（報道発表資料）

2026.6.1

NTT 株式会社

なぜその結果になったのか？
推論根拠を説明できるマルチモーダル XAI 技術を確立

～人と AI、AI と AI の信頼を強化し、
安心できるビジネス意思決定や AI エージェント連携などを実現～

発表のポイント：

- ◆ 大規模視覚言語モデル（LVLM）が段階的な思考による推論（Chain-of-Thought, CoT）を行う際、根拠と結果が一貫していないという重大な課題を発見しました。
- ◆ この課題に対し、画像の情報を維持しながら根拠の情報を最大限活用する理論的な枠組みを導入することで、推論時に任意の LVLM の出力を画像と根拠の双方に忠実に依存させる「根拠強化デコーディング」を確立しました。
- ◆ 本技術により、これまでブラックボックスだった LVLM を追加の学習コストなしで説明可能 AI（eXplainable AI, XAI）として運用でき、ビジネスでの意思決定や AI エージェント連携による複雑な課題解決など、より高い信頼性が求められる幅広いユースケースへの応用が期待できます。

NTT 株式会社（本社：東京都千代田区、代表取締役社長：島田 明、以下「NTT」）は、画像と言語を扱うマルチモーダル AI 基盤モデルによる出力の信頼性を高める新たな推論の仕組みとして「根拠強化デコーディング」技術を確立しました。本技術は、LVLM が CoT を行う際、自身で生成した推論根拠を無視する傾向があるという課題に対して、通常の推論とは異なり、画像による推論と根拠による推論を分割しそれらを重みづけて組み合わせました。これにより、画像と根拠の双方から得られる情報を忠実に活用して回答を出力することを可能にしました。本成果は、2026 年 6 月 3 日から 2026 年 6 月 7 日まで、米国・デンバーで開催されるコンピュータビジョン分野における最難関国際会議 Computer Vision and Pattern Recognition (CVPR) 2026（*1）において発表されます。

1. 背景

近年、大規模言語モデル（LLM）と事前学習済み画像エンコーダを統合した大規模視覚言語モデル（LVLM）の開発が進み、高度なマルチモーダル推論が可能となっています。LVLM はテキストだけでなく画像を直接入力することができ、テキストだけでは解決が難しかった動画像分析や文書読解のような動画像にもとづく複雑なマルチモーダル推論の基盤として活用が進んでいます。テキストのみを入力とする LLM と同様に、LVLM においても視覚情報とテキスト入力から「推論の根拠」を中間的に生成し、根拠を入力系列に加えることで最終的な回答出力を導く Chain-of-Thought（CoT）が推論能力の向上や説明可能な推論手法として有効であると考えられてきました。

しかし、既存の CoT メカニズムは画像と根拠を一つの系列として入力して最終出力を生成するた

め、推論の根拠に含まれている情報を必ず使用して出力するような因果的な構造を持っておらず、根拠の使用をモデル任せにしています。すなわち、CoT による最終出力は根拠の内容にもとづくことが保証されていません（図1）。

実際に、既存の LVLM はマルチモーダル CoT 推論において、生成した推論の根拠の内容を無視して最終的な回答を生成してしまうことが我々の実験と分析によって明らかになりました。例えば、推論の根拠を質問と無関係なものにすり替えてもモデルの最終出力が変わらない場合があります（図2）。この例では、スライド文書の画像に対して無関係なスポーツカーに関する根拠を入力していますが、根拠から期待される誤った回答ではなく、正しい根拠を入力した場合と同じ回答を生成します。このとき、モデルは画像のみから最終出力を生成していると考えられ、推論の根拠は出力の説明として解釈する事ができません。これらの発見は、通常の LVLM による推論は、根拠と最終的な回答の一貫性が制限されており、説明可能な推論を行えないという根本的な課題を示しています。

