ソフトウェア図表理解における視覚言語モデルの多言語性能評価

小原 有以 [1] 西潟 優羽 [1] 宮田 侑佳 [1] 倉光 君郎 [1]
[1] 日本女子大学大学院 理学研究科
m2016026oy@ug.jwu.ac.jp kuramitsuk@fc.jwu.ac.jp

概要

VLM はソフトウェア開発における図表理解への応用が期待されているが，多言語環境下での性能は十分に評価されていない．本研究では，ソフトウェア開発で使用される多様な図表に対する VLM の理解能力を評価するベンチマーク「ModelVista」を拡張し，本研究では，UML 図やモックアップなど，ソフトウェア開発プロセス全体で使用される多様な図表に対する VLM の理解能力を評価するベンチマーク「ModelVista」を，多言語環境に対応させる．本研究では，これらの図表および質問を日本語・英語・韓国語の 3 言語に対応するよう拡張し，多言語条件下での評価を可能にした．主要な VLM を評価した結果，3 言語のうち，日本語で最も高いスコアを示す傾向があることを確認した．

本研究は，ソフトウェア図表理解における VLM の多言語特性を明らかにし，多言語環境を前提とした評価の重要性を示す．

1 はじめに

近年，視覚言語モデル (VLM) は画像とテキストの両方を処理できる AI モデルとして，工場での異常検知や医療レポート生成など，様々な実用タスクに応用されている．ソフトウェア開発においても，VLM が設計図やダイアグラムを理解できれば，設計検証，自動コード生成，ドキュメント作成などの効率化が期待される [1]．

ソフトウェア開発では，要求定義から実装に至るまで，シーケンス図，ユースケース図，モックアップなど多様な図表が使用される [2]．これらの図表に対して，VLM が構造や関係性を踏まえて情報を読み取り，理解した上で適切に回答できる能力については，十分に評価されていない．特に，複数拠点にまたがるチーム協働が行われ，地理的・文化的・言語的な境界を越えた開発環境においては [3]，多言語環境下での VLM の図表理解能力の検証は重要な課題である．

本研究では，ソフトウェア図表に対する VLM の理解能力を評価するベンチマーク「ModelVista」を日本語，英語，韓国語の 3 言語に翻訳・対応して拡張し[1)]，多言語環境下での VLM の性能差異を詳細に分析する．これにより，VLM の多言語環境下でのソフトウェア図表理解能力の現状と課題を明らかにする．

2 ModelVista の構築

本節では，ModelVista の構築について述べる．

2.1 データセット概要

図 1 に ModelVista の概要を示す．ModelVista は，ソフトウェア開発の各段階で使用される 65 種類の図表と 472 問の多肢選択問題から構成される．図表は書籍やオンラインソースから収集し，可読性や記述の一貫性の観点から確認を行い，不明瞭な図表を除外した上で，高品質なものを選定した．

図表の種類は，UML ダイアグラム 13 種 (ユースケース図，オブジェクト図，クラス図，シーケンス図，コミュニケーション図，ステートマシン図，アクティビティ図，パッケージ図，コンポーネント図，配置図，複合構造図，タイミング図，相互作用概要図) と，その他のソフトウェア開発図表 (テーブル，データベース図，業務フロー図，データフロー図，アーキテクチャ図，ER 図，ガントチャート，状態遷移表，CRUD テーブル，システム構成図，画面遷移図，モックアップ) を含む．UML 図 13 種，それ以外の図表 12 種を含む計 25 種類，ソフトウェア開発の全工程をカバーする図表を選定した．

