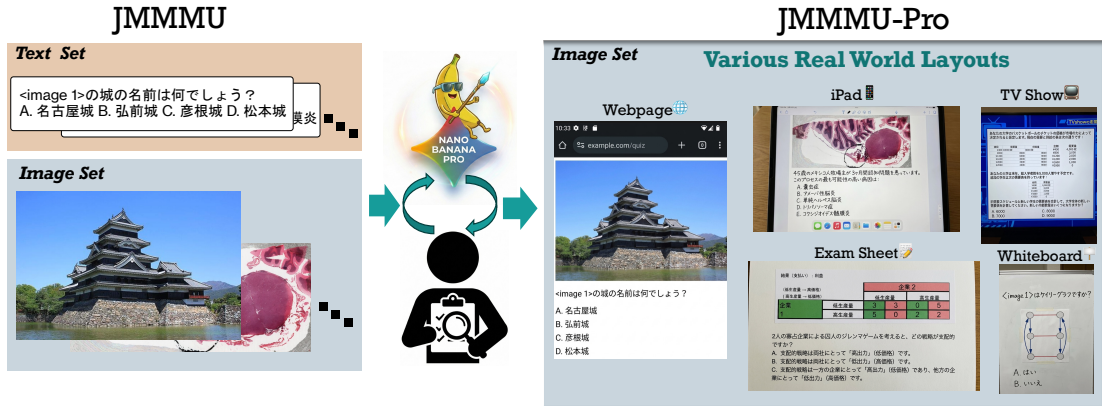


Figure 1: **Building JMMMU-Pro via Vibe Benchmark Construction.** JMMMU-Pro extends JMMMU by embedding each question image and text into a single image. For the construction, we propose Vibe Benchmark Construction, where an image generation model creates questions, followed by human verification and prompt refinement to ensure quality.

084 embedded Japanese Multi-discipline Multimodal
 085 Understanding benchmark. JMMMU-Pro follows
 086 the evolution from MMMU to MMMU-Pro and is
 087 constructed by embedding each of the 1,320 ques-
 088 tion texts and question images from the original JM-
 089 MMU (Japanese MMMU) (Onohara et al., 2025)
 090 tasks into a single composite image. Built on top
 091 of the established JMMMU, it enables an apple-
 092 to-apple comparison between JMMMU-Pro and
 093 JMMMU, which provides a meaningful signal of
 094 a model’s visual cognitive abilities. Consequently,
 095 JMMMU-Pro offers both high usability and highly
 096 informative feedback for model developers.

097 For the construction of JMMMU-Pro, we pro-
 098 pose a new benchmark creation methodology called
 099 **Vibe Benchmark Construction**. In this frame-
 100 work, an image generation model plays the pri-
 101 mary role in producing the visual question, while
 102 humans simply check the outputs and, when nec-
 103 essary, refine the prompts before regenerating the
 104 images, thereby ensuring consistent quality. Previ-
 105 ously, when creating text-embedded benchmarks
 106 (e.g., MMMU-Pro), all questions had to be cre-
 107 ated manually, which was not scalable and re-
 108 sulted in substantial human cost. In contrast, Vibe
 109 Benchmark Construction leverages Nano Banana
 110 Pro (DeepMind, 2025a), a state-of-the-art image
 111 generation model with exceptional photorealism.
 112 Nano Banana Pro can not only generate highly re-
 113 alistic images but also accurately embed Japanese
 114 text within them. Vibe Benchmark Construction
 115 is scalable, requires minimal human effort, and
 116 supports controllable layout diversity. Leveraging
 117 Nano Banana Pro with human verification, approx-
 118 imately 95% of JMMMU-Pro questions were gen-
 119 erated automatically, suggesting its potential as a

guideline for future text-embedded benchmarks. 120

In our experiments, we evaluated a total of 121
 122 15 LMMs, including representative closed-source
 123 LMMs, English-centric open-source LMMs, mul-
 124 tilingual open-source LMMs, and Japanese open-
 125 source LMMs. Our key experimental findings are
 126 summarized as follows:

1. Open-source LMMs struggle substantially on 127
 JMMMU-Pro. When compared with JM- 128
 MMU, most LMMs show a drop in perfor- 129
 mance on JMMMU-Pro. In particular, open- 130
 source LMMs exhibit a decrease ranging up 131
 to 23%. 132
2. Recent strong reasoning-based closed-source 133
 LMMs perform considerably well on 134
 JMMMU-Pro, revealing a substantial gap 135
 between closed-source and open-source 136
 LMMs. 137
3. Through detailed analysis, we find that al- 138
 though a major source of failure is the lack 139
 of Japanese OCR ability, strong OCR alone is 140
 not sufficient to solve JMMMU-Pro. This sug- 141
 gests that solving JMMMU-pro requires im- 142
 proving both OCR capability and the ability to 143
 interpret language and vision in an integrated 144
 manner through visual perception. 145

Our contributions are summarized as follows: 146

- **Construction of JMMMU-Pro:** We ex- 147
 tend JMMMU and introduce JMMMU-Pro, 148
 a benchmark that embeds each question 149
 image and its corresponding text into a single 150
 image, enabling the evaluation of integrated 151
 visual-textual understanding through visual 152
 perception. 153

- **Proposal of Vibe Benchmark Construction:** We propose Vibe Benchmark Construction, a dataset creation framework in which a powerful generation model drives the construction process, while humans only perform checking and minor prompt adjustments. With the continued progress of image generation models, we expect this approach to serve as an efficient and scalable guideline for future benchmark development.
- **Encouraging Future Efforts in the Open-Source Community:** Our results show that open-source LMMs struggle heavily on JMMMU-Pro, highlighting a substantial gap with closed-source LMMs. JMMMU-Pro provides a valuable benchmark that can motivate and guide the open-source community in closing this gap.

2 Related Work

Large Multimodal Models (LMMs). With recent advances in multilingual LLMs (Team et al., 2024; CohereLabs, 2024), both English-centric LMMs with multilingual capabilities (Li et al., 2024a; An et al., 2025; Tong et al., 2024) and fully multilingual LMMs (Yue et al., 2025a; Dash et al., 2025; Microsoft, 2025) have emerged. In parallel, several LMMs specialized for Japanese have also been developed (SB Intuitions, 2025a,b; Baek et al., 2025a; Sasagawa et al., 2025; Turing Inc., 2025). However, these models have not been evaluated on tasks that require solving Japanese questions that rely on integrated visual-textual understanding through visual perception. This highlights the need for a dedicated benchmark that can systematically evaluate such integrated visual-textual understanding capabilities in Japanese.

LMM Benchmarks. Among various recent benchmarks (Li et al., 2024b; Liu et al., 2023, 2024c; Lu et al., 2024; Yue et al., 2024; Miyai et al., 2025), MMMU-Pro (Yue et al., 2025b) is an established benchmark that embeds both the question text and the question image into a single image, challenging models to truly “see” and “read” simultaneously, mirroring how humans naturally process complex scenes in which text and visuals are interleaved. Unlike traditional OCR-related benchmarks (Singh et al., 2019; Liu et al., 2024d) or DocVQA (Mathew et al., 2021), MMMU-Pro requires not only text recognition but also complex reasoning that integrates both visual and textual information, thereby

pushing the capabilities of LMMs beyond standard document understanding. However, MMMU-Pro evaluates only English, leaving abilities in other languages unassessed.

Japanese LMM Benchmarks. The development of Japanese LMM benchmarks remains behind that of English benchmarks. Many existing studies focus primarily on common-sense knowledge and do not adequately cover expert-level domains (Shimizu et al., 2018; Turing, 2024b,a; Inoue et al., 2024; SakanaAI, 2024b,a), despite the rapid progress of LMMs and the importance of evaluating such capabilities. To address these issues, JMMMU (Onohara et al., 2025) was introduced, significantly advancing the landscape of Japanese LMM evaluation. However, JMMMU does not include tasks that require models to interpret both text embedded within images. Benchmarks that include Japanese text within images, such as JDocQA (Onami et al., 2024) and MangaVQA (Baek et al., 2025b), do exist, but they do not require the level of complex reasoning demanded by MMMU-Pro and are therefore insufficient for driving further advances in LMMs. To address this gap, we build upon JMMMU and create JMMMU-Pro, which evaluates a model’s ability to jointly understand different modalities in a more integrated manner and to perform high-level reasoning on such tasks.

QA Construction with Generative Models. In the context of LLM-based and LMM-based QA construction, it is common for humans to manually edit model-generated QA pairs. As a result, it is uncommon to rely solely on iterative prompt adjustments to construct QA data. While several works adopt iterative prompt adjustments in order to eliminate human-induced prompt bias (Shah, 2024; He et al., 2025), these approaches differ fundamentally from our goal, in that our objective is to scalably produce high-quality QA data. In the context of image-generation models, recent works have leveraged powerful image generation models to create the images of VQA benchmarks (Wang et al., 2024b; Xiao et al., 2025). However, these approaches still require substantial additional effort to construct the question texts through manual creation or LMM-based question generation. The most similar work to our concept is LogicOCR (Ye et al., 2025), which uses GPT-Image-1 (OpenAI, 2024b) to embed English question text into images with varied layouts. LogicOCR performs manual verification and discards a portion of the generated samples. Although the approach faces limitations,

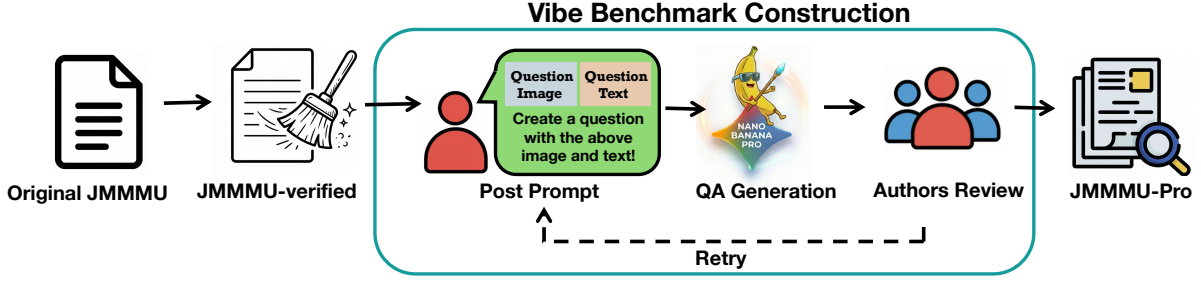


Figure 2: JMMMU-Pro Construction Pipeline.

such as reduced photorealism due to the capabilities of GPT-Image-1 and dataset shrinkage caused by sample filtering, it demonstrates a promising direction of leveraging modern image generation models. We build upon this line of work and we formalize a pipeline, Vibe Benchmark Construction, in which the process of generating VQA embedded in images is primarily driven by image generation models, while humans perform verification and, when necessary, adjust prompts and regenerate the images. By defining this pipeline clearly, we provide an effective guideline for the scalable construction of future text-embedded VQA benchmarks.

