

高精度な日本語マルチモーダル大規模言語モデルの構築にむけたデータセットの検討

田中幹大 朱佩菲 横尾修平

LINE ヤフー株式会社

{mikihiro.tanaka, peifei.zhu, shuhei.yokoo}@lycorp.co.jp

概要

近年、大規模言語モデル (LLM) に視覚情報を統合した、マルチモーダル大規模言語モデル (MLLM) が注目を集めており、その応用範囲は急速に拡大している。しかし、日本語ドメインに特化した MLLM を作る上で、英語のデータに比べて公開データが少ない課題がある。本研究では、高精度な日本語 MLLM を構築するためのデータセットの作成方法について検討し、実験を行った。構築したモデルは、日本語ドメインの画像理解を問うベンチマークにおいて、他のモデルよりも優れた結果を示し、その有効性を実証した。

1 はじめに

大規模言語モデル (LLM) の急速な発展に伴って、画像も入力として扱えるマルチモーダル大規模言語モデル (MLLM) が注目されている。MLLM は視覚と言語の統合的な理解を通じて、より自然で直感的なインタラクションを可能にし、多様なタスクにおいて効果的なサポートを提供することが期待されている。しかし、MLLM の研究は盛んに行われている一方で、日本語ドメインに特化した MLLM を構築する試みはまだ少ない。特に、日本語 MLLM の指示学習において重要となる画像と対話のペアデータは英語に比べてまだ少ない課題がある。そこで、本研究では高精度な日本語 MLLM を学習させるデータセットの構築を目的とする。

日本語 MLLM を構築する上で、日本語ドメインの画像の理解が重要となる。特に、日本固有の物体を認識する能力、日本語の文字を読む能力が重要となる。そこで、本研究ではまず YFCC100M [1] から日本語ドメインの画像を、Common Crawl から日本語や英語の文字が含まれる画像を集めた。これらに対して公開モデルによって生成された対話データを付与

することで、学習データセットを構築した。

実験により、特に YFCC100M から収集した日本語ドメイン画像の対話データがベンチマーク上の性能を大きく向上させる事が分かった。得られたモデルは日本語ドメインの画像理解を問うベンチマークで最も良い性能を示し、構築したデータセットの有用性を確認した。

2 関連研究

MLLM の先駆的な研究の一つとして、LLaVA [2] というモデルがある。CLIP [3] などの大規模な学習を行なった画像エンコーダーと LLM をつなぐことで、少ないデータや計算リソースで様々な物体を含む画像に関する対話能力を獲得した。具体的には、LLM の入力に画像特徴も入れられるようにするために、画像特徴を projector と呼ばれる MLP などの学習可能なモデルによって変換することで両者をつなぐ。より最近では、画像を高解像度で扱う工夫、学習用のデータセットの拡張、より大きな LLM の利用などにより更なる発展を遂げている [4, 5, 6, 7]。

日本語 MLLM を作る試みも近年増えつつある。日本語のデータセットを整備することで LLaVA [2] を元にしたアプローチで学習させたり [8]、モデルマージ [9] を利用したりする試みが行われてきた。これらを評価する上で現在最も一般的に使われている日本語 MLLM のベンチマークは Heron-Bench [10] と呼ばれるもので、日本語ドメインの画像に対してその知識などを問う質問に答えるもので、GPT-4 などで生成文を採点することで評価を行っている。

より最近では JMMMU [11] という、MLLM の評価に一般的に使われる MMMU ベンチマーク [12] を日本語用に拡張したものが提案された。JMMMU は選択肢から解答する形式だが、本研究では VILA-jp [8] と同様に自由形式の対話を行う MLLM の構築を目的とし、選択式への適応は将来課題とする。



図1 本研究で構築した対話データ例を示す。左はYFCC対話、右はOCR対話のデータセット例である。

3 学習データセットの構築

高精度な日本語 MLLM を構築するには、日本ドメインの画像に関する質問と解答からなる対話データを収集する必要がある。本研究では、表1に示す画像ソースを利用し、指示学習用のデータセットを構築した。この節では本研究で構築した対話データ、COCO対話・YFCC対話・OCR対話について順に説明する。