1) https://github.com/KuramitsuLab/ModelVista-3Lang

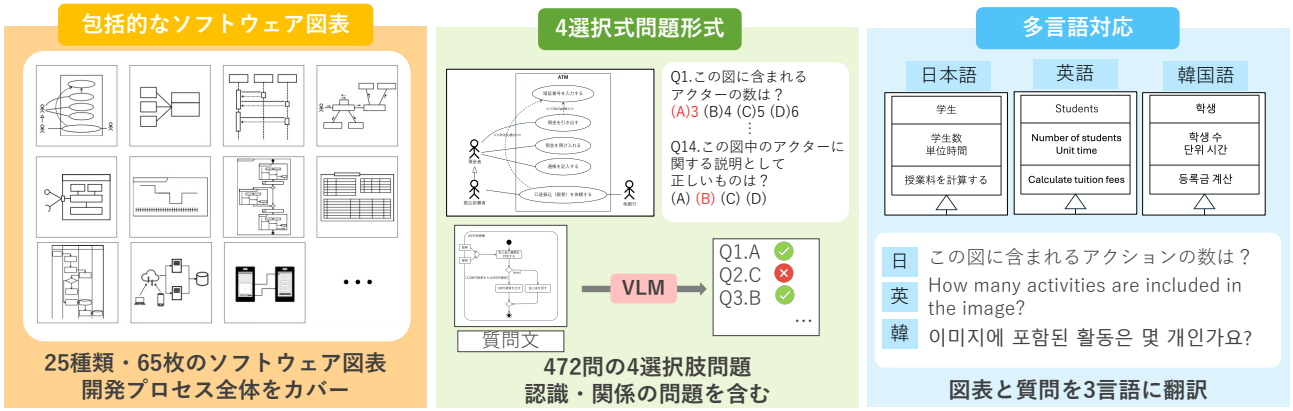

図 1　ModelVista の概要

2.2　問題設計

各質問は 1 文の問いと 4 つの選択肢で構成される．質問作成では，図全体や図中の要素を正しく認識できるか，要素間の関係を正しく把握できるかなどの観点から，図表から抽出可能な情報を洗い出した．その上で，図表の認識（「この図に含まれるライフラインの数は?」）や関係性の理解（「ユーザーから棚に送信されるリクエストは?」）など，様々な観点を問う問題を作成した．作成した質問は，図表の内容を正確に反映しているか，選択肢が適切かを確認し，修正を行った．

2.3　多言語化プロセス

多言語環境下での VLM 性能を評価するため，日本語で作成された ModelVista の全ての問題を，DeepL を用いて英語と韓国語に翻訳した．翻訳プロセスでは以下の点に注意した:

1. 用語の統一: データセット全体で使用される専門用語を統一し，言語間での一貫性を確保した．例えば，ユースケース図における「利用者」を英語では"User"，韓国語では "사용자" と統一した．
2. 図表内テキストの翻訳: 図表に含まれるテキスト要素 (クラス名，メソッド名，ラベルなど) も翻訳した．
3. 質問文の翻訳: 各質問文を対象言語に翻訳し，選択肢も同様に翻訳した．

3　実験設定

VLM の多言語環境下におけるソフトウェア図表理解能力の実態を明らかにするため，以下の実験を行った．

3.1　評価対象モデル

本研究では，主要な VLM として以下の 5 モデルを評価対象とした:

1. GPT-5[2)]
2. Gemini 2.5 Flash Image[3)]
3. Claude Sonnet 4.5[4)]
4. Qwen/Qwen3-VL-8B-Instruct[5)]
5. Llama 3.2-Vision[6)]

これらのモデルは，日本語，英語，韓国語のいずれにも対応しており，多言語比較に適している．

3.2　人間との比較

VLM の性能を相対的に評価するため，ソフトウェアエンジニア 4 名と学生 4 名の計 8 名に同じ問題セットを解答してもらった．専門家は 5 年以上の実務経験を持ち，学生は情報工学を専攻する大学生・大学院生である．翻訳対象として，評価に使用した 5 種類の図表 (ユースケース図，シーケンス図，コミュニケーション図，パッケージ図，モックアップ) に関連する質問を選定した．