3 JMMMU-Pro Benchmark

JMMMU-Pro consists of 1,320 questions whose contents are identical to those in the original JMMMU. Figure 2 illustrates the construction pipeline. We first describe the original JMMMU and then present the core component of our approach: Vibe Benchmark Construction.

3.1 Revisiting the JMMMU Benchmark

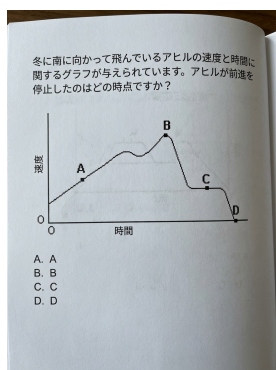
JMMMU (Onohara et al., 2025) consists of 1,320 questions and 1,118 images spanning 28 subjects. The benchmark is intentionally divided into two categories: culture-agnostic and culture-specific subjects. The culture-agnostic subset includes 24 subjects with 720 questions, across five disciplines: (1) Art & Psychology, (2) Business, (3) Health & Medicine, (4) Science, and (5) Tech & Engineering. The culture-specific subset comprises 600 questions across four subjects: (1) Japanese Art, (2) Japanese Heritage, (3) Japanese History, and (4) World History. Due to its comprehensive and rigorous evaluation coverage, JMMMU has become a foundational benchmark for the development of Japanese LLMs (SB Intuitions, 2025b,a; Turing Inc., 2025).

To simplify the interpretation of JMMMU evaluation results, we converted all 50 open-ended questions into multiple-choice questions and revised two samples. We refer to this updated version as JMMMU-verified-2025-12 (see details in Section C). All JMMMU scores reported in this paper are based on JMMMU-verified-2025-12.

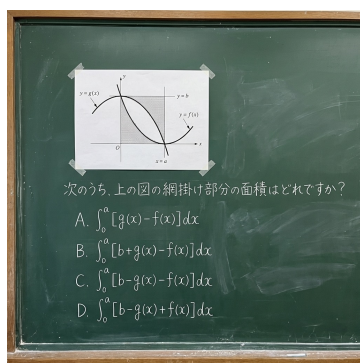
3.2 Definition of Vibe Benchmark Construction

Vibe Benchmark Construction is a methodology in which an image generation model plays the primary role in producing the VQA problem images, while humans only verify the outputs and adjust the prompts when necessary to ensure quality. Although previous VQA benchmarks have used synthetic images generated by image generation models, these models have played only a supplementary role, producing just the visual part while the question text still had to be created separately by humans or by an LMM, incurring additional cost.

In contrast, the key distinction of our proposed Vibe Benchmark Construction is that the VQA creation process is carried out by the image generation model, with humans intervening solely for verification and prompt refinement. This paradigm is particularly effective for text-embedded VQA, where humans cannot easily edit content directly inside the image. By letting the model handle generation and restricting human effort to *adjusting the prompt until a satisfactory image is produced*, the method enables efficient and scalable construction of benchmarks, especially in domains like text-embedded VQA, where dataset creation is difficult. A more detailed comparison with existing work is provided in Section 2.



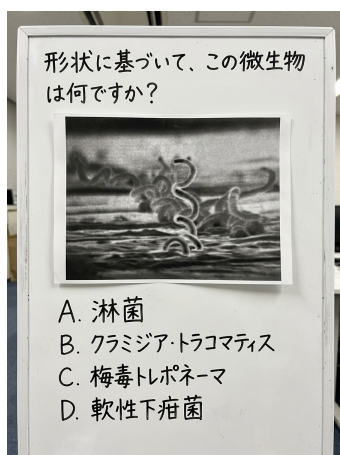
(a) Workbook



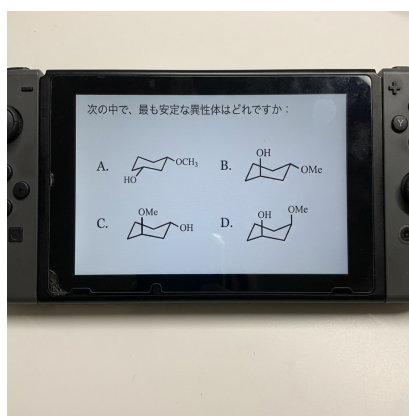
(b) Blackboard



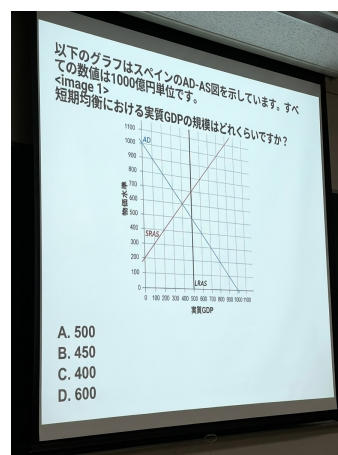
(c) iPad notebook



(d) Whiteboard



(e) Nintendo Switch



(f) Projector

Figure 3: JMMMU-Pro samples.

3.3 Detailed Pipeline of Vibe Benchmark Construction

For image generation, we use Nano Banana Pro via its API interface (gemini-3-pro-image-preview). The image resolution is set to 1K. Below, we describe our prompt design process and the workflow for human checking and regeneration.

Prompt Selection and Image Generation. We first selected a prompt template through preliminary experiments. Specifically, we used the prompt template in Section D and varied the following six components as parameters to generate a diverse set of images (shown in Figure 3): 1. *Background*: chosen from workbook, exam sheet, whiteboard, blackboard, projector, iPad notebook, webpage, Nintendo Switch, and TV quiz show. 2. *Background Color*: selected from white, light green, light yellow, light pink, light gray, and light blue. Note that certain backgrounds have a fixed color (e.g., a whiteboard is always white), and we account for such constraints. 3. *Font*: chosen from

handwritten text, computer font, thick computer font, thin computer font, and manga-style computer font. 4. *Margin*: selected as either small or large. 5. *State*: chosen from photo by smartphone, screenshot by PC, and screenshot by smartphone. 6. *Aspect Ratio*: selected from 9:16, 16:9, 3:4, and 1:1. We show the statistics for the two most controllable factors: *Background* and *Background Color* in Figure A.

In addition, while JMMMU includes image tags such as <image 1> within the question text, we found that Nano Banana Pro does not allow control over these image tags. To address this issue, we attempted to include explicit instructions in the prompt (e.g., “keep the image tag in the question”). However, such instructions significantly degraded the quality of the generated images. We hypothesize that this occurs because Nano Banana Pro internally uses similar image-tag tokens, and explicit instructions about them may interfere with its generation process. Therefore, we intentionally avoid giving any special instructions regarding

image tags in the prompt.

Human Checking and Regeneration. We performed author reviews of the generated images with a custom-built annotation tool. In these reviews, we checked that the generated text and images matched the originals exactly. As mentioned above, controlling image tags in the question text is difficult, so we allowed variations in the tags as long as the generated item remained a valid question.

In the first review round, 71% of the questions passed. The remaining 29% failed primarily because the question image had been replaced with an unrelated image, the text within the image was unreadable, parts of the question text were missing or incorrect, or the generated image was visually unnatural. These examples are shown in Figure B. For these failed cases, we regenerated the images with the same prompt or prompts with minor prompt adjustments. After completing the full set of VQA questions across several rounds, we performed a final cross-check to eliminate inconsistencies in the authors’ evaluation standards.

Manual Construction. We manually created 67 samples that Nano Banana Pro had difficulty generating. These cases had the following characteristics: 1. long question text (16 samples), 2. small or difficult-to-render text within the question image (36 samples), 3. extreme aspect ratios (2 samples), 4. domains that are inherently difficult to generate, such as chemical formulas or musical notation (8 samples), and 5. cases rejected by Nano Banana Pro due to policy constraints (5 samples). These examples are shown in Figure C.

4 Experiments

4.1 Setup

Baseline LMMs. For a more comprehensive evaluation, we assess a diverse set of state-of-the-art LMMs. In particular, for open-source models, we select representative models from three categories, English-centric LMMs, multilingual LMMs, and Japanese LMMs, to ensure that our evaluation accurately captures current progress in each subfield. We mainly use LMMs-Eval (Zhang et al., 2025) for our experiments. We set the temperature to 0 for open-source LMMs (the default setting for closed-source LMMs), and set max_tokens to be configured to be long enough so that the response would not be cut off. We ran with a single A100 80G GPU.

Closed-source LMMs: GPT-5.2 (OpenAI, 2025), Gemini3Pro (DeepMind, 2025b).

