日本語の図表理解のためのマルチモーダルLLMの開発

A Multimodal Large Language Model for Japanese Chart and Table Understanding

木下　彰* Akira KINOSHITA　金箱　裕介* Yusuke KANEBAKO　中田　乙一* Otoichi NAKATA　佐藤　諒* Ryo SATO　中山　佳紀** Yoshiki NAKAYAMA
カビール　シャリアル** Shahriar KABIR　平野　諒司* Ryoji HIRANO　朝井　都* Miyako ASAI　工藤　俊介* Shunsuke KUDO　長谷川　史裕* Fumihiro HASEGAWA

要　旨

近年LLaVAを筆頭に，入力された画像に関するやり取りが可能なマルチモーダルなLLM (LMM) がオープンソースとして公開されている．しかしながら既存のオープンソースLMMのほとんどは，きめ細かな画像認識を必要とするビジネス文書の理解には性能が不十分である．我々はLMMの構成要素であるVision Encoder (VE) 周辺を改良し，その後図表に特化した学習データを生成AIを用いて作成したのち，LMMの訓練を行った．結果，図表を含む日本語文書質問応答データセットであるJDocQA及び図表の読解能力を評価する独自ベンチマークJGraphQAによる評価にて，最先端の性能を示すことを確認した．これは，提案手法がドキュメント画像における図表の読解力を高めるために有効であることを示すものである．本稿の内容は，画像の認識・理解シンポジウムMIRU2025での発表を発展させ，加筆修正したものである．

ABSTRACT

Image-based multimodal LLMs, such as LLaVA, have been released as open source. However, most existing open-source LMMs are insufficient for understanding business documents containing figures, which require fine-grained image recognition. We improved the Vision Encoder (VE), a core component of the LMM, created chart- and table-specific training data using generative AI, and trained the LMM. The LMM was evaluated on JDocQA, a Japanese document QA dataset containing charts and tables, and on JGraphQA, our proprietary benchmark for assessing chart-reading ability, and the results indicated that our model achieved state-of-the-art performance. This demonstrates that the proposed method is effective in enhancing the ability to interpret charts and tables in document images. This article extends and revises the content originally presented at the Meeting on Image Recognition and Understanding (MIRU2025).

*　リコーデジタルサービスBU　AIサービス事業本部　デジタル技術開発センター
Digital Technology Development Center, AI Services Business Division, Ricoh Digital Services BU
**　リコーデジタルサービスBU　AIサービス事業本部　AIインテグレーション統括センター
AI Integration Management Center, AI Services Business Division, Ricoh Digital Services BU

1. はじめに

近年，入力された画像に関するやり取りが可能なマルチモーダルなLLM (LMM) が，オープンソースとして多数公開されている[1,2]．LMMは自然言語に加えて画像を入力し，画像に対する質疑応答が可能である．これらの中でも，2023年に発表されたLLaVAアーキテクチャは多くの派生研究を生んでおり，任意のLLMにVision Encoder (VE) を接続しVisual Instruction Tuningを行うことで，様々なLLMをマルチモーダル化することが可能となっている[1]．一方で，これらの研究の多くは自然画像や英語を含む図表画像を主な対象としており，日本語が書かれた図表画像を対象とした研究はあまり行われていない．そこで本研究では，日本語の図表理解に強いLMMの実現を目的とし，まずは手書きを除く図表読解に特化したモデルの開発を目指した．

まず既存のLMMの課題について説明する．既存のLMMのほとんどは，きめ細かな画像認識を必要とするビジネス文書の理解には性能が不十分であり，画像の入力が性能の向上に寄与できていない可能性も述べられている[3]．この原因の1つとして，VEとして利用されているSigLIPの最大解像度が384x384であることがあげられ[4]，対策として画像をグリッドに分割し，それらを独立してエンコードする方法が提案されている[1]．また別のアプローチとして，タスクに特化した学習データを用いる方法も提案されている[5]．

今回，我々は日本語の文書画像に強いQwen2-VL[2]のVEと2D-RoPE[2]を組み合わせ，日本語の理解力に優れるLlama 3.1 Swallowをマルチモーダル化した．また，LLMを用いてタスクに特化した学習用データセットを作成するとともに，図表の読解能力を評価する独自ベンチマークJGraphQAを構築し，図表を含む日本語文書質問応答データセットであるJDocQA[6]と合わせて評価を行った．

