JaWildText: 日本語文字認識性能評価のための実世界画像データセット

前田 航希 [S,N], 岡崎 直観 [S,N]
[S] 東京科学大学, [N] NII LLMC
{koki.maeda@nlp.,okazaki@}comp.isct.ac.jp

概要

日本語の実世界画像における文字認識と，下流タスクを一体的かつ網羅的に評価できる公開ベンチマークは存在しない．本研究では，**高密度 STVQA・レシート KIE・手書き文字 OCR** を収録した日本語文字認識性能評価のためのデータセット **JaWildText** を構築する．**JaWildText** は 335 枚の実世界画像からなり，純粋な文字認識能力から推論能力までを多面的に測定できる．本稿では，構築したデータセットを用いて公開 VLM および OCR 特化モデルを評価し，最先端のモデルにも日本語読字性能に改善の余地が残されていることを定量的に示す．

1 はじめに

画像と自然言語を理解し汎用的なタスクで推論する視覚言語モデル（Vision Language Model; VLM）の基盤能力の一つに，画像内文字認識（Optical Character Recognition; OCR）がある．画像中の文字は高レベル推論や実用タスクに不可欠であり，Robust Reading Challenge [1] を始め，画像内テキストを扱う画像質問応答（Visual Question Answering; VQA） [2, 3, 4] が盛んに提案されてきた．また近年では VLM を基盤とした OCR 特化モデル [5, 6, 7] も開発されている．しかしながら，これらの多くは英語や中国語の文字を主たる認識対象とし，実世界画像における日本語の文字認識およびその下流タスクには十分に取り組まれていない．

日本語は漢字仮名交じり記法やアルファベットとの混植，読み方向の併存といった固有の表記体系を持つ．加えて実世界画像は，傾きや照明などの撮影条件の変化，背景の写り込みや部分的な遮蔽が生じるため，文字認識および推論を一層困難にする．このような条件下での日本語文字理解性能を包括的に評価するためには，実世界画像を対象とし，文字認識と下流タスクを一貫して扱うことが必要である．多言語文字認識 [8, 9] や文書画像理解 [10] など日本語を対象とした既存の課題も存在するが，表 1 に示すように既存のデータセットだけでは課題の網羅性が不十分である．このような評価基盤の欠落が，VLM 研究の日本語環境への適応を阻害している．

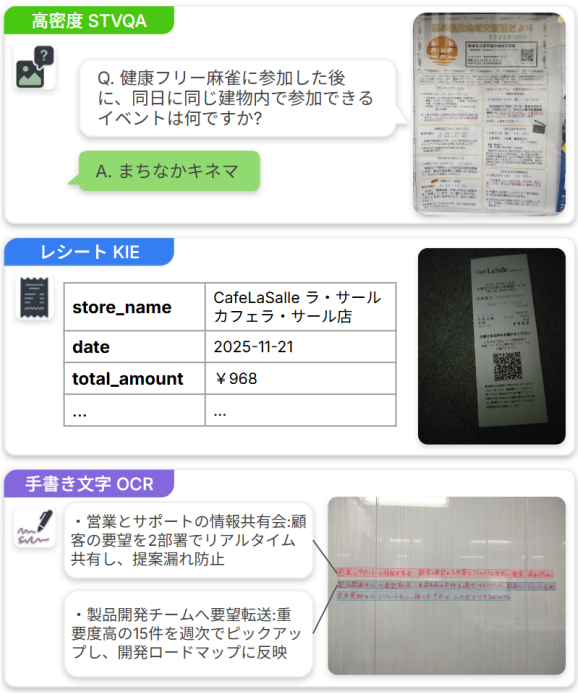

図 1 **JaWildText** は (i) **高密度 STVQA**，(ii) **レシート KIE**，(iii) **手書き文字 OCR** から構成された，文字認識性能評価のための実世界画像データセットである．