3.1 COCO 対話

Heron [10] では、LLaVA [2] の学習に使われていたデータを日本語に翻訳して学習データとしていた。しかし、指示学習では高品質なデータが望ましい一方で、これらのデータは翻訳時に多くのノイズを含む課題があった。本研究では Mixtral-8x22B-v0.1 [18] という LLM を用いて LLaVA [2] と同様に対話データを付与した。画像の詳細な説明文については Pixtral-12B-2409 [19] という MLLM を用いることで、画像から直接生成した。これらを CALM3-22B-Chat [20](以下 CALM3 と記載)で翻訳することで日本語の対話データを収集した。この時、一連の対話全てを CALM3 に入力とすることで、Heron で用いられていた翻訳データよりも品質の高い翻訳ができていることを定性的に確認した。

3.2 YFCC 対話

YFCC100M [1] は撮影場所が日本のデータを多く含んでいる。そこで、これらの画像の一部に対して InternVL2.5-38B [6] という MLLM を用いて詳細な説明文を生成させた。また、日本ドメインの画像に関する論理的な推論能力も獲得させるために、この詳細な説明文を用いて LLaVA [2] と同様に対話データも Qwen2.5-32B [21] という LLM によって生成した。これらのモデルは日本語を扱う能力も有しているた

め、英語を介さずに日本語で直接生成させた。実際に生成した例は図1左である。

3.3 OCR 対話

画像内の文字内容の理解も必要とする対話データの収集も行った。Common Crawl の画像に対して OCR を適用し、日本語や英語の文字が検出された画像に対して、InternVL2.5-38B [6] という MLLM を用いて対話データを収集した。実際に生成した例を図1右に示す。

4 実験

4.1 学習設定

LLaVA [2] と同様に2段階の学習を行なった。1段階目は、LLM と画像エンコーダーを繋ぐ projector のみを、420K の画像テキストペアを用いて1 epoch 学習させた。2段階目は画像エンコーダーの重みは固定し、提案データセットで LLM と projector の重みを1 epoch 学習させた。画像エンコーダーは MLLM において広く用いられている siglip-so400m-patch14-384 [22](以下 SigLIP と記載)、そして LLM は日本ドメインの知識を豊富に持っていると考えられる CALM3 を利用した。高解像度な入力に対応するために、分割した画像を用いて特徴抽出して LLM に入力する、S2-Wrapper [23] を利用した。

4.2 評価方法

VILA-jp [8] と同様に、表2に示す Heron-Bench [10], JA-VLM-Bench-In-the-Wild(以降 JA-VLM-Bench と記載), JA-VG-VQA-500 [9] の3つのデータセットで、LLM-as-a-Judge [24] に基づく GPT-4o による評価と ROUGE-L [25] を用いて評価する。この中で Heron-Bench は最も日本ドメインの知識を問うように設計されており、長い解答が要求される。Heron-Bench

	画像ソース	用途	付与方法	画像数
(本研究で作成)				
COCO 対話	MSCOCO [13]	一般画像の対話	Mixtral-8x22B-v0.1/Pixtral-12B-2409	117K
YFCC 対話	YFCC100M [1]	日本ドメイン画像の対話	InternVL2.5-38B/Qwen2.5-32B	199K
OCR 対話	Common Crawl	文字読み取りの対話	InternVL2.5-38B	168K
(公開データ)				
Japanese Visual Genome (VG) [14]	VG [15]	一般画像の短い対話	人手	98K
Japanese Photo (JP) Instruction [16]	JP [17]	日本ドメイン画像の対話	QwenVL2-7B/Qwen2.5-32B	6K

表 1 指示学習に使用した画像とその用途を示す。

	ドメイン	正解となる文の長さ	評価時の LLM への入力
Heron-Bench	日本ドメイン	複数文	画像情報の記述と正解例
JA-VLM-Bench	日本ドメイン	1 文	正解のみ
JA-VG-VQA-500	一般ドメイン	1 フレーズ	正解のみ

表 2 評価に用いたベンチマークの違いを示す。

では LLM に画像にまつわる情報のテキストも入力することで画像情報も踏まえた評価が行えるが、他のデータセットでの評価では正解の文章と生成文のみが LLM の入力となる。解答が一つに定まらないような質問を含む JA-VG-VQA-500 では評価が正確でない場合があり、Appendix に詳細を示す。

