

Pre-captioning および自己評価報酬を用いた視覚言語モデルによる画像変化記述の精度向上の試み

津守 陽太[†] 御手洗 彰[†] 中尾 恵[†]

[†] 京都大学大学院医学研究科 人間健康科学系専攻 〒606-8507 京都市左京区聖護院川原町 53

E-mail: †tsumori.yota.88k@st.kyoto-u.ac.jp

あらまし 本研究では、視覚言語モデル (VLM) に対して Group Relative Policy Optimization を適用し、画像間の幾何学的変化を推論課程を示しながら自然言語で記述する手法を検討した。推論に先立って入力画像の内容を自然言語で出力する Pre-captioning と VLM 自身の出力に対して自己評価スコアを報酬として付与する自己評価報酬を組み込むことで、画像に加えられた編集の説明精度が向上する傾向を確認した。下顎骨再建計画データセットを用いた評価により、提案手法の有効性を確認した。

キーワード Pre-captioning, 自己評価報酬, 下顎骨再建計画データ

1. はじめに

人工知能 (Artificial Intelligence; AI) は医用画像診断、疾患予測、治療方針決定支援など多岐にわたる臨床応用において、医療の質および効率の向上に寄与する技術として注目されている。一方で、多くの AI は人間の理解を超えた複雑な構造を有し、内部挙動の解析が困難な、いわゆるブラックボックスであることが課題として指摘されており [1], 判断の根拠を明示的に示す説明可能性が求められている。説明可能性は、画像に基づく高度な意思決定を支援する場面において特に重要となる。臨床現場の一例として、手術計画の立案が挙げられる。手術計画では、計画前後の医用画像やシミュレーション画像を比較し、それらの差異に基づいて手術方針を多角的に検討する [2] が、その理由を言語化して説明することは必ずしも容易ではない。具体的には、下顎骨再建計画は術者の経験や勘に依存する部分が大きく、計画過程の標準化が課題となっている。骨片の移動や回転、切断面角度の変更といった操作内容を言語化することで、熟練術者の判断過程を可視化し、将来的な再建計画の標準化や教育支援への応用が期待されている。したがって、AI が画像間の変化内容に加え、判断に至る根拠や理由を自然言語で説明できれば、人間の意思決定を支援する有用な手段となり得る。画像内容を自然言語で説明する手法として、画像とテキストを統合的に扱う視覚言語モデル (Vision-Language Model; VLM) が提案されている。VLM は単一画像の説明 [3] に加え、複数画像を対象とした研究例 [4] も報告されているが、画像間の幾何学的変化を論理立てて説明する能力は十分に達成されていない。この要因として、既存手法では各画像に関する情報を記述しないまま推論を開始

している点が挙げられる。Yaniv Leviathan らは入力プロンプトを繰り返し提示するだけで推論精度が向上することを示しており [5], モデル性能が推論機構そのものだけでなく入力情報の提示方法に大きく依存することを示唆している。すなわち、推論前に入力情報を自然言語で記述することが、推論精度向上に寄与する可能性がある。また、

本研究では、VLM に強化学習を適用し、画像間の変化を推論ステップを明示しながら自然言語で記述させることを目的とする。本手法では、推論ステップに先立ち入力画像の内容を自然言語で記述することで推論精度の向上を狙うとともに、教師データを用いない学習設定により、個々の画像への過剰最適化を回避しつつ、画像間に内在する変化そのものを抽象的に理解・推論する能力の獲得を試みる。

2. 提案手法

2.1 問題設定

本研究では、元の画像とそれに対して編集（幾何学的な変化）が加えられた画像が与えられる状況を想定する。提案モデルは、編集前後の2枚の画像とプロンプトを入力とし、出力として画像編集の内容に関する推論過程および最終的な答えを生成する。従来の教師あり学習のように入力に対する正解文を教師信号として与えるのではなく、本手法では編集内容の候補をプロンプトで提示し、モデル自身に推論過程と最終回答を生成させる。そのため、事前学習された正解文に制約されず、多様な出力が可能となる。

