自然言語・画像情報を活用したロボットアームの階層型バイラテラル制御に基づく模倣学習

小林聖人 [1,2]　Thanpimon Buamanee[1]　浦西友樹 [1]
[1] 大阪大学 [2] 神戸大学
{kobayashi.masato.cmc,yuki.uranishi.cmc}@osaka-u.ac.jp
buamanee.thanpimon@ist.osaka-u.ac.jp

概要

多段階からなる長期タスクのロボット操作では，進行状況に応じたサブゴール（言語指示）の選択と，言語指示条件に適したロボット動作生成を同時に実現することが重要な課題である．本研究では，位置・力制御に基づくバイラテラル制御と自然言語と視覚情報を統合した階層型模倣学習手法（階層型 Bi-VLA）を提案する．提案法は，時系列画像から言語サブゴールを生成する High-Policy と，言語条件付き視覚特徴と位置・力情報からロボットの行動を生成する Low-Policy で構成される．実機の長期タスクにおいて既存法を上回る成功率を達成し，階層化と言語・視覚統合が位置・力制御を伴う長期タスクの安定実行に有効であることを示した．

1　はじめに

バイラテラル制御は，操作者がリーダーロボットを操作しながらフォロワーロボットの環境反力を触覚的に感じ取ることができるため，位置情報と力情報の両方を扱えるデータ収集手法として知られている [1, 2]．特に，バイラテラル制御に基づく模倣学習の Bi-ACT（Bilateral Control-Based Imitation Learning via Action Chunking with Transformer）は，Transformer を用いた Action Chunking とバイラテラル制御を組み合わせることで，対象物に応じた把持力の調整を可能とし，タスクの失敗を軽減した [3]．しかし，Bi-ACT は言語情報を活用していないため，長期タスクにおける適切な動作選択が困難であり，複雑な多段階タスクへの適応に問題があった．

本研究では，この問題を克服するため，言語と視覚情報を活用した階層型バイラテラル制御に基づく模倣学習手法を提案する（図 1）．本研究の目的は，長期タスクにおいて「進行状況に応じたサブゴール（言語指示）の更新」と「言語指示に整合した動作生成」を両立する設計指針を与えることである．本論文の貢献は以下の 3 点である：

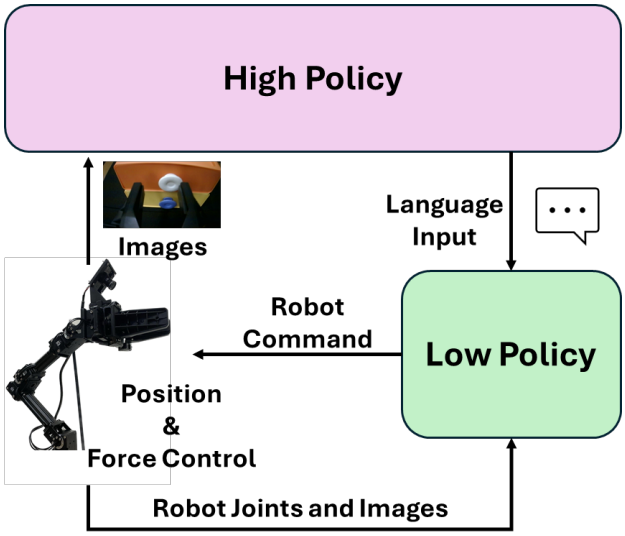

図 1　階層型バイラテラル制御に基づく模倣学習の概要

- 言語サブゴールに基づく階層化：時系列画像からサブゴール文を生成する High-Policy と，その指示に基づき動作を生成する Low-Policy を分離した階層型バイラテラル制御に基づく模倣学習を提案した．
- 言語条件付き視覚統合：Low-Policy において FiLM により言語埋め込みに応じて視覚特徴を動的に変調し，意図に沿った動作生成を実現した．
- 実機長期タスクでの有効性：実機タスクにおいて既存法 Bi-ACT より高い成功率を示し，長期タスクでの有効性を確認した．