これらの課題を解決するため，我々は実世界で撮影された 335 枚の画像から構成される日本語文字理解評価データセット **JaWildText** を構築する．図 1 に示すように，**JaWildText** は (i) 文字密度の高い掲示物を対象とした，シーンテキストを含む画像を用いた質問応答（**高密度 STVQA**），(ii) レシート画像を用いた情報抽出（**レシート KIE**），および (iii) 手書き文字画像の文字認識（**手書き文字 OCR**）という実世界で頻出する三つの類型から構成され，文字認

表 1　**JaWildText** と既存の日本語文字画像データセットとの比較．**JaWildText** は実利用環境における多様な撮影条件の画像を含み，文字認識とその後段タスクを一貫して評価できる．

データセット名	実画像	OCR 注釈	VQA	KIE
DOST [11]	✓	✓	–	–
RRC-MLT [8]	✓	✓	–	–
JPSC1400 [12]	✓	✓	–	–
XFUND [9]	–	✓	–	✓
JDocQA [10]	–	△[3)]	✓	–
Japanese-Mobile-Receipt OCR-1.3K [13]	–	✓	–	✓
JaWildText（提案）	✓	✓	✓	✓

識から下流タスクへの応用までを一貫して評価できる．また，多様な撮影環境の画像を収録しており，理想的な条件下で取得された文書画像と比べ，より実利用に近い難易度で評価できる．

本稿では，**JaWildText** を構築したのちに公開 VLM および OCR 特化モデルを評価し，日本語実世界画像における読字と下流タスク性能の現状と課題を整理する．**JaWildText** は GitHub[1)] および Huggingface Datasets[2)] で公開している．

2　JaWildText

JaWildText は (i) **高密度 STVQA**，(ii) **レシート KIE**，(iii) **手書き文字 OCR** の 3 つのサブセットで構成し，文字密度の高い画像の読解，情報の構造化抽出，複数文の手書き文字認識，という代表的な利用形態を網羅する設計とした．全ての画像および注釈を独自に作成し，Apache 2.0 のライセンスのもとで利用できる．以降では，各サブセットの目的，画像の収集手順[4)]，注釈付けの仕様を順に述べる．

2.1　高密度 STVQA

TextVQA [3] をはじめ，シーンテキストの読解 [14, 15, 16, 17] や文書読解 [2, 18, 19, 20] を要する課題は多く提案されてきた．しかし実環境では，画像内に文字がより密集して配置されることが多い．本サブセットは，文字密度が高く視覚的に複雑な案内板・掲示物等の実環境画像を対象とし，画像内テキストの認識を手がかりとして質問に回答する視覚言語理解の評価を目的とした．

画像の収集　作業者は日本国内の実環境に設置された案内板・掲示物，商品パッケージ等を対象に，カメラおよびスマートフォンを用いて撮影し，102 枚の画像を収集した．

注釈　各画像について，作業者は文字が記載された領域（**文字領域**）を四角形で注釈付けし，各領域に対応する文字列を付与した．注釈単位は文の意味的まとまりを考慮せず，視認できる文字列の連続単位とし，横書きは行単位，縦書きは列単位で付与した．1 画像あたり平均 468 文字のテキストが注釈付けされ，文字の密度が十分に高いサブセットと言える．その後，作業者は画像内テキストの内容理解を前提として回答可能な質問応答を作成した．質問応答は根拠となる 1 つ以上の文字領域（**根拠領域**）を持ち，作業者は文字領域の中から根拠領域を選択した[5)]．

2.2　レシート KIE

レシート画像からの文字認識とキー情報抽出は，SROIE [21] より統一的なタスク設定として整備されてきた [22, 23, 24]．しかし，それらはスキャン画像中の英数字を主な対象とし，文字種の混植や独自の税率体系といった日本のレシートに固有の問題を考慮していない．本サブセットは，実世界で撮影されたレシート画像を対象とし，日本語環境での実利用に即したキー情報抽出の性能評価を目的とした．

