JMMMU-Pro：Vibe Benchmark Construction によるテキスト埋め込み型日本語多分野マルチモーダル理解ベンチマーク

宮井 淳行　小野原 菖太　白 定勳　相澤 清晴
東京大学
{miyai,onohara,baek,aizawa}@hal.t.u-tokyo.ac.jp

概要

本論文では、テキスト埋め込み型の日本語マルチモーダル理解ベンチマークである JMMMU-Pro と、そのスケーラブルな構築手法である Vibe Benchmark Construction を提案する。JMMMU-Pro は JMMMU を拡張し、設問画像と設問文を単一の画像として統合することで、視覚的知覚を通じた視覚・テキスト統合理解を必要とするベンチマークである。本研究で提案する Vibe Benchmark Construction は、画像生成モデルが視覚的質問の候補を生成し、人手による検証を行うとともに、必要に応じてプロンプトを調整して再生成することで品質を担保する。実験の結果、すべてのオープンソース LMMs は JMMMU-Pro に対して著しく性能が低下することが確認された。

1　はじめに

近年、英語における大規模マルチモーダルモデル（LMMs）の成功により [1, 2, 3]、多言語対応 LMMs の開発 [4, 5, 6, 7] や、非英語言語に特化した LMMs の研究 [8, 9] への関心が高まっている。日本語領域においても LMMs の開発は進みつつあるものの [8, 10, 11]、評価用ベンチマークの不足という要因もあり、その進展は英語領域と比較して依然として緩やかである。

既存の日本語ベンチマークにおける主要な制約の一つは、設問画像と設問文がモデルに対して別個のモダリティとして与えられている点にある。現在の LMMs の利用形態を踏まえると、ユーザは日本語テキストと画像の双方を含むスクリーンショットを入力として与えることが一般的である。そのため、幅広い実世界のユースケースを支援するためには、設問画像と設問文の双方が視覚モダリティとして提示される、十分に複雑なタスクにおいて LMMs を評価することが不可欠である。

本論文では、テキスト埋め込み型の日本語多分野マルチモーダル理解ベンチマークである JMMMU-Pro を提案する。JMMMU-Pro は MMMU から MMMU-Pro への進化を踏襲し、既存の JMMMU（Japanese MMMU） [12] に含まれる 1,320 問すべての設問文および設問画像を単一の合成画像として埋め込むことで構築されている。確立された JMMMU を基盤として構築されているため、JMMMU-Pro と JMMMU との間で厳密な条件下での比較が可能であり、これはモデルの視覚的認知能力に関する有意義な指標を提供する。結果として、JMMMU-Pro は高い実用性とともに、モデル開発者に対して極めて情報量の多いフィードバックを提供するベンチマークである。

JMMMU-Pro の構築にあたり、本研究では Vibe Benchmark Construction と呼ぶ新たなベンチマーク構築手法を提案する。本手法では、画像生成モデルが視覚的設問の生成において中心的な役割を担い、人手による作業は生成結果の確認および必要に応じたプロンプトの調整と再生成に限定されることで、一貫した品質を担保する。従来、テキスト埋め込み型ベンチマーク（例：MMMU-Pro [13]）の構築においては、すべての設問を人手で作成する必要があり、スケーラビリティに乏しく、多大な人的コストを要していた。これに対し、Vibe Benchmark Construction は、最先端の画像生成モデルであり高い写実性を有する Nano Banana Pro [14] を活用する。Nano Banana Pro は、高度に現実的な画像を生成できるだけでなく、日本語テキストを画像内に正確に埋め込む能力を有している。Vibe Benchmark Construction は、高いスケーラビリティを備え、人的コストを最小限に抑えつつ、レイアウトの多様性を制御可能である点に特徴がある。Nano Banana Pro と人手による検証を組み合わせることで、JMMMU-Pro に含まれる設問のおよそ 95% が自動生成されており、本手法が今後