2　関連研究

近年，模倣学習の汎化性やロバスト性を高めるために，言語情報をモデルに組み込む試みが盛んに行

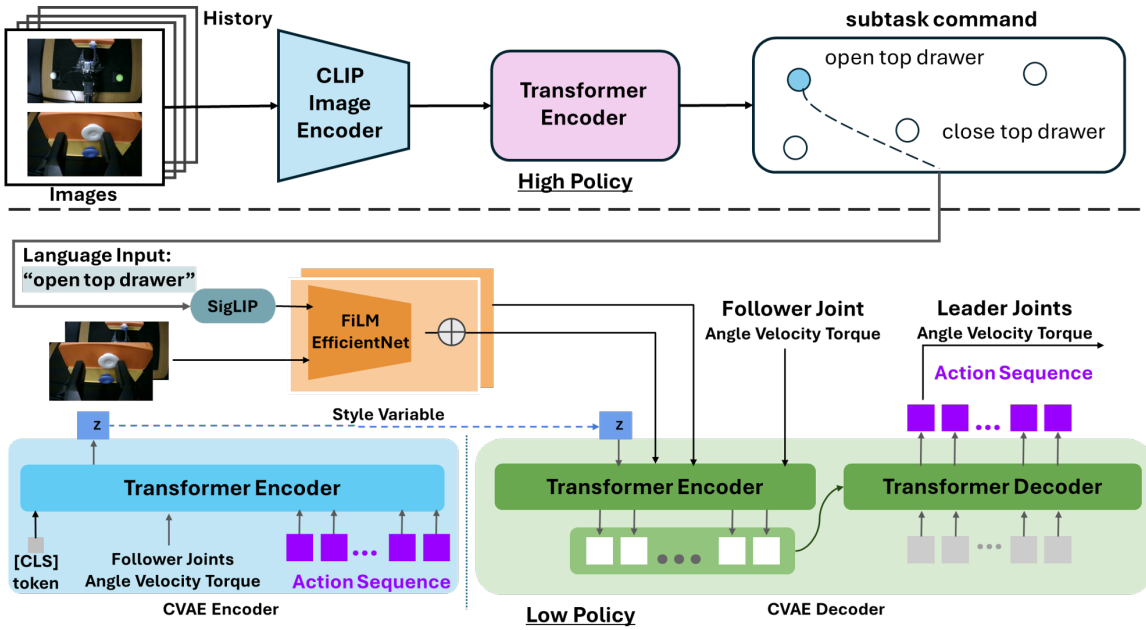

図 2　自然言語・画像情報を活用した階層型バイラテラル制御に基づく模倣学習のアーキテクチャ

われている [4, 5, 6, 7]．言語情報を導入することで，モデルはタスクの目的や状態を抽象的に表現できるようになり，より柔軟に複数タスクへ対応可能となることが示されている [8]．

バイラテラル制御に基づく模倣学習手法として，Bi-ACT[3] は，Transformer を用いた Action Chunking[9] とバイラテラル制御を組み合わせることで，対象物に応じた把持力の調整を実現した．しかし，Bi-ACT は言語情報を活用していないため，タスクの進行状況に応じた動作選択が必要となる長期タスクや多段階タスクへの適用は困難であった．

この課題に対し，Bi-LAT[5] は，Bi-ACT に言語情報を直接入力として追加することで，言語指示に基づいた力加減の調整を可能とした．一方で，Bi-LAT は単一タスク内での言語条件付けに主眼を置いており，複数サブタスクから構成される長期タスクや，タスク進行に応じた動作切り替えには対応していない．さらに，言語情報に応じて視覚特徴を動的に調整することで複数タスクへの対応を目指した Bi-VLA[10] も提案されているが，比較的単純なタスクでの有効性にとどまり，長期タスクにおけるロバストな動作生成には課題が残されている．

以上より，既存の言語統合型バイラテラル模倣学習手法では，長期タスクにおいて必要となる「タスク進行の把握」と「サブタスク間の適切な切り替え」を同時に実現することが困難である．そこで本研究では，タスクの高次な意図と詳細な動作制御を分離する階層型アーキテクチャを導入し，長期タスクにおける適切な動作生成を可能とする階層型 Bi-VLA を提案する．

