JAMMEval: 再アノテーションによる 日本語 VQA 評価データセットの信頼性向上

杉浦一瑳 ♣,‡ 前田 航希 ♢,‡ 栗田修平 †,‡ 小田悠介 ‡ 河原大輔 ♡,‡ 岡崎直観 ♢,‡
♣ 京都大学 ♡ 早稲田大学 ♢ 東京科学大学 † 国立情報学研究所 ‡ NII LLMC
sugiura.issa.q29@kyoto-u.jp

概要

既存の日本語 VQA 評価データセットには，曖昧な質問や回答が誤っている事例，画像を用いずに回答できる事例等があり，これらは評価の有用性・信頼性を低下させている．この問題を解決するため，本研究では 8 つの日本語評価データセットに再アノテーションを施した評価データセット群 **JAMMEval** を構築する．実験では，JAMMEval を用いて最先端のモデルの性能を評価した．また，JAMMEval の一部を用いた再アノテーションの効果検証を行い，再アノテーションによりモデル性能の識別能力が向上することを示した．JAMMEval は公開する．

1 はじめに

評価データセットはモデル開発において重要な役割を果たしている．モデル開発は一般にモデル学習と評価のサイクルで行われるが，評価データセットに不備があるとモデルの性能順位や評価結果が歪められ，評価の信頼性が低下する．その結果，モデル開発者は正しい現状認識を得られず，開発に関する意思決定に悪影響が及ぶ [1, 2, 3, 4, 5]．優れた評価データセットが備えているべき要件として，ラベル誤りや曖昧性などのノイズが少ない，十分な事例数を含む，評価方法がシンプルである，評価を高速に行える，現実的・本質的な能力に対応した意味のあるタスクである，採点が正確である，すぐにスコアが飽和しない，などが提案されている [1, 2, 6]．

大規模視覚言語モデル (VLM) [7, 8, 9] の評価に用いられる画像質問応答 (VQA) 評価データセットのうち，英語では数学や図表を含む多様なドメインにおいてデータセットの継続的な改善がなされている [10, 3, 11]．例えば，大学レベルの幅広い専門分野における知識を問う MMMU [12] には画像を見ずにテキストのみで解けてしまう事例が多く含まれるが，MMMU-Pro [10] はこの問題に対処するため，選択肢の数を増やしたり質問を画像に埋め込むなどの工夫を施し，難易度を高めた．日本語データセットの開発も進んでおり [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]，複数の VLM を包括的に評価するフレームワークも提案されている [24]．これらの評価データセットおよびフレームワークは日本語特化 VLM モデルの開発に活用されている [14, 25, 23, 26]．直近ではデータセットの継続的な改善として．JMMMU [22] の難易度を Nano Banana Pro [27] を用いて向上させた JMMMU-Pro [28] が提案されている．

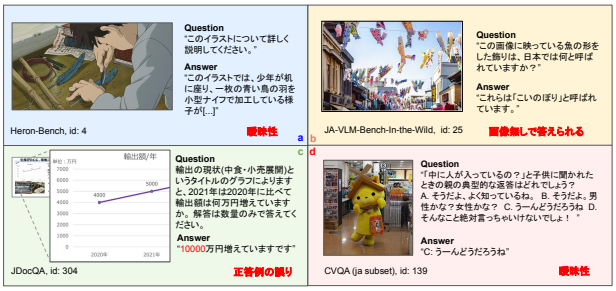

図 1 既存の日本語 VQA 評価データセットに含まれる不適切な事例の例. (a) 答えが一意に定まらず曖昧性のある自由記述式. (b) 画像無しの質問文のみでも高確率で答えられる. (c) 正答例が誤っている. (d) 回答する人によって意見が分かれる.

