衛星画像の時系列変化説明に向けた LVLM の比較

辻本陵 [1] 大内啓樹 [1,2,3] 上垣外英剛 [1] 渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学 [2] 理化学研究所 [3] 国立国語研究所
{tsujimoto.ryo.tq0,hiroki.ouchi,kamigaito.h,taro}@is.naist.jp

概要

衛星画像から時系列変化を説明することは，都市計画や環境モニタリングなどにおいて重要である．しかし，データセットの人手構築には高いコストがかかる．この課題に対処するため，本研究では，衛星画像を用いた時系列変化の説明生成に焦点を当て，Large-scale Vision-Language Models (LVLMs) によって生成された時系列変化説明を比較する．実験では，複数のモデルと Prompting を検証し，生成された説明における正解センテンス中の単語網羅率，説明の忠実性と情報性などを評価した．その結果，網羅性においては LLaVA-1.5，忠実性と情報性においては GPT-4-vision-preview に衛星画像を直接入力する手法が最も有効であることを確認した．

1 はじめに

衛星データは，空間分解能と波長分解能の両面で飛躍的な進歩を遂げ，利用者が入手することのできる衛星データの種類も多くなった．中でも，撮影時期の異なる 2 枚の衛星画像（以下，**二時期衛星画像**）を用いた変化検出は，災害による被害状況の分析 [1] に用いられるなど，実応用上も重要な役割を果たしている．

二時期衛星画像からの変化検出として，従来はエッジ情報に基づいたハイライト [2] や，マルチスペクタルデータに基づいた面積変化量の検出 [3] などが検討されてきた．しかし，このような変化の検出だけではなく，その変化を言語化し，一般利用者にも直感的に理解可能な形式にすることが望ましい．二時期衛星画像の時系列的変化説明の例を図 1 に示す．ここでは，荒れ果てた並木道が住宅地に変わっていることが説明されている．

このような時系列変化説明生成に向けた既存データセットとして，Chenyang ら [4] の Levir-CC[1)] が存在する．しかし，その説明は変化箇所を端的に描写した文となっている．時系列変化説明として，変化箇所の描画だけでは不十分であり，変化の内容や程度における重要度を言語化することが望ましい．たとえば，図 1 では，並木道から住宅地へのインフラ面での変化は，周辺の木々の増減といった自然環境の変化よりも重要度が高いと考えられる．そのため，変化の説明としてもこの点を強調した形で書かれることが望ましい．一方で，詳細な時系列変化説明を含むデータセットを新たに人手で構築するには，コストや所要時間という点で限界がある．

The desolate tree-lined street has been transformed into a bustling residential neighborhood with houses and parked cars.

図 1 衛星画像の時系列変化説明の例．左の衛星画像は変化前，右の衛星画像は変化後を示す．

そこで，本研究では二時期衛星画像の時系列変化説明データセットを効率的に構築することを最終的な目標として，Large-scale Vision-Language Models (LVLMs) によって生成した時系列変化説明の精度比較を行う．評価実験の結果，正解センテンス中の名詞網羅率においては，LLaVA-1.5 [5] に二時期衛星画像を同時入力する手法が最も優れており，説明が事実に基づいているかという点，画像の詳細な情報が含まれているかという点において，GPT-4-vision-preview [6] に二時期衛星画像を同時入力する手法が最も優れていることを確認した．

