ニーズシーズ課題整理SWG公開シンポジウム
2026年4月29日15:15-15:45（発表時間20分、質疑10分）

基盤モデル・生成AIによるロボット作業学習の高度化

奈良先端科学技術大学院大学
松原崇充

本日の講演の趣旨

- 松原の専門は、強化学習・模倣学習の実世界応用
- 基盤モデルそのものに直接寄与する研究は行っていない
- 川村先生から「基盤モデル・生成AIの観点で」とご依頼
- 本日は、自身の研究貢献ではなく，**基盤モデルをどのようにロボット学習に活用しているかという観点で事例をご紹介**

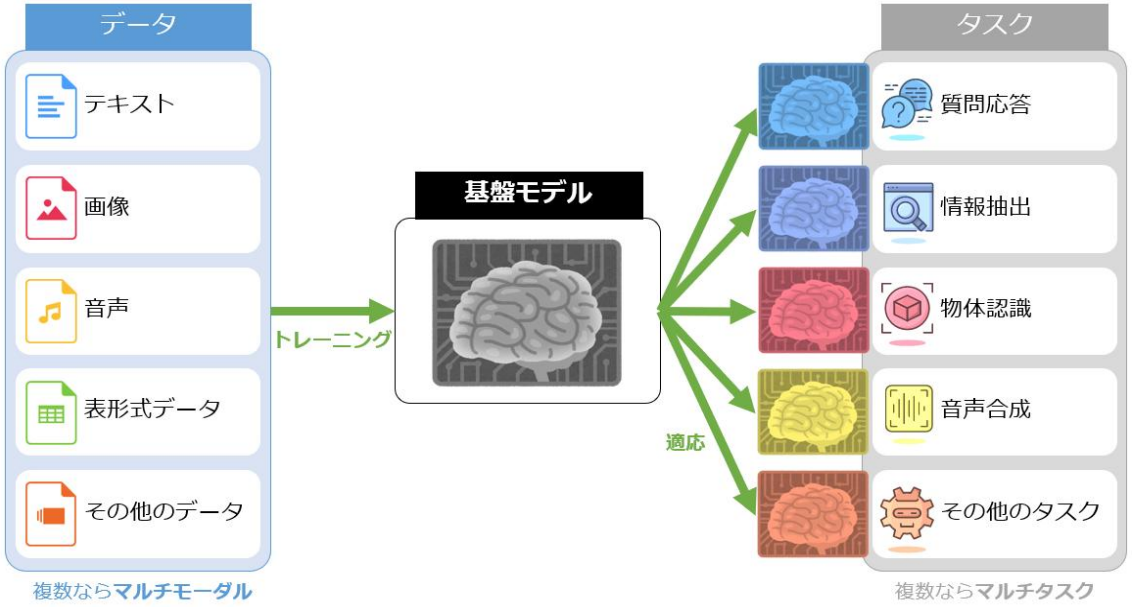
今日の講演内容

1. 基盤モデル概観

2. 基盤モデルを活用したロボット学習事例紹介

基盤モデルとは？

- 基盤モデルの提唱（Foundation Model：ファウンデーションモデル、ファンデーションモデル, 2021年, スタンフォード大）
- 汎用的な目的で大規模データにより事前学習され、少ないデータで様々な下流タスクに応用できるAIモデルのこと。これ以前は、基本的に単一目的のＡＩモデルが都度学習



https://atmarkit.itmedia.co.jp/ait/articles/2302/27/news014.html

例）ChatGPT

基盤モデルのカテゴリ・代表例

1. 大規模言語モデル (LLM: Large-Language Model)
テキストの理解・生成に特化したモデル
ex) GPT-4 / GPT-4o (OpenAI), Gemini 1.5 Pro / Ultra (Google), Claude 3 / 3.5 Sonnet (Anthropic)

LLM → VLM
→ ドメイン特化型
→ VLA

2. 視覚言語モデル（VLM: Vision-Language Model）
テキスト、画像、動画を相互に変換・理解するモデル
ex) GPT-4o (OpenAI), Gemini 1.5 Pro / Ultra (Google), Claude 3.5 Sonnet (Anthropic), LLaVA

3. SAM: Segment Anything Model
画像内のあらゆる物体を自動的に分割・切り出し（セグメンテーション）できる高性能な「コンピュータビジョン」の基盤モデル
ex) SAM / SAM2 / SAM3 (Meta)

4. Depth Anything (V1/V2/V3)
単一の画像（単眼）から高精度な奥行き（深度）情報を推定するAIモデル
ex) Depth Anything V1 / V2 / V3

5. VLA: Vision-Language Action Model
画像・言語の理解に加えて、ロボットやエージェントの行動を生成・制御するモデル
ex) RT-2 (Google DeepMind), OpenVLA, Octo, π0 / Physical Intelligence系モデル

基盤モデルのカテゴリ・代表例

1. 大規模言語モデル (LLM: Large-Language Model)
テキストの理解・生成に特化したモデル
ex) GPT-4 / GPT-4o (OpenAI), Gemini 1.5 Pro / Ultra (Google), Claude 3 / 3.5 Sonnet (Anthropic)

2. 視覚言語モデル（VLM: Vision-Language Model）
テキスト、画像、動画を相互に変換・理解するモデル
ex) GPT-4o (OpenAI), Gemini 1.5 Pro / Ultra (Google), Claude 3.5 Sonnet (Anthropic), LLaVA

