スケッチ指示を用いた VLA モデルによるロボット動作性能の向上に関する研究

TR24012 野田修平　　指導教授：藤吉 弘亘

1. はじめに

言語指示と視覚観測に基づいてロボットの行動を直接生成する Vision-Language-Action（VLA）モデルが注目されている [1]．VLA は，事前に獲得した世界知識を利用することで未知のタスクにおいても汎化可能である．しかし，軌道や回り込み方向，速度変化，停止位置などの動作指示に関する詳細を言語のみで正確に表現することは困難である．

このような言語指示の曖昧さを補うため，視覚的に意図を与えるスケッチ指示を用いる手法が提案されている．RT-Sketch[2] は，手描きのスケッチ指示を目標表現として用いることで，言語目標が曖昧な場合や視覚的外乱が存在する場合でも，空間的な意図を伝達できる可能性を示した．

そこで，本研究では VLA モデルに対してスケッチ指示を導入する．従来の言語指示に加えて，スケッチ指示を用いることで，言語指示による高い汎用性・認識能力を活かしつつ，動作の具体的な意図を補完し，より人の意図をくみ取ったロボット動作の実現と動作性能の向上を目指す．

2. Vision-Language-Action（VLA）モデル

VLA モデルは，視覚情報と言語指示から環境・タスクを理解し，ロボットの状態（関節角など）を条件として，関節角やグリッパなどの行動を直接出力する．これにより End-to-End な制御を実現する．また，汎用的な理解能力と高速な動作生成を両立するため，高レベルの解釈・推論と運動制御を分担させる 2 層構造（dual-system）の VLA も提案されている．

GR00T N1[3] は，視覚・言語モデルと拡散モデルからなる 2 層構造の VLA である． GR00T N1 の構造を図 1 に示す．視覚・言語モデル部分 (System2) では，環境・指示内容を解釈し，観測画像と言語指示をトークン列として入力することで，視覚言語特徴を抽出する．拡散モデル (System1) 部分では，ロボットの各関節角度などの状態情報を実機の構成に合わせた MLP で埋め込み，拡散過程で用いるノイズ付与済み行動と拡散時刻を Action Encoder で埋め込む．これらと視覚・言語モデルで得られた特徴量の cross-attention を求めることで，環境や言語指示を考慮した行動系列を生成する．System1 は 16 ステップ先までの行動を生成し，高頻度に更新することで滑らかな実機制御を可能にする．

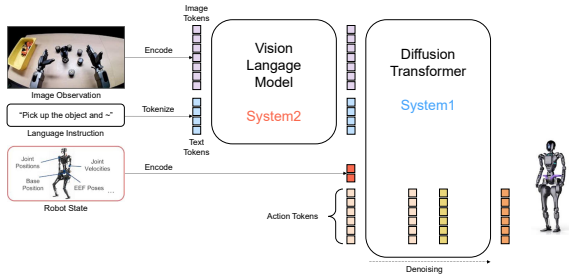

図 1: GR00T N1 のモデル構造

3. 提案手法

本研究では，GR00T N1 をベースとし，Diffusion Transformer にスケッチ指示を入力することで，意図した軌道・速度でのロボット動作を実現する．

3.1 スケッチ入力に対応した VLA モデル

モデル構造を図 2 に示す．視覚言語特徴とロボットの状態ベクトルに加えて，スケッチ指示を動作生成の条件情報として Diffusion Transformer へ入力する．これにより，従来の言語指示のみでは指定が難しい回り込み方向や通過経路などを条件情報として動作に直接反映する．

3.2 スケッチ指示

スケッチ指示は画像上の座標 (x, y) を記録する．その後，変化量（$\Delta x, \Delta y$）および 2 次微分 $(\Delta^2 x, \Delta^2 y)$ を求め，VLA の入力に用いる．これにより，軌道だけでなく動作速度についても反映することを実現する．Sketch Encoder は，スケッチ指示を小規模な MLP で埋め込み，ロボットの状態ベクトルと同様に条件トークンとして Diffusion Transformer へ入力する．

3.3 デモンストレーションデータによるファインチューニング

実機ロボットによるデモンストレーションデータを用いて GR00T N1 をタスクに適応させる．ファインチューニングでは，視覚情報・言語指示・ロボットの状態・スケッチ指示を条件として与え，VLM は固定したまま，条件情報の埋め込み・統合部および Diffusion Transformer を学習する．学習では，教師データである行動系列を生成するように，ノイズ予測誤差を最小化してパラメータを更新する．

