視覚言語モデルと世界モデルの接地

Grounding visual language models and world models

何 振宇 *1 He Zhenyu　吉川 和之 *2 Yoshikawa Kazuyuki　井上 歩紀 *3 Inoue Ibuki

*1東京大学工学部電子情報工学科
Department of Information and Communication Engineering, The University of Tokyo

*2兵庫県立大学工学部電気電子情報工学科電子情報工学コース
Department of Electrical, Electronics and Information Engineering, Electronics and Information Engineering Course,The University of Hyogo

*3東京農工大学工学府知能情報システム工学専攻
Department of Electrical Engineering and Computer Science,Tokyo University of Agriculture and Technology, Tokyo, Japan

VLM を活用した汎化エージェントの学習は，複数タスクの効率的な解決を目指す分野で重要な課題である．従来の RL は，タスクごとに複雑な報酬設計を必要とするため，他タスクへの応用が難しい課題があった．また，視覚言語モデル (VLM) は自然言語でのタスク指定の容易なインターフェースを提供するが，応用にあたってはドメイン間ギャップやデータ不足が障害となっていた，本研究では，VLM の表現と RL 用生成世界モデルの潜在表現を結びつけることによって，ファインチューニング不要で高い汎用性を実現する手法を提案する．VLM を用いてタスクから得た埋め込み表現を潜在表現に変換後，ロールアウトにおける実際の潜在表現との距離を損失関数として方策を学習することによって，ALFRED ベンチマークにおいてタスク解決性能が向上した．

1. 研究背景・目的

自然言語処理（NLP）とロボット制御の融合は，知能エージェントの設計において近年注目を集めている分野である．自然言語を用いたロボット制御は，人間と機械の相互作用を高次元で実現する新たな可能性を提示している．例えば，家庭内のスマートロボットが人間の自然言語命令に基づいて掃除や配膳を自律的に行っている．このようなシステムは，人間が言語を用いてロボットに直感的に命令を伝えることを可能にし，複雑な作業を簡略化する．ロボット制御の汎化能力を向上させるには，複数のタスクを効率的に解決可能な統合的アプローチが不可欠である．例えば，災害救助ロボットが言語指示を受けて瓦礫の中から人を救出する場面や，農業ロボットが複数の作業を同時に管理し，効率的に収穫や植え付けを行うシステムが求められている．この目標を達成するためには，タスク遂行時のロールアウトにおける学習手法が鍵となる．

これには世界モデルと言語表現を結びつける統合的なメカニズムが必要である．世界モデルと言語を結びつける方法として，強化学習（RL）を活用した手法が挙げられる．RL はタスクごとにデータセットと報酬関数を設計することによって，高い精度で特定のタスクを実行することが可能である．しかしながら，従来の自然言語を用いた RL はタスクごとに複雑な報酬関数を設計したり，タスクごとにことなる言語表現を学習する必要があるため，汎用性にかけている．タスクに寄るこうした統合には，マルチモーダルデータの一貫性を保つことや，データの多様性とスケーラビリティを確保する技術が求められる．これらを同時に満たす手法の開発は依然として課題が残されている．

視覚言語モデル（Vision-Language Model, VLM）と生成モデルに基づく世界モデルを統合し，高度な汎化性能を持つエージェントを設計する新たなフレームワークを提案することを目的とする．このフレームワークでは，VLM が提供する言語および視覚表現を潜在空間にマッピングし，生成モデルと統合することで，ロールアウト中の制御方策を効率的に学習することを目指す．特に，従来のタスクごとに設計された複雑な報酬設計を不要とし，ファインチューニングを必要としない高い汎用性を実現する点が本手法の大きな特徴である．

2. 関連研究

2.1 世界モデル

世界モデルとは，「外界（世界）から得られる観測情報に基づき外界の構造を学習によって獲得するモデル」である．なお，ここでの観測とは，画像をはじめ，音声，文書など外界から得られる様々な種類の情報のことを指す．これらを学習することで大規模な外界のモデルを作るというのが世界モデルの重要な点であり，世界モデルを持つことによって，大きく分けて「予測」と「推論」の 2 つが実現できる．