2.2 画像の編集を推論過程を述べながら説明する VLM

図1は、提案モデルの構成および学習、推論時における処理の流れを示している。入力された画像は Vision Encoder に

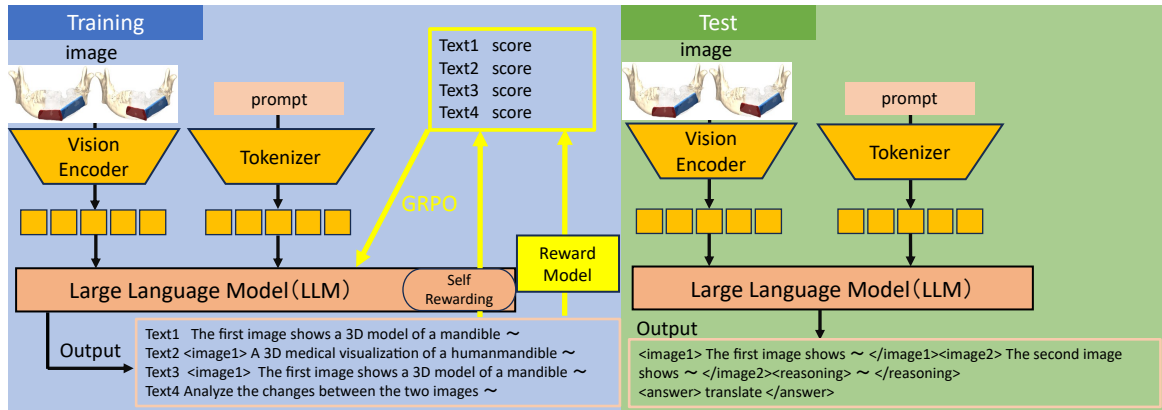


図1 VLMの学習 (GRPO) とデータ処理の流れ.

よってパッチトークンへと変換され、プロンプトはトークナイザーによりテキストトークンに変換される。トークン列はLLMに入力され、LLMが次トークン予測を逐次的に行うことで、出力文が生成される。本研究では、強化学習の手法としてGroup Relative Policy Optimization (GRPO) [6]を採用する。

本研究では、VLMの出力形式としてPre-captioningを導入する。Pre-captioningとは、推論に先立って入力画像の内容を自然言語で記述する手法であり、画像の説明文を生成した後に推論を行わせるものである。<image1>タグの中に編集前の画像の内容、<image2>タグの中に編集後の画像の内容を記述させるようにGRPOの報酬関数を設定する。Pre-captioningの内容が過度に重複するのを防ぐため、両者の語彙集合に基づくJaccard類似度を用いた報酬項を導入した。報酬関数の詳細は次節において詳しく述べる。このように、事前に画像情報に対する説明文の出力を促した後で、推論の生成タスクを実行することで、推論過程の精度向上が期待される。また、外部報酬として設計した報酬関数に加え、VLM自身による自己評価に基づく補助的な報酬 (Self Rewarding: SR) の導入を提案する。これは、モデル自身が生成結果を評価し、その評価値を学習信号として利用するという、自己報酬型学習の考え [7] に着想を得たものである。本手法では、VLMが生成した出力に対して、同一のVLMが自己採点を行い数値的な評価を付与する。その評価値を強化学習における報酬として再びVLMに与えることで、出力の質を自律的に改善することを目指す。

2.3 下顎骨再建手術計画編集ペアデータセットの提案

医用画像における空間的・幾何学的変化に対する視覚言語推論能力を評価するため、先行研究 [8] [9] により整備された下顎骨再建計画データベースを用いた。本データベースは、下顎骨再建術を受けた29症例の頭部および下肢の3次元CT画像上に切断面を設定し、各切除領域に再建計画を付与したものである。切断面の定義は図2(a)に示す。C0は下顎枝前縁、C1は正中とC0の中点として設定される。C5はオトガイ孔位置とし、これを基準としてC3は正中とC5の中点、C4はC3とC5の中点として階層的に定義する。C2はC3と正中に関して対称となる位置に配置する。これらから、(C0, C2), (C0, C3), (C0, C4), (C0, C5), (C1, C2), (C1, C3), (C1, C4),

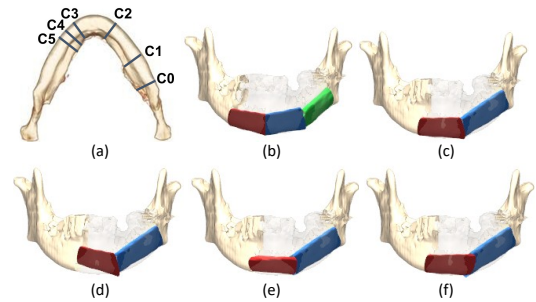


図2 切除面と編集前後の下顎骨再建計画画像 (a) 切除面, (b) (c0, c5) に対する手術計画画像 (c) (c0, c2) に対する手術計画画像, (d) 平行移動, (e) 回転, (f) 切除角度変更