本稿ではまずQwen2-VLのVEを用いてLlama 3.1 Swallowをマルチモーダル化する方法について述べる．次に，LLMを用いた学習データ作成方法について述べる．最後に，JGraphQAの作成方法について述べ，JDocQAによる評価と共に結果を報告する．

2. モデリング

2-1 LLaVAアーキテクチャのAttention可視化

今回はLLaVA-OneVision (LLaVA-OV)[7]をベースラインのモデルとして用いる．LLaVA-OVは入力画像をリサイズするのではなく，マルチクロップされた画像及び入力画像全体をリサイズした画像をVEに入力し，画像トークンに変換することで高解像度画像に対応している．LLaVA-OVでは画像トークンのインデックスは入力画像全体をリサイズした画像トークンから始まり，入力画像を左上から右下にかけてクロップしてVEで変換した画像トークンが並んでいる．そのため，LLMに入力される画像トークンは常に最初のインデックスは画像全体をリサイズした画像トークンとなるが，それ以降のインデックスでは画像の左上からクロップされた画像トークンが順に並ぶ．またクロップ数は入力画像によって異なるため同じインデックスに常に画像の同じ相対位置が現れるわけではない．このように学習されたモデルが回答生成時に画像中のどの領域に着目したかAttentionから可視化を行った．Fig. 1はLLaVA-OVのマルチクロップ戦略によって入力画像が分割して入力された図である．これらの画像に対して文書画像中の文字列を書き出すプロンプトを与え，生成された文字列が画像中のどこに着目したか可視化を行った図をFig. 2に示す．画像中の左上から文字列が書き出されており，文字列に水色のハイライトが重畳されている領域がAttentionの可視化である．この可視化から，Attentionはクロップされた画像を参照せず，画像全体をリサイズした画像（Fig. 2 左上）にAttentionが集中している様子が分かる．これは他の画像でも同様であり，LLaVA-OVの学習戦略ではリサイズされた画像にAttentionが集中する傾向が確認された．リサイズ画像への過剰なAtttentionを解

消するべく，マルチクロップ戦略ではない画像入力方式を採用する必要がある．

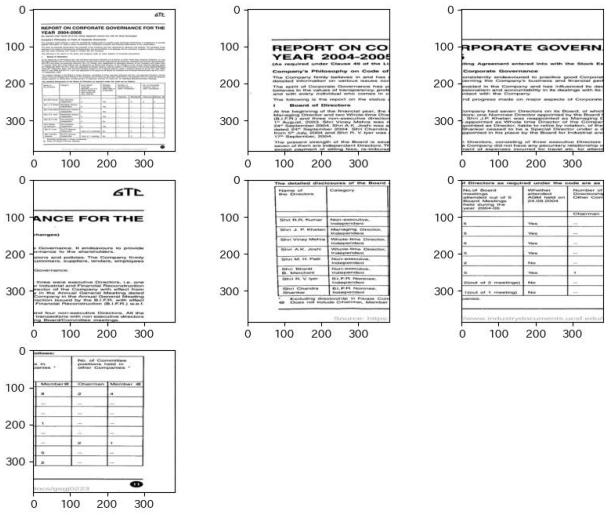

Fig. 1　Multi-Crop Examples of Input Images (from the DocVQA[8] Dataset).

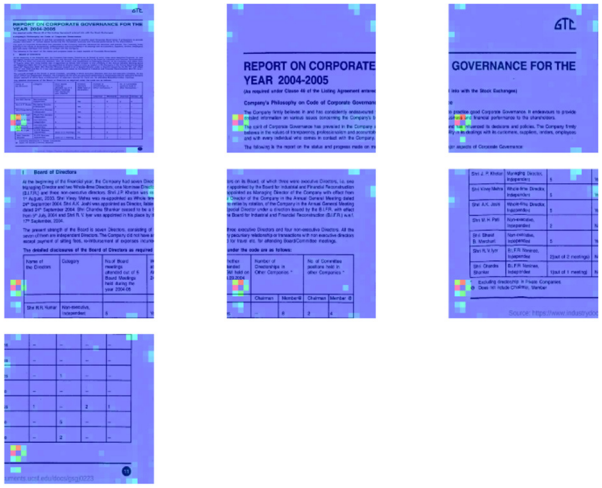

Fig. 2　Attention Visualization in the OCR Task.