画像の収集　作業者は一般的な消費者向けレシート・領収書を対象として撮影した．スキャン画像との差を明確にするために，しわや折れを含むことやレシートを手に持って撮影することを許容した．

注釈　事前に定義したキー集合[6)]を基に，作業者は各キーに対応する値の文字列と文字領域を注釈付けした．その結果，125 枚の画像に対して 5,530 件のキー・値の組が得られた．

2.3　手書き文字 OCR

日本語手書き文字画像を収集したものは ETL 文字データベース [25] や Kuzushiji [26] など多くのデータセットが提案されている [27, 28, 29, 30] が，いずれもが単文字中心であり実利用条件下でページ単位の文字認識を統一的に評価する設定は限定的である．本サブセットは，手書きの日本語文字を対象に複数の文を一度に認識する課題を提供する．

画像の収集　我々は，あらかじめ定めた筆記スタイルおよびテーマの集合をもとに，`gpt-oss-120b` を用いて約 100 文字前後の日本語文を生成し，書字内容

1）https://github.com/llm-jp/jawildtext
2）https://huggingface.co/datasets/llm-jp/jawildtext
3）JDocQA では，OCR 注釈として人手アノテーションではなく，PDF から抽出したテキスト情報が用いられている．
4）共通する撮影時の指示は付録 A に示す．
5）根拠領域数の分布は付録表 5 に示す．
6）キーの一覧は付録表 6 に示す．

表 2　**JaWildText** の統計量．「#領域」はテキストが占める四角形の数を表す．

サブセット	#画像	#領域	#文字	#文字種	縦書 (%)
高密度 STVQA	102	4,456	47,823	1,688	7.9
レシート KIE	125	5,530	43,351	1,067	2.7
手書き文字 OCR	108	600	11,169	1,077	23.3

表 3　**JaWildText** における総画素数の分布．4K 以上の高解像度な画像を中心に構成されている．

サブセット	Full HD<	~ WQHD	~4K	≥4K
高密度 STVQA	2	11	5	84
レシート KIE	0	0	0	125
手書き文字 OCR	0	58	1	49

とした．撮影環境の多様性を確保するため，サンプルごとに紙，タブレット，ホワイトボードなどの筆記媒体を指定した．また，データセットを通して横書きと縦書きが混在するよう書字方向を指定した．改行位置については作業者の裁量に委ねた．生成したテキストを 8 名の作業者に配布し，各作業者は指定された筆記媒体に筆記し，カメラで撮影した．

注釈　各画像について，行ごとに文字列を書きおこし，対応する文字領域を注釈づけした．なお，誤字や脱字などの筆記者に起因する誤りは修正せず，見たままの文字列を注釈した．誤字などで文字が存在しない場合 □ を割り当てた．その結果，108 枚の画像に 600 行，合計 11,169 文字を注釈付けした．

2.4　統計量

表 2 に **JaWildText** の基本統計量を示す[7]．合計で 335 枚の画像に計 102,343 文字が注釈付けされている．また，画像内テキストの量と多様性を重視する狙い通り，1,688 種類の文字を含む．さらに，入力画像の解像度の傾向を把握するため，各サブセットの総画素数の分布を調べた．表 3 に示すように，Full HD（1920x1080）未満，Full HD 以上 WQHD（2560x1440）未満，WQHD 以上 4K（3840x2160）未満，4K 以上で分類した．**レシート KIE** は全例が 4K 以上であり，**高密度 STVQA** も 4K 以上が大半を占める一方で，**手書き文字 OCR** は WQHD 未満の画像を一定数含んでいる．スマートフォン等による撮影環境の高画質化を踏まえ，本データセットでは高解像度の実写画像を優先的に収集した．

JaWildText における平均的な文字サイズの分布を図 2 に示す．**手書き文字 OCR** では文字が相対的に大きく視認性が高い一方，**高密度 STVQA** や **レシート KIE** は文字が小さく，平均 400 文字以上が配置されている．この特徴は画像の情報密度の差を示し，タスク間の難易度の違いを生み出している．

