jh250017

次世代災害対応のための視覚言語モデルの構築

横矢直人（東京大学 大学院新領域創成科学研究科）

世界各地で自然災害が頻発しており、人命や財産に深刻な被害をもたらしている。本研究では、多災害・多センサ・多タスクに対応したリモートセンシング視覚言語データセット DisasterM3 を構築した。能登地震や北海道地すべりなどの日本の災害を含む、5 大陸 36 件の災害事例を対象として、26,988 枚の二時期衛星画像と 123,010 件の高品質な視覚言語インストラクション対を収録し、被災対象認識、構造被害評価、参照分割、物体関係推論、長文災害報告生成までの 9 タスクを整備した。汎用視覚言語モデル（VLM）およびリモートセンシング向け VLM 計 14 種を評価した結果、災害特化コーパスの不足、センサ間ギャップ、被害対象のカウント感度の低さにより、最先端モデルであっても十分な性能を示せないことが分かった。そこで、Qwen2.5-VL-7B や InternVL3-8B など 4 種の VLM を追加学習し、全タスクにおいてベースモデルに対する安定した改善と、クロスセンサ・クロス災害条件下での汎化性能の向上を確認した。

1. 共同研究に関する情報

⑴ 共同利用・共同研究を実施している拠点名

東北大学 サイバーサイエンスセンター

東京大学 情報基盤センター

(2) 課題分野

大規模計算科学課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

横矢直人：研究構想の立案、データ構築および実験環境整備の統括

Bruno Adriano：災害対応の専門知識に基づく助言およびデータセット設計

WANG Junjue:データセットの構築、手法の設計および実験の実施

XUAN Weihao：手法の設計および実装

2. 研究の目的と意義

突発的自然災害は人命に甚大な脅威を及ぼす。インテリジェントな災害対応の実現に向けて、被害評価や救援経路計画などの多様なタスクを支援する、対話型かつ拡張可能なデータセットと視覚言語モデルを構築する（図 1）。能登地震をはじめとする日本の災害事例を含む世界各地の大規模災害事例を対象に、災害前後のリモートセンシング画像（光学および SAR）を収集し、多源・多時期・多センサの災害データを整備する。さらに、災害の緊急対応タスクに即して視覚言語指示データセットを作成し、視覚言語大規模モデル（VLM）を学習させることで、災害緊急対応のためのインテリジェント・アシスタントを構築する。

災害対応に特化したインストラクションチューニング用データセットを整備することで、オープンソースモデルと商用モデルを同一基準で比較・評価し、災害対応タスクにおける推論能力を検証できる。これにより、災害分野における垂直領域（ドメイン特化）モ

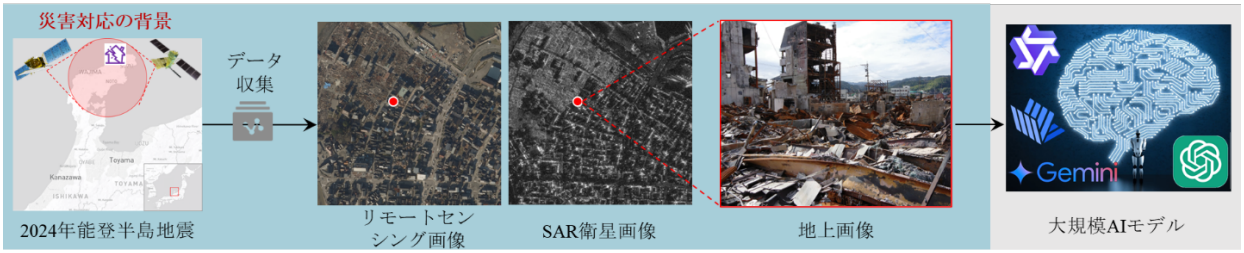

図 1 プロジェクトの目的と意義

デル開発に対して指針となる知見を提供する。さらに、提案する災害 AI アシスタントが防災・危機管理機関における標準的なツールとして普及すれば、マルチモーダル入力に基づいて、迅速かつ対話的な意思決定支援を提供できる。加えて、新たな災害データを継続的に投入することで、データベースとモデルは段階的に高度化し、Society 5.0 の実現に資する基盤として機能する。