4.3 定量評価

表 3 に定量的な結果を示す。提案モデルは日本ドメインの画像の理解を問う Heron-Bench や JA-VLM-Bench で既存のモデルよりも高い性能を達成した。特に性能が向上した Heron-Bench は、GPT-4o と他のモデルの差分が大きく、日本ドメインの画像理解を問う最も難しいベンチマークになっており、日本語 MLLM の発展に重要なベンチマークと考える。一方で、ROUGE-L の性能は向上していないが、これは VILA-jp [8] の論文内でも論じられているように、ROUGE-L が正解と意味は同じだが単語レベルでは異なる文の類似度を正しく評価できないためと考える。また、一般ドメインの画像理解を問う JA-VG-VQA-500 では VILA-jp [8] が上回る評価値となった。これは、Appendix に示すように、提案モデルが解答以上の長い説明を行い、それが減点要素となってしまった事が一因として挙げられる。

表 4 にそれぞれのデータセットを抜いて指示学習を行った時の結果を示す。Heron-Bench においては特に YFCC 対話を抜いた時に精度が大きいことから、日本ドメインの画像に関する対話のデータが重要であることが分かる。一方で OCR 対話の性能への寄与が小さいが、これは文字の読み取りに関する質問が少ないためと考えられ、こういったベンチ

マークの作成も重要な課題であると考えられる。

4.4 定性評価

図 2 に定性的な例を示す。提案モデルは VILA-jp [8] に比べて、図左の例では左右の位置関係は誤っているものの、風神・雷神といった固有名詞を説明できている。図右の例では各標識を正しく認識して説明できており、LLM による評価のスコアも提案モデルが優れている。ここで、右の例で提案モデルでは VILA-jp [8] よりも ROUGE-L の値が低く出ているが、この例からも本来行いたい意味レベルでの評価を ROUGE-L では正確にできず、LLM によるスコアの方が信頼性が高いものであることが示されていると考える。

5 ディスカッション

本研究では、指示学習データに焦点を置いた。しかし、指示学習は画像エンコーダーと LLM が持つ知識をつないで強化することを目的とするため、事前に学習された画像エンコーダーや LLM の性能の問題によって構築した MLLM が日本ドメインの画像を認識できない場合がある。

日本語 MLLM における LLM の重要性

図 2 の左で SigLIP と CALM3 を組み合わせた提案モデルが「風神」と「雷神」を説明できたのは、SigLIP と CALM3 が風神雷神を十分に知っていたためと考えられる。実際に SigLIP でこの画像と“Fujin Raijin-zu”・“Wind God and Thunder God”というテキストの類似度を測ると、95%・99%と高い類似度を示した。また、LLM に「風神雷神図屏風」について説明させると、CALM3 は「風神は大きな袋、雷神

Models	Heron-Bench	JA-VLM-Bench		JA-VG-VQA-500	
	LLM (%)	ROUGE-L	LLM (/5.0)	ROUGE-L	LLM (/5.0)
Japanese InstructBLIP Alpha [26]	14.0	20.8	2.42	–	–
Japanese Stable VLM [27]	24.2	23.3	2.47	–	–
Llama-3-EvoVLM-JP-v2 [9]	39.3	41.4	2.92	23.5	2.96
LLaVA-CALM2-SigLIP [28]	43.3	47.2	3.15	17.4	3.21
VILA-jp [8]	57.2	52.3	3.69	16.2	3.62
提案モデル	63.3	45.0	3.82	15.4	3.45
GPT-4o	87.6	37.6	3.85	12.1	3.58

表 3 日本語 MLLM の評価ベンチマークにおける比較結果を示す。すべての指標は高いほど良い性能であることを示す。LLM は GPT-4o を用いた LLM-as-a-Judge による評価結果を表す。