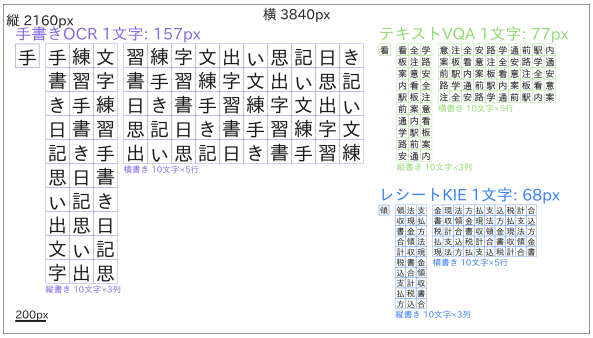

図 2　**JaWildText** におけるタスク別文字サイズの可視化．各タスクの 1 文字あたり平均面積から一辺長（px）を算出し，1 文字，縦書き 10 文字 x3 列，横書き 10 文字 x5 行と 200 px のスケールを 4K キャンバス上に配置している．

3　実験

本節では，構築した **JaWildText** を用いて既存の VLM の文字認識およびそれを前提とした推論能力を評価する．本データセットを視覚言語タスクの評価基盤 llm-jp-eval-mm [31] に実装して実験した[8]．

3.1　実験設定

評価対象モデル　本実験では，パラメータ数が 10B 未満で性能の高い VLM を 4 種類 [32, 33, 34, 35] 選定した．また VLM を基盤とした OCR 特化モデルを 3 種類 [5, 6, 7] 選定し，**手書き文字 OCR** で評価した．

推論設定　ハイパーパラメータとして温度を 0 に，最大生成長を 4096 トークンに設定した．入力画像はデータセットの原画像を RGB に変換して与え，リサイズを含む前処理は各モデルの推論実装に従った．プロンプトはそれぞれ固定の文を使用した[9]．

入出力と評価手法　**高密度 STVQA** は画像と質問が与えられ，モデルは回答文字列を出力する．最終回答の出力を`\boxed {...}` で囲むことを要求し，最初に出現した内容を抽出して解答文字列とした．抽出できない場合は形式エラーとみなした[10]．正解判定は `gpt-5.1-2025-11-13` を評価モデルとし，LLM-as-a-judge により二値で行った．温度 0 として，抽出した解答と正解の同値性を判定させ，それを精度スコア（Accuracy）とした．**レシート KIE** ではレシート画像を入力とし，所定のキー集合に対する値

7）　タスク固有の統計量は付録 B に示す．

8）　実験の詳細は付録 C に示す．

9）　実際に利用したプロンプトは付録表 7 を参照せよ．

10）　各モデルの形式エラー率を集計した表を付録表 8 に示す．

表 4　**JaWildText** における各モデルの評価結果．**太字**は最良の結果であることを示す．

モデル名	サイズ	VQA Acc↑(%)	KIE F1↑	OCR CER↓
VLM				
Gemma 3	4B	15.9	0.241	0.729
InternVL3.5	1B	15.9	0.359	0.362
InternVL3.5	2B	36.4	0.410	0.315
InternVL3.5	4B	41.1	0.422	0.288
InternVL3.5	8B	55.1	0.456	0.268
Qwen3-VL	2B	41.1	0.462	0.252
Qwen3-VL	4B	62.6	0.485	0.250
Qwen3-VL	8B	**74.8**	**0.538**	**0.207**
Sarashina2.2-Vision	3B	65.4	0.399	0.289
OCR 特化モデル				
PaddleOCR-VL	0.9B	—	—	0.352
Deepseek-OCR	3B	—	—	0.511
olmOCR-2	7B	—	—	0.299

