TSUBAME 共同利用　令和6年度　学術利用　成果報告書

利用課題名　大規模マルチモーダルモデルの研究開発および応用
英文：Research, Development, and Application of Large-Scale Multimodal Models

利用課題責任者
栗田修平
First name Surname
Shuhei　Kurita

所属
国立情報学研究所
Affiliation
National Institute of Informatics
URL
https://www.nii.ac.jp/

邦文抄録（300 字程度）

本研究課題では、実世界にて言語指示を理解し環境に応じて動作を生成する、実世界での基盤モデルを目指して研究を行った。具体的には、一人称視点動画を題材にし、言語・視覚を入力として動作系列を抽出およびそれを予測する基盤モデルの作成をおこなった。言語（テキスト）から、あらかじめ抽出された動作系列を、事前学習済みモデルの 3D 言語モデル（3D-LLM）を利用して学習および生成させることで、擬似的な動作生成が可能になった。学習は計算環境 TSUBAME を利用して行われた。得られた成果は、画像系のトップカンファレンスである CVPR2025 にて発表を行った。得られたモデル及びデータセットは、将来的なロボット基盤モデルの作成にも応用できる成果であると考えられる。

英文抄録（100 words 程度）

This research project aimed to develop a foundational model for the real world that understands language instructions and generates actions based on the environment. Specifically, using first-person perspective videos as subject matter, we created a foundational model that extracts and predicts action sequences using language and visual input. By training and generating pre-extracted action sequences from language (text) using a pre-trained 3D language model (3D-LLM), we achieved pseudo-action generation. Training was conducted using the TSUBAME computing environment. The results were presented at CVPR 2025, a top conference in the field of computer vision. The obtained model and dataset can be applied to the creation of future robot base models.

Keywords: Vision and language model, Large Language Model

背景と目的

（課題の背景を記載してください）

ロボット基盤モデル作成にあたって一番の障害となるのは、データ不足の問題である。人手で作られたテレオペデータは、非常に高品質な動作データであるが、公開されているものは非常に少ない。一方で、比較的簡単に収集できる一人称視点動画から動作データを作成できれば、ロボット基盤モデルを作成するうえで強力な手法となりえる。

（現状の問題点等を挙げてください）

本プロジェクトでは、視覚情報（環境情報）・言語指示情報と動作情報を統合して扱うためのデータセットを作成し、作成したデータセットを利用して 3D-LLM をベースとする動作生成モデルを作成する。ロボット基盤モデルのデータ欠乏問題を、一人称視点動画からのデータ抽出によって解決し、既存のロボット基盤モデル学習用データよりも多様な視覚・動作・言語の対応付けデータの作成、および、3D 言語モデル（3D-LLM）を利用して学習および生成させることで、擬似的な動作生成モデルを作成した。得られた成果は、画像系のトップカンファレンスである CVPR2025 にて発表を行った。得られたモデル及びデータセットは、将来的なロボット基盤モデルの作成にも応用できる成果であると考えられる。

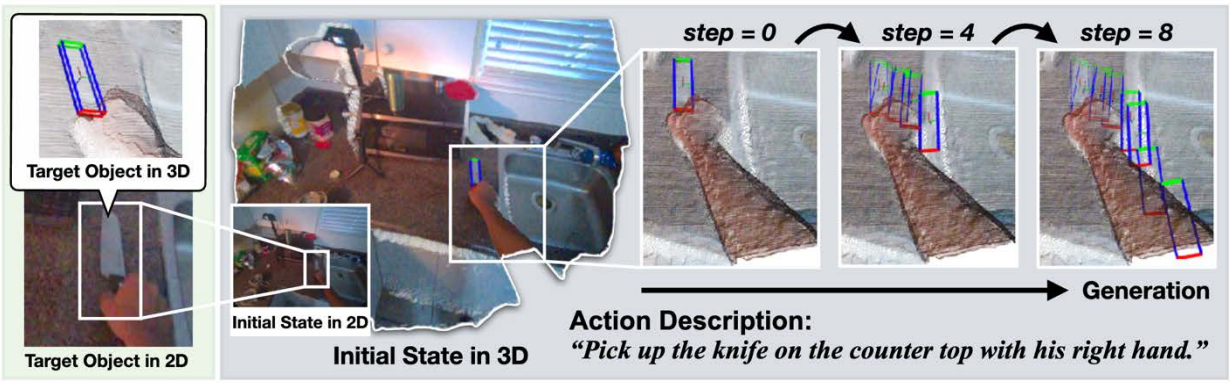

図 1. 今回の研究にて整備した時系列動作データセットの概要.

