オフライン軌道生成による軌道に基づく open-vocabulary 物体操作タスクにおける将来成否予測

○神原元就，杉浦孔明（慶應義塾大学）

open-vocabulary 物体操作タスクにおいて，実行前に失敗を予測し，適切な軌道を再生成することができれば，安全かつ効率的なタスクの実行が可能である．しかし，生成された軌道に基づいてタスクの成功や失敗を事前に予測する手法はほとんど存在しない．そこで本研究では，操作前の一人称視点画像，物体操作指示文，及び軌道に基づいて，物体操作タスクの将来的な成功または失敗を予測する方法を提案する．また，エンドエフェクタの軌道から効果的な特徴を抽出し，画像特徴とアラインメントを行うための Trajectory Encoder を導入する．

1. はじめに

マニピュレータによる正確な物体操作の実行は，ロボットが家事，清掃，農業，及び救助活動などの多様な分野において，高い汎用性と適応性を実現するために極めて重要である．特に，open-vocabbulary 物体操作タスクを実行可能であれば，自然言語によるマニピュレータへの指示が可能であり，利便性が高い．この物体操作において，タスクの実行時，タスクの成否について自動で判定できることはタスク実行の効率化につながる．タスクの成否判定に関する多くの既存手法は，タスクの実行後に成否判定を行っている．一方で，タスクの実行前にタスクの成否を予測できれば，より効率的なタスク実行が可能となる．

本研究は，open-vocabulary 物体操作における物体操作成否予測に着目する．このタスクは，オフライン軌道生成に基づくエンドエフェクタの軌道，物体操作実行前の一人称画像，及び自然言語指示文に基づき物体操作の将来的な成否を予測するタスクである．これは，生成された軌道に基づき，将来的な物体及びエンドエフェクタのインタラクション及びそれらと自然言語指示文とのアラインメントを考慮することが求められるため困難である．タスクの代表例を図 1 に示す．この例では，“pick the apple from the white bowl” という指示文が与えられている．マニピュレータは適切にりんごを白いボウルから把持することができたため，モデルは，物体操作に成功した，と予測することが望ましい．

本研究では，軌道，画像，及び指示文に基づき将来的なタスクの成否を予測する手法を提案する．本手法は，λ-Repformer [1] を拡張し，エンドエフェクタの軌道に基づき将来的な物体操作成否予測を可能にした手法である．これにより，画像および軌道から将来的な物体及びエンドエフェクタ間のインタラクションを考慮できることが期待される．

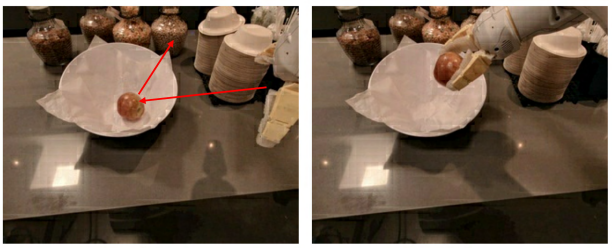
本研究の独自性は以下である．

- エンドエフェクタの軌道から効果的な特徴量を抽出し，画像特徴量とのアラインメントを行うための Trajectry Encoder を導入する．

2. 関連研究

open-vocabulary 物体操作タスクにおいて，マニピュレータの軌道生成の方法として，オンライン軌道生成及びオフライン軌道生成が存在する．オンライン軌道生成手法としては，Transformer ベースモデルや MLLM ベースモデルが挙げられる [2–4]．また，オフライン軌道生成を用いたアプローチとして [5, 6] が挙げられる．

提案手法は long-horaizon タスクにおけるサブタスク計画手法とも関連が深い．いくつかの手法では，サブタスク実行後にタスク成否を判定し，それに基づいたサブタスクの再計画を行う [7–9]．これらの手法では，状況に基づいてタスクの成否判定を行うという点で提案手法と類似している．一方で，提案手法はこれらの手法と異なり，タスク実行前にタスクの成否を予測する．これにより，提案手法を用いた場合，より効率的なタスク実行が可能となる．



(a) (b)

“pick the apple from the white bowl”

図 1: タスク例．図中 (a) 及び (b) は物体操作前後の画像を表す．また，(a) における赤い矢印はエンドエフェクタの動きを示す．

3. 問題設定

本研究は，オフライン軌道生成による軌道に基づくタスク成否予測 (Path-based Task Success Prediction; PTSP) を対象とする．PTSP タスクにおいて，入力は時刻 0 における一人称画像，エンドエフェクタの軌道，及び自然言語指示文とする．また，出力はタスクの予測確率 $P(\hat{y}=1)$ であり，これは指示文に基づく物体操作が適切に実行された確率の予測値を示す．ただし，$\hat{y}$ は物体操作の成否を示す予測ラベルであり，成功を ‘1’ とする．