3. 当拠点の公募型共同研究として実施した意義

本研究は、災害時の迅速な状況把握と意思決定支援を目的として、多災害・多センサ・多時期のリモートセンシング画像および関連テキストを統合し、視覚言語モデル（VLM）の学習・評価基盤を構築するものである。本研究では、データ収集・前処理（幾何補正、雲影などの品質管理、時系列の整合、アノテーション設計）から、複数モデルの大規模ベンチマーク、災害特化 VLM の追加学習・推論に至るまで、計算資源とデータ基盤の両面で高度な要件が求められる。

当拠点の公募型共同研究として実施する意義は、(i) 大規模 GPU 計算資源と高速ストレージを活用し、123,010 件規模の視覚言語インストラクション対に基づく学習・評価を現実的な期間で反復実行できる点、(ii) センサ間ギャップや災害種間汎化など、実運用を想定した条件設定の下で、複数 VLM（汎用モデルおよびリモートセンシング特化モデル）の体系的比較を可能にする点、(iii) 分野横断的連携の観点では、計算機（AI）分野、リモートセンシング分野、ならびに防災・災害分野の専門家が参画し、実運用のシナリオに整合しつつインテリジェント計算の要件を満たすデータセットおよびモデルを共同設計できる点にある。共同利用拠点として、異分野の研究者がデータ仕様、タスク設計、評価指標を共有することで、災害対応 AI の標準化と再現性の確保を促進できる。

以上により、本拠点の計算・データ基盤および共同研究環境を活用することで、災害対応における VLM の限界要因（災害特化コーパス不足、クロスセンサ・ギャップ、被害対象カウント感度の低さ）を定量的に抽出し、改善手法（Qwen2.5-VL-7B 等の追加学習モデル）の有効性を検証するための実験サイクルを確立できた。これは、災害対応に活用可能な AI アシスタントの基盤整備として、当拠点の公募型共同研究の枠組みを活用したことで初めて達成できた成果である。

4. 前年度までに得られた研究成果の概要

該当なし

5. 今年度の研究成果の詳細

今年度は、各分野の専門家との議論と連携を通じて、高分解能な地球観測データに基づく、災害対応向けの視覚言語処理フレームワークを開発した。本フレームワークは、主として、(1) 大規模・多元データに基づくベンチマークの構築と、(2) 視覚言語モデル（VLM）の研究開発の二つから成る。

(1) 大規模・多源データに基づくベンチマークの構築

能登地震をはじめとする日本の災害事

例を含む世界 5 大陸 36 件の災害事例に着目し、xBD、BRIGHT データセットおよび Maxar Open Data Program から災害前後のリモートセンシング画像（光学および SAR）を収集した。収集した画像は 0.8m 解像度に統一し、災害前後の二時期ペアとして空間的に整合させた。さらに、アノテータによる正解作成と GPT-4o による自動生成を組み合わせた品質管理パイプラインにより、視覚言語インストラクション対を体系的に構築した。全体のデータセット構築パイプラインを図 2 に示す。

国際連合衛星センター（UNITAR/UNOSAT）の緊急マッピング指針および FEMA の被害評価基準を参照し、災害評価・対応に必要な 5 つの中核能力に基づいて、VLM（Vision-Language Model）の性能を包括的に評価するための多層的なタスク分類（図 3）を設計した。具体的には、(1) 災害認識：行政が定義する災害種別の識別に加え、主要な被災対象（被災を受ける土地被覆タイプ等）を同定する被災対象認識を含む。(2) 被害カウント：建物および道路の被害程度推定・集計を行う。(3) 局所化：自然言語による参照分割（Referring Segmentation）を通じて、災害関連対象の空間的位置特定を行う。(4) 被害対象の関係推論：異なる対象間の空間的／意味的関係を解析する。(5) 災害報告生成：災害状況を俯瞰的に統合分析するための長文生成タスクとして設計する。災害キャプションでは、複数の被災対象に対する被害レベルを長文（複数文）で記述し、復旧助言では状況に応じた即時対応および中長期対応の提案を提示する。