しかし，既存の日本語データセットは英語データセットと比較して継続的改善が不十分である．多くのデータセットは図 1 に示すような問題を含み，いずれも評価に不適当である．例えば，答えが一意に定まらない曖昧性の高い問題，答えが間違っている問題では，モデルの性能を正しく評価できない．また，画像なしで回答可能な問題があると，VLM が持つ視覚能力の測定の解像度が落ちる．さらに，LLM を評価者としたリッカート尺度による採点手法を採用しているデータセットもあるが，この手法による評価は LLM の持つバイアスの影響を受けやすい．

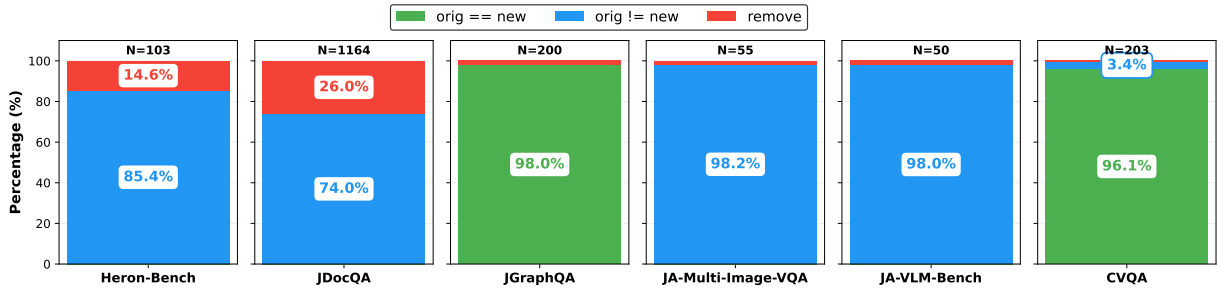

図 2　各データセットの再アノテーションの統計. 各事例について，(i) 変更を加えなかった事例，(ii) 修正を施した事例，(iii) 重複のある画像や情報量が少ない画像のため除去した事例，に分類した.

本研究では既存日本語評価データセットに見られる問題を特定した上で，再アノテーションを実施して信頼性の高い評価データセット群 JAMMEval (**JA**panese **M**ulti**M**odal **Eval**uation Collection) を構築した. さらに JAMMEval を用いたモデル評価と再アノテーションの効果検証を行った.

2　JAMMEval の構築方法

JAMMEval は，再アノテーション元として用いるデータセット (シード・データセット) を収集した上で事例を分析し，各データセットについて再アノテーションを行うことで構築した.

2.1　シード・データセットの収集

シード・データセットには，日本文化や図表理解，ドキュメント理解，OCR 等多様なドメインの評価データセットを用いた. 具体的には，Heron-Bench [13]，JDocQA [15]，CC-OCR [16]，JGraphQA [17]，JA-Multi-Image-VQA [18]，JA-VLM-Bench-In-the-Wild [14]，CVQA [19]，WAON-Bench [20] を用いた. 各データセットの詳細な説明は付録 C に示す.

2.2　既存データセットの事例分析

収集したデータセットの事例を目視で確認したところ，図 1 に示すような評価に不適当な事例が確認された:

- **曖昧性**: 解答が一意に定まらない問題である
- **正答例の誤り**: 正解として用意された答えが誤っている
- **画像無しで答えられる**: 画像を見ずに，外部知識を用いて質問文のみから正しく回答できる

これらは先行研究 [3, 10] でも報告されている評価データセットの典型的な不具合である.

不適当な事例を定量的に分析するために，我々は収集したデータセットの全ての事例の画像・質問・回答の三つ組を，(i) 適切な質問応答であり，変更を加えない事例，(ii) 不適当であり，修正を施した事例，(iii) 重複のある画像や情報量が少ない画像のため除去した事例，に分類した. 分類作業は著者が行った. 作業のためのツールは独自に構築した[1]. 分類結果を図 2 に示す. JDocQA や Heron-Bench などに，曖昧性や正答例の誤りがある事例が多く見られた. 一方でデータセット構築において検証段階を設けた CVQA や JGraphQA では，不適切な事例はほとんど見られず，データセット構築における検証段階の重要性が示唆される.

