マルチモーダル LLM の縦書きテキスト読み取り能力の評価

笹川 慶人 [♣,‡], 栗田 修平 [†,‡], 河原 大輔 [♣,‡]
♣ 早稲田大学, † 国立情報学研究所, ‡ 国立情報学研究所 大規模言語モデル研究開発センター
{kate@fuji.,dkw@}waseda.jp, skurita@nii.ac.jp

概要

文書画像からその内容を理解するタスクはマルチモーダル大規模言語モデル (MLLM) の代表的な応用先である。このタスクを解くには、まず文書に何が書かれているかを読み取る必要がある。特に日本語には縦書きの文書があるため、それに対応する必要があるが、日本語の縦書き文書の読解に焦点を当てた研究は限られている。本研究では、MLLM の日本語縦書きテキスト読み取り能力を評価する。実験のために、日本語の OCR 合成データセットを構築する。また、現実世界の縦書き文書から評価データセットを作る。構築したデータセットでの評価の結果、現在の MLLM は横書きテキストに比べると、縦書きテキストに対する性能は低いことが明らかになった。さらに、合成した日本語縦書き OCR データセットで MLLM を学習した結果、縦書きテキストがうまく読み取れなかったモデルの性能が向上することがわかった。

1 はじめに

マルチモーダル大規模言語モデル (MLLM) の発展は著しく [1, 2, 3, 4]、さまざまなマルチモーダルタスクに応用されている。それらのタスクの一つに文書画像 QA タスクがある。これは入力された文書画像に対する質問に回答するタスクであり、さまざまなデータセットが提案されている [5, 6, 7, 8, 9, 10]。このようなタスクを解くには、MLLM は入力文書画像内のテキストを読む必要がある。英語では OCR タスクを含むさまざまなデータセットが整備されており、現在の MLLM は英語ベンチマークにおいて高精度に英語テキストが読めることが示されている。日本語の文書には、縦書きテキストが含まれることがあり、それらに対応する必要がある。しかし、ほとんどの MLLM の OCR ベンチマークは横書きテキストに焦点を当てており、日本語縦書きテキストが読めるかについては調べられていない。

この問題に対処するため、本研究では MLLM の縦書き日本語テキストの OCR 能力を評価する。そのために、日本語のテキストを描画した合成画像のデータセット (**JSSODa**) を構築する。このデータセットの画像には、横書きと縦書きの両方があり、複数段組み (1-4 段) のテキストを含む。さらに実世界の PDF の縦書きテキストを含むページから、OCR データセット (**VJRODa**) を構築して評価を行う。

実験では、複数のオープン MLLM とクローズド MLLM、さらに合成画像のデータセットで Fine-Tuning したモデルを構築したテストデータセットで評価する。現在の MLLM は横書き日本語テキストよりも縦書き日本語テキストの方が読み取り性能が低いことが予想されていたが、評価の結果これを確認した。また、日本語テキストの合成画像データセットを用いて、縦書きの日本語テキストの読み順がわかっていなかったモデルを Fine-Tuning すると、合成画像と実世界の文書画像両方において、縦書きの日本語テキストの読み取り性能が向上することがわかった。本研究のデータセットとコードは一般公開している[1)]。

2 関連研究

2.1 マルチモーダル LLM

近年の MLLM では Vision Encoder と LLM を Projector で接続するアーキテクチャがよく採用される [11, 12]。Vision Encoder によって入力画像の特徴量を抽出し、Projector に入力することで、LLM が扱うことができる画像トークンに変換し、テキストトークンと一緒に LLM に入力する。Vision Encoder としては CLIP [13] や SigLIP [14] の Vision Transformer [15] が、Projector としては多層パーセプトロンがよく採用される。文書画像を扱うことのできる MLLM についても、様々な研究がされている [16, 17, 18, 19, 20, 21, 22, 23, 24]。中でも、日本語の

