計測自動制御学会東北支部 第 358 回研究集会 (2026.5.25)
資料番号 358-2

視覚フィードバック統合制御に向けた視覚言語行動モデル GR00T の基礎性能評価

Fundamental Performance Evaluation of the Vision-Language-Action Model GR00T: Towards Integrated Visual Feedback Control

○石井創太 *，徳田冬樹 **，林部充宏 *

○ Ishii Souta*, Tokuda Fuyuki**, Hayashibe Mitsuhiro*

*東北大学大学院 工学研究科, **東北大学 未踏スケールデータアナリティクスセンター

*Graduate School of Engineering, Tohoku University,
**Unprecedented-scale Data Analytics Center, Tohoku University

キーワード： 模倣学習 (imitation learning)，視覚言語行動モデル (vision-language-action model)，ロボットマニピュレーション (robot manipulation)，ファインチューニング (fine-tuning)，視覚フィードバック (visual feedback)

連絡先： 〒 980-8579　仙台市青葉区荒巻字青葉 6-6-01 機械知能共同棟 503
東北大学大学院 工学研究科ロボティクス専攻 林部・Zhu/大脇研究室
石井創太，Tel.:022-795-6970, E-mail: ishii.souta.p1@dc.tohoku.ac.jp

1. はじめに

模倣学習は，専門家が実演した状態・行動系列から行動方策を学習し，ロボットにタスクを遂行させる学習枠組みである [1]．近年，深層学習の導入 [2]，Transformer を用いた複雑なタスクへの適用 [3]，大規模データに基づく基盤モデル化 [4] により，模倣学習は大きく発展している．

一方，実機環境における高精度な位置決めや接触を伴う操作では，実行中の誤差を逐次補正する仕組みが依然として重要である．特に，多様なロボット・タスクに対する汎用性と，視覚フィードバックに基づく高精度な制御を両立する枠組みは，なお十分に確立されていない．

本研究の目的は，模倣学習により獲得された汎用的な行動生成能力に，視覚フィードバックによる逐次補正を組み合わせることで，汎用性と高精度性を両立する制御手法を構築することである．その前段階として，本報告では視覚言語行動モデル GR00T[4] の基礎性能を評価し，今後の視覚フィードバック統合に向けた知見を得ることを目的とする．

2. 手法

2.1 GR00T による行動生成

NVIDIA Isaac GR00T[4] は，多様なロボットと多様なタスクから収集された大規模データを用いて事前学習された基盤モデルである．

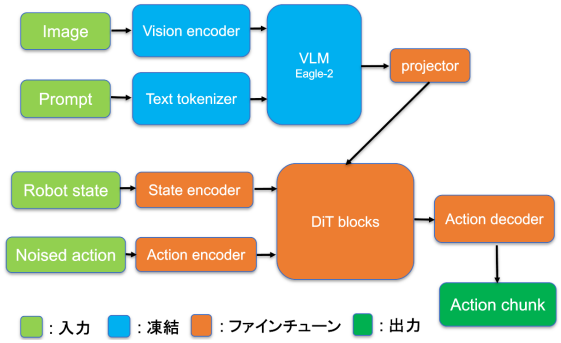

Fig. 1　GR00T のモデル構造 [4]．オレンジ色で示した部分は本報告でファインチューニングした部分であり，これ以外をファインチューニングすることも可能である．

GR00T は，動作指示を表すテキストと現在観測された画像を Vision Language Model (VLM) に入力し，得られた表現を Diffusion Transformer (DiT) に与えることで，言語指示と視覚状態の両方に基づく行動を生成する．具体的には，Eagle-2 VLM で画像・テキストを統合し，State Encoder でロボットの関節角度情報をエンコードして DiT Blocks に入力する．Action Decoder が DiT Blocks の出力をロボット指令値へ変換する．Fig. 1 にモデル構造の略図を示す．

GR00T は任意のロボット・タスクで利用可能であるが，より良いパフォーマンスを得るには，目的のロボット・タスクの下で収集したデータセットによりファインチューニングする必要がある．

2.2　基礎性能検証タスク

本報告では，6 軸マニピュレータにより赤い積み木を把持し，箱の中に入れるタスクで GR00T の基礎性能を評価する．GR00T に与える言語指示は “pick up the red cube and place it in the box” とした．