4. 提案手法

図 2 に提案手法の概要図を示す．提案手法は，主に Trajectory Encoder 及び λ-Representation Encoder によって構成される．モデルへの入力は $\boldsymbol{x} = \{\boldsymbol{x}_{txt}, \boldsymbol{x}_{img}, \boldsymbol{x}_{traj}\}$ とする．ただし，$\boldsymbol{x}_{txt}$, $\boldsymbol{x}_{img}$, 及び $\boldsymbol{x}_{traj} = \boldsymbol{a}_{t\,t=0}^{T}$ はそれぞれ物体操作前の一人称画像，自然言語指示文，及びエンドエフェクタの軌道を示す．ここで，$\boldsymbol{a}_t$ は時刻 t におけるエンドエフェクタの姿勢を示す．提案手法では，まず $\boldsymbol{x}_{txt}$ から言語特徴量 $\boldsymbol{h}_{\mathrm{txt}}$ を獲得する．

4.1 Trajectory Encoder

本手法では，エンドエフェクタの軌道に関する特徴量を抽出するための Trajectory Encoder を導入する．このモジュールは，系列データの一つである脳波に基

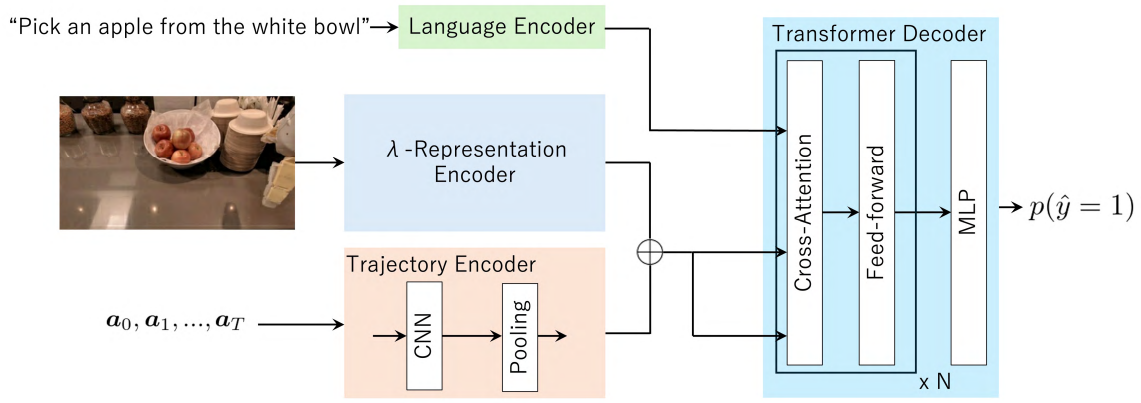


図 2: 提案手法のネットワーク図．図中，CNN 及び MLP は畳み込み層及び線形変換層を表す．

づく分類タスクにおいて良好な結果が報告されている EEG Conformer [10] を参考としている．

本モジュールへの入力は $\boldsymbol{x}_{traj}$ であり，出力は軌道に関する中間特徴量 $\boldsymbol{h}_{\text{traj}}$ である．本モジュールは畳み込み層及びプーリング層から構成される．具体的な処理としては，まず畳み込み層を用いて，$\boldsymbol{x}_{traj}$ をチャネル方向に畳み込む．続いて，プーリング層を用いて，時間方向に畳み込むことで，$\boldsymbol{h}_{\text{traj}}$ を獲得する．

4.2 λ-Representation Encoder

本タスクでは，入力画像について，各物体の位置に関する情報を理解し，軌道が適切な物体についてインタラクションするものであるか否かを予測することが求められる．そのため，入力画像から抽出する視覚特徴量について，物体の詳細な特徴のみでなく，物体同士の空間的な情報を含めた視覚特徴量であることが望ましい．λ representation [1] は，そうした視覚特徴量として適切なものの一つと考えられるため，利用する．本モジュールでは，視覚特徴量の一つである λ representation $\boldsymbol{h}_\lambda$ を抽出する．本モジュールの入力は $\boldsymbol{x}_{img}$ である．また，出力は $\boldsymbol{h}_\lambda$ である．本モジュールにおける処理は，齋藤ら [1] の手法における λ-Representation Encoder モジュールでの処理を参考とする．ただし，Narrative Representation に関して，GPT-4o [11] を用いて生成した説明文を利用して抽出した．得られた $\boldsymbol{h}_\lambda$ に関して，$\boldsymbol{h}_{\text{traj}}$ と結合した後に，$\boldsymbol{h}_{\text{txt}}$ との cross-attention を計算する N 層の Transformer Decoder によって，最終的に物体操作の成否についての予測確率 $P(\hat{y} = 1)$ を出力する．