表 1　JAMMEval の統計量. 複数画像を含む JA-Multi-Image-VQA において「画像数」は画像の組を 1 組として重複除去した件数. † は既存データセットの画像を用いて QA をいちから構築したことを示す.

データセット	シード	画像数	QA 数	問題形式	カテゴリ
Heron-Bench-Pro	[13]	21	88	簡潔回答	知識
JDocQA-Pro	[15]	793	861	簡潔回答	文書
CC-OCR-JA-VQA†	[16]	145	145	簡潔回答	OCR
JGraphQA-Pro	[17]	98	196	簡潔回答	図表
JA-Multi-Image-VQA-Pro	[18]	16	54	簡潔回答	複数画像
JA-VLM-Bench-Pro	[14]	42	49	簡潔回答	知識
CVQA-JA-Pro	[19]	94	202	多肢選択	知識
WAON-Bench-VQA†	[20]	373	373	多肢選択	知識

2.3　データセットの再アノテーション

質問応答の修正作業は，問題事例の分析および先行研究 [1, 2, 6] をふまえて，以下の要件をもとに著者が行った:

- **曖昧性の排除**: 答えが一意に定まること.
- **画像を根拠とした問題**: 質問のみから回答を推測できないこと.
- **性能飽和の予防**: 新たな問題を作成する場合，VLM にとって難易度の高い問題であること.

各データセットに施した特有の処理は以下のとお

1)　アノテーションツールの作業画面を付録図 5 に示す.

りである．

Heron-Bench　LLM によって付与された参照回答の誤り事例を修正した．また，キャプション生成に近い自由記述形式の問題が付与された画像を用いて，回答が一意に定まる VQA を新規に作成した．

JDocQA　複数画像を用いた VQA の事例は，問題の最初の画像のみを対象とした VQA を作成した．

CC-OCR　本来のタスクは画像から文字を読み順で全て抜き出す OCR タスクである．我々は日本語サブセットの画像を用いて，画像中の文字を抜き出す OCR タスクを VQA の形式で作成した．

JGraphQA　質問応答と画像が対応しない事例を除去した[2)]．

WAON-Bench　WAON-Bench は日本文化に関連する 374 クラスの画像分類データセットであり，各クラス 5 枚の画像を含む．各クラスからランダムに 1 枚の画像を選び，画像の内容を問う多肢選択式の VQA を作成した．

JAMMEval の各データセットの統計量を表 1 に示す．質問応答を作りにくい画像や，情報量が乏しい画像については再アノテーションしなかったため，事例数は元のデータセットから減少したが，合計で 1,968 問の VQA を作成した．元のデータセットと区別するため，再アノテーションを行った VQA データセットは “-Pro” を接尾につけ，元のデータセットが VQA でない場合は “-VQA” をつけた．

3　JAMMEval を用いたモデル評価

3.1　評価設定

モデル　評価では，オープンウェイトなモデルとして多言語対応の Qwen3-VL-{2B, 4B, 8B, 32B} [9], InternVL 3.5-{2B, 4B, 8B} [8]，日本語特化モデルの Sarashina2.2-Vision-3B [26] を用いた．クローズドなモデルとして GPT-{4o, 5.1} [29, 30], Gemini 3 Pro [31] を用いた．GPT-4o については，テキストのみ (text-only) での評価も行った．text-only のプロンプトは付録 B に示す．

プロンプト　簡潔回答では以下のプロンプトを用いた．

```
{question}\n 上記の質問に対して、正確かつ簡潔に答えてください。
```

多肢選択式では以下のプロンプトを用いた．

```
{question}\n{choices_str}\n 与えられた選択肢から該当する選択肢のアルファベットだけで答えてください。
```

採点方法　JAMMEval の評価指標は，全てのデータセットを通して正解率（Accuracy）で統一する．簡潔回答については，モデル出力における全角・半角や単位などの軽微な表記ゆれに対応するために LLM-as-a-Judge によって正解・不正解を判定した．具体的には (質問, 正答, モデル出力) の三つ組を GPT-4o に与えて正解・不正解の判定させた．採点用のプロンプトは付録 B に示す．多肢選択については正規表現でアルファベットを抽出し，完全一致により正誤を判定した．