2-2　高解像度画像への対応

マルチクロップ+リサイズの学習戦略におけるAttention偏りの課題を解決，及びマルチクロップによる図表分断を回避するために，任意の解像度をそのまま処理できるNative Dynamic Resolution[9]を採用する．また，ベースとして用いるVEには同じく任意の解像度入力によって学習されているQwen2-VLのVEを画像エンコーダとして採用する．具体的にはQwen2-VLで用いられているVE部分でLLaVAアーキテクチャのVEを置き換え，その後は通常のLLaVAアーキテクチャと同様にProjector (PJ)，LLMと接続しモデルを構成する．またQwen2-VLのVEは画像パッチの位置埋め込みとして通常のRoPEではなく，2D-RoPEを用いている．そのため1次元の位置情報だけではなく縦横2軸の位置情報を別々に回転行列で埋め込むことができ，画像の空間的な構造をそのまま位置情報として学習できる．これにより，ドキュメント画像の文字や図表の位置関係を考慮した高精度な読解が行える．

3.　学習用データセット

データセットについて述べる．商用利用を想定した場合，学習に適した日本語のデータセットがほぼ存在しない．例えばLLaVA-OVでは3.2M枚以上の画像データを利用しているが，これらのデータセットは英語かつ商用利用不可であることが多く，Qwen2.5-VLに関してはデータセット自体が公開されていない．そのため，我々は今回開発するLMMに適した日本語のデータセットを構築する必要がある．我々は既存技術[5]を参考に以下のフローを構築した．

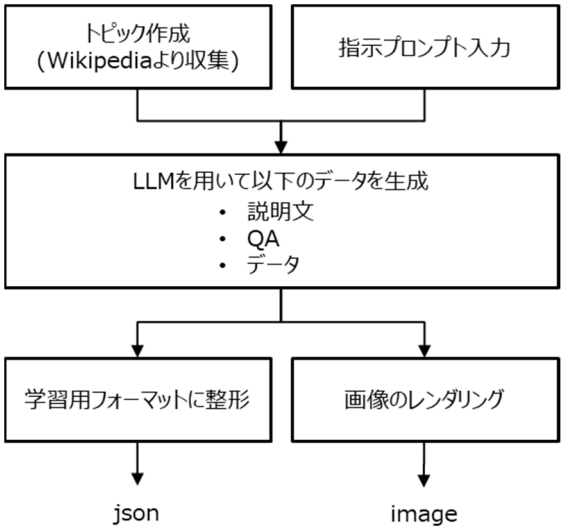

Fig. 3　Dataset Construction Pipeline (Ours).

データの作成フローについて具体的に述べる．まず，Wikipediaの情報を使いLLMが振る舞うための人物像リストを作成する．例えば，企業の経理担当や銀行の投資判断担当などである．その後，作成した人物像リストから1つを選択し，選択した人物が扱うであろうトピックを再度LLMに作成させる．例えば銀行の投資判断担当という人物像をLLMに与えると，LLMは企業のIR情報調査というトピックを返してくる．次に，LLMにトピックに関連する説明文を生成させる．例えば企業のIR情報調査というトピックを用いる場合には，架空の企業のIR情報に関する説明文をLLMに出力させる．併せて，LLMが出力した説明文と合致する数値データをLLM自身に出力させる．前述同様のトピックであれば，売上高の推移などの数値データがタイトルや凡例と共に出力される．さらに，出力された数値データとトピックを用いてデータの内容を問うQAを出力させる．最後に，QA及び説明文のフォーマットを整形してjson形式で出力させ，データを独自に作成したプログラムにてレンダリングする．レンダリングを行うプログラムはLLMが出力する複数のデータフォーマットを受け取り可能であり，データ量，配置，文字サイズ，色など様々な変更に対応している．

我々は上記方法を用いて，キャプション生成等を含めデータセットを合計600万枚程度用意した．Fig. 4に作成した構造化データの例を示す．

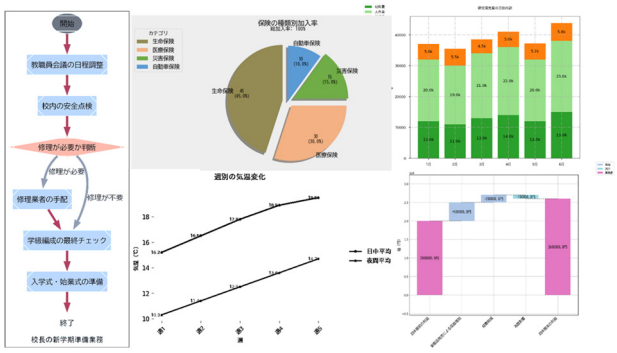

Fig. 4　Examples from Our Dataset.