GPT-4o を活用することで、複数の災害を対象としたデータセットを効率的にアノテーションした。認識・カウント・推論タスクについては、アノテータが正解を作成し、GPT-4o により表現の異なる同義質問を生成して多様性を確保するとともに、選択肢構築のための類似した誤答（ディストラクタ）を生成する。長文の災害報告については、二時期画像と各基本タスクの情報を参照しながら、複数の専門家が災害キャプションと復旧助言を起草し、GPT-4o が文章の推敲・文法誤りの修正を行う。アノテータ訓練、一次アノテーション、クロスバリデーション、専門家検証（10〜20%抽出）、統計分析と GPT-4.1 による意味整合性チェックの 5 段階品質管理を経て、高品質なインストラクション対を整備した。

DisasterM3 データセットの統計的特徴を分析すると、10 種類の災害タイプ（火災 5 件、竜巻 4 件、ハリケーン 7 件、津波 2 件、爆発 2 件、地震 5 件、地すべり 2 件、洪水 4 件、火山 1 件、紛争 4 件）を網羅し、自然災害と人的災害の両方を対象としている。データソースとしては、xBD および BRIGHT データセットから 26 件のイベントを収集し、Maxar Open Data Program から能登地震を含む 10 件の新規イベントを追加した。光学画像には WorldView シリーズの衛星画像を用い、SAR 画像には Capella Space および Umbra のデータを用いた。SAR 画像は振幅データを VV 偏波または HH 偏波で取得し、地形補正の後に[0, 255]に正規化し、光学画像と同一解像度（0.8m）にリサンプリングした。インストラクション対の内訳は、学習用セット（Instruct）が光学画像 17,190 枚・SAR 画像 3,798 枚・インストラクション対 92,968 件、評価用セット（Bench）が光学画像 5,024 枚・SAR 画像 976 枚・インストラクション対 30,042 件である。

タスク別の設計の詳細として、認識タスクでは 13 種の土地利用タイプ（空港、橋、河川、森林、低植生、池、駐車場、港、高架、住宅地、工業地帯、商業地、海）と 12 種の被災対象タイプ（建物、スタジアム、空き地、橋、ダム、道路、港湾施設、貯水タンク、農地、森林、海岸線、鉱業地域）を定義した。