を JSON 形式で出力させる．評価指標にはキー・値の組の一致に基づく F1 スコアを利用した．出力が JSON として解析できない場合は形式エラーとした．**手書き文字 OCR** の入力は手書き日本語を含む画像であり，認識した文字列を出力させる．本タスクでは画像全体を文字認識の対象とし，改行位置も含めて評価した．評価では予測文字列と参照文字列の正規化編集距離に基づく文字誤り率（Character Error Rate; CER）を用いた．

3.2　実験結果

表 4 に **JaWildText** におけるモデル別の評価結果を示す．Qwen3-VL が最も高い性能を示し，同一パラメータ帯のモデルと比較しても良好な結果を達成した．Sarashina や Qwen と比較して InternVL の性能が相対的に低い点は，入力画像を 448px 四方にリサイズする設計上，高密度な画像内テキストの判読が困難になることに要因があると考えられる．また，パラメータ数の増加は読字タスクおよび下流タスク双方の性能向上に寄与した．特に **高密度 STVQA** では，Qwen3-VL で 33.7pt（2B → 8B），InternVL3.5 では 39.2pt（1B → 8B）の向上が見られ，パラメータ数に対するスケール性が確認された．

一方で，**レシート KIE** は最高でも F1 スコアが 0.538 に留まり，実写レシートを対象とした構造化情報抽出が，最先端の VLM にとっても依然として困難であることが示された．**手書き文字 OCR** も最良の CER は Qwen3-VL 8B の 0.207 であり，画像内に複数文が含まれる文字認識タスクにおいては，さらなる性能改善の余地があると言える．

OCR 特化モデルの性能は同一パラメータ帯の VLM の性能を下回り，日本語手書き文字の読字性能において明確な優位性は確認されなかった．

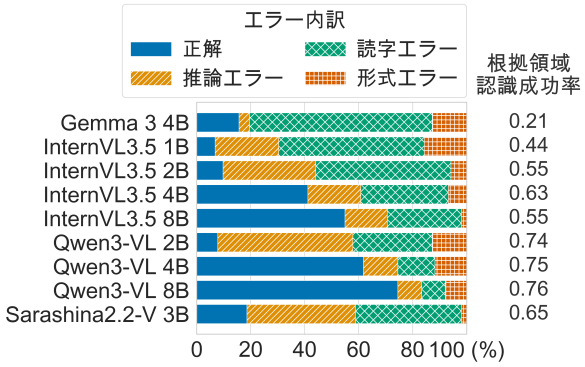

図 3　VLM ごとの**高密度 STVQA** の誤りの内訳（%）と，根拠領域に含まれる文字列の認識成功率．

3.3　誤り分析

高密度 STVQA における誤り要因を明らかにするため，VLM が根拠文字列を正しく認識しているかを検証した．具体的には，VLM に画像中の文字を出力させ，出力に根拠文字列が含まれているかどうかに基づいて分類した．質問応答が誤りであった例のうち，すべての根拠文字列を認識できた場合を**推論エラー**，一部が欠落していた場合を**読字エラー**，解答の抽出に失敗した場合を**形式エラー**と定義した．分類結果と根拠領域の認識成功率を図 3 に示す．Gemma 3 では読字エラーが支配的であった一方，同程度の正解率を示す Qwen3-VL 2B では，多くが推論エラーであったという差が確認された．また，Qwen3-VL はパラメータ数の増加（2B → 8B）に伴って正答率が 41.1% から 74.8% へと大きく向上した一方で，根拠文字列の認識成功率の変化は 0.74 から 0.76 と飽和していた．この結果は，パラメータ数の増加は主に VLM の推論能力向上に寄与することを示す一方で，画像内文字認識の性能改善の難易度の高さを示唆している．

4　おわりに

本稿では実世界画像を対象とした日本語文字認識能力を評価するためのデータセット **JaWildText** を構築した．実験では，文字認識性能・認識した文字に基づく推論性能の両方に課題があり，誤りの性質はモデル毎に異なることを明らかにした．さらなる評価の妥当性の向上を目的として，我々は **JaWildText** を合計 3,000 枚まで増強する予定である．本データセットが日本語環境における文字認識能力の改善を促進する基盤となることを期待する．

謝辞

本研究の成果は，JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（博士後期課程学生支援）JPMJBS2430 の支援，文部科学省の補助事業「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, データ活用社会創成プラットフォーム mdx の利用によって得られたものです.