2.2 GenRL

一般的なエージェントが多様なタスクを異なるドメインで解決することは長年の課題である．また，強化学習（RL）は複雑な報酬設計が必要で，スケールアップが難しい．言語はタスクを自然に指定する手段として有用だが，視覚言語基盤モデルは，エンボディドコンテキストでの適用に課題があり，マルチモーダルデータの不足がエンボディドアプリケーション向けの基盤モデルの開発を妨げていることも挙げられる．GenRL は視覚データのみを使用して，基盤 VLM の表現と RL の生成的世界モデルの潜在空間を接続・整合させるマルチモーダル基盤世界モデル（MFWM）を学習する．タスクプロンプトを潜在ダイナミクス空間のターゲットに翻訳し，想像の中で RL を用いて学習する．これにより，言語アノテーションを必要とせずに，視覚または言語プロンプトから複数のタスクを解決するエージェントを訓練することが可能となる．

連絡先: k.yoshikawa.2002@gmail.com

2.3 GenRL の詳細な手法

本研究は，報酬設計の複雑性を排除し，効率的かつ汎用的なタスク解決を可能にする学習フレームワークを構築することを目的とする．この目的の達成には，現代の VLM および生成世界モデルの高度な統合が不可欠である．初めに，報酬生成を必要としない手法で世界モデルの訓練を行う．具体的には，環境内での動作を的確に表現する潜在空間を構築し，エージェントの動作を予測可能な形式で抽象化する．これにより，報酬設計に依存しない柔軟なモデル構築が可能となる．次に，映像データを利用する VLM を用いて，視覚情報を潜在表現に変換するためのコネクタ（connector）を訓練する．同時に，言語表現を視覚的特徴に整合させるアライナー（Aligner）の学習を行う．これらのモジュールは，視覚と言語の異なるモダリティを統一的な形式で統合する役割を果たし，エージェントが多様なタスクに対応するための基盤を形成する．タスク遂行においては，与えられた初期条件とタスクプロンプトに基づいて，直近の数フレームにおける理想的な潜在状態を予測する．この理想潜在状態は，エージェントが達成すべき目標を指し示す指標として機能する．続いて，理想潜在状態と実際のロールアウト中に得られる潜在状態を比較し，時間的な整合性を保つためにタイムテンポラルマッチング（Time-temporal Matching）を適用する．このプロセスは，目標状態への収束を保証し，適切な行動選択を促進する．さらに，理想潜在状態と実際の潜在状態のペア間でコサイン類似度を最小化するように学習を進める．この学習目標は，エージェントの潜在表現が理想状態に近づくように設計されており，結果としてタスク遂行能力の向上を実現する．具体例として，家庭内での物品探索や，災害現場での救助活動など，多岐にわたる応用シナリオにおいて提案手法の有効性を示すことが期待される．これらの一連のプロセスを通じて，本研究は汎用性の高いエージェント設計の実現を目指す．提案フレームワークは，エージェントが未知のタスクや環境に対しても柔軟に対応できる能力を付与し，人工知能分野における重要な課題解決に寄与するものである．

2.4 Dynalang

Dynalang は, 言語を活用して未来の予測を行うマルチモーダルな世界モデルを学習するエージェントである。このモデルはテキストや画像の表現を予測し, 仮想的なシミュレーションから行動を学習する。これにより, 環境の記述やゲームのルール, 指示など, 多様な種類の言語情報を理解し, タスクの遂行能力を向上させる。さらに,Dynalang はテキストやビデオのデータセットで事前学習が可能で, 行動や報酬のラベルがなくても学習を進められる柔軟性を持っている。

2.5 Socratic Models

Socratic Models（SMs）は, 複数の大規模事前学習モデルに組み合わせ, 言語を介して情報に交換し, 新たなマルチモーダルな能力に引き出すモジュール式のフレームワークである。これにより, 画像キャプション生成やビデオからテキストへの検索などのタスクで最先端の性能に発揮し, エゴセントリックビデオ（ユーザー視点のビデオ）に関する自由形式の質問への回答や, 料理レシピの支援対話, ロボットの認識と計画など, 多様な応用が可能となる。