English-Centric Open-source LMMs: LLaVA-OneVision-7B (Li et al., 2024a), LLaVA-OneVision-1.5-8B (An et al., 2025), InternVL2.5-8B (Chen et al., 2024)

Multilingual Open-source LMMs: Qwen2.5VL-7B, 32B (Bai et al., 2025b), Qwen3VL-8B (Bai et al., 2025a), Phi-4Multimodal (Microsoft, 2025), Pangea-7B (Yue et al., 2025a), Aya Vision-8B (Dash et al., 2025)

Japanese Open-source LMMs: Sarashina2-Vision-8B and 14B (SB Intuitions, 2025a), Sarashina2.2-Vision-3B (SB Intuitions, 2025b), Heron-NVILA-Lite-15B (Turing Inc., 2025)

Inference Prompt. The inference prompt is based on the setup in JMMMU (Onohara et al., 2025) and MMMU-Pro (Yue et al., 2025b). Following MMMU-Pro (Yue et al., 2025b), we evaluate the open-source LMMs with both Direct and CoT prompts (as shown in Section D), and report the higher ones in the overall results. For the closed-source LMMs, they perform reasoning regardless of the prompt types, so we use only the Direct Prompt. Full results are shown in Section E.

4.2 Main Results

We present the experimental results in Table 1. The key findings from these results are as follows.

F1. All open-source LMMs struggle significantly on JMMMU-Pro. Open-source LMMs perform poorly on JMMMU-Pro, with the best model, Qwen2.5-VL-32B, achieving only 56.14, indicating substantial room for improvement. Furthermore, nine models perform less than 32%, close to random guessing. These results highlight that JMMMU-Pro poses a challenging and valuable benchmark for evaluating and advancing open-source LMMs.

F2. Most open-source LMMs exhibit a significant performance drop compared to JMMMU. Most open-source LMMs, except for Qwen2.5-VL-7B, show a substantial decline in accuracy on JMMMU-Pro relative to JMMMU. Moreover, when we compare the CS and CA subsets, we find that models with a clear performance gap between the JMMMU’s two subsets are similarly low on both in JMMMU-Pro. This suggests that their weakness lies in a fundamental lack of vision-side understanding, rather than in the type of question. These results demonstrate that JMMMU-Pro provides valuable feedback to model developers when

Model	JMMMU-Pro (1320)	JMMMU (1320)	CS Pro (600)	CS (600)	CA Pro (720)	CA (720)
Random						
Random Choice	27.05	27.05	26.33	26.33	27.64	27.64
Frequent Choice	27.73	27.73	25.33	25.33	29.72	29.72
Multilingual Open LMMs						
Qwen2.5-VL-32B	56.14	61.89	54.67	62.83	57.36	61.11
Qwen3-VL-8B	47.27	52.88	47.50	55.83	47.08	50.42
Qwen2.5-VL-7B	45.00	47.65	46.67	54.00	43.61	42.36
Phi-4-multimodal	31.82	39.55	28.83	38.00	34.31	40.83
Aya-Vision-8B	26.74	37.73	27.00	40.33	26.53	35.56
Pangea-7B	23.41	37.50	21.67	47.17	24.86	29.44
English-centric Open LMMs						
LLaVA-OV-1.5-8B	31.97	51.74	28.00	53.33	35.28	50.42
LLaVA-OV-7B	27.35	41.14	26.50	43.83	28.06	38.89
InternVL2.5-8B	31.21	41.36	29.00	43.33	33.06	39.72
Japanese Open LMMs						
Sarashina2.2-V-3B	42.88	47.95	54.00	61.50	33.61	36.67
Sarashina2-V-14B	30.68	37.27	32.33	43.17	29.31	32.36
Sarashina2-V-8B	27.88	39.62	27.00	51.00	28.61	30.14
Heron-NVILA-Lite-15B	26.97	50.15	26.67	59.17	27.22	42.64
Closed LMMs						
Gemini3Pro (reasoning high)	87.04	89.77	95.00	95.00	80.42	85.42
GPT-5.2 (reasoning high)	83.33	84.47	88.33	85.50	79.17	83.61

Table 1: **Main Results on JMMMU-Pro and JMMMU.** Overall, open-source LMMs show substantial performance degradation on JMMMU-Pro compared to JMMMU, while closed-source LMMs maintain strong performance, highlighting a significant gap in integrated Japanese visual-textual understanding.

used in comparison with JMMMU.

F3. Closed-source LMMs achieve substantially higher performance on JMMMU-Pro, revealing a substantial gap relative to open-source models. Closed-source LMMs obtain notably high scores on JMMMU-Pro. This indicates that these models already possess the ability to seamlessly integrate visual and textual information and interpret them through visual perception. Importantly, the strong performance of closed-source models does not diminish the value of JMMMU-Pro. Instead, it highlights the crucial role of JMMMU-Pro as a benchmark for guiding the development of open-source LMMs. Given the considerable performance gap between closed-source and open-source LMMs, reducing this gap is an essential goal for the community.

5 Analysis

5.1 Impact of CoT Prompting

We examine the effectiveness of Chain-of-Thought (CoT) prompting on JMMMU-Pro and JMMMU. The results are shown in Figure 4. These results indicate that the effectiveness of CoT varies depending on the model and the evaluation setting for both JMMMU-Pro and JMMMU. For example, on JMMMU-Pro, 8 out of the 13 LMMs

achieve higher performance with CoT prompting, whereas on JMMMU, only 3 models benefit from CoT. Moreover, when examined on a per-model basis, LMMs such as Pangea-7B, LLaVA-OV-1.5-8B, InternVL2.5-8B, and Sarashina2.2-V-3B show different prompt preferences between JMMMU and JMMMU-Pro. These findings suggest that optimal prompting strategies must be tailored to each model and each task, rather than relying on a single prompting approach across settings.

5.2 Correlation with OCR Performance

We hypothesize that the primary cause of performance degradation on JMMMU-Pro is the inability of current LMMs to perform Japanese Optical Character Recognition (OCR). To examine this hypothesis, we compute the correlation between OCR performance and JMMMU-Pro accuracy across several LMMs.

Following the evaluation setting of JMMMU-Pro, we ask each LMM to extract the full text of the question and all answer choices, excluding any text from associated images. OCR accuracy is then calculated by comparing the extracted text with the original text using the Levenshtein distance, which measures the edit distance between two strings. The similarity between the extracted and original

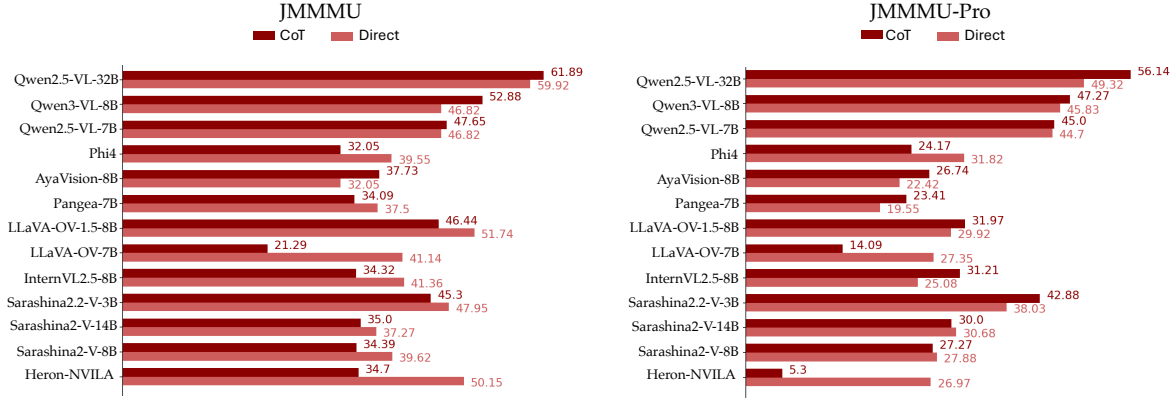


Figure 4: Impact of CoT prompting in JMMMU and JMMMU-Pro.

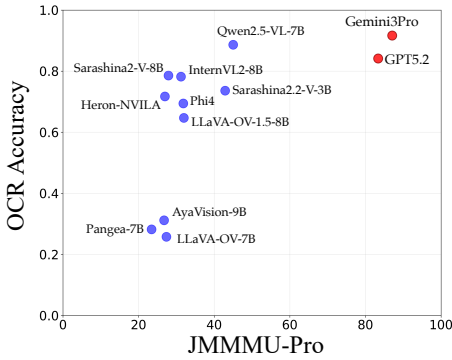


Figure 5: Correlation between OCR accuracy and JMMMU-Pro performance.

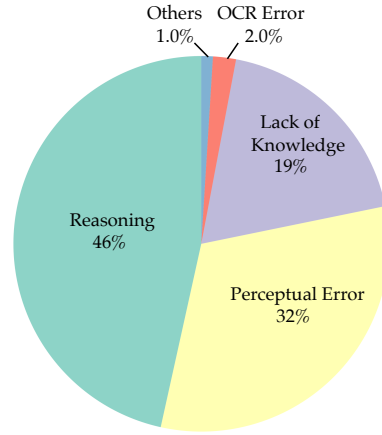


Figure 6: Distribution of 100 annotated Qwen2.5VL-32B errors.

text is computed as:

$$\text{OCR Accuracy} = 1 - \frac{\text{Levenshtein}(\text{text}_1, \text{text}_2)}{\max(\text{len}(\text{text}_1), \text{len}(\text{text}_2))} \quad (1)$$

The results are shown in Figure 5. The correlation coefficient between OCR accuracy and JMMMU-Pro accuracy is 0.593. There is a positive correlation between the two. However, high OCR ability does not necessarily translate directly into high JMMMU-Pro accuracy. For example, while Heron-NVILA and Sarashina2.2-V are comparable for OCR performance, the performance for JMMMU-Pro differs a lot. This indicates that solving JMMMU-Pro demands the ability to interpret and reason over language and vision in an integrated manner through visual perception.

5.3 Error Analysis

We analyze error cases from Qwen2.5VL-32B in the JMMMU-Pro to better understand the error reasons. We select 100 cases where the model makes a mistake on the JMMMU-Pro question but

answers the corresponding JMMMU question correctly. Consistent with MMMU-Pro findings, the errors are broadly categorized into three main types: perception errors, knowledge errors, and reasoning errors. We present representative failure cases of this model in Figure D-Figure G. This result suggests that JMMMU-Pro demands deeper, more integrated visual-textual understanding that goes beyond simple OCR capabilities.