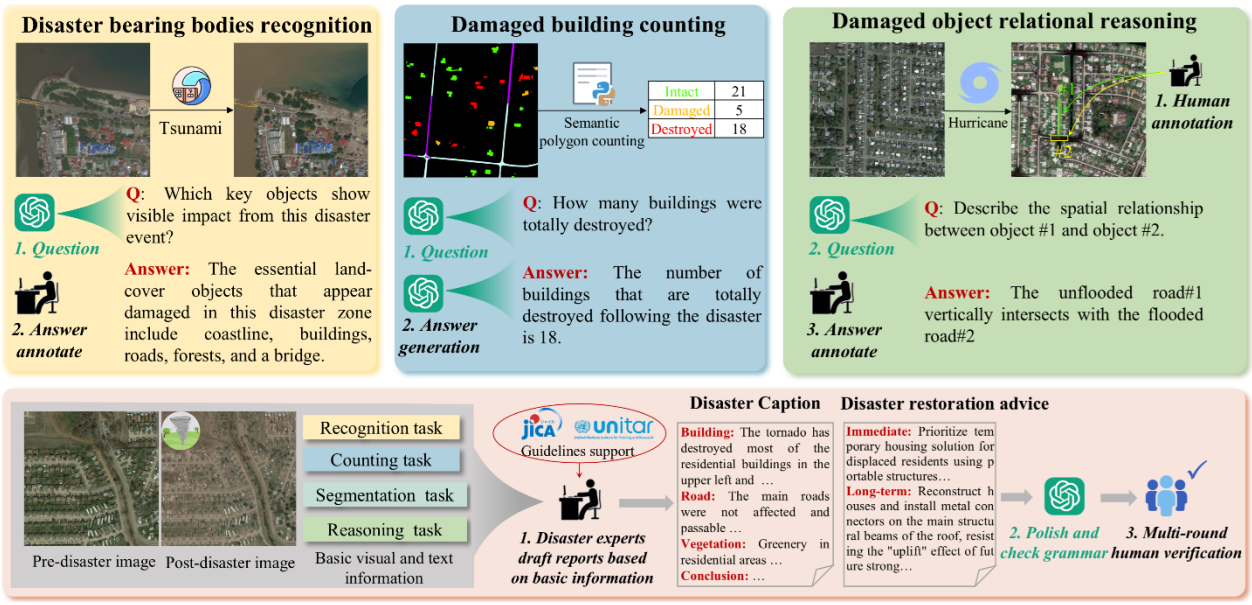

図 2 多段階命令のためのデータセット構築パイプライン

カウントタスクでは、建物の損壊程度別カウント（無傷・損壊・全壊の 3 段階、FEMA 基準準拠）と道路の被害面積比推定（冠水・瓦礫被覆・無傷の 3 タイプ）を設計した。被害カウントの選択肢生成では、正解値に対して±20%および±40%の偏差を持つディストラクタを自動生成し、妥当な誤答候補を構築した。関係推論タスクでは、専門家がバウンディングボックスのアノテーションと空間関係記述を行い、GPT-4o が他の有意な関係および代替選択肢を生成した。長文災害報告タスク（災害キャプションおよび復旧助言）については、UNITAR および FEMA の目標に基づき複数の専門家が起草し、建物・道路・その他の各被災対象について被害レベルを段落単位で記述する構造化された形式を採用した。

（2）視覚言語モデル（VLM）の研究開発

さらに、未ラベルの災害データを継続的に追加し、モデルに擬似ラベル（pseudo-labels）を生成させた。災害分野の専門家の監修の下、パープレキシティの高いラベルを中心に、対話的な修正と精査を行った。確認済みデータは順次データセットに追加し、災害データセットの拡張を進めた。さらに、計算資源制約を踏まえてモデルアーキテクチャの剪定（pruning）および追加学習（fine-tuning）を実施し、拡張後データセットに最適化した。このサイクル型フレームワークにより、データ量とモデル性能の双方を段階的に向上させた。最終的に、DisasterM3 データセットは二時期衛星画像 26,988 枚（光学 22,214 枚、SAR 4,774 枚）、インストラクション対 123,010 件（学習用 92,968 件、評価用 30,042 件）の規模に達した。

本手法の有効性を検証するため、現時点の最先端に位置づけられる汎用モデルおよびリモートセンシング分野の視覚言語モデルを選定し、比較評価を行った。図 3 に示すとおり、追加学習後のモデルは建物損壊評価および空間関係推論の両タスクにおいて、ベースモデルに対して一貫した性能改善を示し、商用モデル（GPT-4.1 等）に匹敵する精度を達成した。さらに、能登地震に関する可視化結果からも、災害シーンを高精度に理解し、適切かつ効率的な応答が可能であることを確認した。

DisasterM3 における詳細なベンチマーク分析では、14 種の先端 VLM（オープンソース：LLaVA-1.5-7B、LLaVA-OV-7B、Kimi-VL-A3B-Instruct/Think、InternVL3-8B/14B/78B、Qwen2.5-VL-3B/7B/32B/72B、リモートセンシ

ング特化：GeoChat-7B、TeoChat-7B、EarthDial-4B、商用：GPT-4o、GPT-4.1）を体系的に比較した。光学―光学 QA タスクでは、商用モデル GPT-4.1 が AVG 42.3%で最高精度を達成し、GPT-4o が AVG 39.3%で続いた。オープンソースモデルでは Qwen2.5-VL-72B が AVG 40.5%、InternVL3-78B が AVG 39.3%で同水準であった。一方、リモートセンシング特化モデルは GeoChat-7B が AVG 23.0%、TeoChat-7B が AVG 22.9%にとどまり、災害ドメインへの転移が困難であることを示した。災害キャプション生成の評価（GPT-4.1 による 5 段階評価）では、追加学習後の Qwen2.5-VL-7B が被害評価精度（DAP）3.76、被害詳細再現率（DDR）3.53、事実正確性（FC）4.41 を達成し、ベースモデル（DAP 1.69、DDR 1.71、FC 1.85）から大幅な改善を示した。これは災害特化コーパスの注入により報告品質が顕著に向上することを示している。