参考文献

[1] Dimosthenis Karatzas, et al. The Robust Reading Competition Annotation and Evaluation Platform. arXiv:1710.06617, 2017.

[2] Minesh Mathew, et al. DocVQA: A Dataset for VQA on Document Images. In **IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 2199–2208, 2021.

[3] Amanpreet Singh, et al. Towards VQA Models That Can Read. In **IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 8317–8326, 2019.

[4] Ling Fu, et al. OCRBench v2: An Improved Benchmark for Evaluating Large Multimodal Models on Visual Text Localization and Reasoning. arXiv:2501.00321, 2025.

[5] Cheng Cui, et al. PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model. arXiv:2510.14528, 2025.

[6] Haoran Wei, et al. DeepSeek-OCR: Contexts Optical Compression. arXiv:2510.18234, 2025.

[7] Jake Poznanski, et al. olmOCR 2: Unit Test Rewards for Document OCR. arXiv:2510.19817, 2025.

[8] Nibal Nayef, et al. ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition – RRC-MLT-2019. arXiv:1907.00945, 2019.

[9] Yiheng Xu, et al. XFUND: A Benchmark Dataset for Multilingual Visually Rich Form Understanding. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 3214–3224, 2022.

[10] Eri Onami, et al. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In **Proceedings of LREC-COLING 2024**, pp. 9503–9514, 2024.

[11] Toru Ishida, et al. Icdar 2019 robust reading challenge on omnidirectional video. In **Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)**, pp. 1488–1493, 2019.

[12] Hideaki Goto, et al. JPSC1400 – Japanese Scene Character Dataset. Dataset (Rev.20201218), 2020.

[13] Sabari Nathan, et al. Japanese-Mobile-Receipt-OCR-1.3K: A Comprehensive Dataset Analysis and Fine-tuned Vision-Language Model for Structured Receipt Data Extraction. TechRxiv (preprint), 2025.

[14] Ali Furkan Biten, et al. Scene Text Visual Question Answering. In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 4291–4301, 2019.

[15] Anand Mishra, et al. OCR-VQA: Visual Question Answering by Reading Text in Images. In **2019 International Conference on Document Analysis and Recognition (ICDAR)**, pp. 947–952, 2019.

[16] Xinyu Wang, et al. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, pp. 10126–10135, 2020.

[17] Oleksii Sidorov, et al. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In **Computer Vision – ECCV 2020**, Vol. 12347 of **Lecture Notes in Computer Science**, pp. 742–758. Springer, 2020.

[18] Minesh Mathew, et al. InfographicVQA. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)**, pp. 1697–1706, 2022.

[19] Jordy Van Landeghem, et al. Document Understanding Dataset and Evaluation (DUDE). In **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**, pp. 19528–19540, 2023.

[20] Ryota Tanaka, et al. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, pp. 13636–13645, 2023.

[21] Zheng Huang, et al. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. arXiv:2103.10213, 2021.

[22] Seunghyun Park, et al. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. In **NeurIPS 2019 Workshop on Document Intelligence**, 2019.

[23] Hongbin Sun, et al. Spatial Dual-Modality Graph Reasoning for Key Information Extraction. arXiv:2103.14470, 2021.

[24] Xuan-Son Vu, et al. MC-OCR Challenge: Mobile-Captured Image Document Recognition for Vietnamese Receipts. In **2021 RIVF International Conference on Computing and Communication Technologies (RIVF)**, pp. 1–6, 2021.

[25] National Institute of Advanced Industrial Science, Technology (AIST), et al. ETL Character Database. Online database, 2014. Collected 1973–1984; accessed 2025-12-18.