生成設定　温度は 0，最大トークン数は各データセットで必要十分な値を設定した．思考モデルの Gemini 3 Pro については思考レベルを low とし，最大トークン数を 1,024 とした．

3.2　評価結果

評価結果を図 3 に示す．多くのタスクおよびモデルシリーズで，モデルの規模と評価スコアに正の相関が見られた．また，日本語特化モデルの Sarashina2.2-Vision-3B は同程度の規模のモデルと比較して Heron-Bench-Pro や CVQA-JA-Pro 等の日本文化関連タスクで強い性能を示し，日本語特化の効果が示唆された．思考モデルの Gemini 3 Pro は全てのタスクにおいて最高性能を示しており，思考モデルの効果が示された．GPT-4o (text-only) は多肢選択式タスクの CVQA-JA-Pro，WAON-Bench-VQA において 45%程度解けており，ランダム予測より高い性能を示した．これは，曖昧性の除去や正答例の修正といった再アノテーションのみでは解消できない，選択肢設計や言語的手掛かりに起因する問題が残存していることを示唆している．この問題への対処は今後の課題とする．

4　再アノテーションの効果検証

再アノテーションによるモデルの性能の識別解像度の変化を検証するために，VLM の学習段階における JDocQA と JDocQA-Pro のスコア推移を調べた．VLM の性能は学習ステップ数や学習データセットによって変化するが，その性能変化の分解能で評価する．学習設定は，InternVL-3 [32] のアーキテクチャを参考として，LLM に Qwen3-4B [33]，画像エンコーダに SigLIP2 [34] を用いた．学習データセットは，(i) FineVision [35]，(ii) FineVision + 独自に開発し