概要

（申請書の概要を基本とし、実情に合わせて変更してください。図や表の利用を図って分かり易く記載して下さい。）。

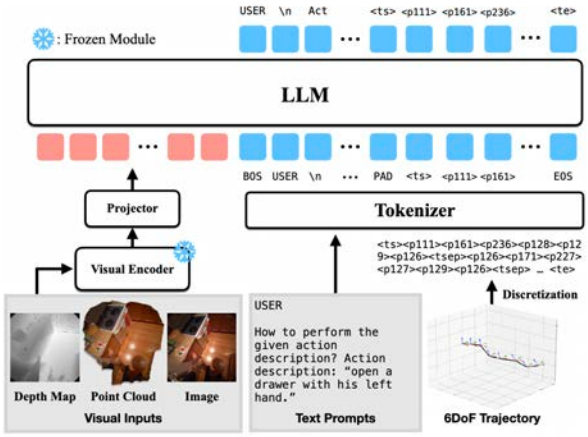

図. 2　3D-LLM を利用した提案手法(PointLLM)

生成したデータセットの概要を図 1 に示す。一人称視点動画が与えられた際に、3D 再構成を行ったうえで、物体の姿勢を 3D bounding box の形式で捉え、その位置および姿勢の変化を時系列的に捉える。本研究では、さらに 3D-LLM を利用した動作生成モデルを作成し、視覚情報に加えて、3D 的な環境情報が有効であることを示した。図 2 のように 3D-LLM を利用して学習を行い、提案したデータセットを利用することで有効な動作軌跡が生成できるか検証した。

結果および考察

（結果と考察を記載してください。図や表の利用を図って分かり易く記載して下さい。）。

表 1 に実験結果を示す。物体位置のみの予測(3DoF)と角度も含めた予測（6DoF）の双方にて、提案手法が優れていることを確認した。さらに、定性的な評価によっても、提案手法、ひいては作成したデータセットが優れていることを確認した。これは、今回の研究で作成したデータが優れていることに加えて、ロボット基盤モデルへの応用可能性を示すものと思われる。

表 1: 3D-LLM(PointLLM)を利用した実験結果

		3DoF Training				*6DoF Training*				
		3D pos.		*2D pos.*		*3D pos.*		*2D pos.*		*3D rot.*
Model	Input	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	GD
Seq2Seq [89]	Image	**0.269**	0.475	0.215	0.372	0.374	0.625	0.325	0.528	0.559
BLIP-2 (2.7B) [50]	Image	0.278	0.464	0.218	0.346	0.280	0.465	0.218	0.344	0.545
BLIP-2 (6.7B) [50]	Image	0.286	0.487	0.224	0.365	0.286	0.475	0.219	0.349	0.543
VILA (3B) [54]	Image	0.500	0.662	0.428	0.546	0.477	0.619	0.390	0.489	0.826
VILA (8B) [54]	Image	0.316	0.512	0.249	0.388	0.293	0.478	0.225	0.347	0.545
Seq2Seq [89]	Image + Depth	0.302	0.541	0.250	0.438	0.341	0.558	0.280	0.452	0.590
BLIP-2 (2.7B) [50]	Image + Depth	0.288	0.481	0.227	0.363	0.275	**0.458**	0.211	0.332	0.543
BLIP-2 (6.7B) [50]	Image + Depth	0.283	0.477	0.218	0.351	0.282	0.469	0.219	0.345	0.542
VILA (3B) [54]	Image + Depth	0.301	0.492	0.229	0.355	0.294	0.480	0.223	0.344	**0.541**
VILA (8B) [54]	Image + Depth	0.298	0.494	0.234	0.368	0.318	0.513	0.253	0.391	0.563
MiniGPT-3D (2.7B) [85]	Point cloud	0.299	0.487	0.236	0.368	0.281	0.467	0.218	0.342	0.544
PointLLM (7B) [97]	Point cloud	0.274	**0.459**	**0.210**	**0.328**	**0.271**	**0.458**	**0.208**	**0.327**	**0.541**

まとめ、今後の課題
（まとめと今後の課題について記載してください。）。

一人称視点の動画を題材に、言語と視覚を入力として動作系列を抽出・予測する基盤モデルを構築した。抽出された動作系列をテキストに基づき、事前学習済みの 3D 言語モデル（3D-LLM）を活用して学習・生成させることで、擬似的な動作生成を実現した。学習は計算環境 TSUBAME 上で行い、その成果は画像分野のトップ国際会議 CVPR2025 にて発表された。本研究で得られたモデルとデータセットは、今後のロボット基盤モデルの開発にも応用可能であると考えられる。特に、今回構築したデータセットの質の高さに加え、ロボット基盤モデルへの発展可能性を示す成果となっている。この成果のロボット実験への応用可能性について、追加の検討を進めている。