1) https://github.com/llm-jp/eval_vertical_ja

文書画像を扱える MLLM としては、Qwen2.5-VL [1] や InternVL3 [2]、Gemma 3 [3] などがある。

2.2 縦書き OCR データセット

自然画像中の文字を読み取るタスク (情景文字認識タスク) 向けの縦書きテキストのデータセットとして SVTD と VTD142 [25] がある。SVTD は合成縦書きテキスト画像データセットで、VTD142 は Web から集められた現実にある縦書きテキスト画像データセットである。折橋ら [26] の研究では、合成した縦書きテキストの画像と、ICDAR2019 の Multi-lingual Scene Text Detection and Recognition [27] コンペティションのデータセットにおける日本語の横書きと縦書きテキスト画像を実験に使用した。CC-OCR [28] は MLLM の OCR 能力を評価するためのデータセットで、日本語縦書きテキストの画像を含む。これらのデータセットの画像はフレーズレベルのテキストしか付与されていないことが多く、文章レベルの評価が行えない。

SynthDoG [29] では、文書のテクスチャを背景画像に合成し、その上からテキストを描画することで画像を合成している。このデータセットの画像は部分的に縦書きの日本語テキストを含むが、その割合は小さく、また、現実に存在しないような不自然な文書画像で評価データに適していない。MangaOCR [30, 31, 32] は漫画における会話や効果音のテキストに関する OCR データセットである。縦書き日本語テキストを含むが、画像が漫画のページに限られてしまう。本研究では、縦書き日本語テキストを文章単位で含む OCR データセットを構築し、MLLM の評価を行う。

3 データセット構築

MLLM の縦書き日本語テキスト OCR 能力を評価するために、二種類のデータセットを構築する。これらのデータセットは、画像とその中に書かれているテキストのペアで構成する。一つ目のデータセットは **J**apanese **S**imple **S**ynthetic **OCR** **Da**taset (**JSSODa**) で、LLM で生成した日本語テキストを画像として描画することで構築する。合成データセットで大規模に作成することができ、モデルの学習にも使える。もう一つは **V**ertical **J**apanese **R**eal-world **OCR** **Da**taset (**VJRODa**) で、実世界の縦書き日本語テキストを含む画像からできている。現実に存在する文書であるため、現実的な設定で評価を行うことができる。

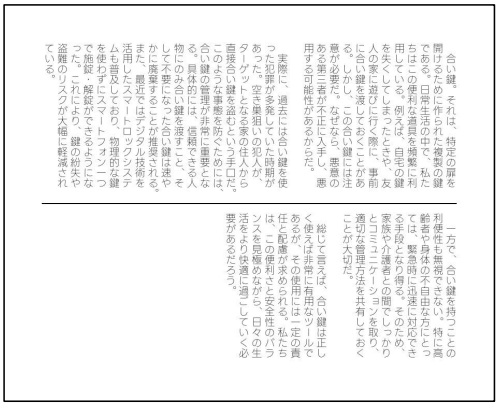

図 1: JSSODa の画像例 (縦書き、2 段組み)

3.1 JSSODa の構築

大規模な評価データを作成するため、日本語テキストを画像に描画することでデータセットを構築する。MLLM は事前学習時に大量のテキストデータを使用して学習するため、データセットの構築に Web 上のテキストを使用すると、それがモデルの学習に使われている可能性があるため、モデルの適切な評価ができなくなる可能性がある。そこで、LLM によって生成した日本語テキストをデータとして使用する。

テキスト生成 LLM で日本語の名詞から、それに関する文章を生成する。日本語の名詞は JUMAN 辞書[2] の名詞を使用した。LLM は llm-jp-3.1-instruct4 [33] を使用した。使用したプロンプトを付録 D に示す。生成した文章のうち、長さが 100 字以下、および 3,000 字以上のものは除去した。

画像の合成 次に、生成した日本語文章を用いて、画像を合成する。縦書き、横書きそれぞれについて、1-4 段組みの計 8 種類のデータを生成する。横書きの場合、文字は左から右に、行は上から下に描画する。複数段組みの場合、左から右にそれぞれの段のテキストを描画する。縦書きの場合は、文字は上から下に、行は右から左に描画する。複数段組みの場合、上から下にそれぞれの段のテキストを描画する。上記の文字の描画順に基づいて、LLM によって生成した文章の文字を Pillow ライブラリ[3] を使用して、一文字ずつ画像に描画することで画像を合成した。フォントは google fonts[4] と free-fonts.jp[5] から集めた約 200 種類の日本語フォントを使用し

2) https://github.com/ku-nlp/JumanDIC
3) https://github.com/python-pillow/Pillow
4) https://fonts.google.com/
5) https://free-fonts.jp/