2)　JGraphQA は画像の代わりに PDF の URL が公開されており，PDF が差し代わっていたことが原因だと考えられる．

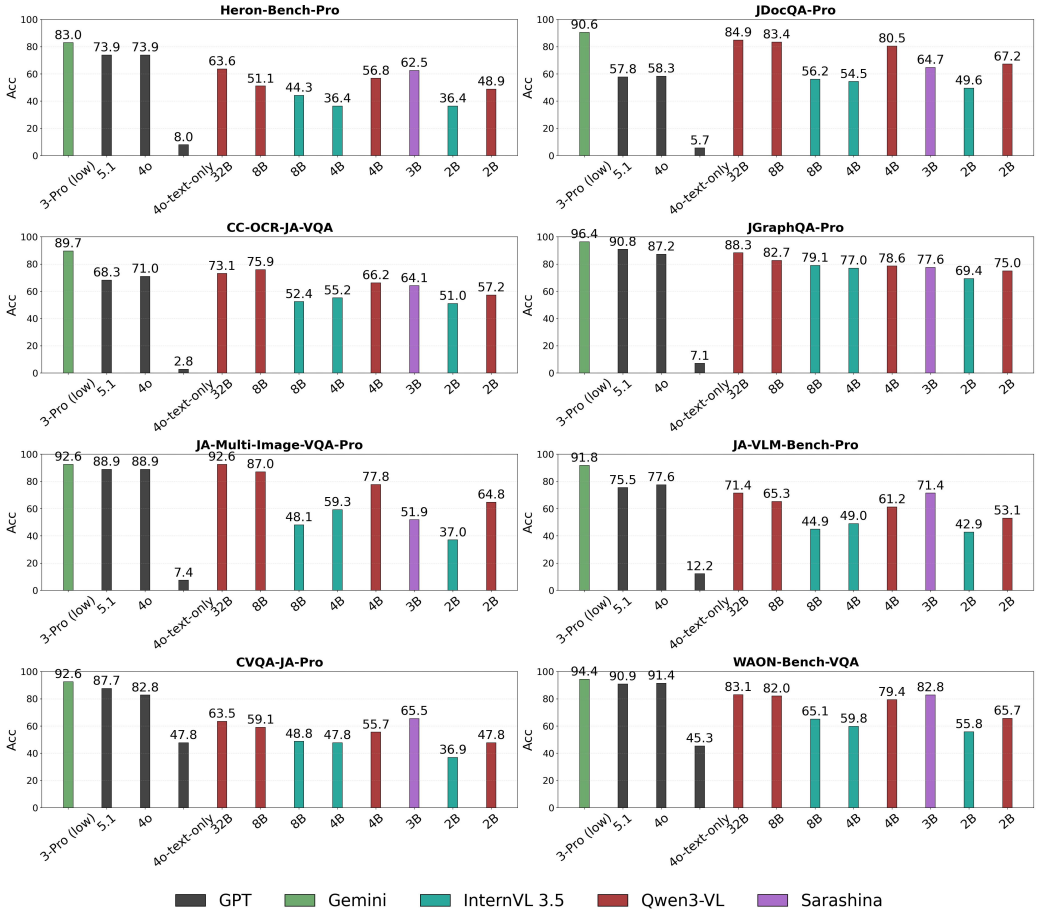

図 3　JAMMEval における各モデルの評価結果.

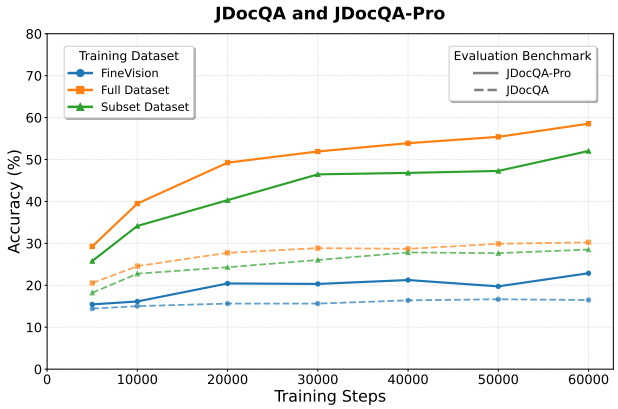

図 4　学習中の JDocQA および JDocQA-Pro のスコア推移. 再アノテーションにより性能の識別能力が向上する.

た日本語データセット Jagle の全部，(iii) FineVision + Jagle の一部，の 3 パターンで検証した．学習は総バッチサイズ 1,024，学習ステップ数は 60,000 ステップ，最大トークン数は 4,096 とした．評価結果の比較を行うため，JDocQA の評価手法も JDocQA-Pro に合わせた．

図 4 に学習中のスコア推移を示す．JDocQA-Pro は JDocQA と比較して評価の分解能が高く，学習ステップや学習データセットの変化に伴うスコアの変動をより明確に捉えられていることが分かる．JDocQA には曖昧性の高い設問が多く含まれており，それらの設問ではモデル性能の変化を十分に捉えられないため，一部の設問が全体スコアを左右する構造となり，結果としてスコアが変化しにくくなっていた可能性がある．

5　おわりに

本研究では，既存の日本語 VQA 評価データセット群の不適当な事例を分析し，再アノテーションを実施することで JAMMEval を構築した．JAMMEval は既存のデータセットの曖昧性や正答例の誤り等の問題を人手で修正しており，より信頼性の高い評価が可能である．JAMMEval を用いたモデル評価では日本語特化モデルの効果や多肢選択式問題の課題が示された．また，効果検証では再アノテーションによってモデル性能の識別能力が向上することが明らかになった．

謝辞

本研究では，産総研及び AIST Solutions が提供する ABCI 3.0 を「ABCI 3.0 開発加速利用」を支援を受けて利用した．また，JAMMEval の構築で用いたシード・データセットの開発者に感謝いたします．

参考文献

[1] Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. FineWeb2: One pipeline to scale them all – adapting pre-training data processing to every language. arXiv preprint arXiv:2506.20920, 2025.

[2] Jason Wei. Successful language model evals. https://www.jasonwei.net/blog/evals, 2025.