LLM → VLM
→ ドメイン特化型
→ VLA

＊後半の事例紹介に関係するので簡単に紹介

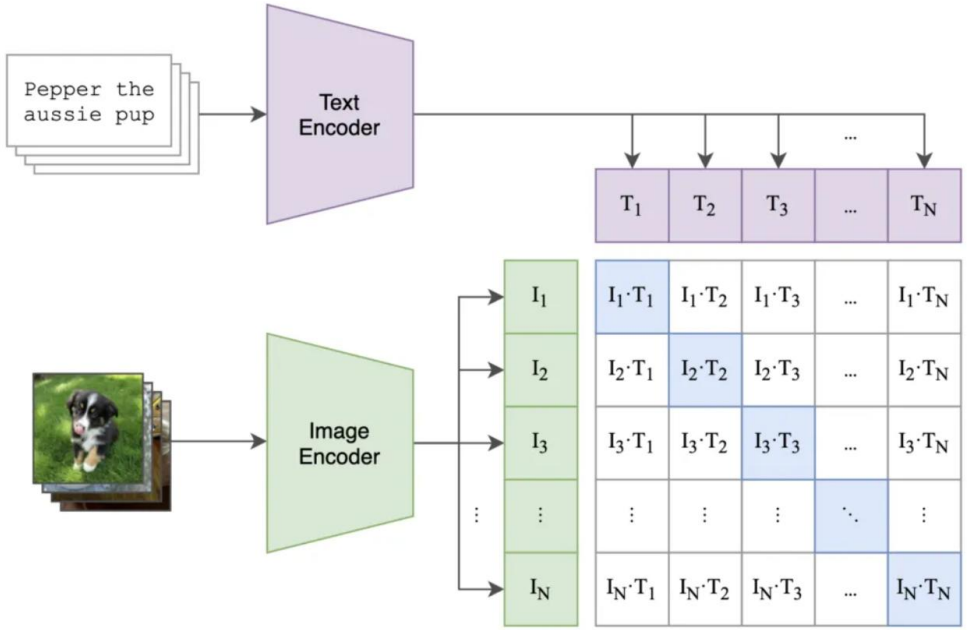
3. SAM: Segment Anything Model
画像内のあらゆる物体を自動的に分割・切り出し（セグメンテーション）できる高性能な「コンピュータビジョン」の基盤モデル
ex) SAM / SAM2 / SAM3 (Meta)

4. Depth Anything (V1/V2/V3)
単一の画像（単眼）から高精度な奥行き（深度）情報を推定するAIモデル
ex) Depth Anything V1 / V2 / V3

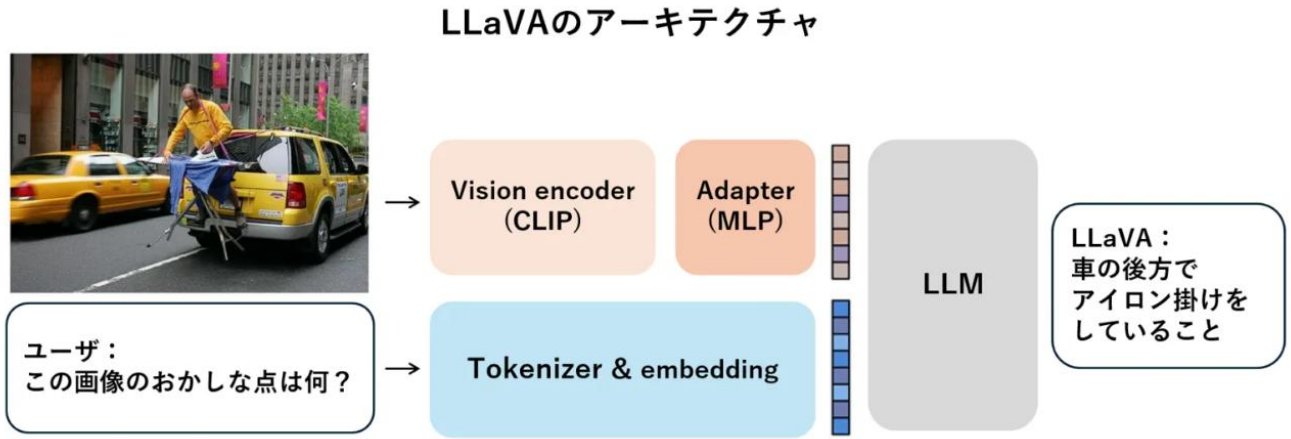
5. VLA: Vision-Language Action Model
画像・言語の理解に加えて、ロボットやエージェントの行動を生成・制御するモデル
ex) RT-2 (Google DeepMind), OpenVLA, Octo, π0 / Physical Intelligence系モデル

CLIP : Contrastive Language-Image Pre-training

- OpenAIが開発した画像とテキストを同時に理解する視覚言語モデル（VLM）
- テキスト・画像それぞれにエンコーダを保有し、画像と言語という異なる次元のデータを、**共通する特徴空間に埋め込む**
- 画像–テキストのペアを使った対照学習で訓練
- Web上の**4億枚の画像とキャプション**のペアデータを用いて最適化



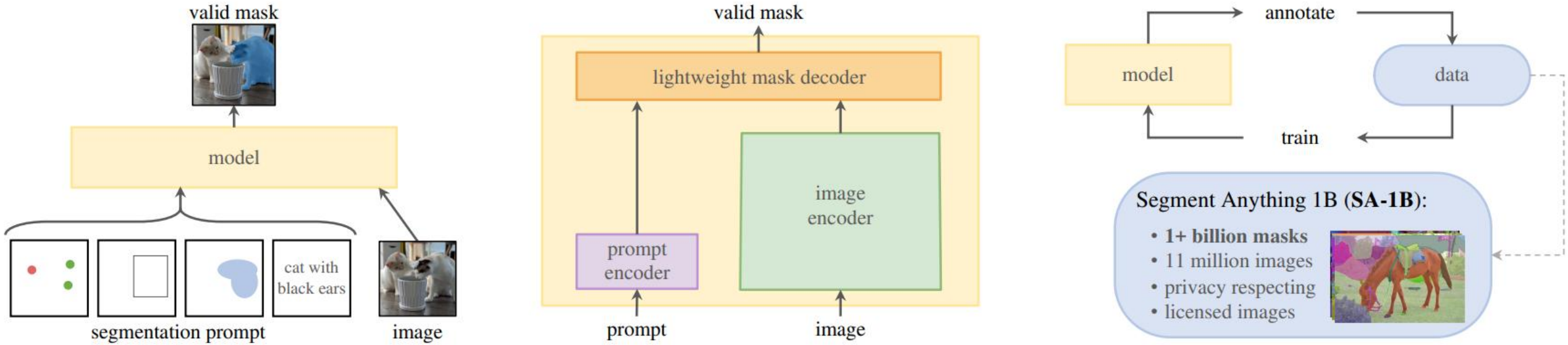
CLIPを画像埋め込みに用いた最近のVLM:



https://arxiv.org/abs/2103.00020

SAM: The Segment Anything Model

- 簡単な指示（プロンプト）で画像のセグメンテーションを実行可能なモデル
- セグメンテーションとは、画像内の物体を1ピクセルごとに分けるタスク
- 1100万枚の画像と、それらに付随した10億以上のマスクからなる莫大なデータセットを使用して訓練