(C1, C5)の8種類の切除領域が構成される。各症例の再建計画は、歯科技工士1名が下顎骨再建システム [8]を用いて作成したもので、腭骨片の本数 (1~3本) および空間配置が症例ごとに決定されている。以上より、29症例×8切除領域=232例の再建計画が得られた。図2(b), (c)には、切除領域(C0, C5)および(C1, C3)に対応する再建例の3次元レンダリングを示す。腭骨片数および配置の違いにより、再建形態が異なることがわかる。本研究では、この232例を基に編集をして追加データを生成した。具体的には、下顎骨と腭骨片の接続点の平行移動、腭骨片の軸まわり回転、および切除角度の変更を適用し、編集前後の対応関係を持つ画像ペアを構築した。図2の(c)の計画画像に対する編集例を図2の(d), (e), (f)に示す。さらに各症例・各切除領域でこれらの操作を12通りの設定で組み合わせてデータ拡張を行い、最終的に2784組の再建ペアデータセットを構築した。

3. 実験

本実験では、提案手法を用いて、編集前後の下顎骨再建手術計画画像に対する操作識別性能を評価する。特に、Pre-captioningおよび自己評価報酬の導入が、画像間差分理解および推論過程生成に与える影響を比較検証する。

3.1 実験条件

本研究では、モデルの挙動が過度に変化することを防ぐため、言語的な推論能力のみを調整した。実験には、複数枚の

画像を同時に入力可能な視覚言語モデルである Qwen3.5-4B を用いた。Qwen3.5-4B の内部構造のうち、テキスト処理を担う自己注意機構 (Self-Attention) および Feed-Forward Network (FFN) を構成する線形射影層に対して LoRA [10] を適用した。具体的には、Self-Attention に対応するクエリ射影、キー射影、バリュー射影、出力射影、および FFN に対応するゲート射影、アップ射影、ダウン射影を学習対象とした。バッチサイズは 2、エポック数は 1、学習率は 5.0×10^{-6} とした。GRPO のパラメータである ϵ は、学習初期には探索を広く行うために大きめの値に設定し、学習が進むにつれて方策が安定するように徐々に小さくなるように調整した。また、KL 正則化項の係数 β については、本研究では参照モデルとの乖離を抑制するよりも、画像編集内容を記述するタスクへの適応を優先するため 0 に設定した。近年の GRPO を用いた視覚言語モデルの追加学習においても、タスク適応を重視する設定として $\beta = 0$ が採用される例 [11] が報告されている。さらに、1 入力あたりの生成数を示すサンプリング数は 16 に設定した。各サンプリングのトークン数の上限は 1024 とした。

提案手法では、GRPO における報酬関数として、出力構造、画像説明品質、推論過程、回答形式、正答性を評価する複数の報酬を導入した。構造報酬では、Pre-captioning ありの場合、`<image1>`、`<image2>`、`<reasoning>`、`<answer>` タグがすべて含まれている場合に 0.5 点を付与し、さらにこれらが `image1` $\rightarrow$ `image2` $\rightarrow$ `reasoning` $\rightarrow$ `answer` の順で出現する場合に追加で 0.5 点を与えた。Pre-captioning なしの場合、`<reasoning>` および `<answer>` タグの存在に 0.5 点、`<reasoning>` が `<answer>` より前に出現する場合に追加で 0.5 点を付与した。画像説明品質報酬では、各 `<image1>` および `<image2>` タグ内の説明文について、単語数が 6 語以上の場合に報酬を与え、総報酬 0.6 を 2 画像へ均等配分した。さらに、2 つの画像説明間の Jaccard 類似度を計算し、類似度が低いほど高報酬となるよう、多様性報酬として最大 0.4 点を付与した。推論報酬では、`<reasoning>` 内の単語数が 12 語以上の場合に 0.6 点を付与し、加えて “therefore”、“hence”、“sum” や四則演算記号など、論理的推論を示す語句・記号が含まれる場合に 0.4 点を与えた。回答形式報酬では、`<answer>` タグが存在する場合に 0.5 点を付与し、さらに回答が単一トークンで構成される場合に追加で 0.5 点を与えた。最後に、`<answer>` 内の予測ラベルが正解ラベルと完全一致した場合に正答報酬として 1.0 点を付与した。これら複数の報酬を組み合わせることで、単なる分類精度のみならず、各画像の特徴記述および推論過程を伴う構造化出力の生成を促進した。また、Base (w/o Precaptioning + 自己評価報酬)、Pre-captioning、Pre-captioning + 自己評価報酬の 3 つの条件で比較実験を行った。