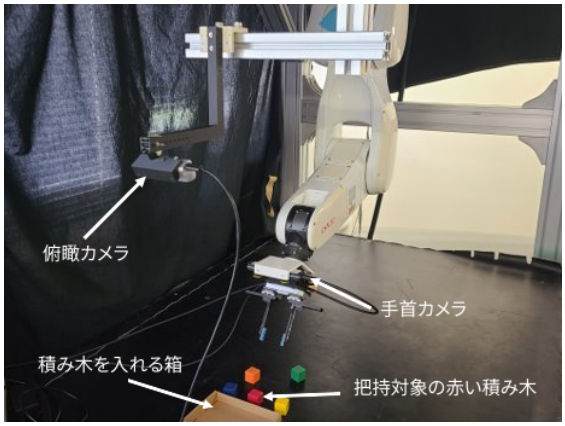

Fig. 2　データセット収集および評価実験の環境セットアップ．

2.3　データセット収集環境

データセット収集環境のセットアップを以下に示す．

- ロボット：6 軸垂直多関節ロボット，Denso VS068
- 俯瞰カメラ：720×1280 pixel, 30 fps, Intel RealSense D435i
- 手首カメラ：540×960 pixel, 30 fps, Stereolabs ZED Mini
- 入力デバイス：3D マウス（ロボット手先の遠隔操作），3Dconnexion SpaceMouse Compact

把持対象の赤い積み木は 2.5 cm × 2.5 cm，積み木を入れる箱は 16 cm × 11 cm である．Fig. 2 にデータセット収集・評価実験の環境を示す．

2.4　データセット収集方法

3D マウスを用いて遠隔操作を行うことにより，35 エピソード分のデータセットを収集した．各エピソードでは，俯瞰・手首カメラの映像とロボットの手先位置・姿勢を記録した．積み木の配置をパターン A，B，C と変え，それぞれ 10，10，15 エピソードずつデータを収集した．把持

対象の赤い積み木がロボットからみて右, 左, 中央にある配置がそれぞれパターン A, B, C である．パターン A, B, C それぞれにおける積み木の配置を Fig. 4 に示す．

2.5　ファインチューニング

収集した 35 エピソードのデータセットを用いて GR00T をファインチューニングした．事前学習済みモデルとして GR00T-N1.6-3B を用い，VLM のパラメータを固定したまま，State Encoder, Action Encorder, Action Decoder, DiT, Projector（VLM の出力を DiT が受け取れる形に変換する層）のみをファインチューニングした．Fig. 1 に凍結部分とファインチューニング部分をまとめた図を示す．

損失関数には GR00T[4] と同様にフローマッチング損失を用いた．時刻 t におけるアームの手先位置・姿勢とグリッパ状態からなる行動チャンク $\mathbf{A_t} = [\mathbf{a_t}, \mathbf{a_{t+1}}, \ldots, \mathbf{a_{t+H-1}}]$（$H = 16$）に対して, フローマッチングタイムステップ $\tau \in [0, 1]$ とサンプルノイズ $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ を用いてノイズ付き行動チャンク $\mathbf{A_t^\tau} = \tau \mathbf{A_t} + (1 - \tau)\boldsymbol{\epsilon}$ を生成し，モデル出力 $\mathbf{V}_\theta(\boldsymbol{\varphi_t}, \mathbf{A_t^\tau}, \mathbf{q_t})$ が脱ノイズベクトル場 $\boldsymbol{\epsilon} - \mathbf{A_t}$ を近似するよう，以下の損失を最小化する [4]．

$$\mathcal{L}_{\mathrm{fm}}(\theta) = \mathbb{E}_\tau \left[\left\| \mathbf{V}_\theta(\boldsymbol{\varphi_t}, \mathbf{A_t^\tau}, \mathbf{q_t}) - (\boldsymbol{\epsilon} - \mathbf{A_t}) \right\|^2 \right] \tag{1}$$

ここで $\boldsymbol{\varphi_t}$ は VLM の出力トークン，$\mathbf{q_t}$ はロボットの状態ベクトルである．推論時には $K = 4$ ステップのデノイジングにより行動チャンクを生成する．