図 2: VJRODa の画像例 (https://warp.ndl.go.jp/info:ndljp/pid/11712522/www.vill.kariwa.niigata.jp/open/info/000000001_0000000609.pdf, 8 ページ, 一部を黒塗りした。)

た。使用する文章のうち、長さの上位 25%を 4 段組み、次の 25%を 3 段組み、次の 25%を 2 段組み、残りを 1 段組みとした。合計画像数は 22,493 枚となった。合成した画像の例を図 1 に示す。

構築したデータセットについて、8 種類の画像数を同じ割合に保ちつつ、学習: 検証: テストデータを 8:1:1 に分割した。

3.2 VJRODa の構築

日本語 PDF 文書のページから縦書きを含むものを自動および人手で抽出し、画像内のテキストを付与する。国立国会図書館 WARP[6] に収録されている日本語の PDF について、それぞれのページを画像化して、以下の処理を行った。

(1) 縦書き日本語テキストを含まない画像の除去 投影プロファイル法による縦書き判別 [34] と Tesseract OCR [35] によって検出された bbox による判別により、自動でフィルタリングした後、人手で縦書きを含む画像を 100 枚収集した。

(2) 画像内のテキストの書き起こし PyMuPDF[7] と PaddleOCR [36] によって、PDF のページからテキストを抽出した。抽出したテキストにおける文字誤りと読み順を人手で修正し、正解のテキストとした。図 2 に画像例を示す。

6) https://warp.ndl.go.jp/
7) https://github.com/pymupdf/PyMuPDF

4 実験

4.1 実験設定

4.1.1 モデルの Fine-Tuning

JSSODa の学習データを用いて、オープン MLLM を学習する。日本語テキストを読む能力のある Qwen2.5-VL7B-Instruct、InternVL3-8B-hf、Gemma 3 12b IT の 3 つのモデルを Fine-Tuning した。これらのモデルは Vision Encoder と Projector、LLM から構成されるモデルで、本研究では全てのパラメータをチューニングした。エポック数は 1、バッチサイズは 32 とし、オプティマイザーには AdamW [37] を使用し、学習率は 2e-5 に設定した。

4.1.2 評価

JSSODa のテストデータと VJRODa において、複数のオープン MLLM とクローズド MLLM を評価した。オープン MLLM としては、Qwen2.5-VL の 7B と 32B パラメータのモデル、InternVL3 の 8B と 38B パラメータのモデル、Gemma3 の 12B と 27B パラメータのモデルを使用した。また、4.1.1 項で述べた 3 つのモデルも評価した。推論時は、Greedy Decoding によって、テキストを生成した。クローズドモデルとしては、GPT-4.1 [38] と GPT-5 [4] を使用した。最大トークン長は JSSODa のテストデータでは 1024 とし、VJRODa では 3072 とした。また、ユーザープロンプトは、モデルの Fine-Tuning 時に使用したものと同じものを使用した。

4.1.3 評価指標

評価指標は Character Error Rate (CER) と BLEU [39] を使用した。スコアの計算の前に、NFKC ユニコード正規化を適用し、空白文字を除去した。

CER は以下の式で定義される:

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

$pred$ はモデルの予測テキストで、ref は正解テキストである。$\mathrm{EditDistance}(p, r)$ は文字列 p と r の編集距離である。CER の値は低い方が性能がよいと考えられる。