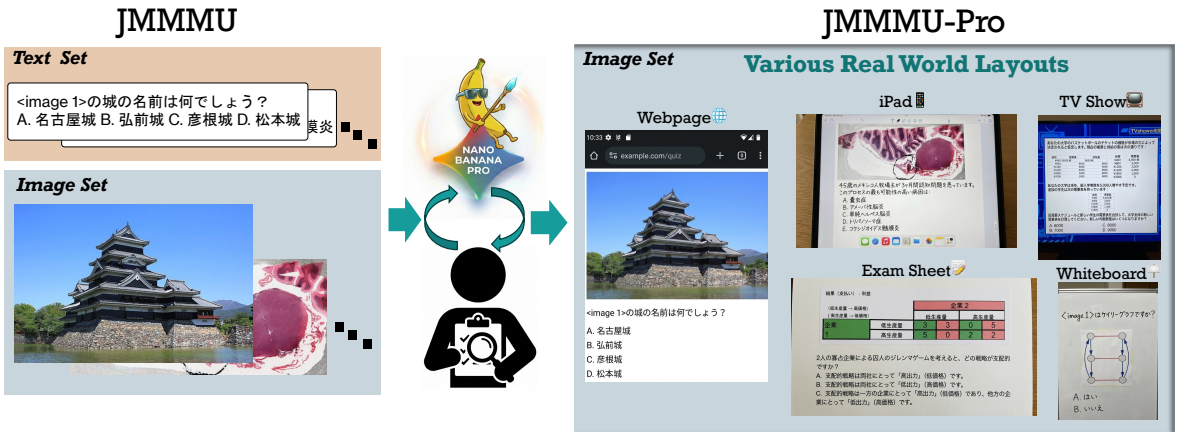


図 1: Vibe Benchmark Construction による JMMMU-Pro の構築

のテキスト埋め込み型ベンチマーク構築に向けた有効な指針となる可能性を示している。

本研究の実験では、計 15 種類の LMMs を評価した。主な実験結果は以下の通りである。(i) オープンソース LMMs は JMMMU-Pro において大きく性能が低下する。(ii) 近年の強力な推論能力を有するクローズドソース LMMs は JMMMU-Pro において比較的良好な性能を示し、クローズドソースとオープンソースの LMMs の間に顕著な性能差が存在することが明らかとなった。(iii) 詳細な分析の結果、失敗要因の大きな一因として日本語 OCR 能力の不足が確認されたものの、高性能な OCR のみでは JMMMU-Pro を解くには不十分であることが分かった。

2 JMMMU-Pro ベンチマーク

2.1 Vibe Benchmark Construction の定義

Vibe Benchmark Construction とは、画像生成モデルが VQA 問題用の画像生成において主導的な役割を担い、人手による作業は生成結果の検証および必要に応じたプロンプトの調整に限定することで品質を担保するベンチマーク構築手法である。従来の VQA ベンチマークにおいても画像生成モデルによる合成画像が用いられてきたが、これらのモデルはあくまで補助的な役割にとどまり、視覚情報のみを生成し、設問文は人手あるいは LMM によって別途作成する必要があったため、追加的なコストが発生していた。これに対し、本研究で提案する Vibe Benchmark Construction の本質的な特徴は、VQA の作成プロセスそのものを画像生成モデルに委ね、人手の介入を検証およびプロンプトの洗練に限定している点にある。このパラダイムは、人手による画像内コンテンツの直接編集が困難なテキスト埋め込み型 VQA において特に有効である。モデルに生成を任せ、人手の作業を 満足のいく画像が得られるまでプロンプトを調整すること に限定することで、データセット構築が困難なテキスト埋め込み型 VQA のような領域においても、効率的かつスケーラブルなベンチマーク構築を可能にする。

2.2 Vibe Benchmark Construction の詳細パイプライン

画像生成には、API インタフェース（`gemini-3-pro-image-preview`）を介して Nano Banana Pro を用いた。画像解像度は 1K に設定した。以下では、プロンプト設計の過程および人手による確認と再生成のワークフローについて説明する。

プロンプト選定と画像生成. まず、予備実験を通じてプロンプトテンプレートを選定した。具体的には、以下の 6 つの要素をパラメータとして変化させることで、多様な画像を生成した。