主なハイパーパラメータを以下に示す．

- 最大学習ステップ数：2000
- グローバルバッチサイズ：32
- 学習率：1.0×10^{-4}（コサインスケジューラ，ウォームアップ比率 0.05）

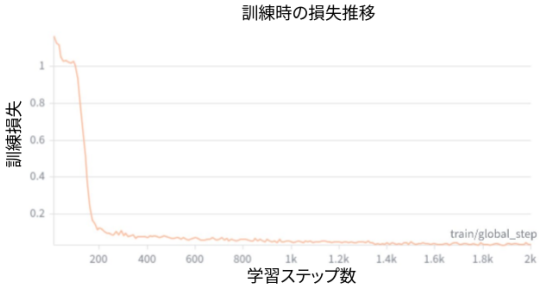

Fig. 3　GR00T ファインチューニング時の損失推移．

- オプティマイザ：AdamW（重み減衰 1.0×10^{-5}）
- 行動チャンク長：$H = 16$
- 推論デノイジングステップ数：$K = 4$
- 使用 GPU：NVIDIA RTX A6000（VRAM 49 GB）

Fig. 3 に訓練時の損失推移を示す．訓練損失は学習ステップの増加とともに速やかに収束し，2000 ステップ付近で安定した．本報告では 2000 ステップ学習後のモデルを用いる．

3.　実験

3.1　実験 1：データセットと同じ環境での評価

実験 1 では，データセットと同様に積み木の配置をパターン A，B，C とし，GR00T に推論させて実機での動作を評価した．パターン A，B，C のそれぞれについて 5 回ずつ試行を行った．Fig. 4 にデータセットの積み木配置と実験 1 の積み木配置の対応を示す．

実験結果として，積み木の配置がパターン A，B，C であるときいずれも GR00T の成功率は 4/5 (80%) であった．それぞれの失敗の原因を示す．パターン A では，グリッパの位置がずれたことにより失敗し，パターン B では，赤い積み木を離す動作のタイミングが早く，箱の中に積

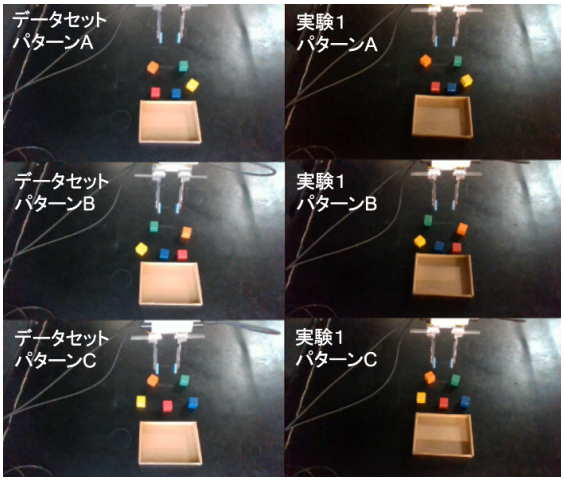

Fig. 4　データセットと実験 1 における積み木配置パターン A，B，C の比較（左：データセット，右：評価実験）.

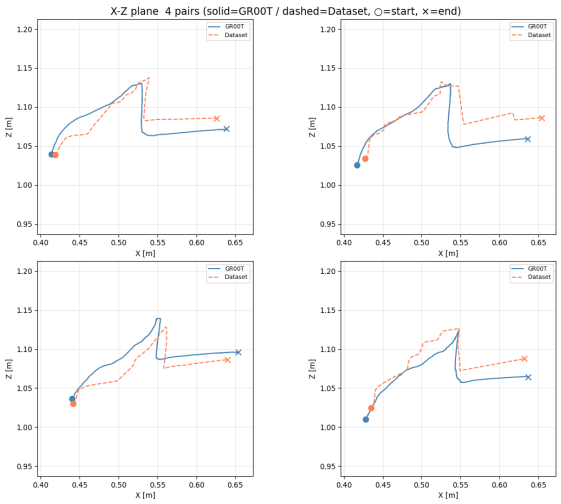

Fig. 5　実験 1, パターン A におけるアーム手先軌跡の比較（青実線：GR00T 出力，橙破線：データセット，○：開始点，×：終了点）．なお、成功動作のみをプロットした.