BLEU スコアの計算には SacreBLEU [40] を使用した。テキストは文字単位に分割した。

MLLM は繰り返し同じトークンを最大トークン長まで生成してしまうことがある。この現象が起こ

表 1: JSSODa における評価結果

モデル	Raw Output		Remove Repetition	
	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
Horizontal Writing				
Qwen2.5-VL-7B	18.0	86.7	16.7	86.5
Qwen2.5-VL-32B	8.65	**99.2**	8.65	**99.2**
InternVL3-8B-hf	12.0	96.8	11.8	96.8
InternVL3-38B-hf	5.06	98.8	5.06	98.8
Gemma 3 12B IT	13.2	90.5	12.5	90.4
Gemma 3 27B IT	10.3	95.1	10.2	95.1
GPT-4.1	**1.27**	98.4	**1.27**	98.4
GPT-5	1.65	98.0	1.65	98.0
Qwen2.5-VL-7B (+FT)	**0.0385**	**99.9**	**0.0385**	**99.9**
InternVL3-8B-hf (+FT)	0.214	99.7	0.214	99.7
Gemma 3 12B IT (+FT)	0.520	99.0	0.520	99.0
Vertical Writing				
Qwen2.5-VL-7B	104	21.8	83.0	20.1
Qwen2.5-VL-32B	109	21.5	88.2	24.3
InternVL3-8B-hf	68.1	58.4	46.5	68.4
InternVL3-38B-hf	37.1	**76.8**	30.3	**80.6**
Gemma 3 12B IT	46.7	61.5	37.4	64.6
Gemma 3 27B IT	**30.0**	75.7	**26.6**	77.7
GPT-4.1	40.1	70.0	38.7	69.8
GPT-5	48.7	67.6	48.7	67.6
Qwen2.5-VL-7B (+FT)	**0.175**	**99.8**	**0.175**	**99.8**
InternVL3-8B-hf (+FT)	0.593	99.4	0.593	99.4
Gemma 3 12B IT (+FT)	1.78	97.4	1.78	97.4

ると、極端に悪いスコアになりやすくなり、スコアが参考にならない可能性がある。この問題に対処するため、出力テキストのうち、繰り返し部分を除去した場合のスコアも報告する。正規表現を使って、文字列全体のうち、最後の 10 回以上連続する文字列を、最初の 1 回の部分を除いて除去する。

4.2 結果

JSSODa での評価結果 表 1 に JSSODa のテストデータにおける評価結果を示す。“Raw Output” は MLLM の出力をそのまま評価した場合の結果で、“Remove Repetition” は繰り返し部分を除去した場合の結果である。“(+FT)” は 4.1.1 項で述べたモデルを表す。結果より、どのモデルも横書きのテキストはある程度読めていたが、それと比較すると縦書きはどのモデルでも性能が悪かった。特に、Qwen2.5-VL は縦書きのテキストでも、横書きの読み順で読んでしまうことがあった。InternVL3、Gemma 3、GPT-4.1、GPT-5 は文字の読み順はある程度分かっていたが、横書きよりも文字認識の誤りが多くなった。InternVL3 と Gemma 3 では、モデルのパラメータ数が大きくなるほど、縦書きテキストにおけるスコアが良くなっていた。GPT-5 はテキストを全く出力しない場合があった。推論部分のみで最大トークン長まで生成しきってしまっている可能性がある。また、繰り返しを除去するとスコアが良くなっており、MLLM はトークンを繰り返し生成していることがわかる。さらに、合成データでモデルを学習すると、縦書きにおけるスコアが向上した。

表 2: VJRODa における評価結果

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

VJRODa での評価結果 表 2 に VJRODa における評価結果を示す。どのモデルもスコアは悪く、実世界の文書画像の縦書きテキストは不得意であることがわかる。合成データで学習した場合、Qwen2.5-VL-7B の性能は “Raw Output” と “Remove Repetition” の両方で向上していたが、他の 2 モデルの大幅な性能向上は見られなかった。これは、Qwen2.5-VL-7B は日本語の縦書きの読み順があまりわかっていなかったが、他の 2 モデルはある程度理解していたからだと考えられる。縦書き日本語テキストの読み順がわからないモデルの性能向上に、JSSODa が有効であることが示唆される。一方、実世界の文書の縦書きテキストを読む能力を向上させるには、実世界の OCR データセットが必要だと考えられる。

5 おわりに

本研究では、MLLM の縦書き日本語テキストの OCR 能力について評価を行った。そのために、横書きと縦書きの日本語テキストの合成画像データセットと、実世界の PDF から集めた、縦書き日本語テキストを含む文書画像のデータセットを構築し、評価を行った。評価の結果、現在の MLLM は縦書きの日本語テキストの読み取り性能がよくないことがわかった。今後は、日本語のさまざまな文書画像に対応できるモデル構築のための手法を探求したい。

謝辞

本研究は、文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです。本研究成果は、データ活用社会創成プラットフォーム mdx [41] を利用して得られたものです。

参考文献

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

[2] 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, 2025.

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

[4] OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, aug 2025. Accessed: 2025-09-24.

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

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

[7] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. VisualMRC: Machine reading comprehension on document images. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, No. 15, pp. 13878–13888, May 2021.

[8] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A dataset for document visual question answering on multiple images. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, No. 11, pp. 13636–13645, Jun. 2023.