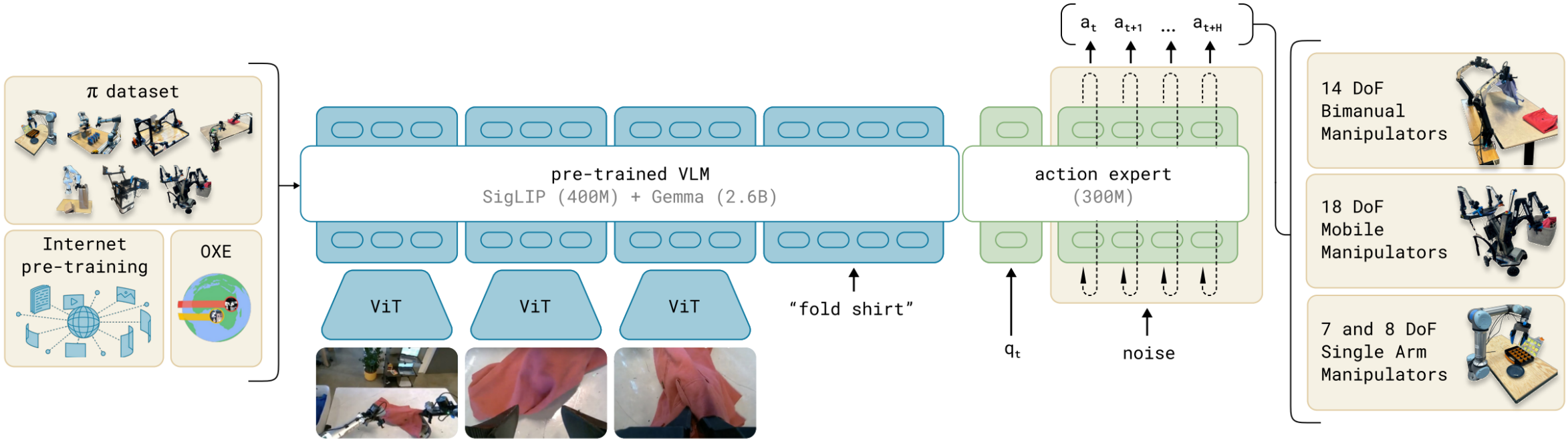


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

https://arxiv.org/abs/2304.02643

π0：A Vision-Language-Action Flow Model for General Robot Control

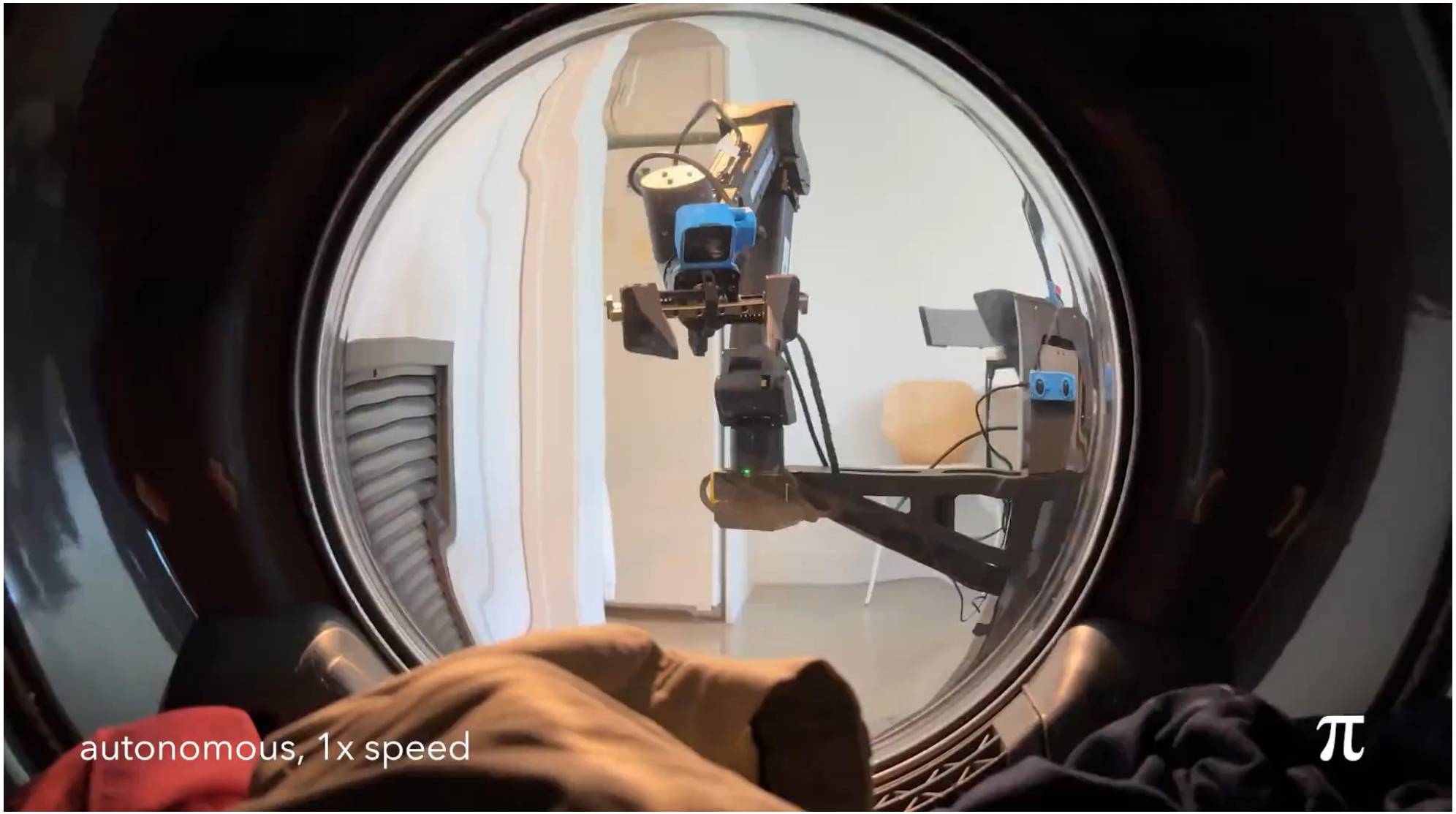
- Physical Intelligence が開発した VLA モデル
- 事前学習済みの VLM モデルをベースに、ロボットアームの移動量などの行動ベクトル値を出力する出力層（action expert）を追加して VLA モデルに拡張、テキスト＆画像でロボット制御指示が可能
- 7種類のロボット構成、68のタスク、10,000時間分の操作データに加え、過去のPJで収集された大量ロボット操作データ含む最大規模の事前データセットを利用