み木が入らなかったため失敗, パターン C では, グリッパが赤い積み木を把持できる位置に来たときに, グリッパが閉じられず失敗した．この実験により, GR00T はデータセットと同一の環境下においてロボットの動作生成が可能であることを確認した.

また，パターン A の条件下でアーム手先の軌跡を比較した結果を Fig. 5 に示す．GR00T の出力軌跡（実線）はデータセットの軌跡（破線）と完全には一致しないものの，動作の傾向は等しかった.

3.2　実験 2：データセットと異なる環境での評価

実験 2 では，データセットとは異なる積み木の配置（パターン D，E，F）の下で，GR00T に推論させて実機での動作を評価した．パターン D，E，F のそれぞれについて 5 回ずつ試行を行った．Fig. 6 にデータセットの配置（パターン A–C）と実験 2 の配置（パターン D–F）を示す.

実験結果として，パターン D では成功率 4/5 (80%)，パターン E，F では 5/5 (100%) であった．パターン D では, グリッパを閉じるタイミングが早く, 赤い積み木の把持に失敗した．この実験結果から, GR00T はデータセットと異なる積み木の配置においても，ロボットの動作生成が可能であることがわかった.

4.　考察

4.1　失敗の原因分析

実験 1 と実験 2 の失敗した試行の原因を分析すると，グリッパの開閉タイミングミスが 3 回, 手先位置のズレが 1 回であった.

グリッパの開閉タイミングについては, GR00T が出力するグリッパコマンドにばらつきがあることが原因であると考えられる．グリッパは開 (0)・閉 (1) の 2 値制御であるが，GR00T は連続値を出力するため，閾値をまたいだときに開閉を切り替えるように実装している．Fig. 7 に失敗試行におけるグリッパコマンドの時系列を示す．GR00T の出力するグリッパコマンドにばらつきが大きいことが確認できる．したがって，閾値等のパラメータ調整が重要である.

手先位置のズレについては，GR00T が出力する手先位置が目標位置とわずかにずれることで

Fig. 6　データセットと実験 2 における積み木配置の比較（左：データセットのパターン A–C，右：評価実験のパターン D–F）．

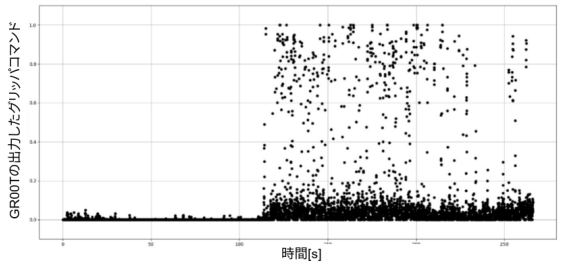

Fig. 7　グリッパ開閉失敗試行における GR00T のグリッパコマンド出力のを時系列順に並べた図の 1 例．

生じる失敗であり，視覚フィードバックを統合することでこのような失敗を低減できると考えられる．

5.　おわりに

本報告では，視覚フィードバックとの統合制御に向けた前段階として，視覚言語行動モデル GR00T の基礎性能評価を行った．35 エピソードのデータセットを用いたファインチューニングにより，6 軸マニピュレータを用いた積み木の把持・配置タスクにおいて，データセットと同一の環境および異なる配置環境のいずれにおいても 80%以上の成功率を達成した．

一方，グリッパの開閉タイミングの不安定さや手先位置のズレに起因する失敗も見られ，精密な動作の実現は依然として課題である．今後は力センサの導入によるグリッパ接触検知と，視覚フィードバックの統合によりこれらの課題に取り組む予定である．

6.　謝辞

本研究は, 公益財団法人電気通信普及財団 2025 年度研究調査助成の支援を受けて実施されました．

参考文献

1) P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal: Learning and Generalization of Motor Skills by Learning from Demonstration, *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 763–768 (2009)

2) T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel: Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation, *arXiv preprint arXiv:1710.04615* (2018)

3) T. Z. Zhao, V. Kumar, S. Levine, and C. Finn: Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, *arXiv preprint arXiv:2304.13705* (2023)

4) J. Bjorck, F. Castañeda, et al.: GR00T N1: An Open Foundation Model for Generalist Humanoid Robots, *arXiv preprint arXiv:2503.14734* (2025)