2 関連研究

Simone ら [7] は，複数の画像キャプション生成モデルから出力されたキャプションを融合すること

1) https://github.com/Chen-Yang-Liu/RSICC

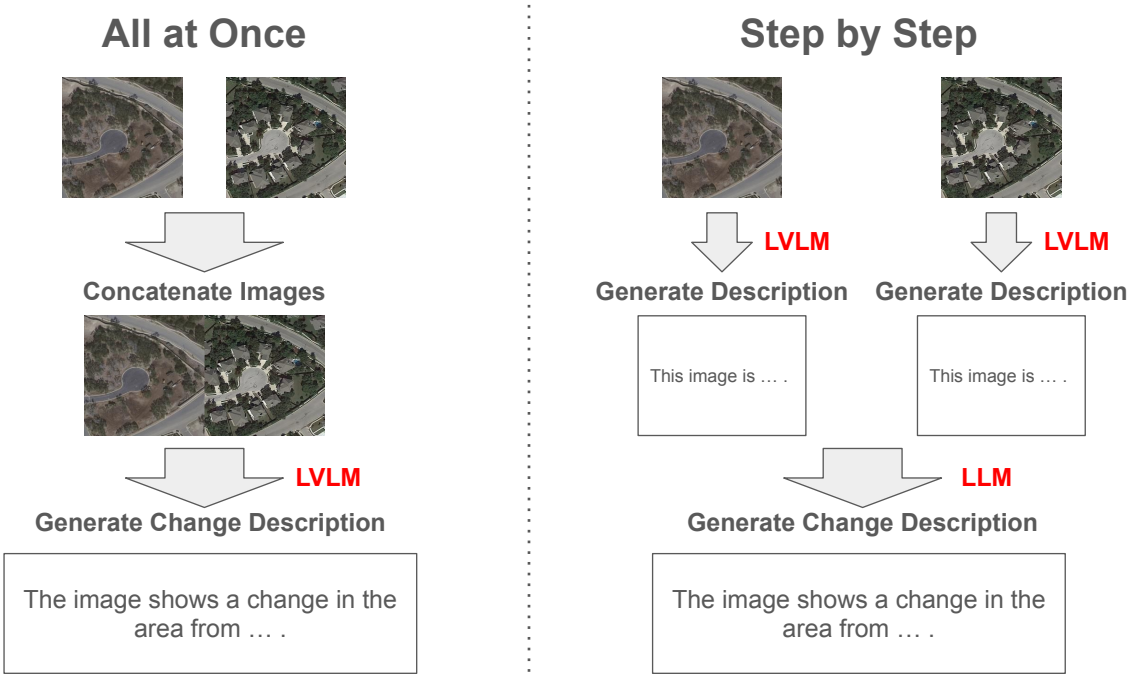

図 2　2 種類の Prompting による二時期衛生画像からの時系列変化説明生成

で，より詳細なキャプションを生成する手法を提案している．この手法では，5 つの異なるモデルによって画像キャプションを生成し，画像テキストマッチングによりキャプションをランキングした後，上位 2 つのキャプションを GPT-3 によって融合することで，より詳細なキャプションを生成する．

Zilun ら [8] は，LVLMs に衛星画像を説明させるための RSICap データセットと評価ベンチマークを提案している．

以上のように従来研究でも，画像キャプション生成タスクにおいては Large Language Model (LLM) は利用され，衛星画像の認識においても LVLMs は利用されてきた．本研究では，二時期衛星画像の時系列変化説明生成タスクにおける，LVLMs の有効性を比較する．

3　方法

図 2 に示すように，本手法では 2 種類の Prompting による時系列変化説明生成を提案する．具体的なプロンプト例は付録表 3, 4 を参照のこと．

3.1　All at Once Prompting

LVLMs の中には，複数枚の画像入力に対応していないものも存在する．このアプローチでは，画像ペアを左右に連結して 1 枚の画像として LVLM に入力する．なお指示文には以下の要素を含める．

- 画像は 2 枚の衛星画像を左右に連結したものであること
- 二時期衛星画像であること
- 時系列的変化について説明する旨

3.2　Step by Step Prompting

LVLM は，個別の衛星画像に対するキャプショニング能力は高い [8]．このアプローチでは，まず LVLM を用いて，時系列変化の前後にある各衛星画像に対するキャプションを生成する．そして，これらのキャプションを LLM に入力することで時系列変化を説明させる．なお，各衛星画像に関するキャプション生成の際には，より詳細な説明を期待して，以下の Spatial Concepts[9] を含むようにキャプションに制約を与える．

- 場所
- 空間エンティティ
- パス
- 位相関係
- 方向関係
- オブジェクトの特性
- 参照フレーム
- モーション

表 1 人手評価基準

二時期衛星画像	評価軸	スコア	時系列変化説明の例
	忠実性	1	The changes in the area include the grass being replaced by a paved road, which is visible in the right image. This transformation likely occurred due to urbanization or infrastructure development, leading to the conversion of **the grassy field into a paved road**.
		5	In the image on the left, there is a stretch of land with healthy green vegetation on either side of a well-defined white road, and some brownish patches indicate areas where the vegetation is sparse or dead. By contrast, in the image on the right, the overall color of the terrain is duller.
	情報性	1	The change in the image is the construction of a new road or street, which is visible in the right-side image.
		5	In the image, there is a noticeable change in the area, with the left image showing a dirt road surrounded by trees, while the right image displays a neighborhood with a street intersection and houses. The changes that have occurred include the transformation of **the dirt road into a paved street**, the addition of houses and street infrastructure, and the presence of multiple cars parked or driving through the intersection. **The trees in the left image have been replaced by houses**, and the overall appearance of the area has shifted from a rural setting to a more urban, residential neighborhood.