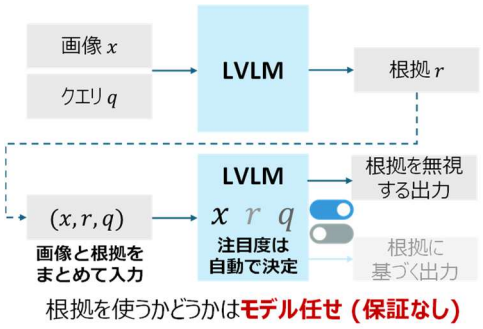

図1　LVLM における CoT 推論

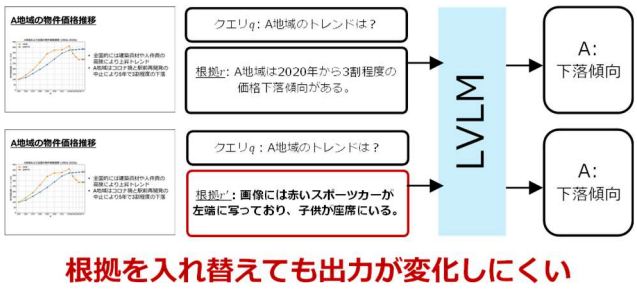

図2　LVLM の根拠への非依存性

2. 研究成果の概要

本研究成果では、この課題を解決するため、既存の LVLM の推論方法を見直し、追加のデータセットやコストのかかる再学習を必要としない、プラグアンドプレイ型の推論時デコーディング技術「根拠強化デコーディング」を確立しました。

根拠強化デコーディングは、LVLM が次のトークンを予測する確率を、画像に条件付けられた分布と推論の根拠に条件付けられた分布に分離し、これらをかけ合わせることで画像から得られる情報と根拠から得られる情報を調和させて回答を出力します（図3）。この方式では画像と根拠が別々に LVLM に入力されるため、根拠の使用を保証することができます。具体的には、マルチモーダル CoT を根拠条件付き分布の対数尤度を報酬とした KL ダイバージェンス制約付きの報酬最大化問題として定式化し、この問題を閉形式で解くことで、推論時のみの計算で LVLM が画像情報と根拠情報の双方に明示的に基づく最適な次トークン予測を実現しています（図4）。

3. 技術のポイント

① マルチモーダル CoT の KL ダイバージェンス制約付き報酬最大化問題としての定式化：

本研究では、まず通常のマルチモーダル CoT が画像と根拠を同時に条件付けた単一の次トークン予測分布を利用しており、これが必ずしも根拠の内容に確実にもとづいて回答を生成していないことに着目します。そこで、画像と根拠の両方にもとづいて推論を行うためにこの推論プロセスを新たな最適化問題として再定式化しました。具体的には、推論の根拠に条件付けられた予測確率を「報酬」として最大化しつつ、画像に条件付けられた予測確率から大きく逸脱しないように制約（KL

ダイバージェンス制約）を遵守するようにトークン生成を行います（図4上）。

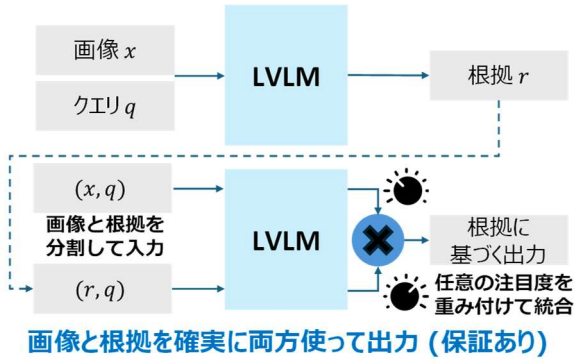

図3　根拠強化デコーディングの概要

報酬最大化問題としての CoT

$$\max_{\pi} \mathbb{E}_{\pi}[\log p(y_i \sim \pi | y_{<i}, r, q)] - \frac{1}{\lambda} \mathbb{D}_{\mathrm{KL}}[\pi || p_\theta(y_i | \boldsymbol{y}_{<i}, x, q)]$$

報酬：根拠条件付き確率の対数尤度　　制約：画像条件付き確率とのKLダイバージェンス

最適解：追加学習無しで計算可能

$$\pi^\star = \frac{1}{Z_\theta} p_\theta(y_i | \boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i | \boldsymbol{y}_{<i}, r, q)^\lambda$$