3　手法

3.1　提案手法の概要

図 2 に提案手法のアーキテクチャを示す．提案手法は，High-Policy と Low-Policy からなる 2 層の階層型アーキテクチャにより構成される．

High-Policy は，時系列の画像観測を入力として受け取り，タスクの進行状況を理解して適切な自然言語コマンドをサブゴールとして生成する．具体的には，複数フレームの画像を CLIP[11] Image Encoder でエンコードし，Transformer Encoder で時系列の視覚特徴を処理する．処理された特徴から，最も類似度の高い自然言語コマンドを予測し，Low-Policy に伝達される．

Low-Policy は，High-Policy から与えられた自然言語コマンド，現在の画像観測，関節情報（角度，角速度，トルク）を統合し，ロボット動作を生成する．自然言語コマンドは SigLIP[12] により埋め込みベクトルに変換され，画像は EfficientNet[13] により視覚特徴が抽出される．FiLM（Feature-wise Linear Modulation）[14] により，言語埋め込みに基づいて視覚特徴が動的に変調される．変調された視覚特徴，関節情報，言語情報が Transformer Encoder に入

力され，さらに Transformer Decoder によりのアクション列が予測される．この階層型アーキテクチャにより，タスクの高次な意図と動作制御を分離して学習し，長期タスクにおける適切な動作生成を実現する．

3.2 データ収集

データ収集にはバイラテラル制御を用いた．バイラテラル制御システムでは，リーダーロボットとフォロワーロボットに対して位置と力制御を用いているため，操作者がリーダーロボットを操作しながら，フォロワーロボットの環境反力を触覚的に感じ取ることができる．これにより，対象物に応じた適切な力加減を学習できる．

データ収集時には，バイラテラル制御を用いてリーダーおよびフォロワーロボットアームの関節角度，角速度，トルクを 1000Hz，オーバーヘッドカメラとグリッパーカメラの 2 つの RGB カメラ画像を 100Hz で記録する．オーバーヘッドカメラは，作業空間全体を俯瞰する視点を提供し，グリッパーカメラは，把持対象物の詳細な情報を提供する．言語指示は，タスクの目的を明確に表現する自然言語文として英語で記述した（例：「open top drawer」「pick right ball」「place on top drawer」など）．各エピソードは複数のサブタスクから構成され，各サブタスクに対応する言語指示が時系列に沿ってラベル付けされた．これにより，モデルがタスクの意図を理解し，適切な動作を生成できるようになる．

3.3 High-Policy

High-Policy は，時系列の画像情報からタスクの進行状況を理解し，適切な自然言語コマンドをサブゴールとして生成する．モデルアーキテクチャは，事前学習済みの CLIP を画像エンコーダーとして使用し，時系列の画像特徴を Transformer エンコーダーで処理する．処理された特徴から，最も類似度の高い自然言語コマンドを予測する．このモデルは，タスクの高次な意図を理解し，Low-Policy に対して適切な自然言語指示を生成する役割を担う．

3.4 Low-Policy: Bi-VLA モデル

Low-Policy として，言語情報を統合した Bi-VLA モデルを採用している．Bi-VLA モデルは，Bi-ACT[3] を拡張し，言語情報と視覚情報を効果的に統合する．具体的には，関節情報（角度，角速度，トルク），

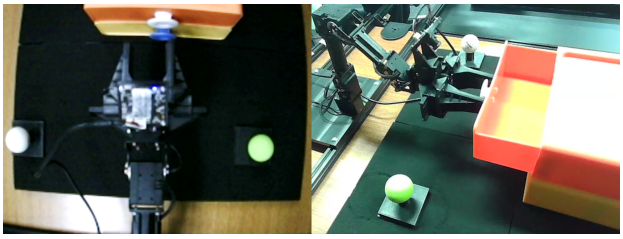

図 3 実験環境

オーバーヘッドカメラおよびグリッパーカメラで取得した RGB 画像，さらに High-Policy から与えられた自然言語コマンドをモデルに入力し，次の時刻の関節情報を推測する．