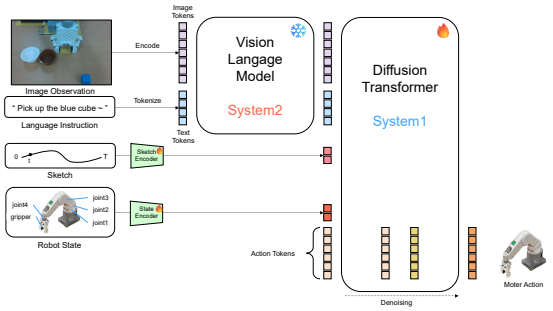

図 2: 提案手法のモデル構造

3.4 データセット作成

提案手法を学習・検証するため，RealSense D435 によって撮影した RGB 動画とロボットの動作ログを同期して記録し，各デモデータに対して言語指示とスケッチ指示を付与することでデータセットを作成する．ロボットは MyPalletizer 260-M5（4 軸＋グリッパ）を使用し，各関節角およびグリッパ開閉量を時刻情報付きで記録する．データセットの作成環境を図 3 に示す．

各デモデータは観測画像列（30fps）と状態・行動（関節角・グリッパ）から構成され，両者を対応付けることで，学習時に同一時刻の各データの参照を可能にする．タスクはピックアンドプレースとし，「青色のキューブを白色のカップに入れる」といった基本的な指示から，「青色のキューブを茶色のカップの前を通って手前側から白色のカップに入れる」などの複雑な指示まで 16 通り設定する．タスクを表す言語指示を各デモデータに付与し，さらに同一な配置に対しても，複数経路のスケッチ指示を作成する．また，1 つのモデルでスケッチ指示の有無の評価を行うため，スケッチ指示を含むデータと含まないデータの 2 種類を学習用に各 300 セット，評価用に 30 セット用意する．

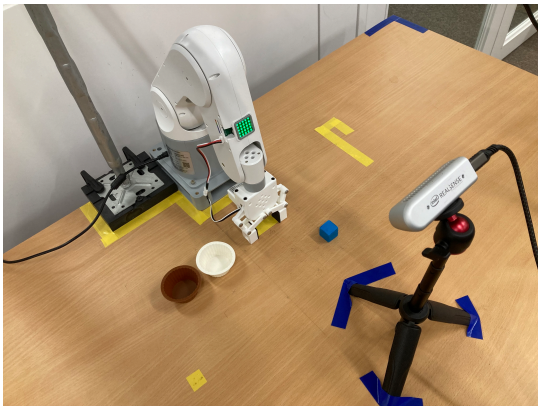

図 3: データセットの作成環境

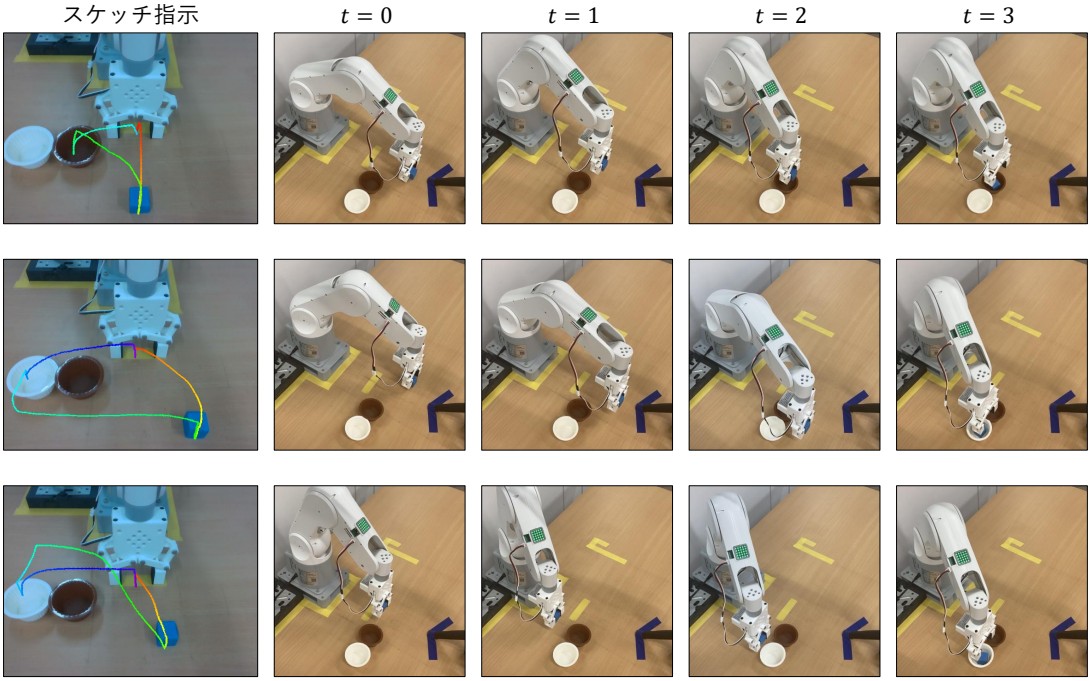

図 4: 定性的評価：スケッチ経路の妥当性

4. 評価実験

作成したデータセットを用いた提案手法の実験を行い，スケッチ指示を用いない場合と用いる場合で比較を行う．学習条件はバッチサイズ 4，学習ステップ 100000，最適化手法 AdamW，学習率 1e-4 とする．定量的評価として実機制御を行った際のタスク成功率を比較し，定性的評価として動作結果を観察することで，経路・速度の妥当性を確認する．