4. 評価用ベンチマーク

図表の読解能力に関するベンチマークであるJGraphQAについて述べる．日本でも図表を含む日本語文書質問応答データセットであるJDocQAや日本文化に基づいた専門的タスクに対するベンチマークであるJMMMU[10]が存在する．一方，JDocQAは文章中心で一部に図表が含まれたベンチマークであり，またJMMMUは，石垣島全体が写った画像を入力し，画像の島で見られる有名な動物を問うなど，図表の読解に加えて日本文化に関する知識の有無をベンチマークしており，図表の読解能力に絞ったベンチマークは存在しない．そのため我々は，図表の読解能力に絞ったベンチマークを作成した．

作成したベンチマークについて具体的に述べる．我々はIRを中心に円グラフ，棒グラフ，折れ線グラフ，表の4種類を合わせて100枚の画像として抽出した．その後，各画像に対して2問のQAを付与し計200問から成るQAベンチマークを作成した．付与したQAはChartQA[11]の設計思想を参考にTable 1の3種類とした．

Table 1　Examples from JGraphQA (QA).

タイプ	質問例
検索	2021/1の法人の預金等残高は何億円か
視覚	一番左の緑のバーの値はいくつか
構成	2020/1から2022/1まで個人の預金等残高は何億円増加したか

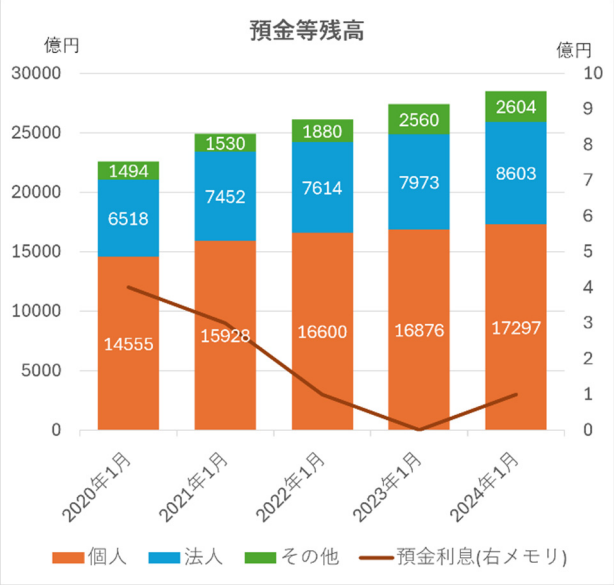

Fig. 5　Example from JGraphQA (Image).

評価方法は，指示追従性の能力がスコアに反映されることをなるべく排除し，図の理解度合いを最適に評価するために，作成した回答がモデルの出力に含まれているかどうかで評価することとした．

5. モデルの学習

学習はLLaVA-OVに従い，第1段階，第1.5段階，第2段階から成る3段階の学習方法を採用する．各段階における学習パラメータもLLaVA-OVと同等とした．第1段階では，プロジェクタの学習にのみ焦点を当て，LLMの単語埋め込み空間に視覚的特徴をうまく整合させることを目標とする．この段階では，簡単な画像とテキストのペアデータを用いて学習し，画像の視覚情報を言語モデルが理解できる形式に変換する．第1.5段階では，LMMに高品質な知識を学習させることを目的とし，前段より高解像度かつ説明テキスト量も多いデータを用いる．また，自然画像だけでなくOCRや図表などの構造化データも含める．この段階では，全てのパラメータを凍結解除し，より広い範囲のデータで学習する．最終段階では第1.5段階と同様，全てのパラメータを凍結解除しLMMに図表の読み取りタスクをはじめ多様な視覚タスクを解決する能力を獲得するためのインストラクションチューニングを実施する．

6. 評価

6-1　小型モデルを用いた改善効果の可視化

構築したモデル及び学習用データセットの効果に関して確認を行った．具体的には構築したデータセットの一部を用いてLLaVA-OVと同じ手法で学習することによって日本語における図表の読解能力の向上を確認した．評価はモデルアーキテクチャをLLaVA-OV（llava-onevision-qwen2-7b-si）のまま学習用データセットのみ変更した場合と，学習用データセットに加えてモデルアーキテクチャを変更した場合で評価を行った．評価結果をTable 2に示す．

Table 2　Model & Dataset Effects on Score.