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In **NeurIPS**, 2024.

[4] Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. Do large language model benchmarks test reliability? arXiv preprint arXiv:2502.03461, 2025.

[5] Sang Truong, Yuheng Tu, Michael Hardy, Anka Reuel, Zeyu Tang, Jirayu Burapacheep, Jonathan Perera, Chibuike Uwakwe, Ben Domingue, Nick Haber, and Sanmi Koyejo. Fantastic bugs and where to find them in ai benchmarks. arXiv preprint arXiv:2511.16842, 2025.

[6] Liang Hu, Jianpeng Jiao, Jiashuo Liu, Yanle Ren, Zhoufutu Wen, Kaiyuan Zhang, Xuanliang Zhang, Xiang Gao, Tianci He, Fei Hu, Yali Liao, Zaiyuan Wang, Chenghao Yang, Qianyu Yang, Mingren Yin, Zhiyuan Zeng, Ge Zhang, Xinyi Zhang, Xiying Zhao, Zhenwei Zhu, Hongseok Namkoong, Wenhao Huang, and Yuwen Tang. FinSearchComp: Towards a realistic, expert-level evaluation of financial search and reasoning. arXiv preprint arXiv:2509.13160, 2025.

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.

[8] InternVL Team. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265, 2025.

[9] Qwen Team. Qwen3-VL technical report. arXiv preprint arXiv:2511.21631, 2025.

[10] 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.

[11] Siddharth Joshi, Haoli Yin, Rishabh Adiga, Ricardo Monti, Aldo Carranza, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Darren Teh, David Schwab, Fan Pan, Haakon Mongstad, Jack Urbanek, Jason Lee, Jason Telanoff, Josh Wills, Kaleigh Mentzer, Luke Merrick, Parth Doshi, Paul Burstein, Pratyush Maini, Scott Loftin, Spandan Das, Tony Jiang, Vineeth Dorna, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. DatBench: Discriminative, faithful, and efficient vlm evaluations. arXiv preprint arXiv:2601.02316, 2026.

[12] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In **CVPR**, 2024.

[13] Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. Heron-Bench: A benchmark for evaluating vision language models in japanese. arXiv preprint arXiv:2404.07824, 2024.

[14] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. **Nature Machine Intelligence**, 2025.

[15] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese document question answering dataset for generative language models. In **LREC-COLING**, 2024.

[16] 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. CC-OCR: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. arXiv preprint arXiv:2412.02210, 2024.

[17] Akira Kinoshita. JGraphQA. https://huggingface.co/datasets/r-g2-2024/JGraphQA, 2025.

[18] Inoue Yuichi, Akiba Takuya, and Makoto Shing. Llama-3-EvoVLM-JP-v2. https://huggingface.co/datasets/SakanaAI/JA-Multi-Image-VQA, 2024.

[19] David Orlando Romero Mogrovejo, Chenyang Lyu, et al. CVQA: Culturally-diverse multilingual visual question answering benchmark. In **NeurIPS (Datasets and Benchmarks Track)**, 2024.

[20] Issa Sugiura, Shuhei Kurita, Yusuke Oda, Daisuke Kawahara, Yasuo Okabe, and Naoaki Okazaki. WAON: Large-scale and high-quality japanese image-text pair dataset for vision-language models. arXiv preprint arXiv:2510.22276, 2025.

[21] 前田航希, 長谷川騎平, 栗田修平, 小田悠介, 徳久良子, 岡崎直観. 日本の文化常識・日常生活知識理解のための視覚言語ベンチマーク mecha-ja の構築. 情報処理学会 第 263 回自然言語処理研究会 研究報告 (2024-NL-263), 2025.

[22] 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.

[23] Stockmark. BusinessSlideVQA. https://github.com/stockmarkteam/business-slide-questions, 2025.

[24] 前田航希, 杉浦一瑳, 小田悠介, 栗田修平, 岡崎直観. llm-jp-eval-mm: 日本語視覚言語モデルの自動評価基盤. 言語処理学会第 31 回年次大会 (NLP2025), 2025.