画像条件付き分布　　根拠条件付き分布　　根拠への注目度

図4　報酬最大化問題としての再定式化

② 追加学習不要のプラグアンドプレイ実装：

上記の最適化問題は LVLM を追加学習することによって解くことができますが、訓練データセットや計算機環境などのコストが非常に大きくなってしまいます。そこで、本研究ではこの最適化問題の最適解となる分布が、画像に条件付けられた分布と根拠に条件付けられた分布の積で表現される分布と等しいことを数学的に証明しました（図4下）。これにより、実際の実装ではモデルが出力するロジットの重み付き和を計算するだけで済むため、追加学習を一切行う必要がなく、既存のあらゆる LVLM にそのまま組み込める（プラグアンドプレイな）極めて実用性の高い手法となっています。

実験においても、様々な LVLM に対して根拠強化デコーディングを適用することで一貫して推論性能（正答率など）を大幅に向上させることに成功しています。さらに、推論の根拠としてより高品質なテキスト（例: GPT-4 によって生成された根拠）を与えた場合、根拠強化デコーディングの優位性はさらに増幅されることが確認され、LVLM が根拠の内容を忠実に解釈・活用できていることが実証されました。

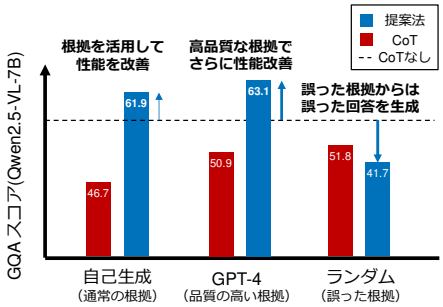

図5　根拠介入実験結果

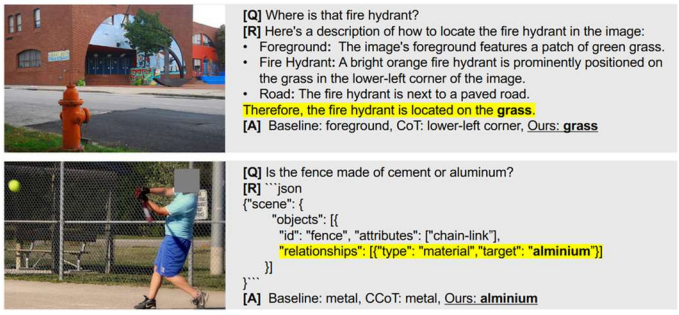

図6　提案法による推論の例

4. 今後の展開

本研究成果は、画像と推論の根拠の双方を明示的に使用して最終回答を生成する新しい推論の枠組み「根拠強化デコーディング」技術を確立し、様々な LVLM に対して追加学習を行うことなく根拠への忠実度と推論性能を大きく高められることを確認しました。この技術はこれまでブラックボックスだった LVLM の推論過程に解釈性を与えられる可能性を示唆しています。これによって、医療画像診断や人間の意思決定に係る重大なケースを扱う対話エージェントなど、より確実で信頼性の

高い推論システムが求められる分野への LVLM の社会実装が加速することが期待されます。NTT は今後も AI の信頼性の改善や、多数の AI を連携させる AI コンステレーション（*2）の具現化につながる次世代の技術開発に貢献していきます。

発表について：

本成果は、2026 年 6 月 3 日～7 日に開催されるコンピュータビジョン分野における最難関国際会議 CVPR 2026（Computer Vision and Pattern Recognition）にて、下記のタイトル及び著者で発表されます。

著者： 山口 真弥、千々和 大輝（コンピュータ＆データサイエンス研究所）、西田 光甫（人間情報研究所）

【用語解説】

※1. CVPR 2026

コンピュータビジョンに関するトップレベルの国際会議。 https://cvpr.thecvf.com/

※2. AI コンステレーション

https://www.rd.ntt/cds/ai-constellation/

■本件に関する報道機関からのお問い合わせ先
NTT 株式会社
サービスイノベーション総合研究所
広報担当
問い合わせフォームへ