1. 背景（*Background*）: workbook、exam sheet、whiteboard、blackboard、projector、iPad notebook、webpage、Nintendo Switch、TV quiz show から選択。*2.* 背景色（*Background Color*）: white、light green、light yellow、light pink、light gray、light blue から選択。なお、特定の背景には固定色が存在する（例：whiteboard は常に白）ため、そのような制約を考慮している。*3.* フォント（*Font*）: handwritten text、computer font、thick computer font、thin computer font、manga-style computer font から選択。*4.* 余白（*Margin*）: small または large。*5.* 状態（*State*）: photo by smartphone、screenshot by PC、screenshot by smartphone から選択。*6.* アスペクト比（*Aspect Ratio*）: 9:16、16:9、3:4、1:1 から選択。

人手による確認と再生成. 生成された画像に対しては、独自に構築したアノテーションツールを用いて著者によるレビューを実施した。レビューでは、生成されたテキストおよび画像が元の設問内容と正確

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

表 1: **JMMMU-Pro での結果**。オープンソース LMMs は JMMMU と比較して JMMMU-Pro において大幅な性能低下を示す一方で、クローズドソース LMMs は高い性能を維持しており、日本語における視覚・テキスト統合理解能力に関して両者の間に顕著なギャップが存在することが明らかとなった。

に一致しているかを確認した。前述のとおり、設問文中の画像タグの厳密な制御は困難であるため、生成結果が妥当な設問として成立している限り、タグの表記揺れは許容した。

最初のレビューでは、全体の 71% の設問が採用された。残りの 29% は、設問画像が無関係な画像に置き換えられている、画像内テキストが判読不能である、設問文の一部が欠落または誤っている、あるいは生成画像が視覚的に不自然であるといった理由により不採択となった。不採択となった設問については、同一のプロンプト、または軽微に調整したプロンプトを用いて再生成を行った。すべての VQA 設問の生成が完了した後、著者間の評価基準の不一致を排除するため、最終的なクロスチェックを実施した。

手動構築. Nano Banana Pro による生成が困難であった 67 件の設問については、人手により作成した。これらの設問は、以下の特徴を有していた。設問文が長いもの（16 件）、設問画像内の文字が小さい、または描画が困難なもの（36 件）、極端なアスペクト比を有するもの（2 件）、化学式や楽譜など、生成自体が本質的に困難なドメインに属するもの（8 件）、ポリシー制約により Nano Banana Pro によって生成が拒否されたもの（5 件）。

3 実験

3.1 実験詳細

評価する LMMs. 包括的な評価を行うため、最先端の多様な LMMs を対象として評価を実施した。特にオープンソースモデルについては、英語中心の LMMs、多言語対応 LMMs、日本語対応 LMMs の 3 つのカテゴリから代表的なモデルを選定することで、各サブフィールドにおける最新の進展を適切に反映した評価となるよう配慮している。実験には主として LMMs-Eval [15] を用いた。オープンソース LMMs に対しては temperature を 0 に設定し（クローズドソース LMMs についてはデフォルト設定を使用）、応答が途中で打ち切られないよう、max_tokens は十分に長く設定した。実験は、単一の A100 80G GPU を用いて実行した。

推論プロンプト. 推論プロンプトは、JMMMU [12]

および MMMU-Pro [13] の設定に基づいて設計した。MMMU-Pro [13] に従い、オープンソース LMMs については Direct プロンプトおよび CoT（Chain-of-Thought）プロンプトの両方を用いて評価を行い、全体結果にはより高いスコアを採用した。一方、クローズドソース LMMs はプロンプトの種類に依らず推論を行うため、Direct プロンプトのみを用いて評価した。

3.2 結果

実験結果を 表 1 に示す。これらの結果から得られた主な知見を以下にまとめる。

F1. すべてのオープンソース LMMs は JMMMU-Pro において著しく苦戦する。オープンソース LMMs は JMMMU-Pro において全体的に低い性能を示しており、最も高い性能を達成した Qwen2.5-VL-32B でさえスコアは 56.14 にとどまっており、依然として大きな改善の余地がある。また、ほとんどのオープンソース LMMs は、JMMMU と比較して JMMMU-Pro において大幅な精度低下を示した。これらの結果は、JMMMU-Pro を JMMMU と併用することで、モデル開発者に対して有益なフィードバックを提供できることを示している。