[25] 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 (System Demonstrations)**, 2025.

[26] SB Intuitions. Sarashina2.2-Vision-3B. https://huggingface.co/sbintuitions/sarashina2.2-vision-3b, 2025.

[27] Google DeepMind. Gemini 3 Pro Image (Nano Banana Pro). https://deepmind.google/models/gemini-image/pro, 2025.

[28] Atsuyuki Miyai, Shota Onohara, Jeonghun Baek, and Kiyoharu Aizawa. JMMMU-Pro: Image-based japanese multi-discipline multimodal understanding benchmark via vibe benchmark construction. arXiv preprint arXiv:2512.14620, 2025.

[29] OpenAI. GPT-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[30] OpenAI. GPT-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1, 2025.

[31] Google DeepMind. Gemini 3 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, 2025.

[32] 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. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

[33] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

[34] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786, 2025.

[35] Luis Wiedmann, Orr Zohar, Amir Mahla, Xiaohan Wang, Rui Li, Thibaud Frere, Leandro von Werra, Aritra Roy Gosthipaty, and Andrés Marafioti. FineVision: Open data is all you need. arXiv preprint arXiv:2510.17269, 2025.

[36] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. arXiv preprint arXiv:1906.02569, 2019.

[37] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In **NeurIPS (Datasets and Benchmarks Track)**, 2024.

[38] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, et al. Humanity's Last Exam. arXiv preprint arXiv:2501.14249, 2025.

[39] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In **ACL (Findings)**, 2022.

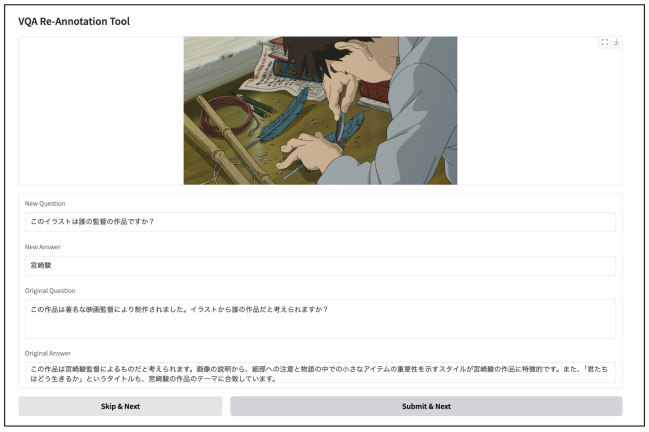

図 5　VQA 再アノテーションツールのインターフェース．各事例について，(Image, Original Question, Original Answer) があらかじめ表示される．アノテータは画像に対する (New Question, New Answer) を入力した上で Submit ボタンを押すか，良い VQA を作成できない場合は Skip ボタンを押して次の事例に進む．

表 2　再アノテーションの例．質問と回答の冗長性を排除する．答えが一意に決まらない曖昧性のある QA は別の QA を作成する．

シード	元の事例		新しい事例	
	Question	Answer	Question	Answer
Heron-Bench	この作品は著名な映画監督により制作されました。イラストから誰の作品だと考えられますか？	この作品は宮崎駿監督によるものだと考えられます。画像の説明から、細部への注意と物語の中での小さなアイテムの重要性を示すスタイルが宮崎駿の作品に特徴的です。[...]	このイラストは誰の監督の作品ですか？	宮崎駿
JDocQA	この前小瀬鵜飼をはじめてみましたが、春夏秋冬問わず行っていますか。参加したい場合はどのようにしたらいいですか。\ n 解答は自由に記述してください。	小瀬鵜飼は夏の風物詩となり、申し込みする場合は関遊船事務所に連絡します。料金は貸し切りまたは乗合でかわり、実際に体験することができるパックもあります。	小瀬鵜飼は何月何日に開幕しますか?	5月11日