3.2 下顎骨再建手術計画ペアデータセット精度評価

学習データには、編集前後の下顎骨再建ペア画像と、表 1 のプロンプトを組み合わせたデータを 2592 件、テストデータには 192 件用いた。結果を表 2 に示す。まず、Base では全体的に各クラス間の混同が大きく、特に `rotate` は `cut_plane` や `translate` に、`translate` も `rotate` および `cut_plane` に

表 1 入力プロンプト

```
# Instruction
You will receive TWO images.
Respond in EXACTLY these 4 sections and in this order:
<image1> A concise description of the 1st image only. </image1>
<image2> A concise description of the 2nd image only. </image2>
<reasoning> Step-by-step reasoning that uses the descriptions to solve the task. </reasoning>
<answer> The final answer ONLY. </answer>
Do not add any other text outside these tags.

# Input
Observe the two images (front view) taken before and after the editing operation in a mandibular reconstruction surgery plan, and determine which type of edit was performed, choosing from the following three options:
translate: parallel translation of a segment (change in position)
rotation: roll rotation around the segment's long axis (change in cross-sectional orientation)
cut_plane: rotation along the normal direction of the cut plane (change in the angle of the cut plane)
Output only one of the following: translate / rotation / cut_plane.
```

それぞれ 15 件ずつ誤分類されており、3 クラス間の境界が曖昧であった。次に、Pre-captioning では、`cut_plane` の正解数が 34 件から 60 件、`translate` も 33 件から 52 件へ増加した。一方で、`rotate` は正解数が 20 件から 0 件となり、うち 44 件が `cut_plane` に誤分類されており、両者を頻繁に混同していることがわかる。さらに、Pre-captioning + 自己評価報酬では、`rotate` の正解数が 0 件から 11 件へ改善し、`rotate` が `cut_plane` に誤分類される数が 44 件から 36 件へ減少した。`cut_plane` は 60 件から 56 件へ、`translate` は 52 件から 47 件へ低下したものの、全体として大きな性能低下はみられなかった。

表 3 に `translate` 編集に対する生成例を示す。Base ではセグメントの位置変化ではなく向きの変化に着目し `rotate` と誤分類した。一方、Pre-captioning を導入した場合には、“red segment has been translated downward” 記述が付与され、平行移動を正しく認識できている。表 4 に、`rotate` 編集に対する生成例を示す。Pre-captioning のみを用いた条件では、“the red region has been tilted” や “rotating a cut plane” といった記述から、モデルが赤色セグメントの角度変化を `cut_plane` の傾き変化と解釈し誤分類していたことがわかる。一方、Self-Rewarding を追加した条件では、“rotated around its long axis” や “cross-sectional shape and orientation have changed” と記述されており、編集領域の回転軸や断面方向の変化を適切に捉えられている。

以上の結果から、Pre-captioning は画像間に与えられた編集を言語的特徴として整理しやすくし、`cut_plane` や `translate` の識別性能工場に寄与したと考えられる。一方、Pre-captioning 導入後に `rotate` の性能が極端に低下したのは、`rotate` と `cut_plane` の視覚的特徴が類似していたためと考えられる。特に、本研究では前面画像 (front view) のみを入力としたため、腭骨片軸周りの回転による断面形状の変化が、切断面角度の変化と類似した見え方になった可能性が高い。これに対し、自己評価報酬導入後に `rotate` の正解数が増加したことから、推論過程や回答の妥当性をモデル自身に評価させる仕組みが特定ラベルへの偏りを抑制したと考えられる。ただし、`cut_plane` や `translate` の制度がわずかに低下していることから、自己評価報酬は特定クラスへの過度な収束を防ぐ代わりにクラス

表2 各設定における混同行列. PC, SR, Rot., Cut, Trans., Unp. は, それぞれ pre-captioning, self rewarding, rotate, cut_plane, translate, unparsed を表す.