5. 実験設定

実験では，open-vocabulary 物体操作タスクにおける大規模標準データセットである RT-1 データセット [3] に基づき構築したデータセットを用いた．PTSP タスクでは，物体操作前の画像，エンドエフェクタの軌道，自然言語指示文及びタスクの成否のラベルが含まれたデータセットが必要である．RT-1 データセットは実機による open-vocabulary 物体操作タスクに関する代表的なデータセットであり，PTSP タスクで用いられる上記のデータを含む．それゆえに，実験では RT-1 データセットからエピソードを抽出して構築したデータセットを用いた．一方で，RT-1 データセットに含まれるサンプルに，ラベル誤りであるものが複数存在した．具体的には，物体操作に成功しているにも関わらず Failure ラベルが付与されているものがあった．それらのサンプルについては，指示文をランダムに付与することでネガティブサンプルとした．構築したデータセットには，合計 13,915 サンプルが含まれていた．これらについて，11,915 サンプル，1,000 サンプル，及び 1,000 サンプルに分割し，それぞれ訓練集合，検証集合，及びテスト集合とした．SP-RT-1 データセットにおいて，データ収集に用いられたマニピュレータのアームは 7 次元の行動空間を持つ．そのため，$\boldsymbol{a}_t$ は，これにエンドエフェクタの開閉に関する 1 次元を含めた 8 次元のベクトルであった．

実験は，24GB の VRAM を持つ NVIDIA GeForce RTX 4090, 64GB RAM, 及び Intel Core i9-13900KF を用いて行った．提案手法の訓練時間及び特徴量抽出を除く推論時間はそれぞれ約 1.5 時間および 0.78ms/サンプルであった．

手法	精度 [%]
齋藤ら [1]	74.9 ± 0.79
提案手法	**83.4** ± 0.65

表 1: 比較実験における定量的結果

6. 実験結果

6.1 定量的結果

表 1 に定量的結果を示す．実験は 5 回行い，表にはその平均値及び標準偏差を示した．実験において，ベースライン手法を齋藤らの手法 [1] とした．この手法は，PTSP タスクと関連が深い SPOM タスク [1] において優れた結果を示した手法であるため採用した．一方で，SPOM タスクでは入力として物体操作前後の画像及び自然言語指示文を用いていたが，実験にあたり，齋藤らの手法の入力としては物体操作前の画像及び自然言語指示文のみを用いた．また，評価尺度として精度を用いた．結果より，ベースライン手法および提案手法の精度はそれぞれ 74.9%及び 83.4%であった．このことから，提案手法はベースライン手法を 8.5 ポイント上回った．これより，提案手法が軌道を適切に考慮し将来的な物体操作の成否を予測できたことがわかった．

6.2 Ablation study

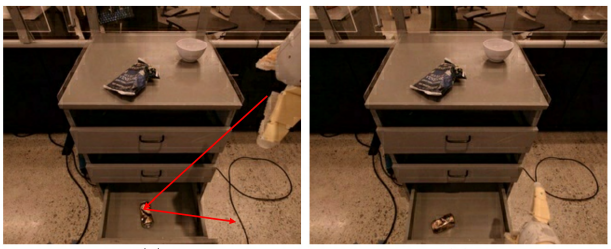
Ablation study として，Trajectory Encoder を取り除くことによる性能への影響を調査した．表 2 に 5 回の実験の平均と標準偏差を示す．表において，モデル (i) は Trajectory Encoder を線形関数に置き換えたものを示す．表より，モデル (i) 及びモデル (ii) の精度はそれぞれ 83.2%及び 83.4%であった．このことから，Trajectory Encoder を用いた場合の方が 0.2 ポイント高いという

モデル	条件	精度 [%]
(i)	提案手法 w/o Traj. Enc.	83.2 ± 0.48
(ii)	提案手法	**83.4** ± 0.65

表 2: Ablation study の定量的結果

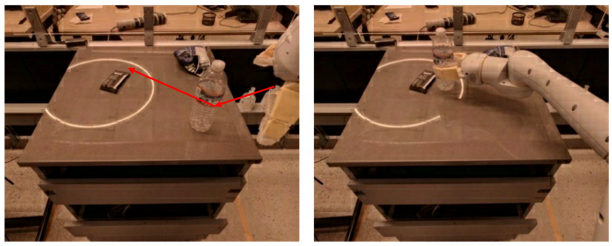
(a) (b)