3.5 推論

実行時には，階層型アーキテクチャにより，High-Policy と Low-Policy が協調して動作する．推論プロセスは以下のように進行する．

High-Policy が，現在の状態 s_t（時系列の画像情報）を入力として受け取り，タスクの進行状況を評価する．High-Policy は，時系列の履歴情報を用いて，タスクの全体的な進行状況を理解し，適切な自然言語コマンド c_t をサブゴールとして生成する．この自然言語コマンドは，Low-Policy に対する高次な指示として機能する．

その後，Low-Policy（Bi-VLA モデル）が，High-Policy から与えられた自然言語コマンド c_t，現在の関節情報（角度，角速度，トルク），画像を組み合わせて，次のアクション列を予測する．Bi-VLA モデル内部では，SigLIP[12] により自然言語コマンド c_t が埋め込みベクトル e_l に変換され，EfficientNet[13] により画像から視覚特徴 f_v が抽出される．FiLM[14] により言語埋め込み e_l に基づいて視覚特徴 f_v が変調され，変調された視覚特徴と関節情報が Transformer エンコーダーに入力され，アクション列が予測される．

最終的に，予測されたアクション列の最初のアクション a_t がバイラテラル制御システムに送信され，フォロワーロボットが制御される．このプロセスが繰り返されることで，タスクが段階的に実行される．High-Policy は定期的に自然言語コマンドを更新し，タスクの進行状況に応じた適切な指示を生成し続ける．

表 1　各手法のタスク成功率

手法	成功率（%）
Bi-ACT	40.0
提案手法（階層型 Bi-VLA）	**60.0**

4　実験

4.1　実験設定

実験環境として，4 自由度＋グリッパー（計 5 モータ）のロボットアーム 2 台を用いたバイラテラル制御システムを構築した．観測として，オーバーヘッドカメラとグリッパーカメラの 2 つの RGB カメラ画像，および各ロボットアームの関節角度，角速度，トルクを記録した．

4.2　タスク内容

図 3 に実験環境を示す．評価タスクとして，「2 つのボールを引き出しに入れる」タスク（2ball_putInDrawer）を用いた．このタスクは，複数の対象物を順序立てて操作する必要があり，タスクの進行状況に応じた適切な動作選択が求められる．具体的には，以下の 8 つのサブタスクから構成される：(1) 上段の引き出しを開ける（言語コマンド：open top drawer），(2) 右側のボールを把持する（pick right ball），(3) 上段の引き出しに配置する（place on top drawer），(4) 上段の引き出しを閉める（close top drawer），(5) 下段の引き出しを開ける（open bottom drawer），(6) 左側のボールを把持する（pick left ball），(7) 下段の引き出しに配置する（place on bottom drawer），(8) 下段の引き出しを閉める（close bottom drawer）．なお，デモ収集は操作者による遠隔操作を要しコストが高いため，本設定は「少数デモで長期タスクを成立させる」実運用に近い条件を意図し，5 回のデモンストレーションを収集した．

4.3　評価指標

評価指標として，タスク成功率（Success Rate）を用いた．タスクが成功したかどうかは，2 つのボールがともにそれぞれの引き出し（上段と下段）の中に入っており棚がしまっているかどうかで判定した．各手法について，10 エピソードを実行し，成功率を計算した．

4.4　実験結果

表 1 に実験結果を示す．提案手法（階層型 Bi-VLA）は，Bi-ACT（成功率 40%）を上回り，成功率 60%を達成した．特に，階層型アーキテクチャにより，High-Policy がタスクの進行状況を理解し適切な自然言語コマンドを生成することで，Low-Policy がより適切な動作を生成できるようになったと考えられる．また，FiLM による言語情報と視覚情報の統合により，言語指示に応じた視覚特徴の動的な調整が可能となり，タスクの意図に沿った動作生成が実現された．さらに，SigLIP と EfficientNet の組み合わせにより，言語埋め込みと効率的な視覚特徴抽出が実現され，多様な情報を適切に処理できるようになった．