[26] Tarin Clanuwat, et al. Deep Learning for Classical Japanese Literature. arXiv:1812.01718, 2018.

[27] Stefan Jäger, et al. Two On-Line Japanese Character Databases in UNIPEN Format. In **Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR 2001)**, pp. 566–571, 2001.

[28] Kaoru Matsumoto, et al. Collection and Analysis of On-line Handwritten Japanese Character Patterns. In **Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR 2001)**, pp. 496–500, 2001.

[29] Masaki Nakagawa, et al. Collection of on-line handwritten Japanese character pattern databases and their analyses. **International Journal on Document Analysis and Recognition**, Vol. 7, No. 1, pp. 69–81, 2004.

[30] Tomohisa Matsushita, et al. A Database of On-Line Handwritten Mixed Objects Named Kondate. In **2014 14th International Conference on Frontiers in Handwriting Recognition (ICFHR)**, pp. 369–374, 2014.

[31] 前田航希ほか. 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 (NLP), March 2025.

[32] Gemma Team. Gemma 3 Technical Report. arXiv:2503.19786, 2025.

[33] Weiyun Wang, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv:2508.18265, 2025.

[34] Shuai Bai, et al. Qwen3-VL Technical Report. arXiv:2511.21631, 2025.

[35] SBIntuitions. Sarashina2.2-Vision. https://huggingface.co/sbintuitions/sarashina2.2-vision-3b, 2025.

表 5　高密度 STVQA における回答の根拠領域数の分布．

根拠領域数	質問数	割合 (%)	根拠領域数	質問数	割合 (%)
1	19	18.6	4	10	9.8
2	35	34.3	5+	21	20.6
3	17	16.7			**平均**: 3.32

表 6　レシート KIE におけるキー集合，インスタンス数および欠損率．

グループ	キー名	値型	インスタンス数	欠損率 (%)
ヘッダ	store_name	文字列	125	0.0
ヘッダ	store_address	文字列	81	35.2
ヘッダ	receipt_id	文字列	123	1.6
ヘッダ	date	日付	125	0.0
ヘッダ	time	時刻	124	0.8
ヘッダ	total_amount	数値	125	0.0
ヘッダ	tax_amount	数値	113	9.6
明細 (line_item)	item_name	文字列	367	—
明細 (line_item)	item_price	数値	364	—
明細 (line_item)	item_quantity	数値	80	—

A　データ収集時の指示

本節では，**JaWildText** を構成するデータ収集および注釈付けにおいて，全サブセットで共通して適用した制約条件と作業指示の詳細を記述する．

写真撮影時に共通する制約　画像の多様性を確保するため，同一被写体に対する複数回の撮影を禁止した．また，背景物，指，反射光の映り込みを積極的に許容するとともに，撮影場所については屋内および屋外の両方を含め，撮影時間についても日中および夜間で変化を持たせるよう指示した．プライバシー保護の観点から，顔，車両番号，クレジットカード情報など，個人を特定し得る情報は撮影段階で除外した．撮影時には対象物が鮮明に写るよう作業者に指示し，強いブレ，歪み，またはボケによって画像内文字の判読が困難な画像は除外した．さらに，収集した画像について，対象物以外に個人が特定可能な他者の写り込みがないかを確認し，該当する場合は除外した．

注釈時の制約　空白記号，数値，アルファベットは半角文字に，カタカナは全角文字に統一した．また，レシートなどに見られる半角カタカナや濁点の分離表記についても，注釈時には全角カタカナとして統一した．

質問応答の作成にあたり，作業者に対して，質問が以下の条件を満たすよう指示した．

1. 画像中の文字情報の読解を前提とすること
2. 20〜100 文字程度の疑問文であること
3. 複数文から構成される場合，出力形式の指定を除き，最後の文のみを疑問文とすること
4. 回答が一意に定まり，画像外の知識や作業者の主観に依存しないこと