(i) “pick rxbar blueberry”



(a) (b)

(ii) “pick orange can from bottom drawer and place on counter”



(a) (b)

(iii) “move water bottle near rxbar chocolate”

図 3: 定性的結果．各例において，(a) 及び (b) は物体操作前後の画像をそれぞれ表す．また，(a) における赤い矢印はエンドエフェクタの動きを示す．

ことが分かった．これより，Trajectory Encoder によって，軌道から効果的な特徴量を抽出し物体操作の成否に利用できていたと言える．

6.3 定性的結果

図 3 に定性的結果を示す．この図において，(i), (ii), 及び (iii) はそれぞれ True Positive, True Negative, 及び False Negative の例を示す． 図 3(i) において，与えられた指示文は “pick rxbar blueberry” であり，マニピュレータは rxbar を適切に把持したため，このエピソードにおけるラベルは Success であった．提案手法はこれについて適切に Success であると予測した一方で，ベースライン手法は Fail と誤った予測を行った．

また，図 3(ii) に示されたエピソードでは，“pick orange can from bottom drawer and place on counter” という指示文が与えられた．一方で，マニピュレータは缶を把持することに失敗し，引き出しの中から取り出すことができなかった．したがって，このエピソードのラベルは Fail であった．提案手法はこのエピソードに関しても適切に Fail と予測できた．ベースライン手法は，誤って Success と予測した．これらのことから，提案手法はこれらのエピソードに関して，エンドエフェクタの軌道及び画像内における物体の位置から，物体操作の成否について適切に予測することができたと言える．

一方で，図 3(iii) に提案手法が予測を誤った例を示した．このエピソードでは，指示文は “move water bottle near rxbar chocolate” であった．マニピュレータは適切に rxbar をペットボトルに近づけたため，ラベルは Success であった．一方で，提案手法は Fail と予測した．このエピソードでは複数物体の空間的な位置を適切に把握する必要があり，さらに軌道とそれらの位置をマッピングする必要があったため，困難であったと考えられる．

7. おわりに

本研究では，open-vocabulary 物体操作において，自然言語指示文，物体操作前の一人称視点画像，及びオフライン生成されたエンドエフェクタの軌道に基づき，物体操作の将来的な成否を予測するタスクを扱った．また，入力された軌道について効果的な特徴量を抽出するためのモジュールである Trajectory Encoder を導入した．結果として，ベースライン手法を上回るタスク予測精度を達成した．

謝辞

本研究の一部は，JSPS 科研費 23K28168， JST ムーンショット， NEDO， JSPS 特別研究員奨励費 JP23KJ1917 の助成を受けて実施されたものである．

参 考 文 献

[1] 齋藤大地，神原元就，九曜克之，杉浦孔明，“マルチモーダル LLM および視覚言語基盤モデルに基づく大規模物体操作データセットにおけるタスク成功判定，” 第 38 回人工知能学会全国大会資料，pp.3O1OS16b02–3O1OS16b02，2024.

[2] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, et al., “VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models,” CoRL, pp.540–562, 2023.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, et al., “RT-1: Robotics Transformer for Real-World Control at Scale,” arXiv preprint arXiv:2212.06817, 2022.

[4] M. Shridhar, L. Manuelli, and D. Fox, “CLIPort: What and Where Pathways for Robotic Manipulation,” CoRL, pp.894–906, 2022.

[5] R. Korekata, et al., “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, pp.3865–3872, 2023.

[6] M. Kambara and K. Sugiura, “Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks,” ICIP, pp.2601–2605, 2022.

[7] M. Shirasaka, et al., “Self-Recovery Prompting: Promptable General Purpose Service Robot System with Foundation Models and Self-Recovery,” ICRA, 2024.

[8] D. Driess, F. Xia, M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, et al., “PaLM-E: An Embodied Multimodal Language Model,” ICML, vol.202, pp.8469–8488, 2023.

[9] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, et al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances,” CoRL, pp.287–318, 2023.

[10] Y. Song, Q. Zheng, B. Liu, and X. Gao, “EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization,” IEEE TNSRE, vol.31, pp.710–719, 2023.

[11] J. Achiam, S. Adler, S. Agarwal, et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.