2.6 SayCan

SayCan は言語モデルとロボットの行動可能性（アフォーダンス）に組み合わせることで, 自然言語の指示にロボットが理解し, 実行可能な行動に変換する手法である。具体的には, 言語モデルが指示に関連するスキルの確率に評価し, ロボットの行動モデルがそのスキルを実行する成功確率に評価する。これらの情報に組み合わせることで, ロボットは与えられた指示に最も適した行動に選択し、実行する。

2.7 InternVideo2

InternVideo2 はマルチモーダルなビデオ理解のための新たなビデオ基盤モデル（ViFM）であり, ビデオ認識やビデオとテキストの関連タスク, ビデオ中心の対話において最先端の成果に達成している。このモデルは, マスク付きビデオモデリング, クロスモーダルなコントラスト学習, 次のトークン予測に統合した漸進的な学習アプローチに採用し, ビデオエンコーダーのパラメータ数に 60 億にまで拡張している。これにより, 長い文脈の理解や推論能力に向上し,60 以上のビデオやオーディオのタスクで優れた性能に示している。

2.8 ALFRED

自然言語指示を基に，家庭内タスクを実行するための行動シーケンスを学習するためのベンチマークであり，現実世界のアプリケーションと研究ベンチマークのギャップを縮めることを目指す．ALFRED は視覚的かつ物理的に現実的なシミュレーション環境であり，言語から行動への翻訳を学習するモデルの開発を促進する．また，このベンチマークは実世界の言語駆動型ロボティクスに向けた多くの課題を捉えており，これらの課題を克服するモデルの開発が期待される．

2.9 ALFWorld

TextWorld はテキストベースの環境であり，ALFRED は視覚的な環境での指示に従うためのデータセットである．抽象的なテキストベースのポリシーを学習し，それを具体的な視覚環境でのタスク実行に活用するエージェントを開発する．また，新しい環境やタスクに対して，エージェントがゼロショットで一般化できる能力を向上させる目的で，ALFWorld というシミュレーターを導入し，TextWorld と ALFRED を統合した環境を提供する．

3. 課題

本章では，GenRL における課題を述べる．

3.1 時間的依存性

現状は 8 フレームのタイムステップでしか生成できておらず，時間的な柔軟性を高めることで性能向上が見込まれる．例として，Stickman エージェントが地面に横たわって実行を命じられている場合，立ち上がって実行状態に到達するためのステップ数が VLM が認識する時間スパン（たとえばこの場合は 8 フレーム）を超え，報酬に不整合が生じる可能性がある．

3.2 静的なタスク

GenRL がプロンプトから推測するターゲット配列が静的な場合でもわずかに動いていることが多いため，一部の静的タスクでは，他の方法が GenRL よりも優れている場合がある．これに対処するため，静的プロンプトのターゲットシーケンス長を 1 に設定することもできるが，元論文ではささいな問題として認識し，メソッドの単純性と一般性を維持することを選択している．

4. 提案手法

本章では，本研究の提案手法を述べる．まず ALFWorld 上で random にエージェントを動かし，データを収集する．次に世界モデルを作り，同じデータで LLM の特徴量を世界モデルの潜在表現にマップする学習を行う．最後に方策を LLM の特

徴量から得られた世界モデル上の潜在表現の trajectory で評価して学習する．

5. 実験・考察

5.1 実験設定

5.1.1 使用する環境

本研究では，シミュレーション環境として ALFWorld を採用した．ALFWorld は，エージェントの行動とタスク環境間の複雑な相互作用を精密に再現可能であるため，多様なタスクを通じた評価が可能である．この環境は，視覚と言語の情報を統合する学習モデルの検証に最適なプラットフォームであり，提案手法の有効性を体系的に検証するための基盤を提供する．