4 評価実験

各手法を比較するため，様々な評価方法を用いて検証を行った．

4.1 実験設定

モデル All at Once Prompting では，二時期衛星画像を入力する LVLM として，LLaVA-1.5 および GPT-4-vision-preview を利用した．Step by Step Prompting では，LLaVA-1.5 で各衛星画像に関するキャプションを生成し，各キャプションを入力する LLM として LLaVA-1.5，GPT-3.5-turbo および GPT-4-turbo を利用した．

データセット 画像ペアと対応した時系列変化に関する 5 つの説明文を含む Levir-CC を用いた．Levir-CC は 6815/1333/1929 の訓練/検証/テスト用データを含み，半数は変化がないペアである．本実験では，テスト用データから抽出した 100 ペアの二時期衛星画像を用いて，各手法で生成された時系列変化説明を比較した．

出力の制約 文長を制限するために，1 節で説明がなされるように指示文に制約を与えた．

4.2 評価方法

自動評価 生成された説明に対して，Levir-CC における変化説明文中の名詞の網羅率を測る．Levir-CC に含まれる正解センテンスから spaCy ライブラリで名詞を抽出し，生成文に含まれる割合を網羅率とした．

人手評価 生成されたキャプションに対して，忠実性と情報性を測る．表 1 は各評価軸の実例である．忠実性は説明に嘘が含まれていないかを評価し，忠実性 1 の例では「芝生から舗装路への変化」という誤りを含んでいる．情報性は画像の詳細な情報（特徴）が記述されているかを評価し，情報性 5 の例では，「未舗装の道路から舗装路への変化」や「木々が家に置き換えられた」など，オブジェクトの位置や関係性についての記述を含む．本実験では，英語話者のアノテーター 1 名に各評価軸について 1 から 5 までの 5 段階で回答してもらった．

4.3 実験結果

表 2 は，各手法による衛星画像の時系列変化説明に対する評価結果を示している．網羅率に関しては，Step by Step Prompting を利用した LLaVA-1.5 が最も優れていた．忠実性と情報性に関しては，All at Once Prompting を利用した GPT-4-vision-preview が最も優れていた．

4.4 考察

低スコアの要因分析 GPT-4-vision-preview に Step by Step Prompting を適用して生成した時系列変化説

表 2　評価実験の結果

Prompting	モデル	網羅率 [%]	忠実性	情報性	平均単語数
All at Once	LLaVA-1.5	22.08	2.78	2.54	89.15
Step by Step	LLaVA-1.5	**29.04**	2.91	3.21	132.92
All at Once	GPT-4-vision-preview	16.61	**3.51**	**3.26**	94.24
Step by Step	GPT-3.5-turbo	14.37	3.02	1.37	31.06
Step by Step	GPT-4-turbo	9.60	2.97	1.41	29.99

In comparing the left and right images of the same area, one can observe that a number of changes have taken place over time. In the right image, which represents the area after the changes, **additional roads have been constructed**, the landscape appears more developed with more houses built throughout the area, trees have been removed, and there is an overall increase in the infrastructure and human habitation, suggesting residential or suburban development.

図 3　忠実性 1 の出力例

In the left image of the area before the change, the land appears largely undeveloped with clear lots and minimal construction activity, whereas the right image shows significant **development with completed houses, additional roads, and swimming pools**, indicating a transition from an undeveloped to a developed residential neighborhood.