6 Conclusion

This paper introduces JMMMU-Pro, a text-embedded Japanese Multi-discipline Multimodal Understanding Benchmark, along with Vibe Benchmark Construction, a scalable methodology for creating such a benchmark. Our experiments show that all open-source LMMs face significant difficulty on JMMMU-Pro, highlighting its importance as a benchmark that can inspire future progress in the open-source community.

564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614

Limitations

Extension to Other Languages. Although the proposed benchmark construction framework is, in principle, language-agnostic, this work focuses exclusively on Japanese. It remains unclear whether the same approach can be directly extended to other languages, as its effectiveness largely depends on the capabilities of the underlying image generation model (e.g., NanoBanana Pro). Nevertheless, we argue that focusing on Japanese alone is already valuable, given the current lack of rigorous benchmarks for evaluating integrated visual–textual understanding in Japanese and the growing demand for high-quality Japanese LMM evaluation.

Unconstructable Cases in Vibe Benchmark Construction. While Vibe Benchmark Construction enables scalable and cost-effective benchmark creation, it does not successfully handle all samples. In certain cases, such as those involving extremely long texts, small or complex visual details, it remains difficult for image generation models and requires manual construction. This highlights that human oversight remains a necessary component and that fully automated benchmark construction is still challenging.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as i can, not as i say: Grounding language in robotic affordances. In *CoRL*.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, and 1 others. 2025. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*.

Jeonghun Baek, Akiko Aizawa, and Kiyoharu Aizawa. 2025a. Harnessing pdf data for improving japanese large multimodal models. In *ACL Findings*.

Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. 2025b. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding. *arXiv preprint arXiv:2505.20298*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng

Li, and 45 others. 2025a. Qwen3-v1 technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-v1 technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

CohereLabs. 2024. *c4ai-command-r7b-12-2024*. Accessed: 2025-11-29.

Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, and 1 others. 2025. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*.

Google DeepMind. 2025a. *Gemini 3 pro image (nano banana pro)*. Web page. Accessed: 2025-11-29.

Google DeepMind. 2025b. *Gemini3*. Web page. Accessed: 2025-12-10.

Zeyu He, Saniya Naphade, and Ting-Hao Kenneth Huang. 2025. Prompting in the dark: Assessing human performance in prompt engineering for data labeling when gold labels are absent. In *CHI*.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *ICML*.

Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. 2024. Heronbench: A benchmark for evaluating vision language models in japanese. In *CVPR workshop*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *TMLR*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. 2024c. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

668	Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge . Accessed: 2025-11-29.	719
669		720
670		
671		
672	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In <i>NeurIPS</i> .	
673		
674	Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024c. Mmbench: Is your multi-modal model an all-around player? In <i>ECCV</i> .	
675		
676		
677		
678		
679	Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024d. Ocr-bench: on the hidden mystery of ocr in large multimodal models. <i>Science China Information Sciences</i> , 67(12):220102.	
680		
681		
682		
683		
684		
685	Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In <i>ICLR</i> .	
686		
687		
688		
689		
690	Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021. Docvqa: A dataset for vqa on document images. In <i>WACV</i> .	
691		
692		
693	Microsoft. 2025. Phi-4-multimodal-instruct . Hugging Face Model Repository. Accessed: 2025-11-29.	
694		
695	Atsuyuki Miyai, Jingkan Yang, Jingyang Zhang, Yifei Ming, Qing Yu, Go Irie, Yixuan Li, Hai Helen Li, Ziwei Liu, and Kiyoharu Aizawa. 2025. Unsolvable problem detection: Robust understanding evaluation for large multimodal models. In <i>ACL</i> .	
696		
697		
698		
699		
700	Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. Jdocqa: Japanese document question answering dataset for generative language models. In <i>LREC-COLING</i> .	
701		
702		
703		
704	Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. 2025. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In <i>NAACL</i> .	
705		
706		
707		
708		
709		
710	OpenAI. 2024a. Gpt-4o.	
711	OpenAI. 2024b. Image generation api . Accessed: 2025-02-06.	
712		
713	OpenAI. 2025. Introducing gpt-5 . Web page. Accessed: 2025-11-29.	
714		
715	OpenAI. 2025. Introducing gpt-5.2 . Web page. Accessed: 2025-12-13.	
716		
717	SakanaAI. 2024a. Ja-multi-image-vqa . Accessed: 2025-11-29.	
718		
	SakanaAI. 2024b. Ja-vlm-bench-in-the-wild . Accessed: 2025-11-29.	719
		720
	Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. 2025. Constructing multimodal datasets from scratch for rapid development of a Japanese visual language model. In <i>NAACL: Human Language Technologies (System Demonstrations)</i> .	721
		722
		723
		724
		725
		726
	SB Intuitions. 2025a. Sarashina2-vision-14b . Accessed: 2025-11-29.	727
		728
	SB Intuitions. 2025b. Sarashina2.2-vision-3b . Accessed: 2025-11-29.	729
		730
	Chirag Shah. 2024. From prompt engineering to prompt science with human in the loop. <i>arXiv preprint arXiv:2401.04122</i> .	731
		732
		733
	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 <i>COLING</i> .	734
		735
		736
		737
	Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In <i>CVPR</i> .	738
		739
		740
		741
	NAVER Cloud HyperCLOVA X Team. 2025. Hyperclova x think technical report. <i>arXiv preprint arXiv:2506.22403</i> .	742
		743
		744
	Qwen Team and 1 others. 2024. Qwen2 technical report. <i>arXiv preprint arXiv:2407.10671</i> , 2(3).	745
		746
	Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In <i>NeurIPS</i> .	747
		748
		749
		750
		751
		752
	Turing. 2024a. Llava-bench-in-the-wild (japanese) . Accessed: 2025-11-29.	753
		754
	Turing. 2024b. Llava-bench-ja . Accessed: 2025-11-29.	755
	Turing Inc. 2025. Heron-nvila-lite-15b . Accessed: 2025-11-29.	756
		757
	Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. <i>arXiv preprint arXiv:2409.12191</i> .	758
		759
		760
		761
		762
		763
	Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024b. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. <i>arXiv preprint arXiv:2406.14194</i> .	764
		765
		766
		767
		768

769 Yisong Xiao, Xianglong Liu, QianJia Cheng, Zhenfei
770 Yin, Siyuan Liang, Jiapeng Li, Jing Shao, Aishan
771 Liu, and Dacheng Tao. 2025. Genderbias-vl: Bench-
772 marking gender bias in vision language models via
773 counterfactual probing: Y. xiao et al. *IJCV*, pages
774 1–24.

775 Maoyuan Ye, Jing Zhang, Juhua Liu, Bo Du, and
776 Dacheng Tao. 2025. Logicocr: Do your large multi-
777 modal models excel at logical reasoning on text-rich
778 images? *arXiv preprint arXiv:2505.12307*.

779 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng,
780 Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang,
781 Weiming Ren, Yuxuan Sun, and 1 others. 2024.
782 Mmmu: A massive multi-discipline multimodal under-
783 standing and reasoning benchmark for expert agi.
784 In *CVPR*.

785 Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim,
786 Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kan-
787 tharuban, Lintang Sutawika, Sathyanarayanan Ra-
788 mamoothy, and Graham Neubig. 2025a. Pangea: A
789 fully open multilingual multimodal llm for 39 lan-
790 guages. In *ICLR*.

791 Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang,
792 Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu,
793 Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Gra-
794 ham Neubig. 2025b. Mmmu-pro: A more robust
795 multi-discipline multimodal understanding bench-
796 mark. In *ACL*.

797 Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu,
798 Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuan-
799 han Zhang, Jingkang Yang, Chunyuan Li, and 1 oth-
800 ers. 2025. Lmms-eval: Reality check on the evalu-
801 ation of large multimodal models. In *NAACL Find-*
802 *ings*.

803 Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu,
804 Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan
805 Welker, Azyaan Wahid, Quan Vuong, Vincent Van-
806 houcke, Huong Tran, Radu Soricut, Anikait Singh,
807 Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi,
808 Grecia Salazar, and 35 others. 2023. Rt-2: Vision-
809 language-action models transfer web knowledge to
810 robotic control. In *CoRL*.

Appendix

In this appendix, we include the dataset distribution in Section A, failure cases of image generation and manual construction examples in Section B, the verified JMMMU in Section C, detailed prompt examples in Section D, the full results in Section E, and failure cases of LMM inference in Section F.

A Dataset Distributions

We show the statistics for the two most controllable factors: Background and Background Color in Figure A.

B Image Generation Failures and Manual Construction Examples

B.1 Image Generation Failures in Nano Banana Pro

Figure B presents examples of image generation failures observed in Nano Banana Pro. Nano Banana Pro can occasionally produce failures such as those shown in the figure. We attribute these failures to the inherent diversity of outputs produced by generative models. Therefore, it is crucial to manually inspect and filter such outputs to ensure correctness.

B.2 Manual Construction Examples

Figure C shows examples that were manually constructed. Images with these characteristics were inherently difficult to generate automatically using Nano Banana Pro. Therefore, we found that not all images can be created through the Vibe Benchmark Construction pipeline.

C Verifying the original JMMMU Benchmark

We first verify the existing JMMMU benchmark by correcting its samples and refining its evaluation protocol.

Sample Corrections. Although most questions in JMMMU are multiple-choice, 50 questions in the culture-agnostic subset, which translated directly from MMMU, are open-ended. Prior work (SB Intuitions, 2025b) has noted that these open-ended questions introduce additional complexity in interpreting evaluation results, and thus converted them into multiple-choice format for their experiments. Following this approach, we also convert all open-ended questions in JMMMU into multiple-choice questions. We provide the problem statement and the correct answer to an LLM (e.g., GPT-5 (OpenAI, 2025)) and instruct it to propose plausible false choices. We then manually verify the generated choices to ensure quality, correcting cases where false options might unintentionally match the correct answer due to numerical precision, ambiguity, or formatting issues. Additionally, we identified and corrected errors in two samples (test_Japanese_Art_120, validation_Agriculture_1), where the answer or question text contained mistakes.