災害種別ごとの詳細分析では、地すべりで全手法が比較的高い精度を達成した一方（追加学習 InternVL3-8B: 56.9%）、地震（同: 26.5%）、竜巻（同: 31.1%）、爆発（同: 26.1%）では精度が低下した。これは、地すべりが主に農村部で発生し対象物が限定されるのに対し、地震等の災害は高密度都市部で発生し複雑なシーン構成が推論を困難にするためである。建物損壊カウントに関するバイアス分析では、InternVL シリーズは密度の増加に伴い精度が一旦低下した後に回復する U 字型のパターンを示し、周辺範囲（50 棟未満または 200 棟以上）で高い精度を達成する一方、GPT-4 モデルでは建物密度と精度が逆相関を示すなど、モデルアーキテクチャによって異なるバイアスが存在することが判明した。光学―SAR 設定では全体的に性能が低下するものの、追加学習後の Qwen2.5-VL-7B が AVG 29.9%を達成し、ベースモデル（AVG 22.6%）から 7.3 ポイントの改善を維持した。これはセンサ間汎化において災害特化追加学習が有効であることを実証している。

(3) 計算機利用上の工夫

計算機利用上の工夫として、DisasterM3 のデータ構築と VLM 追加学習で得られた知見を活用し、学習・推論パイプラインの再利用性を高めることで、都市理解・都市レジリエンス分野への展開も進めた。DynamicVL では、米国 42 都市を対象に 2005 年から 2023 年までの長期多時期衛星画像（NAIP、1.0m 解像度）14,871 枚と、学習用 63,771 件・評価用 8,682 件の計 72,453 件のインストラクション対からなる DVL-Suite を構築し、基本変化分析・変化速度推定・環境評価・参照変化検出・地域変化キャプション・密集時系列キャプションの 6 タスクで 18 種の MLLM を評価した。提案手法 DVLChat（7B パラメータ）は Qwen2.5-VL を基盤に変化検出用と QA 用の 2 つの LoRA アダプタを統合し、商用モデル o4-mini（AVG 34.1%）に匹敵する AVG 33.3%を達成した。CityVLM では、日本全国 60 都市を対象にリモートセンシング画像 20,589 枚と街路画像 110 万枚、80 万件の QA ペアからなる CitySet を構築し、地理空間推論から経済評価・持続可能な都市開発レポート生成まで 4 段階の多視点タスクを整備した。提案する CityVLM（7B パラメータ）は、地理空間・擬似時間モデリング（GPTM）により RS 画像と SV 画像の特徴を効果的に統合し、OA 83.4%を達成して 72B～78B パラメータの大規模モデル（GPT-4o 含む）を大きく上回る性能を示した。これらの研究は、災害対応の基盤技術を都市レジリエンスの定量評価へと発展させるものであり、本プロジェクトの成果の波及効果を示している。

DynamicVL における DVL-Bench の詳細として、DVL-Bench は 3,469 枚の多時期地球観測画像に対し、セグメンテーション指示 1,391 件、時系列分析 QA ペア 5,854 件、包括的キャプション 1,437 件を人手検証により構