品質管理　注釈付けは各作業者が単独で実施し，その後，全件について別の作業者による検査を行った．その際，規約の解釈に関して判断が分かれた場合には，合議によって修正を行った．

作業者への報酬　画像の撮影，筆記，および注釈付けはデータ収集を専門とする業者に委託し，作業者には適切な報酬を支払った．

B　JaWildText の詳細な統計量

表 5 に**高密度 STVQA** における根拠領域数の分布を示す．1 画像あたり平均 3.32 個の根拠領域を持ち，80% 以上が複数の根拠領域の文字を認識した上で推論を必要とする設問である．

また，表 6 に**レシート KIE** におけるキー集合・それぞれのインスタンス数・欠損率を示す．多くのレシートに共通する内容をキーとして事前に定義した．レシート画像中にキーに対応する値が表記されていなかった場合，null として扱った．また，明細行も全ての行が 3 つのキー・値の組を持つとは限らないことに留意されたい．各レシートの明細行は複数あるため，欠損率を計算しなかった．

表 7　JaWildText における推論時のプロンプト．高密度 STVQA タスクの<question> は，画像に対応付けられた質問文で置換される．

タスク	プロンプト
高密度 STVQA	<question>\n 画像を参照して回答してください 。推論過程は出力しても構いませんが、最終回答は必ず \boxed {...} で囲み、ボックス内には最終回答のみを 1 つだけ記載してください。
レシート KIE	レシート画像からキー情報を抽出し、JSON 形式で返してください。 フィールド: store_name, store_address, receipt_id, date, time, total_amount, tax_amount, line_items[]. 値は画像の文字をそのまま出力してください（推測・正規化・整形しない）。無い項目は null（None）にしてください。line_items は {"item_name": "", "item_price": "", "item_quantity": ""} の配列で返してください。
手書き文字 OCR VLM	画像内の文字をすべて読んでください。 改行されている部分には必ず\n を挿入してください 。
手書き文字 OCR PaddleOCR-VL olmOCR-2	Read all characters in this image. Preserve line breaks with \n .
手書き文字 OCR DeepseekOCR	Free OCR.

表 8　JaWildText におけるモデル別の形式エラー率．

モデル名	サイズ	VQA(%)	KIE(%)
Gemma 3	4B	12.1	11.1
InternVL3.5	1B	15.9	5.6
InternVL3.5	2B	5.6	4.0
InternVL3.5	4B	7.5	4.8
InternVL3.5	8B	1.9	4.0
Qwen3-VL	2B	13.1	4.0
Qwen3-VL	4B	11.2	4.0
Qwen3-VL	8B	8.4	4.8
Sarashina2.2-Vision	3B	1.9	19.0

C　実験の詳細

プロンプト　推論に利用したプロンプトを表 7 に示す．全ての VLM は全てのタスクを通して同一のプロンプトを使用している．OCR 特化モデルは VLM のためのプロンプトを入力したときに適切な文字列を出力しない傾向があり，性能計測が困難であった．そのため，モデルごとに固有のプロンプトを使用した．

モデル出力の正規化　評価の再現性と表記揺れへの頑健性のため，タスクごとに以下の正規化を適用する．**高密度 STVQA** は \boxed {⋯} の内容を抽出し，抽出文字列に対して空白の正規化のみを行い，採点を行う LLM への入力とした．評価プロンプトでは，表記揺れや同義表現を認めたうえで判定させた．**レシート KIE** および**手書き文字 OCR** は値比較の際に Unicode 正規化 (Normalization Form Compatibility Composition; NFKC) を行ったうえで，前後の空白を除去し，連続する空白を圧縮した．

形式エラー率　表 8 に，各モデルの形式エラー率を示す．**高密度 STVQA** では 1.9〜15.9%，**レシート KIE** では 4.0〜19.0% の形式エラーが生じた．これは，日本語文字認識や推論の性能のみならず，日本語の指示に基づいてタスク形式に従った出力を生成する能力においても課題があることを示唆している．