Revision of the Evaluation Procedure. We also revised the evaluation procedure in JMMMU. First, we found that the answer-parsing algorithm used in JMMMU often fails on recent reasoning models, whose outputs tend to be longer. In particular, when models list all options as part of their reasoning, the parser may incorrectly extract the predicted

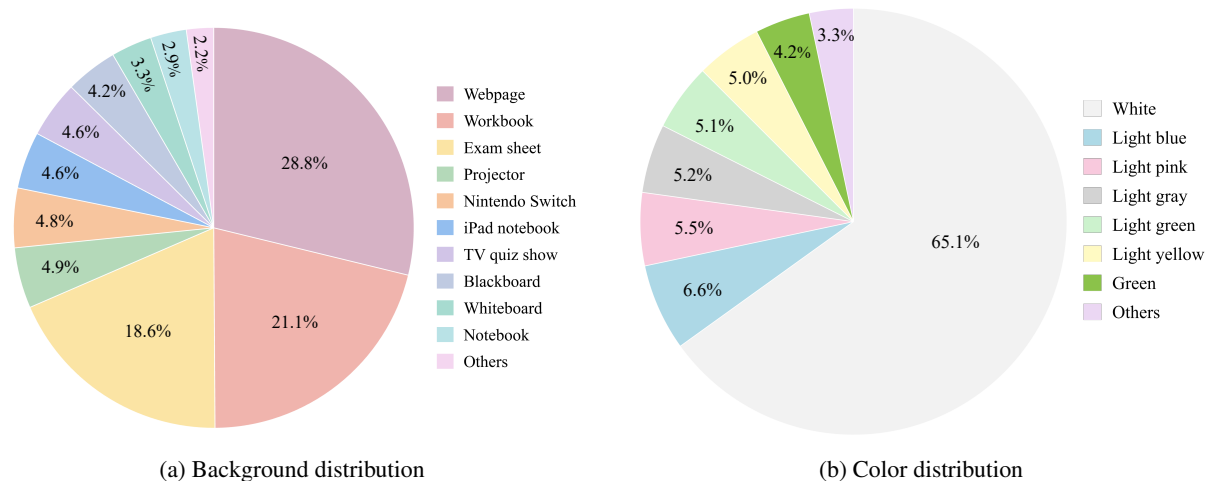


Figure A: Distribution of background and its color in the JMMMU-Pro benchmark.

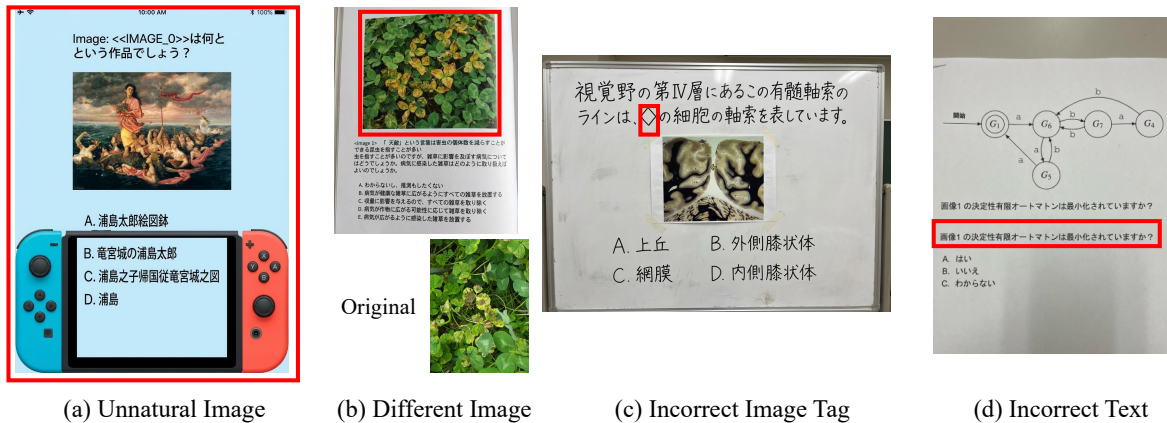


Figure B: Failure examples for Nano Banana Pro.

choice. To address this, we modify the parser to ignore such auxiliary option listings. In addition, when parsing fails, JMMMU selects an answer at random. We remove this random selection step, as it obscures whether a model has genuinely failed to produce a valid option. Our modification instead marks such cases explicitly as incorrect, allowing a clearer distinction between invalid output and an actual wrong answer.

To avoid confusion in the community, we name this corrected version of the benchmark JMMMU-verified-2025-12. All JMMMU scores reported in this paper are based on JMMMU-verified-2025-12.

D Prompt

D.1 LMM Inference Prompt

Below, we present the prompt used for LMM inference.

Inference Prompts: Direct

JMMMU:
 “与えられた選択肢の中から最も適切な回答のアルファベットを直接記入してください。” (Answer with the option’s letter from the given choices directly.)

JMMMU-Pro:
 “あなたの役割は画像で与えられた問題を解くことです。与えられた選択肢の中から最も適切な回答のアルファベットを直接記入してください。” (Your role is to solve the question provided in the image. Answer with the option’s letter from the given choices directly.)

Inference Prompts: CoT

JMMMU:
 “次の選択式問題に答えてください。あなたの回答の最終行は、次の形式にしてください：'答え: \$LETTER' (引用符なし)。ここで \$LETTER は選択肢のいずれかの文字です。答える前に step by step で考えてください。” (Answer the following multiple choice question. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of options. Think step by step before answering.)

JMMMU-Pro:
 “あなたの役割は画像で与えられた問題を解くことです。あなたの回答の最終行は、次の形式にしてください：'答え: \$LETTER' (引用符なし)。ここで \$LETTER は選択肢のいずれかの文字です。答える前に step by step で考えてください。” (Your role is to solve the question provided in the image. The last line of your response should be of the following format: 'Answer: \$LETTER' (without quotes) where LETTER is one of options. Think step by step before answering.)

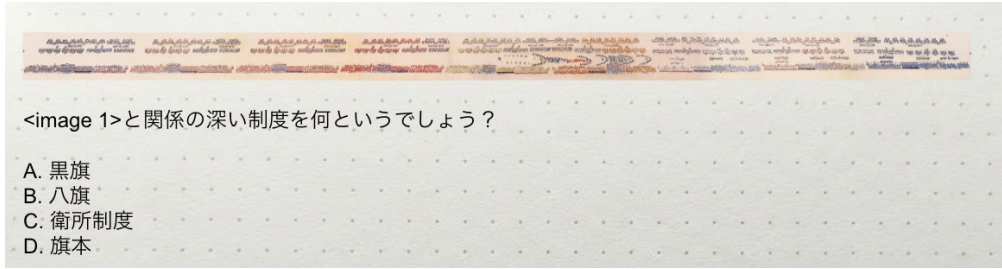
Inference Prompts: OCR Task

OCR Task Prompt:
 “画像内の選択式(多肢選択式)問題から、導入説明文を含む質問文全体と、対応する選択肢を抽出して出力してください。関連画像のテキストや問題番号は除外してください。OCRのみを行い、問題を解こうとしないでください。Formatは、'Question: , Options: ' (引用符なし)の形で出力してください” (Extract and output the full text of the question, including any introductory descriptions, as well as the corresponding answer choices from the multiple-choice question in the image. Exclude any text from associated images or the question number. Perform OCR only; do not attempt to solve the question. Please output in the format: 'Question: , Options: ' (without quotes).)

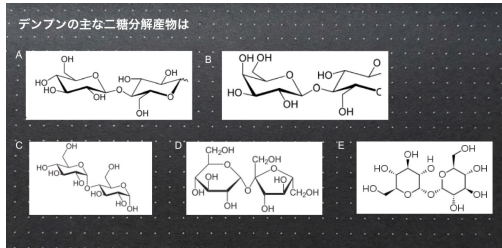
D.2 Image Generation Prompt
 Below, we present the prompts used to generate images with Nano Banana Pro. The base prompt

875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893

894
895
896
897
898



(a) Extreme aspect ratio



(b) Difficult domain for generation



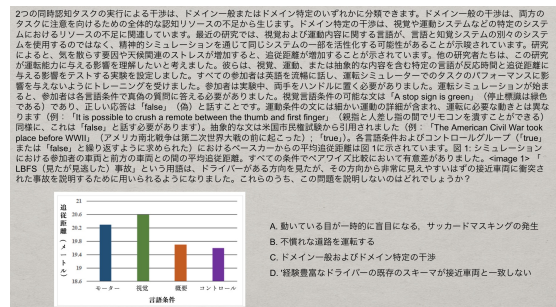
(c) Rejected sample by the policy

表8は、20店舗の小売チェーンにおける総売上額（千ドル単位）、1日の平均売上数、顧客訪問の平均滞在時間（分単位）を含む3か月間の売上データを示しています。<image 1>表に示された情報に基づき、次の文の真偽を判定してください。文：5月の平均売上額が最も低い店舗の総売上額は、その月の平均売上額が最も高い店舗よりも高い。

店舗	1月	2月	3月	4月	5月	6月
Ashford	\$2000	\$2000	\$2000	4000	2000	1000
Birmingham	\$1000	\$1000	\$1000	4000	2000	3000
Cardiff	\$2000	\$2000	\$2000	4000	2000	1000
Dagen	\$2000	\$2000	\$2000	4000	2000	1000
Edlington	\$2000	\$2000	\$2000	4000	2000	1000
Farnham	\$2000	\$2000	\$2000	4000	2000	1000
Gloucester	\$2000	\$2000	\$2000	4000	2000	1000
Harrington	\$2000	\$2000	\$2000	4000	2000	1000
Inverness	\$2000	\$2000	\$2000	4000	2000	1000
Jackson	\$2000	\$2000	\$2000	4000	2000	1000
Killingly	\$2000	\$2000	\$2000	4000	2000	1000
Leamfield	\$2000	\$2000	\$2000	4000	2000	1000
Malvern	\$2000	\$2000	\$2000	4000	2000	1000
Newfield	\$2000	\$2000	\$2000	4000	2000	1000
Oxton	\$2000	\$2000	\$2000	4000	2000	1000
Princesville	\$2000	\$2000	\$2000	4000	2000	1000
Queensland	\$2000	\$2000	\$2000	4000	2000	1000
Riverhead	\$2000	\$2000	\$2000	4000	2000	1000
Southland	\$2000	\$2000	\$2000	4000	2000	1000
Tandee	\$2000	\$2000	\$2000	4000	2000	1000

A. 真
B. 偽

(d) Small text within question image



(e) Long question text

Figure C: Manual construction examples.

is designed for the single-image setting, while for the multiple-image setting, the prompt is modified depending on whether images are included in the options.