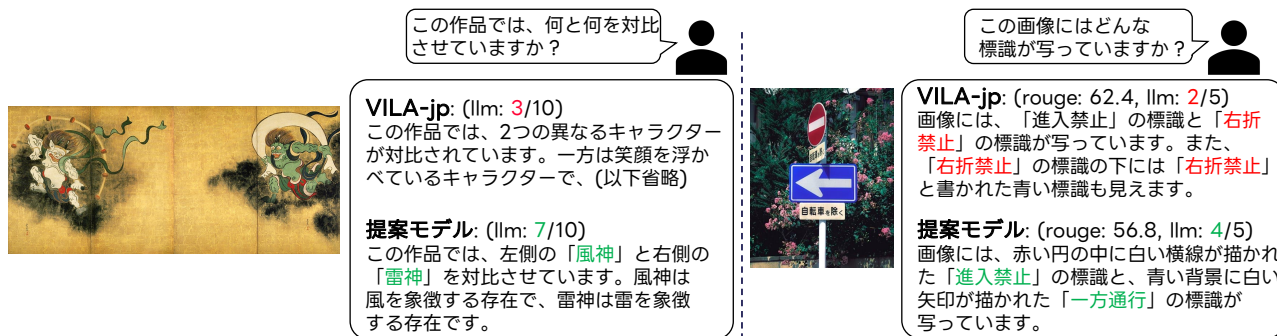


図 2 左は Heron-Bench, 右は JA-VLM-Bench での生成結果の比較例を示す。

Models	Heron-Bench	JA-VLM-Bench
	LLM Average (%)	LLM (/5.0)
提案 w/o COCO 対話	60.9	3.62
提案 w/o YFCC 対話	58.0	3.40
提案 w/o OCR 対話	62.3	3.40
提案 w/o Japanese VG	61.9	3.54
提案 w/o JP Instruction	60.1	3.7
提案	63.3	3.82

表 4 それぞれのデータセットを除いた時の性能の変化を示す。

は太鼓を叩いている」という外観的特徴を捉えた説明ができた。一方で、VILA-jp が LLM として用いた llm-jp-3-13b-instruct は、「風神は口を開けて恐ろしい顔つきで」といった記述のみを生成し、風神の見た目の特徴の理解が曖昧であった。VILA-jp が風神雷神を説明できなかったのは LLM の知識不足のためである可能性が高い。LLM 自身が持つ日本ドメインの知識を深めることで、日本語 MLLM の性能に大きく寄与するものと考えられる。

日本語 MLLM における画像エンコーダーの重要性

本研究では画像エンコーダーに SigLIP [22] を用いたが、これは WebLI [29] という海外の画像テキストペアを中心に学習している。そのため、日本固有の画像の理解に弱い面があり、実際に Appendix に一例を示すように、提案モデルは Heron-Bench でも日本特有の基本的な物体をまだまだ誤認識してい

る。これは SigLIP の学習データに起因する課題であるため、VILA-jp でも同様の間違いが起きている。VILA-jp では web から収集した数千万枚の日本ドメインの画像とテキストペアを用いているが、LLM を用いた学習はコストが高いためそれぞれのステップでの学習は 1epoch にとどまり、新しい日本ドメインの知識をうまく獲得できていない課題がある。より効率的に日本ドメインに強い画像特徴抽出を行うには、画像エンコーダーも日本ドメインに強いものを構築していくのが望ましいと考えられる。

6 おわりに

本研究では、日本ドメインの画像理解を問うベンチマークで、既存の公開モデルを上回る性能を達成した。特に、YFCC100M を用いて日本ドメインの画像に関する対話のデータが性能に最も寄与した。しかし、提案モデルでも基本的な日本ドメインの物体の誤認識が多く起こるを確認した。これは、画像エンコーダーが海外のデータがメインで学習されたものを用いているためと考えられる。日本語 MLLM の更なる精度向上には、データの拡張に加えて、日本ドメインに強い画像特徴抽出方法が将来課題として重要と考える。