築した。基本変化分析（BCA）では植生から非植生・建物への土地被覆遷移が支配的であり、急速な都市化を反映している。変化速度推定（CSE）では、2010 年から加速し 2017 年頃にピークに達した後 2018 年以降は減速する米国の都市開発の非線形パターンが確認された。DVLChat のアーキテクチャは LISA を拡張し、入力テキストの先頭プレフィックス（[QA]または[SE]）に基づいて VQA 用 LoRA と変化検出用 LoRA にルーティングするタスク固有のメカニズムを導入した。変化検出では＜SEG＞トークンの埋め込みを SAM バックボーンとマスクデコーダに入力し、ピクセルレベルの変化マスクを生成する。キャプション評価では、地域変化キャプション(RCC)で AVG 3.98、密集時系列キャプション（DTC）で AVG 3.40 を達成し、InternVL3-78B（RCC AVG 3.92、DTC AVG 3.33）を上回った。参照変化検出では IoU 29.06%を達成し、専門モデル ChangeMamba（IoU 32.41%）に迫る性能を示した。

CityVLM の研究では、日本全国 60 都市のリモートセンシング画像（GSD 0.5m）と対応する街路画像（Google Map、サンプリング間隔 10m）を共登録して統合し、CitySet を構築した。CitySet は地理空間推論（494.1K 件）・社会対象分析（226.5K 件）・経済評価（61.8K 件）・持続可能な開発レポート（20.6K 件）の 4 レベルの合計 80 万件以上の QA ペアから構成される。社会対象分析では、レストラン・コンビニ・公園・学校・スーパー・郵便局・鉄道駅・消防署・病院の 9 種の施設を対象とした。経済評価では ODIAC による CO2 排出量、Global High-Resolution Population Denominators Project による人口密度、VIIRS 夜間光強度の 3 指標を統合し、RS 画像のみでは捉えられない経済活動の定量化を可能にした。CityVLM の地理空間・擬似時間モデリング（GPTM）では、SV 画像を k 最近傍グラフ（k=6）と最小全域木に基づく探索で空間的に一貫した順序に並べ替え、フーリエ特徴マッピングによる位置埋め込みを SV トークンに付加した。アブレーション実験では GPTM により OA が 78.2%から 83.4%に向上し、他の位置符号化手法（Sinusoidal: 80.3%、Learnable: 81.8%、RoPE: 81.9%）を上回った。さらに、東京都本土を対象としたヒートアイランド効果の応用事例では、農村部（27.2° C）から商業地区（37.8° C）まで地域特性に応じた自然環境改善策を自動生成できることが確認された。

6. 進捗状況の自己評価と今後の展望

JHPCN の強力な計算基盤の支援により、2025 年度には顕著な成果を得ることができた。当初計画では、(a) 災害特化 VLM データセットの構築、(b) 複数 VLM のベンチマーク評価、(c) 災害特化 VLM の追加学習と性能検証を設定した。(a)については、36 件の災害事例から 26,988 枚の二時期画像と 123,010 件のインストラクション対を整備し、計画を上回る規模を達成した。(b)については、汎用・RS 向け VLM 計 14 種を体系的に評価し、災害ドメインにおける 3 つの主要課題（災害特化コーパスの不足、センサ間のギャップ、カウント感度の低さ）を定量的に明らかにした。(c)については、Qwen2.5-VL-7B 等 4 種の VLM を追加学習し、QA タスクでベースモデルから最大約 10 ポイントの改善(AVG 41.7%)、参照分割で mIoU 40.8 ポイントの向上(50.5%)を達成し、商用モデル GPT-4.1（AVG 42.3%）に匹敵する水準に到達した。また、2025 年の「AI 地震対応チャレンジ」において優勝を達成した。成果は計算機分野のトップ国際会議（NeurIPS 2025×2 件、EMNLP 2025×1 件）およびリモートセンシング分野の主要国際誌（ISPRS JPRS, IF 12.2）にて発表した。さらに、災害データセットの構築を通じて蓄積した、都市理解および多時期解析に関する知見を活かし、動的都市変化分析（DynamicVL）

および多視点統合による持続可能な都市開発支援（CityVLM）に関する研究を展開し、災害対応から都市レジリエンスの向上へと研究の適用範囲を拡張した。

7. 今後も本課題への継続申請を行い、災害対応エージェントの研究開発を加速する。具体的には、災害 QA（問答）システムの能力を拡張・高度化し、説明可能性（可説明性）を備えた高精度な推論と、広域・リアルタイムな応答を両立する災害支援基盤の実現を目指す。