各図表タイプ，各モデル，各言語における正答率を計算し，以下の観点から分析を行った:

- 人間と VLM の性能差
- 言語による性能差
- 図表タイプによる性能差

2)　https://platform.openai.com/docs/models/gpt-5
3)　https://aistudio.google.com/models/gemini-2-5-flash-image
4)　https://www.anthropic.com/claude/sonnet
5)　https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct
6)　https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

4　実験結果と考察

4.1　人間と VLM の性能比較

表 1 に，5 種類の図表タイプにおける人間と VLM の正答率を示す．表 1 では，

：各図種別および言語における VLM の正答率の最高値

赤字　：各図形タイプにおける正答率の最高値

を表す．

4.2　3 言語間の性能比較

全問題における正答率を見ると，評価した 5 つの VLM のうち 4 つの VLM において，日本語が最も高い正答率を示した．GPT-5 では日本語 71.05%，Gemini 2.5 Flash Image では日本語 70.02%，Claude Sonnet 4.5 では日本語 70.84%, Llama 3.2-Vision は 40.12%となり，いずれも英語・韓国語を上回った．

一方，今回アンケートを取った図表タイプ別に見ると，必ずしも全ての図表タイプで日本語が最高正答率を示したわけではない．5 つの図表タイプのうち，ユースケース図，シーケンス図，モックアップでは英語が最高値を示す場合が多く，パッケージ図で日本語と韓国語が同率で最高値を示した．

4.3　図表タイプ別の性能分析

一方，VLM の正答率は図表タイプと言語によって大きく変動した．正答率が最高値を示したのは GPT-5 によるモックアップ理解 (英語，100.0%) であり，人間の最高値 (93.75%) を上回った．また，パッケージ図では複数のモデルが 90%以上の正答率を達成し (Gemini 2.5 Flash Image 日本語 92.3%，Claude Sonnet 4.5 日本語 92.3%)，この図表タイプにおいては人間に匹敵する正答率を示した．

しかし，図表タイプ間の正解率の差は顕著であった．ユースケース図とモックアップでは 70%〜100%の正答率を示すモデルが複数あった一方，コミュニケーション図では多くのモデルが 50%以下の正答率に留まった．特に GPT-5 は日本語で 25.0%という低い正答率を示し，Llama 3.2-Vision も全言語で 16.67%〜29.17%と著しく低い正答率であった．これは，オブジェクト間のメッセージ交換を表現するコミュニケーション図の構造的複雑さが，現在の VLM にとって困難なタスクであることを示唆している．

4.4　人間と VLM の性能比較

全ての図表タイプにおいて，人間 (エンジニア・学生) の平均正答率が VLM を大きく上回る傾向が見られた．専門家の正答率は 82.5%〜97.44%の範囲にあり，特にパッケージ図 (97.44%) とコミュニケーション図 (92.71%) で高い正答率を示した．学生も 75.0%〜94.23%と高い正答率を維持しており，ソフトウェア図表の理解において人間が全体として優位性を持つことが確認された．

4.5　全体の正答率での日本語優位性の要因分析

全体で 4 モデルで日本語が最も高い正答率を示した傾向には，以下の要因が考えられる．

日本のソフトウェアシステムを題材としたデータセットの言語背景: ModelVista は日本のソフトウェア開発の書籍や Web サイトを参照して作成されており，図表の内容自体が日本のソフトウェアシステムを題材としている．例えば，図書館の貸出システムや駅の改札システムなど，日本特有の業務フローやシステム構成で，日本のソフトウェア開発実践に基づいた図表が多く含まれている．このような文化的・地域的な文脈が，日本語での評価において有利に働いた可能性がある．VLM は訓練データを通じて各国のシステムを学習しているため，日本語の質問と日本のシステムを題材とした図表の組み合わせが，最も自然な文脈として理解されやすかったと考えられる．実際，GPT-5，Gemini 2.5 Flash Image，Claude Sonnet 4.5 Qwen3-VL の 4 モデル全てで全体正答率が日本語で最も高く，この傾向を裏付けている．