4.1 定量的評価

実機制御におけるタスク成功率を用いて，スケッチ指示の有無による性能差を比較する．ここでタスク成功とは，物体（青色キューブ）を把持し，指示された目標カップに投入できた場合である．カップに投入出来なかった場合や，目標カップに投入できた場合でも，他のカップへの接触や，指示された経路で動作しなかった場合はタスク失敗とする．タスク 1 は直線的に移動するシンプルな経路，タスク 2 は回り込む経路，タスク 3 ではより遠回りの経路や他のカップ位置も考慮した経路とする．各タスクについてスケッチ指示あり／なしの条件でそれぞれ 20 回ずつ検証する．

表 1: 実機実験におけるタスク成功回数

	タスク 1	タスク 2	タスク 3
スケッチあり	20/20	17/20	3/20
スケッチなし	16/20	8/20	0/20

表 1 より，全てのタスクにおいて，スケッチ指示を用いた場合の成功回数が向上した．これにより，スケッチ指示が軌道の意図（直線移動や回り込み方向）を明示し，スケッチ指示なしの場合よりも正確な動作を実現できていると言える．一方で，タスク 3 ではスケッチ指示を用いた場合でも成功回数が 20 回中 3 回のみであった．タスク 3 は，遠回り経路や他のカップ位置の考慮といった複数の制約を同時に満たす必要があり，タスク 1・2 と比較して要求される軌道の多様性が高い．このため，学習データにおけるタスク 3 のバリエーション不足や，スケッチ表現の分解能（点列密度・速度情報）不足により，モデルが安定して意図通りの回避・経路選択を生成できなかったと考えられる．

4.2 定性的評価

経路の差が分かりやすい設定としてタスク 4 を用意し，動作を観察することでスケッチ指示の有無による挙動の差を確認する．まず，スケッチなし条件では，把持から投入までの一連の動作において，目標へ向かう途中で手先が迷うように揺らぐ，直線的に接近して他のカップへ接触する，あるいはカップ手前で停止位置が定まらないといった挙動が観察された．特にタスク 4 のような回り込み動作では，回り込み方向の選択が安定せず，目標カップへ到達できない例が見られた．一方でスケッチあり条件では，移動方向や回り込み方向が明確となり，目標へ向かう経路が安定する傾向が確認できた．

次に，経路の妥当性について評価する．入力したスケッチ指示に対して，手先の移動経路が沿っているかを確認する．図 4 に，可視化したスケッチと，物体把持後からゴールまでの実機の経路の対応例を示す．この結果から，スケッチ指示に沿う経路で動作する様子が確認できた．また，近い経路でスケッチ指示の密度（150 ステップと 300 ステップ）を変えて入力した場合の動作速度の変化も確認できた．

5. おわりに

本研究では，言語指示に基づく VLA モデルにスケッチ指示を時系列の条件情報として入力し，軌道・速度といった具体的な動作意図を行動生成へ反映する手法を提案した．評価では，オフライン指標（MSE）において提案手法の誤差が増加した一方，実機ではスケッチ指示に沿う経路で動作する傾向や，スケッチ指示の点列密度の違いに応じた速度変化が確認できた．今後は，スケッチ指示パターンごとのデータ不足を解消するためのデータ拡充，タスク追加を行い，より高い汎化性能の獲得を実現する．また，別の VLA モデルやロボット実機を使った実験を行い，更なる動作性能の向上を目指す．

参考文献

[1] B. Zitkovich, et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control”, CoRL, 2023.

[2] P. Sundaresan, et al., “RT-Sketch: Goal-Conditioned Imitation Learning from Hand-Drawn Sketches”, CoRL, 2024.

[3] J. Bjorck, et al., “GR00T N1: An Open Foundation Model for Generalist Humanoid Robots”, arXiv preprint, arXiv:2503.14734, 2025.

研究業績

野田修平，平川翼，山下隆義，藤吉弘亘，“Transformer モデルを用いたスケッチ指示による把持位置推定”，日本ロボット学会学術講演会，2024.