No.	モデル	学習用データセット	Score (JGraphQA)
1	LLaVA-OV	-	0.540
2	LLaVA-OV	Ours	0.625
3	Ours	Ours	0.810

また，改善の程度を確認するため既存技術[12]を用いて可視化を行った．Fig. 6に例を示す．Table 2に記載の3モデルについて同一パラメータ設定で可視化を行い，各モデルが質問の応答時に画像中のどの領域に注意を向けるかを比較した．

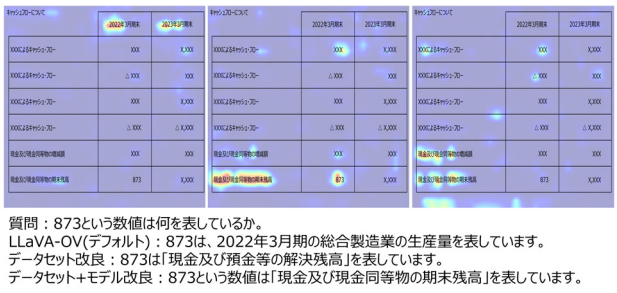

Fig. 6　Attention Map Visualization (Left: LLaVA-OV, Center: Dataset Improvement, Right: Dataset and Model Improvement).

公開されているLLaVA-OVの応答は質問に対して適切な回答を得られず，注目領域も質問に関連性の低い領域に集中していた．一方，LLaVA-OVを独自に構築したデータセットで学習したモデルは，質問の回答に必要な領域に注意を向けていたものの，文字情報の正確な抽出には至らなかった．これに対して提案モデルは，回答に必要な領域に適切に注目し，かつ正確な応答を出力した．

Table 2に示した定量改善に対応して，注目領域でも妥当なものになっていることを確認できた．

6-2　既存モデルとの比較

作成したモデルに対してJGraphQAによる評価を行った．またJGraphQAは質問数200と数が少ないため，質問数1,175問のJDocQAによる評価も併せて行った．JGraphQAはlmms-eval[13]のフレームワークを用いて評価し，JDocQAはllm-jp-eval-mm[14] (v0.4.0) のフレームワークを用いてYes/No，Factoid，Numerical，Open-endedを平均したOverallスコアを記載した．回答を導くのに参照する必要のあるページが複数にわたるサンプルについては，concatして1枚の画像として処理した．公開モデルは，各フレームワークに付随している評価スクリプトを動作させることによってスコアを算出した．評価スクリプトが用意されていないLLaVA-OVはJDocQA未実施とした．評価結果をTable 3に示す．我々の提案する手法はLlama-3.1-Swallowに限らず，複数のモデルにて改善効果を確認できた．また，Llama-3.1-Swallow-70B-Instructを用いたLMMにLlama-3.1-Swallow-70B-v0.1のウェイトを少量マージすることで更なる性能向上を確認できた．こちらはTable 3のNo. 11に示した通り．同等モデルの中で最高性能を示している．本手法はVEとLLMを選択可能であるため，今後高性能な国産のLLMが開発された際には，本手法を用いてマルチモーダル化することでより高性能な国産のLMMが開発できると考える．

Table 3　Effect of Improving the Model and Dataset.

No.	Model (Backbone LLM)	JGraphQA	JDocQA
1	Ours (Llama-3.1-Swallow-70B)	0.885	0.2965
2	Ours (Llama-3.1-Swallow-8B)	0.860	0.2797
3	Qwen2.5-VL-72B	0.860	0.2340
4	Qwen2.5-VL-7B	0.855	0.2536
5	Llama-3.2-90B-Vision	0.700	0.2312
6	Llama-3.2-11B-Vision	0.620	0.1752
7	llava-onevision-qwen2-72b-si	0.480	-
8	llava-onevision-qwen2-7b-si	0.335	-
9	Ours (Qwen2-72B)	0.895	0.3270
10	Ours (Qwen2-7B)	0.835	0.2820
11	Ours (merged No. 1& Llama-3.1-Swallow-70B-v0.1)	0.905	0.3011