5.1.2 実験設計

本研究の実験プロセスは，以下の 3 つの主要な段階に分けられる．それぞれの段階はデータの収集と世界モデルの構築，モデルの学習から構成されている．

第一段階では，ALFWorld 内でランダムに行動するエージェントを用いて，データ収集を行う．このデータには，エージェントの行動履歴，環境の状態遷移，およびそれに伴う変化が含まれているため，後続のモデル訓練における基盤となる．

第二段階では，収集されたデータを用いて世界モデルを構築する．この世界モデルは，環境の動的特性を潜在空間で効率的に表現することを目的とし，エージェントが内部的に環境の挙動を予測できる基盤を提供する．このモデルの構築により，タスク達成に必要な洞察が促進される．

第三段階として，同じデータセットを使用して，大規模言語モデル（LLM）の特徴量を世界モデルの潜在表現にマッピングする学習を行う．このプロセスでは，LLM から得られる言語表現を世界モデルの潜在空間に整合させることで，視覚と言語の統合的な情報処理を実現する．この整合性により，言語によるタスク指示が世界モデル上で適切に解釈され，実行可能となる．

5.1.3 評価方法

提案手法の性能評価には，複数の指標を用いた包括的なアプローチを採用した．評価の中心は，LLM の特徴量から生成された潜在表現に基づく trajectory 上での方策学習とその性能である．この評価は，エージェントが環境内で目標を効果的に達成できるかを確認することを目的とする．具体的には，LLM から得られる特徴量を用いて生成された潜在表現を基に，エージェントの方策を学習させる．この際，タスク達成率，環境内での行動効率，潜在表現の予測精度を主要な評価指標として採用する．これらの指標に基づき，提案手法が従来手法に対してどの程度優れた性能を示すかを定量的に比較分析した．さらに，提案手法の汎用性を検証するため，異なるタスク構造や環境設定における一貫性も評価対象とした．この汎用性の検証は，提案手法がさまざまなタスク環境でどの程度適応可能かを確認するものであり，より広範な応用可能性を示す重要な指標となる．本評価プロセスを通じて，提案手法がエージェントのタスク遂行能力をいかに向上させるかを詳細に分析し，その理論的および実践的な有効性を明らかにした．

5.2 結果と考察

6. まとめ

本研究では，世界モデルと言語モデルの接地についての論文サーベイを行うとともにマルチモーダルデータセット，マルチモーダルベンチマーク，視覚言語モデル（VLM）のサーベイを行い,Dynalang,Socratic Models,SayCan,InternVideo2,GenRL,ALFRED,ALFWorld 等のマルチモーダル世界モデルや視覚言語モデル，マルチモーダルベンチマークについて理解を深めた。

その後、これらのサーベイを踏まえて GenRL の可能性に注目し、探求するためシミュレーション環境として ALFWorld を用いて研究を行った。

7. 謝辞

本研究は，東京大学 世界モデル・シミュレータ寄付講座，2024 年度「世界モデル　Deep Learning 応用講座」の講座の一環として行われました．このような場を提供してくださった松尾豊教授，及び講師の方々に深く感謝を申し上げます．

参考文献

・[Mohit,2020]Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté,Yonatan Bisk,Adam Trischler,Matthew Hausknecht:ALFWORLD: ALIGNING TEXT AND EMBODIED ENVIRONMENTS FOR INTERACTIVE LEARNING(2020)

・[Mobit,2020]Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, Dieter Fox:ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks(2020)

・[Pietro,2024]Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, Sai Rajeswar:GenRL: Multimodal-foundation world models for generalization in embodied agents(2024)

・[Michael,2022]Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, Andy Zeng:Do As I Can, Not As I Say: Grounding Language in Robotic Affordances(2022)

・[Andy,2022]Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, Pete Florence:Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language(2022)

・[Jessy,2023]Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, Anca Dragan:Learning to Model the World with Language(2023)

・[Yi,2024]Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, Limin Wang:InternVideo2:

Scaling Foundation Models for Multimodal Video Understanding(2024)