参考文献

- [1] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. **Commun. ACM**, Vol. 59, No. 2, pp. 64–73, 2016.
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In **ICML**, 2021.
- [4] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **CVPR**, 2024.
- [5] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
- [6] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. **arXiv preprint arXiv:2412.05271**, 2024.
- [7] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. **arXiv:2409.12191**, 2024.
- [8] Constructing multimodal datasets from scratch for rapid development of a japanese visual language model. **arXiv:2410.22736**, 2024.
- [9] Evolutionary optimization of model merging recipes. **arXiv:2403.13187**, 2024.
- [10] 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:2404.07824**, 2024.
- [11] Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Kazuki Egashira, Jeonghun Baek, Xiang Yue, Graham Neubig, and Kiyoharu Aizawa. JMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark. **arXiv:2410.17250**, 2024.
- [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 Wenhua Chen. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In **CVPR**, 2024.
- [13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In **ECCV**, 2014.
- [14] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In **COLING**, pp. 1918–1928, 2018.
- [15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yanis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. **IJCV**, Vol. 123, pp. 32–73, 2017.
- [16] Kendamarron/japanese-photo-instruction, 2024.
- [17] Thepioneer/japanese-photos, 2024.
- [18] mistralai/mixtral-8x22b-v0.1, 2024.
- [19] mistralai/pixtral-12b-2409, 2024.
- [20] Ryosuke Ishigami. cyberagent/calm3-22b-chat, 2024.
- [21] Qwen Team. Qwen2.5: A party of foundation models, September 2024.
- [22] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **ICCV**, 2023.
- [23] Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. When do we not need larger vision models? **ECCV**, 2024.
- [24] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **NeurIPS**, Vol. 36, pp. 46595–46623, 2023.
- [25] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, 2004. ACL.
- [26] Makoto Shing and Takuya Akiba. Japanese instructblip alpha, 2023.
- [27] Makoto Shing and Takuya Akiba. Japanese stable vlm, 2024.
- [28] Aozora Inagaki. cyberagent/llava-calm2-siglip, 2024.
- [29] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In **ICLR**, 2023.

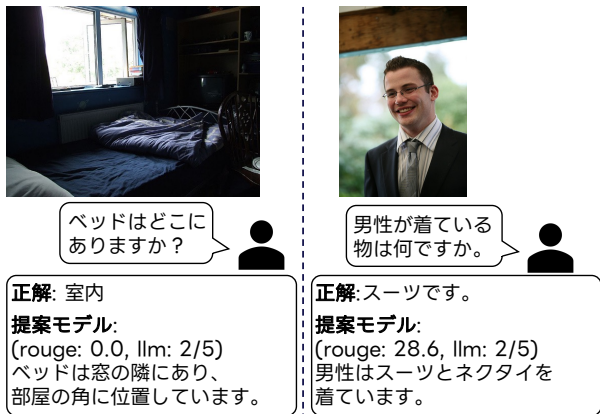


図3 JA-VG-VQA-500 の評価方法のリミテーションを示す。いずれも生成結果は正しいが低い評価値となってしまう。



図4 Heron-Bench における失敗例を示す。いずれも日本特有の物体の認識に失敗している。

A JA-VG-VQA-500 の評価方法のリミテーション

図3に、JA-VG-VQA-500 の評価方法のリミテーションを示す。提案モデルが生成した文章は正確に質問に解答できているが、正解として付与されたラベルと一致しないために ROUGE-L や LLM による評価値が低くなってしまっている。LLM の評価値が低くなるのは、Heron-Bench とは違って画像情報が与えられていないため、解答例以上に詳しい説明部分が誤りとして判定されてしまうためである。

B 画像エンコーダー起因と考えられる MLLM の失敗例

図4に、失敗例を示す。それぞれ、提案モデルも VILA-jp も共に日本特有の物体である、シーサー・油そば・みかん・こたつの説明に失敗している。こ

Models	Heron-Bench	JA-VLM-Bench
	LLM Average (%)	LLM (/5.0)
VILA-jp (SigLIP+llm-jp)	57.2	3.69
提案 (SigLIP+llm-jp)	56.6	3.62
提案 (SigLIP+CALM3)	63.3	3.82

表5 提案データセットで、VILA-jp と同じモデル構成にした時の結果を示す。llm-jp は llm-jp-3-13b-instruct を表す。

これは両者とも日本ドメインの画像理解に弱い面を持つ SigLIP を画像エンコーダーとして用いているためと考えられる。

C LLM を変えた時の実験

表5に VILA-jp と同じ LLM を使って提案データセットで学習させた時の実験結果を示す。学習には1ノードの 8xA100 (80GB) で合計30時間要した。同じデータセットで学習させたときに CALM3-22B-Chat の方が llm-jp-3-13b-instruct よりも精度が高いため日本語 MLLM においても LLM の性能が重要であることがわかる。VILA-jp とは同じモデル構成で学習データが異なるが、学習に3つのステップを要し、特に2つ目の学習ステップで8ノードの 8xA100 (40GB) で130時間を要する VILA-jp よりも少ない計算リソースで同等の精度を達成している。