https://arxiv.org/abs/2410.24164

π0 : A Vision-Language-Action Flow Model for General Robot Control

https://arxiv.org/abs/2410.24164



https://www.physicalintelligence.company/blog/pi0

VLA開発に向けたデータ工場も登場

- 中国Agibot（ハード+データ）と米国Physical Intelligence（AI）が提携

袋詰め

陳列

電子レンジの使用

冷蔵庫の整理

https://36kr.jp/340772/

今日の講演内容

1. 基盤モデル概観

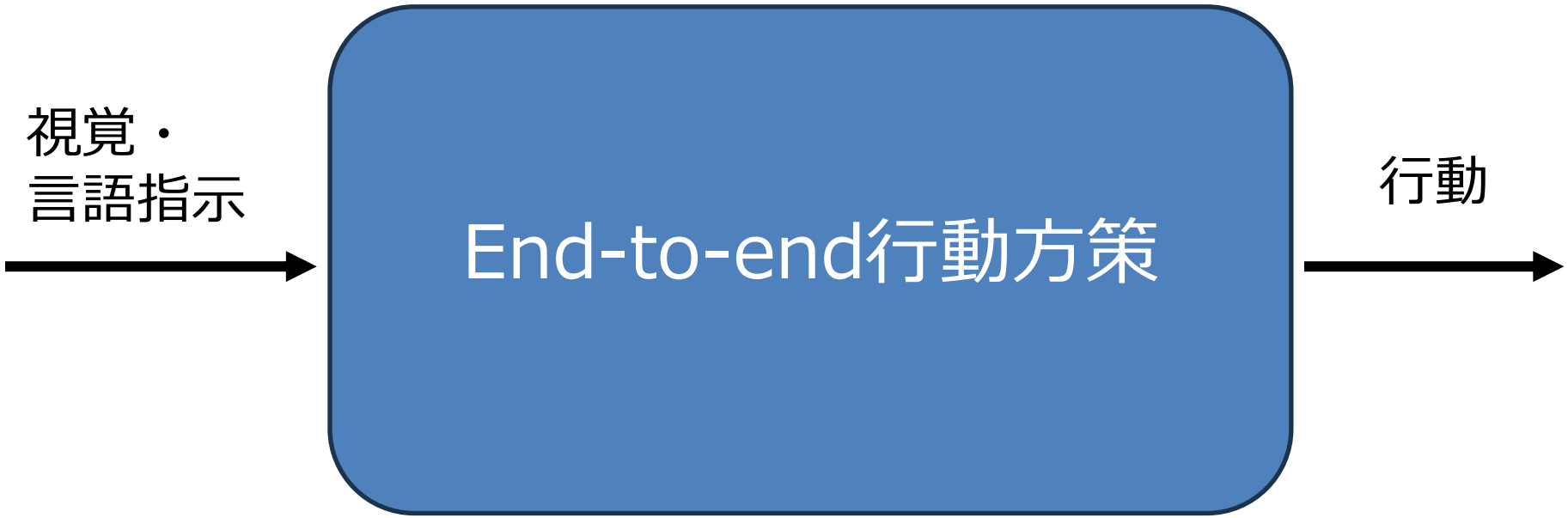
2. 基盤モデルを活用したロボット学習事例紹介

ロボット学習と基盤モデル

- 深層学習の登場以降、従来のパイプライン構成から、センサから行動へのend-to-end学習に大きな期待
- しかし、その実用化を阻む本質的な制約は、**ロボットデータの不足**
- 多くの基盤モデルは、そのまま行動方策（閉ループ制御のコントローラ）としては成立しない
- **では、基盤モデルを「行動方策」ではなく、ロボット学習のデータ依存性/収集コストを低減するために活用できないか？**