[9] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multipage DocVQA. **Pattern Recognition**, Vol. 144, p. 109834, 2023.

[10] Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. JDocQA: Japanese document question answering dataset for generative language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 9503–9514, Torino, Italia, May 2024. ELRA and ICCL.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 34892–34916. Curran Associates, Inc., 2023.

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 26296–26306, June 2024.

[13] 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 Marina Meila and Tong Zhang, editors, **Proceedings of the 38th International Conference on Machine Learning**, Vol. 139 of **Proceedings of Machine Learning Research**, pp. 8748–8763. PMLR, 18–24 Jul 2021.

[14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 11975–11986, October 2023.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **International Conference on Learning Representations**, 2021.

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

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

[18] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for document ai with unified text and image masking. In **Proceedings of the 30th ACM International Conference on Multimedia**, MM '22, p. 4083–4091, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mPLUG-DocOwl: Modularized multimodal large language model for document understanding, 2023.

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

[21] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 3096–3120, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[22] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5817–5834, Vienna, Austria, July 2025. Association for Computational Linguistics.

[23] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2-4KHD: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 42566–42592. Curran Associates, Inc., 2024.

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

[25] Chankyu Choi, Youngmin Yoon, Junsu Lee, and Junseok Kim. Simultaneous recognition of horizontal and vertical text in natural images. In Gustavo Carneiro and Shaodi You, editors, **Computer Vision – ACCV 2018 Workshops**, pp. 202–212, Cham, 2019. Springer International Publishing.

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

[27] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, and Jean-Marc Ogier. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition — RRC-MLT-2019. In **2019 International Conference on Document Analysis and Recognition (ICDAR)**, pp. 1582–1587, 2019.

[28] 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, 2024.

[29] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free document understanding transformer. In **European Conference on Computer Vision (ECCV)**, 2022.

[30] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "Manga109" with annotations for multimedia applications. **IEEE MultiMedia**, Vol. 27, No. 2, pp. 8–18, 2020.

[31] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using Manga109 dataset. **Multimedia Tools and Applications**, Vol. 76, No. 20, pp. 21811–21838, 2017.

[32] Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. MangaVQA and MangaLMM: A benchmark and specialized model for multimodal manga understanding, 2025.

[33] LLM-jp. LLM-jp: A cross-organizational project for the research and development of fully open japanese LLMs. **CoRR**, Vol. abs/2407.03963, , 2024.

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

[35] R. Smith. An overview of the Tesseract OCR engine. In **Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)**, Vol. 2, pp. 629–633, 2007.

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

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

[38] OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, apr 2025. Accessed: 2025-09-24.

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[40] Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[41] Toyotaro Suzumura, Akiyoshi Sugiki, Hiroyuki Takizawa, Akira Imakura, Hiroshi Nakamura, Kenjiro Taura, Tomohiro Kudoh, Toshihiro Hanawa, Yuji Sekiya, Hiroki Kobayashi, Yohei Kuga, Ryo Nakamura, Renhe Jiang, Junya Kawase, Masatoshi Hanai, Hiroshi Miyazaki, Tsutomu Ishizaki, Daisuke Shimotoku, Daisuke Miyamoto, Kento Aida, Atsuko Takefusa, Takashi Kurimoto, Koji Sasayama, Naoya Kitagawa, Ikki Fujiwara, Yusuke Tanimura, Takayuki Aoki, Toshio Endo, Satoshi Ohshima, Keiichiro Fukazawa, Susumu Date, and Toshihiro Uchibayashi. mdx: A cloud platform for supporting data science and cross-disciplinary research collaborations. In **2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)**, pp. 1–7, 2022.

[42] Haoran Wei, Yaofeng Sun, and Yukun Li. DeepSeek-OCR: Contexts optical compression, 2025.