F2. クローズドソース LMMs は JMMMU-Pro において著しく高い性能を達成し、オープンソースモデルとの大きなギャップを示す。クローズドソース LMMs は JMMMU-Pro において顕著に高いスコアを達成した。これは、これらのモデルがすでに視覚情報とテキスト情報をシームレスに統合し、視覚的知覚を通じて解釈する能力を備えていることを示している。重要なのは、クローズドソースモデルの高い性能が JMMMU-Pro の価値を損なうものではない点である。むしろ、オープンソース LMMs の開発を導くベンチマークとして、JMMMU-Pro が果たす重要な役割を強調する結果となっている。クローズドソースとオープンソース LMMs の間には依然として大きな性能差が存在しており、このギャップを縮小することはコミュニティの課題である。

3.3 OCR 性能との相関

複数の LMMs を対象に、OCR 性能と JMMMU-Pro における正解率との相関を算出する。

MMMU-Pro の評価設定に従い、各 LMM に対して、関連する画像内のテキストを除外した上で、設問文およびすべての選択肢の全文を抽出するよう指

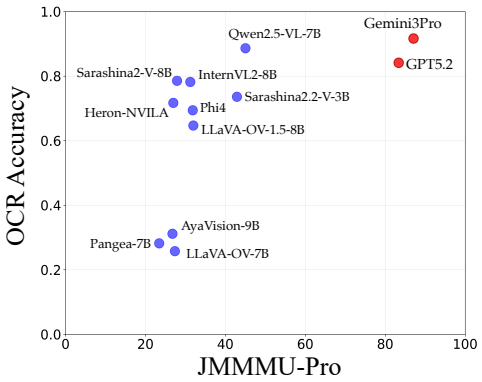


図 2: OCR 精度と JMMMU-Pro 性能との相関

示した。OCR 精度は、抽出されたテキストと元のテキストとを比較し、2 つの文字列間の編集距離を測る Levenshtein 距離を用いて算出する。抽出テキストと元テキストとの類似度は、以下の式により計算される。

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

結果を 図 2 に示す。OCR 精度と JMMMU-Pro における正解率との相関係数は 0.593 であり、両者の間には正の相関が確認された。しかしながら、高い OCR 能力が必ずしも JMMMU-Pro における高い性能に直結するわけではない。例えば、Heron-NVILA と Sarashina2.2-V は OCR 性能において同程度であるにもかかわらず、JMMMU-Pro における性能には大きな差が見られる。

この結果は、JMMMU-Pro を解くためには、単なる OCR 能力にとどまらず、視覚的知覚を通じて言語情報と視覚情報を統合的に解釈し、推論する能力が求められることを示している。

4 おわりに

本論文では、テキスト埋め込み型の日本語多分野マルチモーダル理解ベンチマークである JMMMU-Pro と、その構築を可能にするスケーラブルな手法 Vibe Benchmark Construction を提案した。実験の結果、すべてのオープンソース LMMs は JMMMU-Pro において著しい困難を示し、JMMMU-Pro がオープンソースコミュニティにおける今後の発展を促す重要なベンチマークであることが明らかとなった。

参考文献

[1] OpenAI. Gpt-4o, 2024.

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

[3] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. Accessed: 2025-11-29.

[4] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. **arXiv preprint arXiv:2409.12191**, 2024.

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

[6] Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. Pangea: A fully open multilingual multimodal llm for 39 languages. In **ICLR**, 2025.

[7] 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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Li Ying Meng, Xuancheng Ren, Xin yi Ren, Sibo Song, Yu-Chen Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxing Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jingren Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. **arXiv preprint arXiv:2511.21631**, 2025.

[8] SB Intuitions. Sarashina2.2-vision-3b, 2025. Accessed: 2025-11-29.

[9] NAVER Cloud HyperCLOVA X Team. Hyperclova x think technical report. **arXiv preprint arXiv:2506.22403**, 2025.