ロボット学習のための基盤モデル活用に向けて

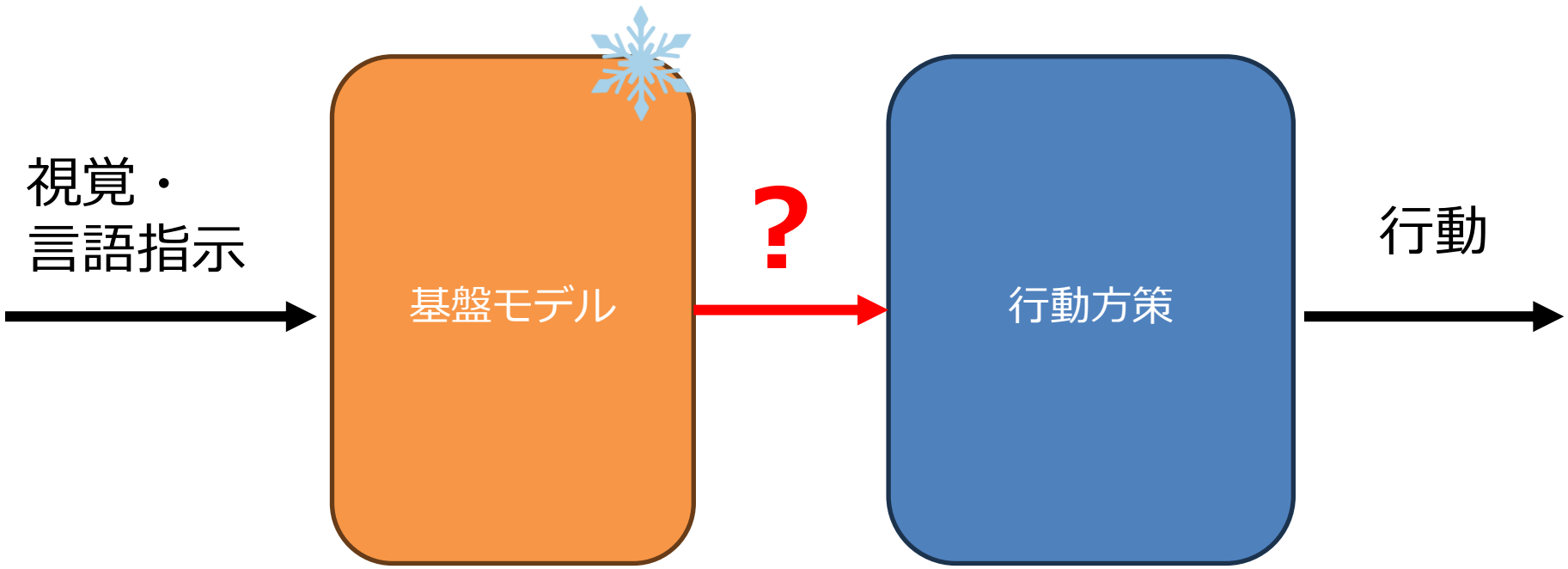
- 基盤モデルをロボット学習のデータ依存性/収集コストの低減に活用する



タスクや環境の複雑さ・多様性に応じて大量の学習データが要求

ロボット学習のための基盤モデル活用に向けて

- 基盤モデルをロボット学習のデータ依存性/収集コストの低減に活用する



基盤モデルと行動方策に分離することで方策学習を簡略化できる可能性
ただし、基盤モデルの出力設計が重要

ロボット学習のための基盤モデル活用事例を紹介

- VLMを用いた動作生成 **[Anarrossi, IEEE Access 2025]**
  - VLMを使って、言語指示と視覚情報に応じた動作を生成する

End-to-endな模倣学習問題から物体・指示認識を分離

- SAMを用いた農作業自動化のための模倣学習 **[Hattori+, arXiv2026]**
  - 画像中のセグメンテーションを容易化する基盤モデルを模倣学習に利用

End-to-endな模倣学習問題から前処理（セグメンテーション）の負担軽減

- LLMを用いた言語指示に基づくマルチロボット制御 **[Yano+, IROS2025]**
  - 言語指示に応じたマルチロボットの協調制御
  - 言語の理解・解釈と、マルチロボ協調制御学習を分離

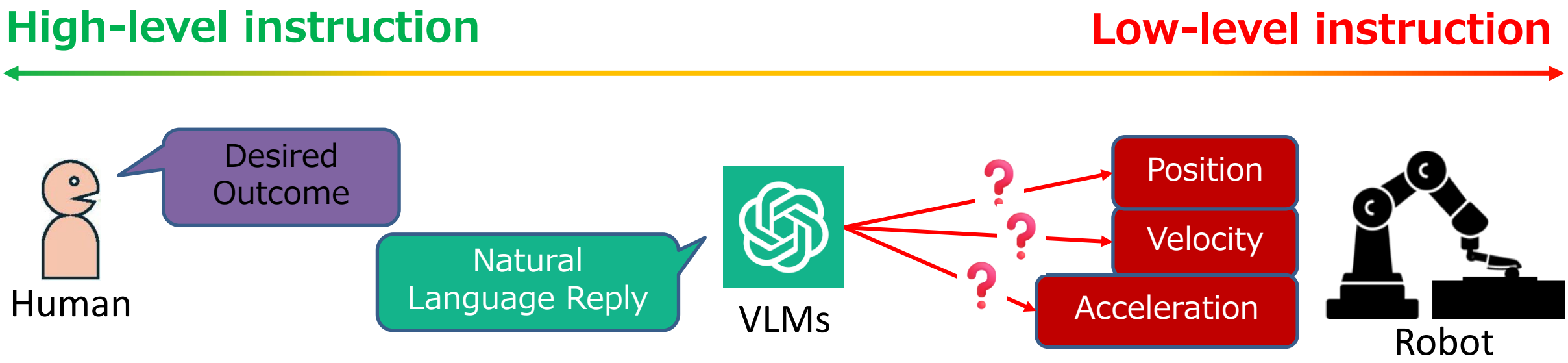
言語指示に基づくマルチロボット協調学習問題から言語認識を分離

- VLMを用いた言語と視覚に基づくリーダー・フォロワー制御 **[Despature+, arXiv2026]**
  - 視覚と言語指示に応じた対象物体の特定
  - 言語の理解・解釈と、マルチロボ協調制御学習を分離