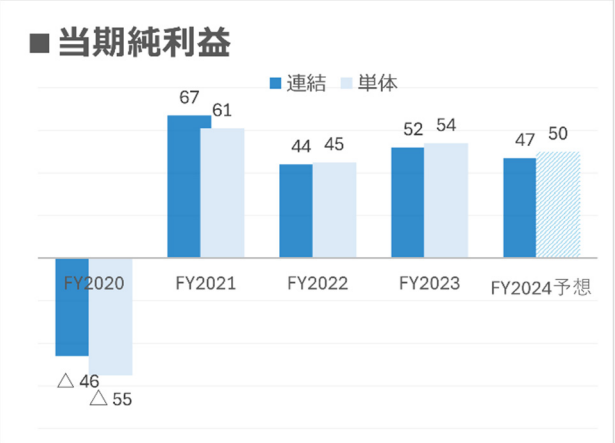

Fig. 7　A Sample Image Used for Evaluation.

Table 4　QA Results.

質問	単体が連結の値を上回りだしたのはFYいつからか？
Qwen2.5-VL-72B	単体が連結の値を上回りだしたのはFY2023からです．グラフを見ると，FY2023では単体の値（54）が連結の値（52）を上回っています．それ以前の年では，連結の値が単体の値を上回っていました．
Ours (Llama-3.1-Swallow-70B)	単体が連結の値を上回りだしたのはFY2022からです．

6-3　追加学習の効果比較

作成したモデルに対して企業のデータを含めて追加学習を行う場合，VE，PJ，LLMのどの部分を学習対象に含めるべきか評価を行った．評価は我々のモデル（8B）及びQwen2.5-VL (7B) に対して，モデルの学習対象を変えて追加学習を行い，性能の向上幅を確認することによって行った．評価対象は企業データから抽出した結合セル部に関する抜き出し課題と学習データとは別の企業文書の図表に対する質問応答課題とした．評価結果をFig. 9に示す．双方のモデルにおいて性能を向上させることができ，特にLLMの学習を行うことで大きな改善を見いだせることが分かった．

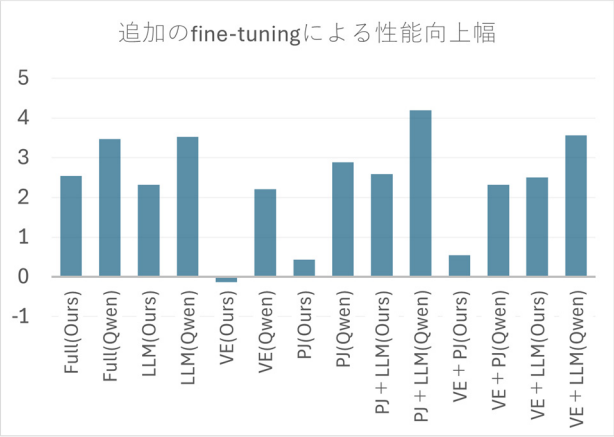

Fig. 8　Impact of Additional FT.

6-4　量子化の効果比較

作成したモデルに対して量子化を行い，使用メモリ量と，性能について評価を行った．適用した量子化手法は8bit量子化がLLM.int8[15]，4bit量子化がNF4[16]である．Table 5は作成したモデルに対して量子化を行った場合のメモリ使用量を示している．4bit量子化に関してはモデル全体に適用した場合と，LLM部分のみに適用した場合の2通りを作成した．Table 5より，70Bモデル（Table 3のNo. 1）では，8bit化によりメモリ使用量が51.4%，4bit化で27.2%となり，大幅な削減を確認した．また，全レイヤを4bit量子化した場合と，LLMのみを4bit量子化した場合では，使用メモリ量は1GBほどしか変化がなく，特にLLMが大きくなればなるほど，VEを量子化して得られるメモリ使用量削減効果は相対的に小さくなる．これはLLMに比べてVEのパラメータ数が圧倒的に少ないことに起因している．なお，8Bモデル（Table 3のNo. 2）でも同様の傾向を確認した．

次に，Table 6では作成した70Bモデル（Table 3のNo. 1）に対して4bit量子化したときの性能の比較を行った結果を示している．量子化は，全レイヤを対象としたものと，LLM部分のみを対象としたものの2種類を用意した．なお，用いた評価データはJGraphQAベンチマークで，そのスコアを示している．Table 6より，全レイヤを対象として量子化を行うと，性能が大幅に低下し，元のモデルの性能を維持できないことが分かった．一方で，LLM部分のみを量子化した場合は，性能低下を大幅に抑制することができた．これに関して，Self-Attentionの核心部分の量子化による情報損失が非常に大きいことなどが報告されており[17]，それに類似したことが起きていると思われる．また，学習の際に最後にLLM部分を調整しているため，LLMに入力される視覚情報がVEの量子化によって変わってしまうことで正しく計算できなくなることが考えられる．

以上から，本モデルではLLM部分のみを量子化することがメモリ使用量を抑えつつ，性能も維持しやすい結果であったと言える．

Table 5　VRAM Usage: Quantized vs Full-Precision.