図 4　情報性 1 の出力例

明のうち，忠実性と情報性がそれぞれ 1 と評価されたものを図 3, 4 に示す．忠実性 1 のケースでは存在しない変化に関する記載があり，情報性 1 のケースでは変化の具体的な内容に関する記載がない．また，低スコア出力事例は，その多くがほぼ変化のない二時期衛生画像であった．このため，忠実性や情報性が低い説明を生成する要因としては，目立った時系列変化がない場合にも変化を説明しようとしていることが考えられる．よって，今後は変化がない場合の例外処理を指示文に含める必要がある．

Prompting の比較　LLaVA-1.5 では，All at Once Prompting と Step by Step Prompting の両方で説明生成を行い，全ての評価軸において Step by Step Prompting のほうが優れていることが判明した．このことから，Step by Step Prompting の有効性が示された．また，自身の推論情報を用いて最終的な時系列変化説明を生成するという点で，視覚タスクにおいても Chain of Thought Prompting [10] が有効であることが確認できたと考える．

単語数のスコアへの影響　LLaVA-1.5 に Step by Step Prompting を適用して生成された説明の平均単語数は 132.92 [tokens] である．一方で，GPT-4-turbo に Step by Step Prompting を適用して生成された説明の平均単語数は 29.99 [tokens] であり，前者の 4 分の 1 程度であった．また，Step by Step Prompting を適用した GPT-3.5-turbo および GPT-4-turbo は，忠実性は他の手法と同等でありながらも，網羅性および情報性においてスコアが極端に低下している．従来の画像キャプションタスクでは，被写体の簡潔な説明が求められるが，今回の時系列変化説明タスクにおいては，より詳細な出力が求められる．このため，網羅率，情報性には出力の単語数が大きく影響をしていると考えられる．

5　おわりに

本研究では，LVLM を用いて，二時期衛星画像の時系列変化に関する説明を生成した．また，評価実験により，網羅性においては LLaVA-1.5，忠実性と情報性においては GPT-4-vision-preview の有効性を確認した．分析により，GPT だと出力単語数が他のモデルと比べて顕著に少ない傾向にあり，これが網羅性を損なう原因になっていることが明らかになった．今後の展開としては，指示文の調整がある．また，地理情報システムを利用した，より情報性の高い時系列変化説明生成の手法などが考えられる．

謝辞

本研究は JSPS 科研費 JP22H03648 の助成を受けたものです.

参考文献

[1] 宇宙航空研究開発機構 (JAXA). 東日本大震災後の人工衛星の防災活用について, 2021. https://earth.jaxa.jp/ja/earthview/2021/03/12/1720/index.html.

[2] 笹川啓, 田代ゆかり, 石塚麻奈, 柴田光博. 二時期の空中写真と衛星画像による自動変化抽出手法の開発. 国土地理院時報, Vol. 134, No. 4, pp. 33–42, 2021.

[3] 橋本直之, 齋藤裕樹, 山本修平, 牧雅康, 本間香貴. 農家水稲圃場における uav によるマルチスペクトル空撮画像を用いた追肥に伴う葉面積変化の検出. 日本作物学会紀事, Vol. 90, No. 2, pp. 211–221, 2021.

[4] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. **IEEE Transactions on Geoscience and Remote Sensing**, Vol. 60, pp. 1–20, 2022.

[5] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[6] OpenAI. Gpt-4 technical report, 2023.

[7] Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. Improving image captioning descriptiveness by ranking and llm-based fusion, 2023.

[8] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark, 2023.

[9] James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. SemEval-2015 task 8: SpaceEval. In Preslav Nakov, Torsten Zesch, Daniel Cer, and David Jurgens, editors, **Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)**, pp. 884–894, Denver, Colorado, June 2015. Association for Computational Linguistics.

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

A プロンプト例

表 3 All at Once Prompting

入力文	This image is a concatenation of two satellite images placed side by side. - Both images show the same area. - The left image shows the area before the change over time, while the right image shows it after. Please describe where and what kind of changes in a clause, don't use bullet-points.

表 4 Step by Step Prompting

LVLM への入力文	Please provide a detailed description of the image. The description should includes the following spatial concepts - Places: toponyms, geographic and geopolitical regions, locations. - Spatial Entities: entities participating in spatial relations. - Paths: routes, lines, turns, arcs. - Topological relations: in, connected, disconnected. - Orientational relations: North, left, down, behind. - Object properties: intrinsic orientation, dimensionality. - Frames of reference: absolute, intrinsic, relative. - Motion: tracking objects through space over time.
LLM への入力文	description of the area before the change: DESCRIPTION description of the area after the change: DESCRIPTION Please describe where and what kind of changes occurred in the area, in a clauce.