言語指示に基づくマルチロボット協調学習問題から物体認識を分離

VLMを用いた動作生成

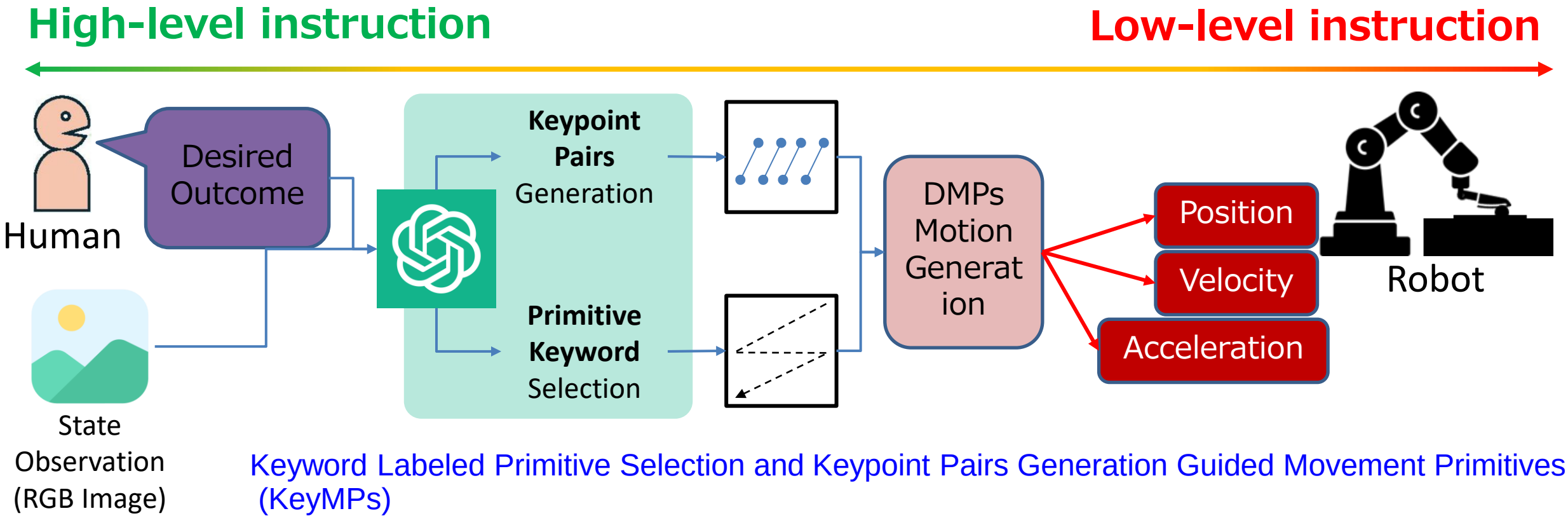
VLMは、視覚や言語指示からロボットの低レイヤの動作情報生成は苦手



KeyMPs：視覚言語指示に基づく動作生成

[Anarrossi, IEEE Access 2025]

- 経由点と動作特徴から一連動作を出力する軌道生成器（Sequential DMPs）を事前に模倣学習
- VLMの出力を、1)二次元動作経由点（Keypoints）と、2) 動作特徴キーワード（Keyward-based primitive）の生成に限定して接続
- 画像と言語指示から状況に応じた多様な動作生成を実現



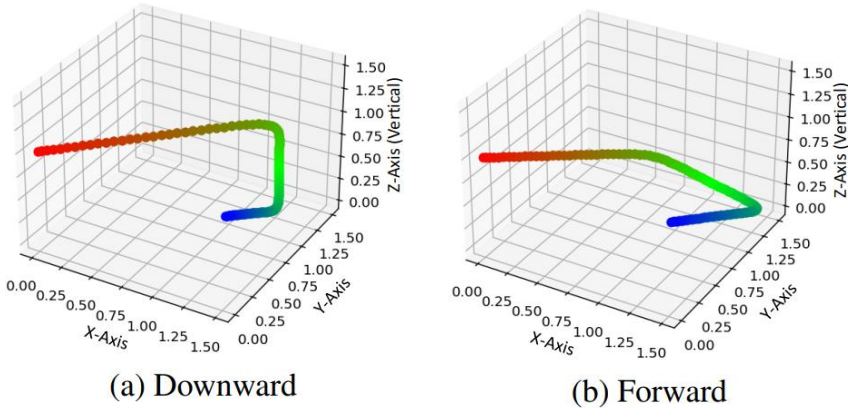
Keyword Labeled Primitive Selection and Keypoint Pairs Generation Guided Movement Primitives (KeyMPs)

KeyMPs：包丁による食材の切断タスク

[Anarrossi, IEEE Access 2025]