E Full Results

In Table A and Table B, we present detailed results for both the Direct Prompt and the CoT Prompt.

F Failure Cases

We show representative failure cases of the state-of-the-art open-source LMM, Qwen2.5-VL-32B, in Figure D, Figure E, Figure F, and Figure G. These examples show that the model exhibits various errors in JMMM-Pro, indicating that JMMM-Pro requires deeper integrated visual-textual understanding beyond mere OCR capability.

Base prompt

Your role is to create an image-based question. You must not derive or provide the answer. Please create an image that looks as if the pictures, text, and options were actually placed or written on a surface, such as a notebook, a sheet of paper, a webpage, or other backgrounds, and then captured either by being photographed with a phone camera or taken as a smartphone/PC screenshot.

****Critical Instruction****

1. Insert the images exactly as they are.
2. Do not change the character of any text in the Image or the Question. Keep the original character exactly as it appears.
3. Do not derive the answer. Your task is to create an image-based question.
4. Make sure the pasted images blend naturally into the notebook background while still retaining a subtle "pasted" feel. For cases such as a blackboard or a notebook, keep a slight pasted effect. For things like a webpage, a projector screen, or printed material, make the pasted images blend in more naturally without an obvious pasted look, while still keeping the boundary between the Question and the Image clearly distinguishable.
5. Please make the font of the text within the images follow the original font in the given image as closely as possible. The font for the question and option text should follow the instructions provided below.
6. Paste the image without making any edits.

****More Detailed Conditions****

1. The image should reflect {state}.
2. The font used in the question and options should be {font}.
3. The background should be {background}.
4. The background color should be {color}.
5. The marginal space should be {margin}.

Image:

As attached

Question:

{question}

{options}

Prompt for multiple images (wo option images)

Your role is to create an image-based question. You must not derive or provide the answer. Please create an image that looks as if the pictures, text, and options were actually placed or written on a surface, such as a notebook, a sheet of paper, a webpage, or other backgrounds, and then captured either by being photographed with a phone camera or taken as a smartphone/PC screenshot.

****Critical Instruction****

1. Insert the images exactly as they are.
2. Do not change the character of any text in the Image or the Question. Keep the original character exactly as it appears.
3. Do not derive the answer. Your task is to create an image-based question.
4. Make sure the pasted images blend naturally into the notebook background while still retaining a subtle "pasted" feel. For cases such as a blackboard or a notebook, keep a slight pasted effect. For things like a webpage, a projector screen, or printed material, make the pasted images blend in more naturally without an obvious pasted look, while still keeping the boundary between the Question and the Image clearly distinguishable.
5. Please make the font of the text within the images follow the original font in the given image as closely as possible. The font for the question and option text should follow the instructions provided below.
6. Paste the image without making any edits.
7. When there are multiple images, place them from left to right, one by one.

****More Detailed Conditions****

1. The image should reflect {state}.
2. The font used in the question and options should be {font}.
3. The background should be {background}.
4. The background color should be {color}.
5. The marginal space should be {margin}.

Image:

As attached

Question:

{question}

{options}

Prompt for image options

Your role is to create an image-based question. You must not derive or provide the answer. Please create an image that looks as if the pictures, text, and options were actually placed or written on a surface, such as a notebook, a sheet of paper, a webpage, or other backgrounds, and then captured either by being photographed with a phone camera or taken as a smartphone/PC screenshot.

****Critical Instruction****

1. Insert the images exactly as they are.
2. Do not change the character of any text in the Image or the Question. Keep the original character exactly as it appears.
3. Do not derive the answer. Your task is to create an image-based question.
4. Make sure the pasted images blend naturally into the notebook background while still retaining a subtle “pasted” feel. For cases such as a blackboard or a notebook, keep a slight pasted effect. For things like a webpage, a projector screen, or printed material, make the pasted images blend in more naturally without an obvious pasted look, while still keeping the boundary between the Question and the Image clearly distinguishable.
5. Please make the font of the text within the images follow the original font in the given image as closely as possible. The font for the question and option text should follow the instructions provided below.
6. Paste the image without making any edits.
7. For the <image> tags in options (e.g., <image 1>, <image 2>), please replace each tag with the actual image. The images correspond to the tags in order, with the first image being <image 1>, the second <image 2>, and so on.

****More Detailed Conditions****

1. The image should reflect {state}.
2. The font used in the question should be {font}.
3. The background should be {background}.
4. The background color should be {color}.
5. The marginal space should be {margin}.

Image:

As attached

Question:

{question}

{options}

Model	JMMMU-Pro (1320)	JMMMU (1320)	CS Pro (600)	CS (600)	CA Pro (720)	CA (720)
Random						
Random Choice	27.05	27.05	26.33	26.33	27.64	27.64
Frequent Choice	27.73	27.73	25.33	25.33	29.72	29.72
Multilingual Open LMMs						
Qwen2.5-VL-32B	49.32	59.92	51.67	63.50	47.36	56.94
Qwen3-VL-8B	45.83	46.82	47.00	56.33	44.86	38.89
Qwen2.5-VL-7B	44.70	46.82	50.17	57.83	40.14	37.64
Phi-4-multimodal	31.82	39.55	28.83	38.00	34.31	40.83
Aya-Vision-8B	22.42	32.05	23.83	40.67	21.25	24.86
Pangea-7B	19.55	37.50	23.00	47.17	16.67	29.44
English-centric Open LMMs						
LLaVA-OV-1.5-8B	29.92	51.74	26.33	53.33	32.92	50.42
LLaVA-OV-7B	27.35	41.14	26.50	43.83	28.06	38.89
InternVL2.5-8B	25.08	41.36	23.83	43.33	26.11	39.72
Japanese Open LMMs						
Sarashina2.2-V-3B	38.03	47.95	40.17	61.50	36.25	36.67
Sarashina2-V-14B	30.68	37.27	32.33	43.17	29.31	32.36
Sarashina2-V-8B	27.88	39.62	27.00	51.00	28.61	30.14
Heron-NVILA-Lite-15B	26.97	50.15	26.67	59.17	27.22	42.64
Closed LMMs						
Gemini3Pro (reasoning high)	87.04	89.77	95.00	95.00	80.42	85.42
GPT-5.2 (reasoning high)	83.33	84.47	88.33	85.50	79.17	83.61

Table A: Results with the direct prompt.

Model	JMMMU-Pro (1320)	JMMMU (1320)	CS Pro (600)	CS (600)	CA Pro (720)	CA (720)
Random						
Random Choice	27.05	27.05	26.33	26.33	27.64	27.64
Frequent Choice	27.73	27.73	25.33	25.33	29.72	29.72
Multilingual Open LMMs						
Qwen2.5-VL-32B	56.14	61.89	54.67	62.83	57.36	61.11
Qwen3-VL-8B	47.27	52.88	47.50	55.83	47.08	50.42
Qwen2.5-VL-7B	45.00	47.65	46.67	54.00	43.61	42.36
Phi-4-multimodal	24.17	32.05	22.00	31.50	25.97	32.50
Aya-Vision-8B	26.74	37.73	27.00	40.33	26.53	35.56
Pangea-7B	23.41	34.09	21.67	36.17	24.86	32.36
English-centric Open LMMs						
LLaVA-OV-1.5-8B	31.97	46.44	28.00	46.83	35.28	46.11
LLaVA-OV-7B	14.09	21.29	14.33	18.00	13.89	24.03
InternVL2.5-8B	31.21	34.32	29.00	39.67	33.06	29.86
Japanese Open LMMs						
Sarashina2.2-V-3B	42.88	45.30	54.00	59.00	33.61	33.89
Sarashina2-V-14B	30.00	35.00	30.50	44.50	29.58	27.08
Sarashina2-V-8B	27.27	34.39	25.33	40.83	28.89	29.03
Heron-NVILA-Lite-15B	5.30	34.70	1.00	38.33	8.89	31.67

Table B: Results with the CoT prompting.