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

[11] Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. Constructing multimodal datasets from scratch for rapid development of a Japanese visual language model. In **NAACL: Human Language Technologies (System Demonstrations)**, 2025.

[12] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. Jmmmu: A japanese massive multi-discipline multimodal understanding benchmark for culture-aware evaluation. In **NAACL**, 2025.

[13] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In **ACL**, 2025.

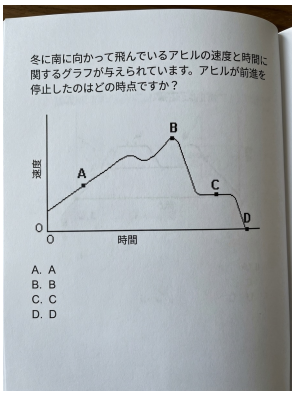
[14] Google DeepMind. Gemini 3 pro image (nano banana pro). Web page, 2025. Accessed: 2025-11-29.

[15] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In **NAACL Findings**, 2025.

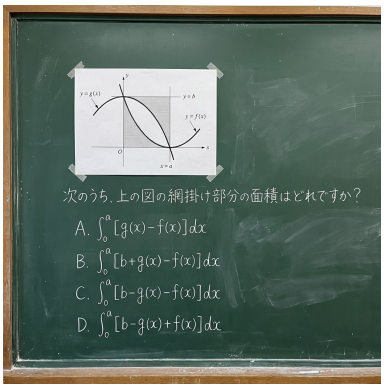
参考情報

A　JMMMU-Pro のサンプル例

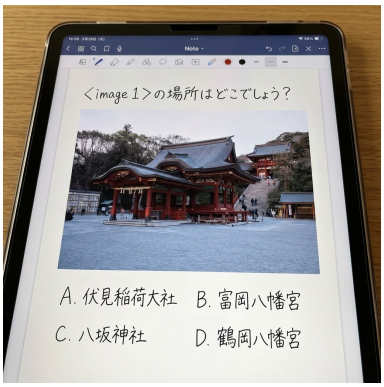
図 A に JMMMU-Pro のサンプル例を示した。このように多様な背景の画像を作ることが可能である。



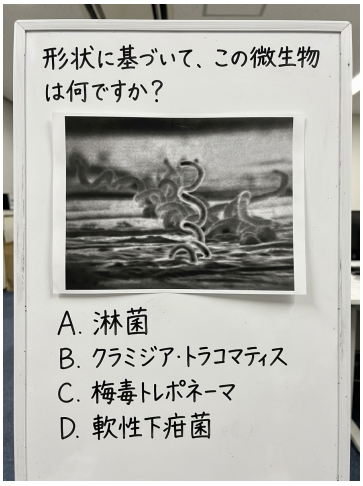
(a) Workbook



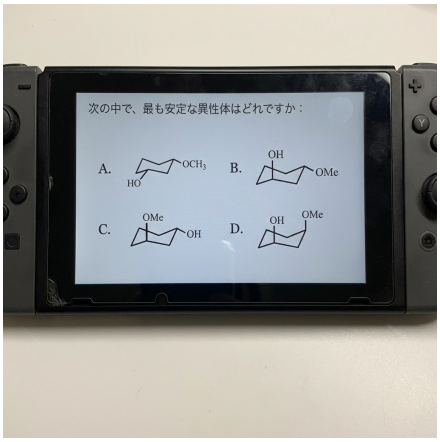
(b) Blackboard



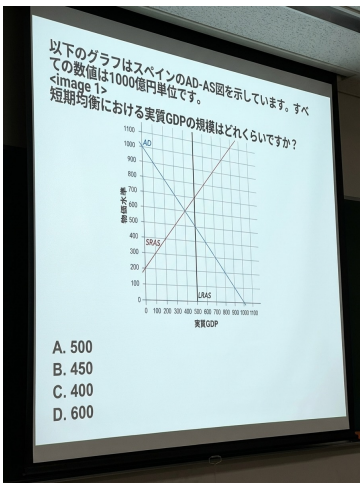
(c) iPad notebook



(d) Whiteboard



(e) Nintendo Switch



(f) Projector

図 A: JMMMU-Pro のサンプル例