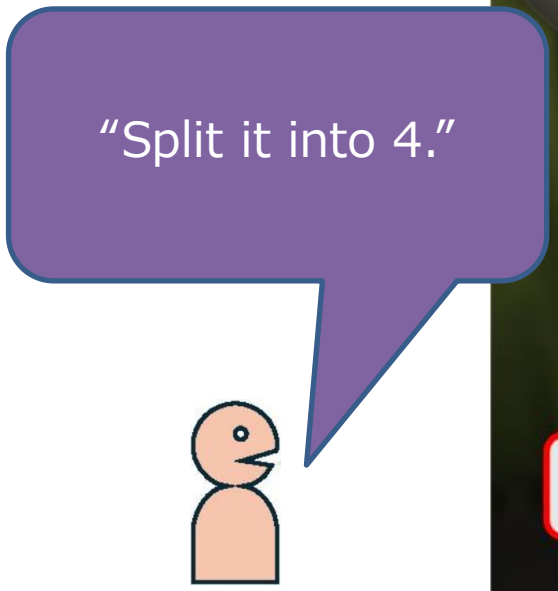
画像と言語指示に応じた動作生成の精度を実験検証

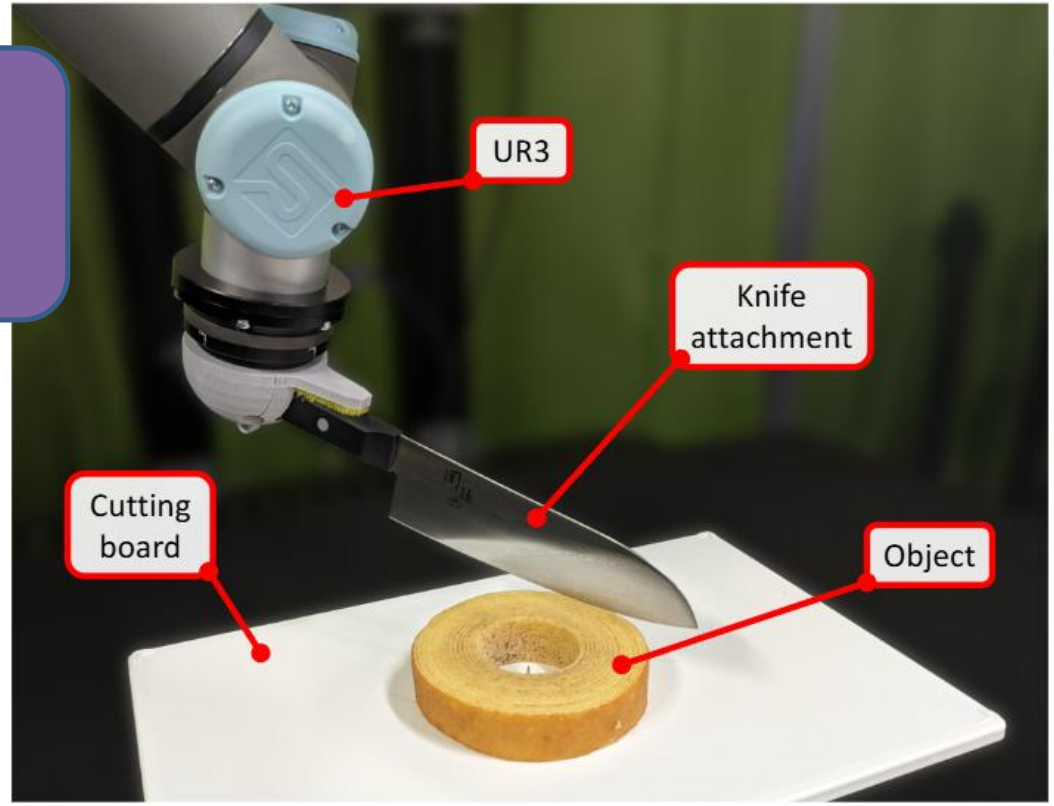
→ 食材に応じた切り分け方を選択実行できるか？



(a) Downward　(b) Forward

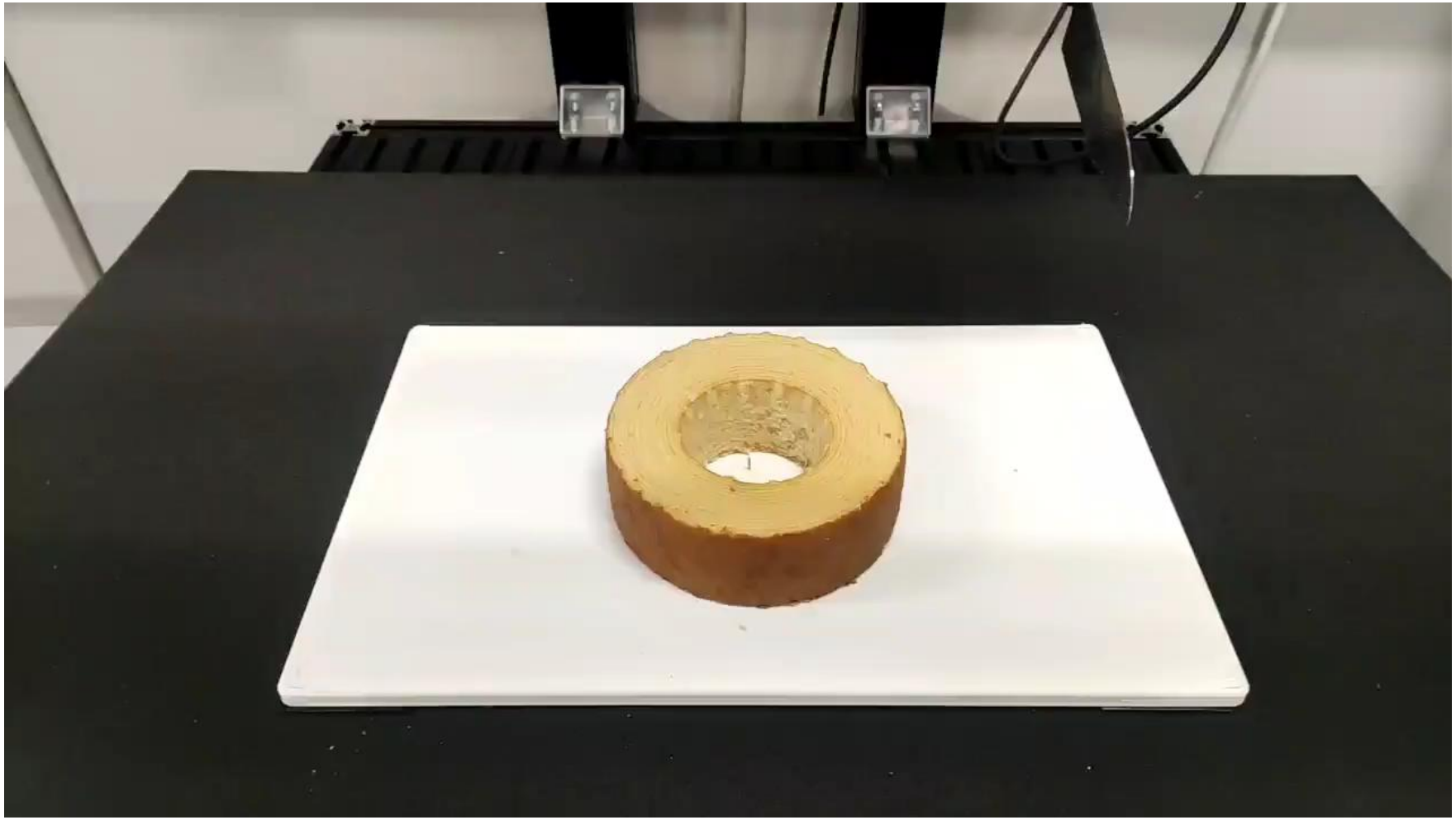
キーワード付き
特徴動作を模倣学習





包丁による食材の切断タスク：実験結果

[Anarrossi, IEEE Access 2025]



“Split it into 4.”