Base					PC					PC+SR				
G.T.\Pred.	Rot.	Cut	Trans.	Unp.	G\P	Rot.	Cut	Trans.	Unp.	G\P	Rot.	Cut	Trans.	Unp.
Rot.	20	20	24	0	Rot.	0	44	20	0	Rot.	11	36	17	0
Cut	14	34	16	0	Cut	0	60	4	0	Cut	3	56	5	0
Trans.	15	15	33	1	Trans.	0	12	52	0	Trans.	10	7	47	0

表3 Base, Pre-captioning における translate 編集の生成結果比較

Base
Model completion <reasoning> ~~ The segment orientation appears changed, indicating a rotation edit. ~~ </reasoning> <answer> rotate </answer> Extracted answer : rotate Gold answer : translate
Pre-captioning
Model completion <image1> A 3D medical visualization of a human mandible (lower jaw). </image1> <image2> The 2nd image shows the same scene, but the red segment has been translated downward. </image2> <reasoning> ~~ The segment position changes without orientation change, indicating translation. ~~ </reasoning> <answer> translate </answer> Extracted answer : translate Gold answer : translate

表4 Pre-captioning, Pre-captioning + Self-Rewarding における rotate 編集の生成結果比較

Pre-captioning
Model completion <image2> ~~ The red segment's angle has changed relative to the bone structure. ~~</image2> <reasoning> ~~ 3. Evaluate 'rotate': A roll rotation would spin the block around its own long axis. 4. Evaluate 'cut_plane': The red region has been tilted, which is consistent with rotating a cut plane. </reasoning> <answer>cut_plane</answer> Extracted answer : cut_plane Gold answer : rotate
Pre-captioning + Self-Rewarding
Model completion <image2> ~~The highlighted region has been rotated around its long axis, changing its cross-sectional orientation while remaining in the same position. </image2> <reasoning> ~~ - It has not moved to a new location, so it's not a 'translate'. - Its cross-sectional shape and orientation have changed; it appears rotated around its own long axis. ~~</reasoning> <answer>rotate</answer> Extracted answer : rotate Gold answer : rotate

間のバランスを均衡化する方向に作用したといえる。

4. おわりに

本論文では, 画像間に加えられた編集を推論ステップを明示

しながら自然言語で説明することを目的とし, VLM に対する GRPO による強化学習を基盤として, そこに Pre-captioning および自己評価報酬の導入を提案した. 評価実験の結果, Pre-captioning を導入することで画像特徴の比較が容易となり, 画像間に加えられた編集を言語的に整理しやすくなることを確認した. さらに, 自己評価報酬を併用することで, 出力形式の安定化および誤出力の抑制が可能であることを示した.

文 献

- [1] A. B. Arrieta, N. D'iaz-Rodr'iguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-L'opez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges. *Information Fusion*, Vol. 58, pp. 82–115, 2020.
- [2] M. Nakao, M. Hosokawa, Y. Imai, N. Ueda, T. Hatanaka, T. Kirita, and T. Matsuda. Volumetric fibular transfer planning with shape-based indicators in mandibular reconstruction. *IEEE Journal of Biomedical and Health Informatics*, Vol. 19, No. 2, pp. 581–589, 2015.
- [3] A. Radford, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [4] H. et al. Liu. Mibench: Evaluating multimodal large language models over multiple images. In *EMNLP*, 2024.
- [5] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*, 2025.
- [6] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, and M. Zhang. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [7] X. Zhao, Z. Kang, A. Feng, S. Levine, and D. Song. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*, 2025.
- [8] M. Nakao, S. Aso, Y. Imai, et al. Automated planning with multivariate shape descriptors for fibular transfer in mandibular reconstruction. *IEEE Transactions on Biomedical Engineering*, Vol. 64, No. 8, pp. 1772–1785, 2017.
- [9] K. Nagai, M. Nakao, N. Ueda, et al. Enumerated sparse extraction of important surgical planning features for mandibular reconstruction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5519–5522. IEEE, 2020.
- [10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations (ICLR)*, 2022.
- [11] Wenhao Yu, et al. Doctinker: Explainable multimodal large language models with rule-based reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.