一方で，引き出しの把持動作の精度や，複雑なタスクの進行状況の理解には改善の余地がある．特に，引き出しを開けた直後に閉める動作を繰り返す失敗は，High-Policy がタスクの進行状況を適切に評価できていない可能性を示唆している．また，より長期的なタスクや，より複雑な言語指示への対応も今後の課題である．以上より，実験結果から提案手法の有効性が確認された．

5　おわりに

本研究では，自然言語と視覚情報を活用した階層型バイラテラル制御に基づく模倣学習手法を提案した．提案手法は，自然言語コマンドをサブゴールとして出力する High-Policy と，言語情報を統合した Low-Policy（Bi-VLA モデル）からなる階層型アーキテクチャにより，タスクの高次な意図と動作制御を分離して学習する．さらに，SigLIP による言語埋め込みと EfficientNet による視覚特徴抽出，FiLM による統合により，自然言語と視覚情報を効果的に融合する．実機実験により，提案手法の有効性が確認された．

今後の課題として，以下の点が挙げられる：(1) より多様なタスクへの適用：現在は 2 つのボールを引き出しに入れるタスクに限定されているため，より多様なタスクへの適用を検討する．(2) より長期的なタスクへの対応：現在の設定では限定的な長期タスクにしか対応できていないため，より長期的なタスクへの対応を検討する．これらの課題に取り組むことで，より汎用的で実用的な自然言語統合型ロボット制御システムの実現が期待される．

謝辞

本研究は，JST 国家戦略分野の若手研究者及び博士後期課程学生の育成事業（BOOST） 次世代 AI 人材育成プログラム（若手研究者支援）JPMJBY24C8 の支援を受けたものです.

参考文献

[1] Toshiaki Tsuji. Mamba as a motion encoder for robotic imitation learning. **IEEE Access**, Vol. 13, pp. 69941–69949, 2025.

[2] Nozomu Masuya, Hiroshi Sato, Koki Yamane, Takuya Kusume, Sho Sakaino, and Toshiaki Tsuji. Variable-speed teaching–playback as real-world data augmentation for imitation learning. **Advanced Robotics**, Vol. 39, No. 10, pp. 550–565, 2025.

[3] Thanpimon Buamanee, Masato Kobayashi, Yuki Uranishi, and Haruo Takemura. Bi-act: Bilateral control-based imitation learning via action chunking with transformer. In **2024 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)**, pp. 410–415. IEEE, 2024.

[4] Lucy Xiaoyang Shi, Zheyuan Hu, Tony Z Zhao, Archit Sharma, Karl Pertsch, Jianlan Luo, Sergey Levine, and Chelsea Finn. Yell at your robot: Improving on-the-fly from language corrections. **arXiv preprint arXiv:2403.12910**, 2024.

[5] Takumi Kobayashi, Masato Kobayashi, Thanpimon Buamanee, and Yuki Uranishi. Bi-lat: Bilateral control-based imitation learning via natural language and action chunking with transformers, 2025.

[6] Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. **Advances in Neural Information Processing Systems**, Vol. 33, pp. 13139–13150, 2020.

[7] Hongkuan Zhou, Zhenshan Bing, Xiangtong Yao, Xiaojie Su, Chenguang Yang, Kai Huang, and Alois Knoll. Language-conditioned imitation learning with base skill priors under unstructured data. **IEEE Robotics and Automation Letters**, Vol. 9, No. 11, pp. 9805–9812, 2024.

[8] 高城頌太, 谷口尚平, 中野聡大, 上田亮, 谷中瞳, 松尾豊. 大規模言語モデルの活用による効率的なロボット制御の学習. 言語処理学会 第 29 回年次大会 発表論文集, pp. 2460–2465, 2023.

[9] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. **arXiv preprint arXiv:2304.13705**, 2023.

[10] Masato Kobayashi and Thanpimon Buamanee. Bi-vla: Bilateral control-based imitation learning via vision-language fusion for action generation, 2025.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[12] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In **Proceedings of the IEEE/CVF international conference on computer vision**, pp. 11975–11986, 2023.

[13] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[14] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.