A　再アノテーション

Gradio [36] を用いて開発したアノテーションツールのインターフェースを図 5 に示す．また，再アノテーションの例を表 2 に示す．

B　プロンプト

以下に WAON-Bench-VQA の誤答選択肢の生成において用いたプロンプトを示す．

あなたは画像 VQA の 4 択問題を作る AI です。
与えられた画像と正解クラスから **誤りの選択肢 (distractors)** を 3 つだけ生成してください。
制約:
- distractor は正解と紛らわしいが誤りのもの
- 出力は JSON のみ
- 返すのは "distractors": ["...","...","..."] のみ
- question や correct answer は生成しない
正解クラス: "{correct}"
返す形式: {{ "distractors": ["...", "...", "..."] }}

GPT-4o (text-only) の評価には，CharXiv [37] を参考に，以下のプロンプトをシステムプロンプトに追記し，モデルに質問文のみを与えて答えを出力させた．

Randomly guess a reasonable answer based on the question only. If the question asks for a number, you can randomly guess a number within a reasonable range. If the question asks for a term, you can randomly guess a term that is relevant to the question.

簡潔回答形式の評価では，[38] で用いられた採点プロンプトに軽微な修正を行った以下のプロンプトを用いた．

Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below.
When judging equivalence, allow variations in script or notation that convey the same meaning (e.g., '2 羽' and ' 二羽' should be considered equivalent).
Treat the following cases as correct: - The extracted answer includes additional context (e.g., series name, author name, location, broader category) while still containing the correct_answer exactly or as its unambiguous, specific instance. (For example, "富嶽三十六景 江戸日本橋" is correct if the correct_answer is "江戸日本橋".) - The extracted answer is more specific than the [correct_answer] while remaining consistent with it. - The extracted answer is an alternate name, synonymous phrasing, or another commonly accepted way to refer to the same concept, object, place, or artwork. - The extracted answer omits information that is not essential to the correctness of the question. - Allow minor variations in spacing, capitalization, or script, as long as the core correct_answer is unambiguously present.
[question]: {question}
[response]: {response}
Your judgement must be in the format and criteria specified below:
extracted_final_answer: The final exact answer extracted from the [response]. Put the extracted answer as 'None' if there is no exact, final answer to extract from the response.
[correct_answer]: {correct_answer}
reasoning: Explain why the extracted_final_answer is correct or incorrect based on [correct_answer], focusing only on if there are meaningful differences between [correct_answer] and the extracted_final_answer. Do not comment on any background to the problem, do not attempt to solve the problem, do not argue for any answer different than [correct_answer], focus only on whether the answers match.
correct: Answer 'yes' if extracted_final_answer matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer 'no' otherwise, i.e. if there if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect.
confidence: The extracted confidence score between 0% and 100% from [response]. Put 100 if there is no confidence score available.

C　シード・データセットの説明

Heron-Bench [13] はジブリ映画や日本の建造物など，日本に関連した画像を用いた VQA データセットである．

JDocQA [15] は，日本の公的機関によって公開された文書の画像を用いた VQA データセットである．

CC-OCR [16] は，マルチシーンテキスト読解，多言語テキスト読解，文書パース，およびキー情報抽出の 4 つのトラックで構成された OCR データセットである．

JGraphQA [17] は，ChartQA [39] を参考にして構築された，日本の IR 資料中に存在する円グラフ，折れ線グラフ，棒グラフ，表の 4 種類からなる計 100 枚の画像に対して，各画像について 2 つの QA ペアが付与されたデータセットである．

JA-Multi-Image-VQA [18] は，VLM の複数画像に対する質問応答能力を測るために構築されたデータセットである．

JA-VLM-Bench-In-the-Wild [14] は日本の文化や日本国内の物体を含む画像を用いた VQA データセットである．

CVQA [19] は 39 の国・言語ペアにまたがる 1 万件以上の質問から構成される，文化的多様性を考慮した多言語 VQA ベンチマークである．

WAON-Bench [20] は 8 つのカテゴリ，374 クラスで構成された日本文化に関連する画像を用いた画像分類データセットである．