量子化	対象レイヤ	使用メモリ量（8Bモデル）	使用メモリ量（70Bモデル）
なし	なし	16.27GB	132.86GB
8bit	All	9.12GB	68.39GB
8bit	LLM	---	---
4bit	All	5.54GB	36.15GB
4bit	LLM	6.52GB	37.23GB

Table 6　Performance: Quantized vs Full-Precision.

モデル	スコア
量子化なし	0.885
4bit量子化（ALL）	0.175
4bit量子化（LLM）	0.840

7. 結論

日本語の図表を読解し内容について回答可能なマルチモーダルLLMの開発を行った．まず，細かな文字の内容が重要な図表画像への対応を強化するために，VEの高解像度画像対応を行った．また日本でよく使われる図表表現への対応のために，LLMを用いて日本の図表学習用データセットを生成し学習を行った．ここで，日本語への対応を強化するため，学習用データセットは概ね日本語とした．また，マルチモーダル化に利用するLLMは日本語で継続事前学習され日本語への対応が強化されたと考えられるLlama 3.1 Swallowを採用した．

最後に，効果を正しく計測するために図表に特化した評価用ベンチマークの作成を行い，開発したモデルが図表の読解能力に関して最先端の性能を示すことを確認した．

謝辞

本モデルの開発は，経済産業省とNEDO（国立研究開発法人新エネルギー・産業技術総合開発機構）が実施する，国内の生成AIの開発力強化を目的としたプロジェクト「GENIAC (Generative AI Accelerator Challenge)※1」の支援を受けて実施したものです．

※1　GENIAC

主に生成AIのコア技術である基盤モデルの開発に対する計算資源の提供，データエコシステムの構築，生成AIの利活用に向けた先進事例創出に関する支援等が行われます．

参考文献

1) H. Liu et al.: Improved Baselines with Visual Instruction Tuning, *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296-26306 (2024).
2) P. Wang et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, arXiv, arxiv.org/abs/2409.12191 (2024).
3) Y. Ma et al.: MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations, *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 95963-96010 (2024).
4) X. Zhai et al.: Sigmoid Loss for Language Image Pre-Training, *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11975-11986 (2023).
5) W. Zhang et al.: Multimodal Self Instruct: Synthetic Abstract Image and Visual Reasoning Instruction Using Language Model, *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 19228-19252 (2024).
6) 大南英理ほか: JDocQA: 図表を含む日本語文書質問応答データセットによる大規模言語モデルチューニング, Proc. 言語処理学会第30回年次大会 (NLP2024), pp. 691-698 (2024).
7) B. Li et al.: LLaVA-OneVision: Easy Visual Task Transfer, arXiv, arxiv.org/abs/2408.03326 (2024).
8) M. Mathew, D. Karatzas, C. V. Jawahar: DocVQA: A Dataset for VQA on Document Images, *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2200-2209 (2021).
9) M. Dehghani et al.: Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution, *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2252-2274 (2023).

10) S. Onohara et al.: JMMMU: A Japanese Massive Multi-discipline Multimodal Understanding Benchmark, *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 932-950 (2025).
11) A. Masry et al.: ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning, *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2263-2279 (2022).
12) X. Xing et al.: Where do Large Vision-Language Models Look at When Answering Questions?, arXiv, arxiv.org/abs/2503.13891 (2025).
13) K. Zhang et al.: LMMs-Eval: Reality Check on the Evaluation of Large Multimodal Models, *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 881-916 (2025).
14) 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, Proc. 言語処理学会第31回年次大会 (NLP2025), pp. 1303-1308 (2025).
15) T. Dettmers et al.: LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale, *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 30318-30332 (2022).
16) T. Dettmers et al.: QLORA: Efficient Finetuning of Quantized LLMs, *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 10088-10115 (2023).
17) Y. Li et al.: Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer, *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, pp. 34451-34463 (2022).