図表内テキストと質問文の言語整合性: 図表に含まれる日本語テキストが，VLM の視覚的理解を補助している可能性がある．英語や韓国語への翻訳では，図表内テキストも翻訳されているものの，元の文脈やニュアンスが一部失われることがある．

5　関連研究

5.1　VLM の図表理解

VLM の図表理解能力を評価する研究として，ChartQA[4] や PlotQA[5] が，データ可視化グラフの理解を評価するベンチマークを提供している．また，DiagramNET[6] や StructVLM[7] など，図表理解に特化したアーキテクチャの提案も進んでいる．し

評価対象	言語	全体	ユースケース図	シーケンス図	コミュニケーション図	パッケージ図	モックアップ
エンジニアの平均（4 人）	日	–	89.39	82.5	92.71	97.44	91.67
学生の平均（4 人）	日	–	91.92	78.19	75.0	94.23	93.75
GPT-5	英	70.76	90.9	70.0	66.7	76.9	100.0
	日	71.05	81.3	63.3	25.0	76.9	83.3
	韓	66.10	81.8	60.0	37.5	84.6	66.7
Gemini 2.5 Flash Image	英	60.17	84.8	63.33	58.3	84.6	75.0
	日	70.02	78.1	60.0	54.2	92.3	66.7
	韓	60.17	66.7	46.7	54.2	84.6	66.7
Claude Sonnet 4.5	英	68.01	87.9	63.3	50.0	76.9	75.0
	日	70.84	78.1	63.3	41.7	92.3	83.3
	韓	60.81	72.7	40.0	62.5	84.6	66.7
Qwen 3-VL	英	64.83	81.8	46.7	41.7	76.9	83.3
	日	57.91	65.6	43.3	45.8	61.5	75.0
	韓	51.69	51.5	50.0	29.2	69.2	58.3
Llama 3.2-Vision	英	35.81	48.48	20.0	29.17	15.38	33.33
	日	40.12	48.48	32.37	29.17	72.32	40.21
	韓	38.98	45.45	36.67	16.67	76.92	41.67

各図種別および言語における VLM の正答率の最高値
赤字　各図形タイプにおける正答率の最高値

表 1　人間および各 VLM の図表タイプ別正答率（言語比較）

かし，これらはソフトウェア開発で用いられる構造的な図表を対象としていない．

5.2　ソフトウェア図表を対象としたベンチマーク

ソフトウェア図表理解のベンチマークとして，Bates ら [8] が UML 図からのコード生成タスク，Nguyen ら [9] が構造情報抽出手法，Silva ら [10] が UMLBench を提案している．Guo ら [11] はユーザーストーリーからの UML クラス図導出，Kumar ら [12] は UML 図へのフィードバック生成，Hassan ら [13] は動的意味理解の重要性を示している．しかし，これらは特定の図表タイプに限定されていたり，単一言語での評価に留まっている．

既存のベンチマークと比較して，ModelVista は以下の点で特徴を持つ：

- **多様な図表タイプ**：UML13 種，業務図表 10 種を含む計 23 種類
- **複数言語での評価**：日本語，英語，韓国語の 3 言語で評価

5.3　多言語評価ベンチマーク

多言語性能評価に関する研究として，MTVQA[14] や xGQA[15] が画像に対する質問応答タスクを多言語で評価するベンチマークを提供しているが，自然画像が対象であり，専門的な図表は含まれていない．本研究では，ソフトウェア図表を対象として 3 言語で評価を行い，VLM の言語による性能差を分析する．