10:28 3月20日 水曜

食品	食品を食べた人の数	食品を食べ体調を壊した人の数
冷製チキン	86	34
ポテトサラダ	54	38
卵サンドイッチ	76	40
フルーツパイとクリーム	32	12
チーズ	48	12

<image 1> の表によると、食中毒の発生原因として最も可能性が高い食品はどれですか？

A. 冷製チキン
B. ポテトサラダ
C. 卵サンドイッチ
D. フルーツパイとクリーム

Qwen2.5VL-32B (JMMMU CoT) ✓

問題分析 この問題では、与えられた表に基づいて、食中毒の発生原因として最も可能性が高い食品を特定する必要があります。表には以下の情報が含まれています： - **食品を食べた人の数**：各食品を食べた人数。 - **食品を食べ体調を壊した人の数**：各食品を食べた後に体調を壊した人数。食中毒の発生可能性を評価するためには、**体調を壊した人の割合**（すなわち、体調を壊した人数 ÷ 食べた人数）を計算し、その値が最も高い食品を選択するのが合理的です。 ### ステップバイステップの解法 1. **各食品について体調を壊した人の割合を計算する**： - **冷製チキン**： $\frac{34}{86}$ - **ポテトサラダ**： $\frac{38}{54}$ - **卵サンドイッチ**： $\frac{40}{76}$ - **フルーツパイとクリーム**： $\frac{12}{32}$ - **チーズ**： $\frac{12}{48}$ 2. **割合を具体的な値に変換する**： - **冷製チキン**： $\frac{34}{86} \approx 0.395$ - **ポテトサラダ**： $\frac{38}{54} \approx 0.704$ - **卵サンドイッチ**： $\frac{40}{76} \approx 0.526$ - **フルーツパイとクリーム**： $\frac{12}{32} = 0.375$ - **チーズ**： $\frac{12}{48} = 0.25$ 3. **割合を比較して最も高い値を選ぶ**： - 冷製チキン：約 0.395 - ポテトサラダ：約 0.704 - 卵サンドイッチ：約 0.526 - フルーツパイとクリーム：0.375 - チーズ：0.25 これらの値を比較すると、**ポテトサラダ**の割合が最も高く、約 0.704 です。 ### 答え 食中毒の発生原因として最も可能性が高い食品は **ポテトサラダ** です。 答え：B

Qwen2.5VL-32B (JMMMU-Pro, CoT)

✗ Reasoning Error

問題の理解 表には以下の情報が記載されています： - **食品**：各食品の名前 - **食品を食べた人の数**：その食品を食べた人数 - **食品を食べ体調を壊した人の数**：その食品を食べた後に体調を壊した人数 食中毒の発生原因として最も可能性が高い食品を特定するためには、どの食品が最も多くの人々に影響を与えたかを判断する必要があります。そのため、各食品について「体調を壊した人の数」を比較します。 ### 表のデータ |食品|食品を食べた人の数|食品を食べ体調を壊した人の数||-----|-----||冷製チキン|86|34||ポテトサラダ|54|38||卵サンドイッチ|76|40||フルーツパイとクリーム|32|12||チーズ|48|12| ### ステップバイステップの分析 1. **冷製チキン**： - 食べた人数：86人 - 体調を壊した人数：34人 2. **ポテトサラダ**： - 食べた人数：54人 - 体調を壊した人数：38人 3. **卵サンドイッチ**： - 食べた人数：76人 - 体調を壊した人数：40人 4. **フルーツパイとクリーム**： - 食べた人数：32人 - 体調を壊した人数：12人 5. **チーズ**： - 食べた人数：48人 - 体調を壊した人数：12人 ### 比較 各食品について「体調を壊した人の数」を比較すると： - 冷製チキン：34人 - ポテトサラダ：38人 - 卵サンドイッチ：40人 - フルーツパイとクリーム：12人 - チーズ：12人 最も多くの人々が体調を壊した食品は **卵サンドイッチ** で、40人が体調を壊しています。 ### 答え 最も可能性が高い食品は **卵サンドイッチ** です。 答え：C

Analysis

The cause of the error is judging based solely on the actual number of people who became ill. In epidemiological investigations to identify the source of food poisoning, comparisons should not be made by the "number of people" alone, but by the "attack rate" (incidence rate)—which represents the percentage of people who developed symptoms among those who ate the food.

Figure D: Failure case: Reasoning error.



Qwen2.5VL-32B (JMMMMU, CoT) ✓

問題分析 この問題では、スクリーニング乳房MRIの画像から左乳房の所見の病因を特定する必要があります。与えられた画像は以下のもの：1. **差分処理画像** (Image 1) : MRIの差分画像は、異なる時点での信号強度の変化を強調します。2. **T2強調画像** (Image 2) : T2強調画像では、水分含有量が高い領域が明るく表示されます。3. **T1強調軸画像** (Image 3) : T1強調画像では、脂肪組織や血液などの信号強度が高く表示されます。### 画像の観察ポイント 1. **差分処理画像 (Image 1)** : 左乳房に明確な異常所見が認められる。- 強い信号強度の領域が存在し、これは病変の可能性を示唆しています。2. **T2強調画像 (Image 2)** : 同じ領域で明るい信号が確認される。- T2強調画像では、水分含有量が高い病変（例えば、嚢胞性病変や炎症性病変）が明るく表示される傾向があります。3. **T1強調軸画像 (Image 3)** : 左乳房の同じ領域で暗い信号が観察される。- T1強調画像では、脂肪組織は明るく、病変が暗い信号を示す場合、その病変は脂肪組織と比べて信号が低下していることを意味します。### 病因の候補について検討 ##### A. **磁化率アーチファクト** : 磁化率アーチファクトは金属や不均一な組織構造による信号の歪みを引き起こします。- しかし、この症例では明確な金属の存在やアーチファクトの特徴（ストライプ状のパターンなど）は見られません。- **可能性低い**。 ##### B. **血腫** : 血腫は出血による病変で、T1強調画像では初期には高信号を示し、その後低信号になることがあります。- ただし、T2強調画像でも高信号を示すことが一般的です。- この症例では、T1強調画像で低信号、T2強調画像で高信号という特徴が見られます。血腫の可能性はありますが、他の選択肢も考慮する必要があります。 ##### C. **脂肪壊死** : 脂肪壊死は、脂肪組織の壊死により生じる病変で、通常は外傷後に発生します。- MRI上では、T1強調画像で低信号、T2強調画像で高信号を示すことが特徴です。- この症例の画像所見は脂肪壊死の典型的な特徴に一致します。 ##### D. **シリコン肉芽腫** : シリコン肉芽腫は、シリコン注入による異物反応で形成される病変です。- MRI上では、T1強調画像で低信号、T2強調画像で高信号を示すことが一般的です。- ただし、シリコン肉芽腫は通常、シリコン注入の歴史がある患者で見られるため、その背景情報が必要です。- この症例では、シリコン注入の有無が不明であるため、確定は難しい。 ### 結論 画像所見 (T1強調画像で低信号、T2強調画像で高信号) は、**脂肪壊死**の特徴に最もよく一致します。また、他の選択肢 (磁化率アーチファクト、血腫、シリコン肉芽腫) については、画像所見や背景情報から脂肪壊死が最も可能性が高いと考えられます。 答え: **C**

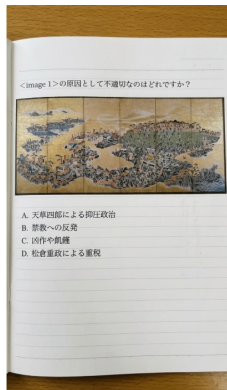
Qwen2.5VL-32B (JMMMMU-Pro, CoT) ✗ Perceptual Error

問題の理解 この問題では、スクリーニング乳房MRIの3種類の画像が提示されています：1. **差分処理画像 (image 1)** 2. **T2強調画像 (image 2)** 3. **T1強調軸画像 (image 3)** これらの画像から、左乳房の所見に基づいて病変の診断を選択する必要があります。選択肢は以下の通りです：- A. 磁化率アーチファクト - B. 血腫 - C. 脂肪壊死 - D. シリコン肉芽腫 ### 画像の分析 ##### 1. 差分処理画像 (image 1) - 差分処理画像は、異なる時点でのMRI画像の差分を示すもので、動的な変化を捉えます。- 画像には明確な高信号領域が見られます。これは病変部位の活性化や増殖性病変を示唆する可能性があります。 ##### 2. T2強調画像 (image 2) - T2強調画像では、水分含有量が高い組織が高信号を示します。- 画像には高信号の領域が見られます。これは病変部位が水分を多く含んでいることを示唆しています。 ##### 3. T1強調軸画像 (image 3) - T1強調画像では、脂肪組織が高信号を示し、病変が低信号になることが多いです。- 画像には低信号の領域が見られます。これは病変部位が脂肪組織よりも信号が低いことを意味し、病変の性質を特定する手がかりとなります。 ### 病変の特徴に基づく考察 1. **磁化率アーチファクト (A)** : 磁化率アーチファクトは金属や磁化率の異なる物質によるMRI画像の歪みを指します。- 画像からは明確なアーチファクトの特徴 (例：ストライプ状のパターンなど) は見られません。したがって、この選択肢は適切ではないと考えられます。 2. **血腫 (B)** : 血腫は出血による病変で、T2強調画像では高信号を示します (血液中の水分子が多いため)。- T1強調画像では初期は低信号ですが、時間が経過すると信号が変化します。- 提示された画像では、T2強調画像で高信号、T1強調画像で低信号の領域が見られるため、血腫の可能性は高いです。 3. **脂肪壊死 (C)** : 脂肪壊死は脂肪組織の壊死による病変で、通常はT1強調画像で高信号を示します。- 提示された画像では、T1強調画像で低信号の領域が見られるため、脂肪壊死の可能性は低いです。 4. **シリコン肉芽腫 (D)** : シリコン肉芽腫はシリコン注入による異物反応で形成される病変で、MRI画像では特有の信号パターンを示します。- 提示された画像からはシリコン肉芽腫の典型的な特徴 (例：境界が不明瞭な低信号領域など) は見られません。したがって、この選択肢も適切ではないと考えられます。 ### 結論 画像の特徴を総合的に考慮すると、T2強調画像での高信号とT1強調画像での低信号は血腫の典型的な特徴に一致します。したがって、最も適切な診断は **血腫** です。 答え: B

Analysis

The main cause of the error was a misinterpretation of the signal intensity of the left breast lesion. On the T1-weighted image (Image 3), the lesion does not show low signal intensity; instead, it appears **high signal (bright)**, similar to the surrounding fat, suggesting that its contents are fat. This finding is typical of **fat necrosis (oil cyst)**, which characteristically shows high signal intensity on T1-weighted images due to its fatty content. In contrast, a hematoma rarely shows such uniform signal intensity matching that of surrounding fat. Therefore, the correct answer is C

Figure E: Failure case: Perceptual error.