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

A Limitations

A.1 視覚的に多様な学習用合成データ

JSSODa データセットは白背景に黒テキストという、シンプルな文書画像であり、ノイズなども加えられていない。画像や表、図なども挿入されていない。実世界の文書画像に対応できるようにするためには、より視覚的に多様な画像を含むデータセットを整備する必要がある。

A.2 文書画像内のテキストの読み順が複数考えられる場合の評価方法

現実の文書画像では、読み順が複数通り考えられる場合がある。例えば、画像のキャプションのテキストを読む順番は複数通り考えられる。VJRODa データセットでは、一つの読み順しかアノテーションされておらず、CER、BLUE といった指標では、正しい読み順でも、スコアが良くない可能性がある。理想的には、全ての正解の読み順を付与し、スコアが一番よくなるものを選択するのがよいと考えられるが、コストが大きくなってしまう。効率的、効果的な MLLM の OCR 能力の評価をすることが今後の課題である。

A.3 VJRODa の画像の多様性

VJRODa の画像のソースは主に公的機関の Web から得られる PDF となっている。そのため、文書画像のドメインが限られており、視覚的多様性に欠ける可能性がある。また、このデータセットの画像には表やグラフなどといった画像は含まれていないため、それらのラベルの日本語テキストが縦書きになっていた場合に、それを認識することができるかなどの評価を行うことができない。

B OCR 特化 MLLM の評価

本研究で提案したデータセットにおいて、最新の OCR 特化 MLLM の性能を調べるため、DeepSeek-OCR [42] と PaddleOCR-VL [43] を評価する。DeepSeek-OCR においては、Gundam (Dynamic Resolution) モードを使用し、プロンプトは “`<image>\nFree OCR.`” とした。max new tokens は JSSODa では 1024、VJRODa では 3072 に設定した。PaddleOCR-VL では、レイアウト解析を行なって MLLM に入力して結果を得るというパイプライン処理を行い、予測された読み順に従ってテキストを結合し、回答として採用した。max new tokens は設定していない。他のモデルとは、プロンプトや max new tokens などの設定が異なることに注意する必要がある。

JSSODa での結果を表 3 に示す。DeepSeek-OCR は横書きテキストの読み取り性能は良かったが、縦書きテキストの読み取り性能は悪かった。PaddleOCR-VL では DeepSeek-OCR ほど縦書きテキストでの性能は悪化しなかった。VJRODa での結果を表 4 に示す。DeepSeek-OCR はスコアが低く、縦書きの日本語テキストに対する性能は悪いということがわかる。PaddleOCR-VL は表 2 のモデルと合わせて、全てのモデルで一番スコアがよかった。

表 3: JSSODa における OCR 特化 MLLM の評価

Models	Raw Output		Remove Repetition	
	CER(↓)	BLEU(↑)	CER(↓)	BLEU(↑)
Horizontal Writing				
DeepSeek-OCR	3.12	98.5	31.2	98.5
PaddleOCR-VL	8.68	94.4	8.68	94.4
Vertical Writing				
DeepSeek-OCR	138	16.4	137	16.5
PaddleOCR-VL	13.2	93.6	9.96	96.6

表 4: VJRODa における OCR 特化 MLLM の評価

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

C 事例分析

表 5: 図 2 の事例におけるモデルの出力例

Qwen2.5-VL-7B (全出力を表示)
国民年金\n ねんきん特別便について\nQ1\n ねんきん特別便に関するよくある質問\nA1\n 年金特別便の一年金記録のお知らせ\nQ2\nA2\n お問い合わせ先\nTEL [REDACTED]\n2008.10.10 広報かりわ
Qwen2.5-VL-7B (+FT) (出力の一部を表示)
民年金\n\n ねんきん特別便について\n\n 社会保険庁ではみなさまにご自分の年金加入記録を確認していただくため平成十九年十二月から平成二十一年十月にかけて、年金加入記録をお送りしています。\n\n ご自分の年金加入記録を確認し、間違いの有る・無しを問わず必ず回答してくださいますようお願いします。\n\n ねんきん特別便に関するよくある質問\n\nQ1\n\n 先日、ねんきん特別便が届きました。「年金記録のお知らせ」の内容を確認したところ、結婚前に旧姓で加入していた厚生年金の記録が漏れています。どのようにしたらいいですか？…

VJRODa におけるモデルの出力例を表 5 に示す。Qwen2.5-VL では一部のテキストしか出力していなかったが、Fine-Tuning することで、縦書きの読み順でテキストを出力するようになった。

D プロンプト

JSSODa を作る際、日本語の名詞から文章を LLM に生成させた際のプロンプトを表 6 に示す。JSSODa の学習データでモデルをチューニングする際と、JSSODa テストデータと VJRODa でモデルを評価する際に使用したプロンプトを表 7 に示す。

表 6: 日本語の名詞から文章を生成する際のプロンプト

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

表 7: 学習時と評価時のプロンプト

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