6　むすびに

本研究では，ソフトウェア図表理解のためのベンチマーク ModelVista を拡張し，VLM の多言語性能を評価した．主な知見は以下の 3 つである．まず，言語性能は図表タイプとモデルの組み合わせにより複雑に変化することが明らかになった．主要モデル（GPT-5，Gemini 2.5 Flash Image，Claude Sonnet 4.5）では日本語が全体的に高い正答率を示したが，図表タイプごとには英語や韓国語が優位なケースも存在した．次に，図表タイプによる性能の大きな差が観察された．モックアップやシステム構成図では高い性能を示す一方，コミュニケーション図では著しく低い正答率となり，視覚的構造の明確さが VLM の理解能力に大きく影響することが示された．さらに，人間と VLM の比較では，全体として人間が優位であったが，特定の図表タイプでは VLM が人間に匹敵する性能を示した．今後は，VLM の誤答パターンを詳細に分析し，図表理解能力向上に向けた知見を得たい．

謝辞

本研究は，JSPS 科研費 JP23K11374 の助成を受けたものです．本研究を進めるにあたり，有意義なアドバイスを下さった NTT 株式会社コンピュータ＆データサイエンス研究所の倉林利行氏，秋信有花氏に感謝いたします．

参考文献

[1] Averi Bates, Ryan Vavricka, Shane Carleton, Ruosi Shao, and Chongle Pan. Unified modeling language code generation from diagram images using multimodal large language models. **Machine Learning with Applications**, Vol. 20, p. 100660, 2025.

[2] Sebastian Baltes and Stephan Diehl. Sketches and diagrams in practice. In **Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering**, pp. 530–541. ACM, 2014.

[3] Darja Šmite, Nils Brede Moe, Aivars Šāblis, and Claes Wohlin. Software teams and their knowledge networks in large-scale software development. **Information and Software Technology**, Vol. 86, pp. 71–86, 2017.

[4] 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 **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2625–2643, 2022.

[5] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In **Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision**, pp. 1527–1536, 2020.

[6] Xinyu Chen, Hao Wang, and Yang Liu. Diagramnet: A unified framework for diagram understanding. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 12345–12355, 2024.

[7] Yifan Zhang, Jing Li, and Kai Chen. Structvlm: Enhancing vision-language models with structural understanding. In **Advances in Neural Information Processing Systems**, Vol. 37, 2024.

[8] Averi Bates, Brayden Siegel, Andrew Shearer, Yutong Liang, Daniel Patterson, Wei Wu, and Zhichao Zhao. Unified modeling language code generation from diagram images using multimodal large language models. **Machine Learning with Applications**, Vol. 19, p. 100660, 2025.

[9] Tuan Nguyen, Wei Chen, and Jihwan Kim. Overcoming vision language model challenges in diagram understanding. **arXiv preprint arXiv:2502.04389**, 2025.

[10] Thiago Silva, Maria Santos, and Carlos Oliveira. Umlbench: A comprehensive benchmark for uml diagram understanding. In **Proceedings of the 46th International Conference on Software Engineering**, pp. 234–245. ACM, 2024.

[11] Jingyuan Guo, Song Wang, and Christoph Treude. Llm-based class diagram derivation from user stories: An empirical study. In **2024 IEEE International Conference on Software Maintenance and Evolution (ICSME)**, pp. 457–468. IEEE, 2024.

[12] Anish Kumar, Priya Sharma, and Raj Patel. Automated feedback on student-generated uml diagrams using large language models. In **2024 IEEE Frontiers in Education Conference (FIE)**, pp. 1–8. IEEE, 2024.

[13] Mohammed Hassan, Chen Li, and Douglas Schmidt. Behavioral augmentation of uml class diagrams using large language models. **arXiv preprint arXiv:2506.00788**, 2025.

[14] Fangyu Liu, Guy Emerson, and Nigel Collier. Mtvqa: Benchmarking multilingual text-centric visual question answering. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 8577–8588, 2023.

[15] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Iryna Gurevych, and Andreas Rücklé. xgqa: Cross-lingual visual question answering. In **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2497–2511, 2022.