Qwen2.5VL-32B (JMMMU, CoT) ✓

この問題では、<image 1>の内容を理解し、その背景にある原因について考えます。まず、画像の内容を分析しましょう。### 画像の内容分析: 画像は、戦闘や集団行動が描かれた歴史的な場面を示しています。多くの人物が集まり、建物や旗が見えます。これは戦争や蜂起のような大規模な社会的・政治的事件を示唆しています。こうした描写は、歴史上の「天草四郎時貞の乱」（島原の乱）と一致すると考えられます。### 島原の乱の背景: 島原の乱（1637年～1638年）は、江戸時代初期に九州の島原半島で起こった農民・町人・キリシタン（キリスト教徒）による大規模な蜂起です。主な原因としては以下が挙げられます: 1. 禁教への反発** 徳川幕府によるキリスト教の弾圧政策に対する反発が大きなき要因でした。キリシタンたちは信仰の自由を求めて蜂起しました。2. 凶作や飢饉** 当時の九州地方では連続した凶作や飢饉が発生しており、人々の生活は困難していました。経済的な苦境も蜂起の背景にありました。3. 松倉重政による重税** 松倉重政が九州の領地で行った過酷な課税政策が、農民たちの不満を高めました。特に島原藩では重税が課せられ、人々の負担が増加していました。### 選択肢の検討: A. 天草四郎による抑圧政治** 天草四郎時貞は島原の乱の指導者ですが、彼自身が「抑圧政治」を行ったわけではありません。むしろ、彼は農民やキリシタンの不満を代表して蜂起を主導した人物です。この選択肢は不適切です。B. 禁教への反発** 島原の乱では、キリシタンたちが禁教政策に反発して蜂起したことが大きな要因の一つです。この選択肢は適切です。C. 凶作や飢饉** 当時の九州地方では連続した凶作や飢饉が発生しており、経済的な困窮が蜂起の背景にありました。この選択肢は適切です。D. 松倉重政による重税** 松倉重政の重税政策が農民たちの不満を高め、蜂起の一因となりました。この選択肢は適切です。### 結論: 不適切な選択肢はAです。天草四郎時貞は蜂起の指導者であり、彼自身が抑圧政治を行ったわけではありません。答え:A

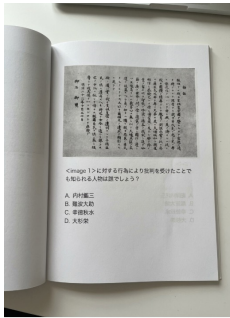
Qwen2.5VL-32B (JMMMU-Pro, CoT) ✗ Lack of Knowledge Error

問題分析 画像には、戦いや混乱が描かれている様子が見受けられます。この画像が示す出来事の原因として不適切なものをを選択する必要があります。そのため、各選択肢について検討し、画像と関連性がないものを特定します。---### 選択肢の内容確認 1. **A. 天草四郎による抑圧政治** - 天草四郎は、江戸時代にキリシタン迫害を背景に起こった「天草・島原の乱」に関連する人物です。この乱では、キリシタンや農民が幕府の抑圧政策に反発して蜂起しました。画像が戦いや混乱を描いている場合、この選択肢は適切である可能性があります。2. **B. 禁教への反発** - 禁教（キリスト教の禁止）は、江戸時代の幕府政策であり、特に「天草・島原の乱」の背景にもなっています。キリシタンたちが禁教政策に反発して蜂起したことが歴史的に知られています。この選択肢も適切である可能性があります。3. **C. 凶作や飢饉** - 凶作や飢饉は、農民が生活困窮に陥り、反乱や暴動を引き起こす要因となることがあります。歴史上、多くの反乱や飢饉を背景に起こっています。この選択肢も適切である可能性があります。4. **D. 松倉重政による重税** - 松倉重政は、江戸時代初期の武将で、大坂夏の陣後に大坂城代を務めました。彼は厳しい統治を行い、重税を課したことで不評を買ったとされています。しかし、松倉重政の活動は主に大坂周辺に限定されており、全国的な影響は限定的でした。また、彼の死後すぐにその政策も廃止されています。したがって、この選択肢は他の選択肢に比べて歴史的な影響力が弱いと言えます。---### 画像の内容との関連性 画像には戦いや混乱が描かれています。このような状況を引き起こす要因としては、**A. 天草四郎による抑圧政治**、**B. 禁教への反発**、**C. 凶作や飢饉**が考えられます。これらは歴史上、大きな反乱や戦争を引き起こす典型的な要因です。一方、**D. 松倉重政による重税**は、地域的な影響が限定的であり、全国的な戦いや混乱を引き起こすほどの影響力はないと考えられます。そのため、この選択肢は他の選択肢に比べて不適切である可能性が高いです。---### 結論 画像の内容と照らし合わせた結果、最も不適切な選択肢は**D. 松倉重政による重税**です。答え:D

Analysis

The primary cause of this error is a lack of knowledge regarding Amakusa Shiro. While the JMMMU model correctly identified that Amakusa Shiro Tokisada, the leader of the Shimabara Rebellion, did not personally carry out "oppressive politics," the JMMMU-Pro model failed to apply this correct knowledge in its judgment.

Figure F: Failure case: Lack of Knowledge error.



Qwen2.5VL-32B (JMMMU, CoT) ✓

問題分析 画像には以下の要素が含まれています：1. **手書き文書**：古風な日本語で書かれた文書が写真として表示されています。2. **質問文**：「<image 1>に対する行為により批判を受けたことでも知られる人物は誰でしょう？」とあります。3. **選択肢**：- A. 内村鑑三 - B. 難波大助 - C. 幸徳秋水 - D. 大杉栄 ### 手書き文書の内容について 手書き文書は古風な日本語で書かれており、内容を正確に読み取ることは難しいですが、以下のような特徴が見受けられます：- 文章の形式や用語から、歴史的な文書であることが伺えます。- 「明治二十三年十月三十日」という日付が記載されており、これは西暦1890年10月30日を指します。 ### 質問の解釈 質問は、「<image 1>に対する行為により批判を受けたことでも知られる人物」を求めています。このため、次のステップでは各選択肢の人物について検討し、その行動や時代背景を考慮して解答を導き出します。 --- ### 各選択肢の検討 #### A. 内村鑑三 - **概要**：内村鑑三は明治時代のキリスト教伝道者であり、思想家でもありました。彼は「内村鑑三事件」で知られており、これは1903年に起こった事件です。しかし、この事件は1903年であり、手書き文書の日付（1890年）とは時期が異なります。 - **結論**：この人物は今回の質問に関連性が低いと考えられます。 #### B. 難波大助 - **概要**：難波大助は明治時代の政治家であり、自由民権運動の指導者の一人でした。彼は自由民権運動の活動を通じて政府に批判的な立場を取っていましたが、具体的な「批判を受けた行為」については明確ではありません。 - **結論**：この人物も今回の質問と直接関連する可能性は低いです。 #### C. 幸徳秋水 - **概要**：幸徳秋水は明治時代の社会主義思想家であり、日本における社会主義運動の先駆者です。彼は1910年に起こった「大逆事件」で逮捕され、死刑判決を受けました。この事件は、彼が政府批判的な思想を広めたことによるものです。 - **結論**：幸徳秋水は政府批判的な思想活動を行ったことで知られていますが、手書き文書の日付（1890年）と時期が一致しません。 #### D. 大杉栄 - **概要**：大杉栄は明治時代の社会主義思想家であり、労働運動の指導者でもありました。彼は1890年10月30日に「大杉栄事件」で逮捕されました。この事件は、彼が政府批判的な演説を行ったことによるもので、手書き文書の日付（1890年10月30日）と一致しています。 - **結論**：大杉栄は政府批判的な行動により逮捕されたことで知られ、手書き文書の日付とも一致しています。したがって、この人物が最も適切な答えです。 - **答え** 手書き文書の日付（1890年10月30日）と一致し、政府批判的な行動により批判を受けたことでも知られる人物は **大杉栄** です。 答え：D

Qwen2.5VL-32B (JMMMU-Pro, CoT) ✗ OCR Error

問題の内容を整理 - 提示された画像には「勅語」が記載されています。これは明治天皇が発布した「教育勅語」であり、日本において非常に重要な歴史的な文書です。 - 問題では、「教育勅語」に関連して批判を受けた人物について尋ねています。 ### 各選択肢の人物について検討 #### A. 内村鑑三 - 内村鑑三は明治時代のキリスト教伝道者であり、思想家でもあります。 - 彼は「教育勅語」に対して批判的な立場を取ったことで知られています。特に、彼は「教育勅語」が神道と国家主義を強調し、個人の自由や宗教的多様性を抑圧している点を問題視しました。 - 内村鑑三は「教育勅語」を批判する文章を書いたことがあり、これが彼の思想活動の一環として広く知られています。 #### B. 難波大助 - 難波大助は明治時代の政治家であり、衆議院議員を務めた人物です。 - 彼は「教育勅語」に関する直接的な批判活動についてはあまり知られていません。そのため、この人物が「教育勅語」に対して批判を受けたという直接的な関係は薄いと考えられます。 #### C. 幸徳秋水 - 幸徳秋水は日本の社会主義運動の先駆者であり、思想家・政治家です。 - 彼は「教育勅語」に対して批判的な立場を取ったことはありますが、その批判は主に「教育勅語」が階級差別や個人の自由を抑圧するものであるという点に焦点を当てていました。しかし、幸徳秋水が「教育勅語」に対して批判を受けたという直接的なエピソードは、内村鑑三ほど有名ではありません。 #### D. 大杉栄 - 大杉栄は日本の社会主義運動の指導者であり、労働運動にも深く関わっていました。 - 彼は「教育勅語」に対して批判的な立場を取ったことはありますが、その批判は主に「教育勅語」が労働者階級を抑圧するものであるという点に焦点を当てていました。しかし、大杉栄が「教育勅語」に対して批判を受けたという直接的なエピソードは、内村鑑三ほど有名ではありません。 ### 結論 最も「教育勅語」に対して批判的で、その批判が広く知られている人物は **内村鑑三** です。 答え：A

Analysis

The main cause of the error was a the failure to read an important word. Compared to JMMMU, JMMMU-Pro extracts less information from the image, which leads to an incorrect answer.

Figure G: Failure case: OCR error.