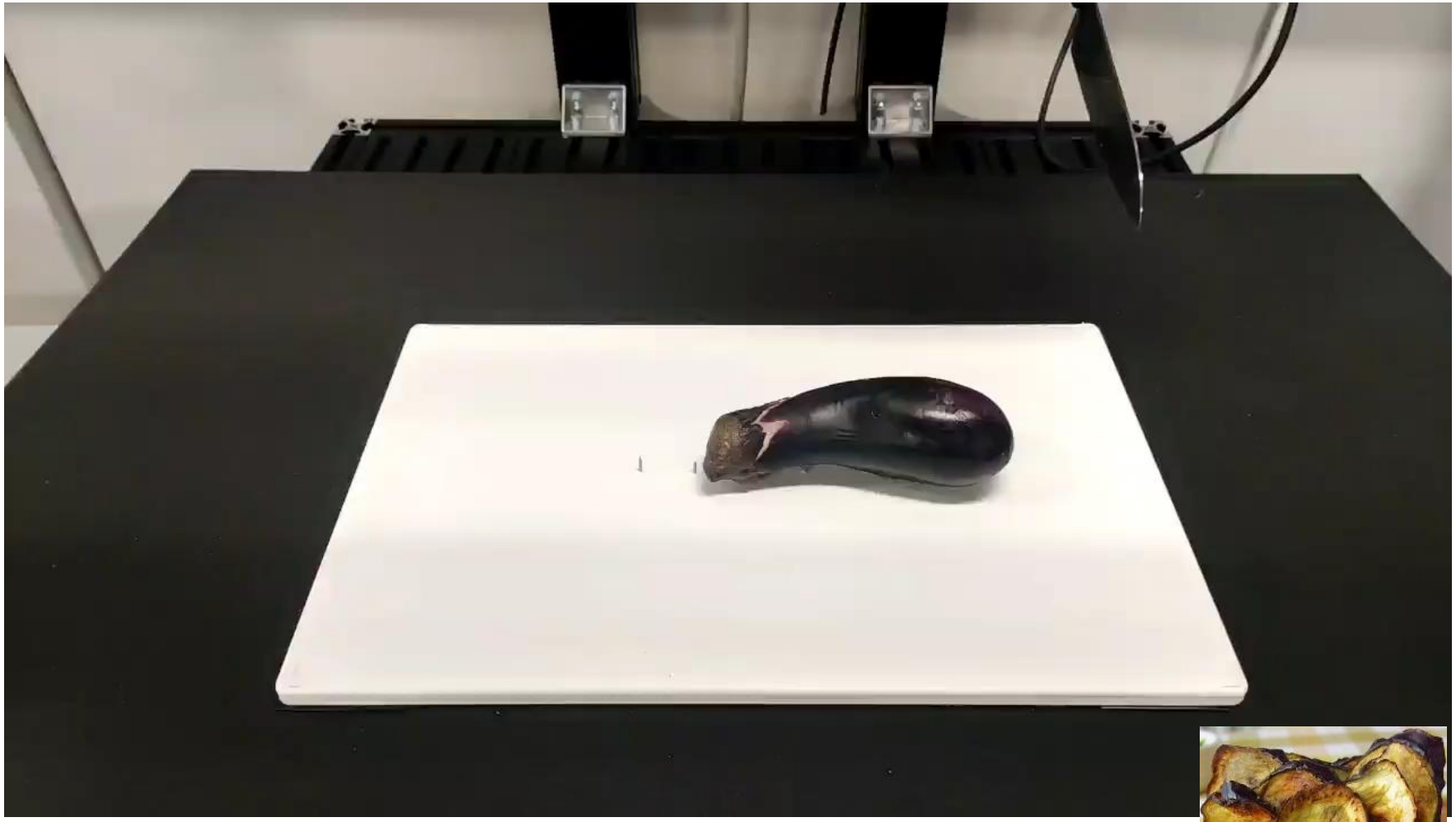
包丁による食材の切断タスク：実験結果

[Anarrossi, IEEE Access 2025]

“I have 3 guests, cut a few thin slices of the chiffon cake for them.”

包丁による食材の切断タスク：実験結果

[Anarrossi, IEEE Access 2025]

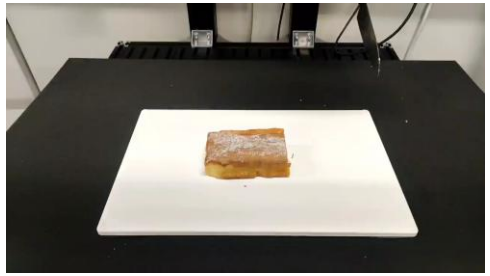


“I want to make wide chips out of this.”

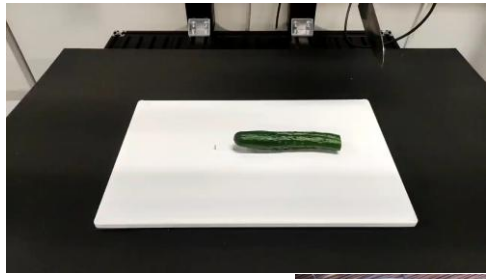
包丁による食材の切断タスク：実験結果

[Anarrossi, IEEE Access 2025]

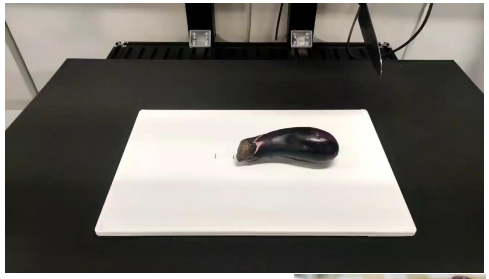
“I have 3 guests, cut a few thin slices of the chiffon cake for them.”



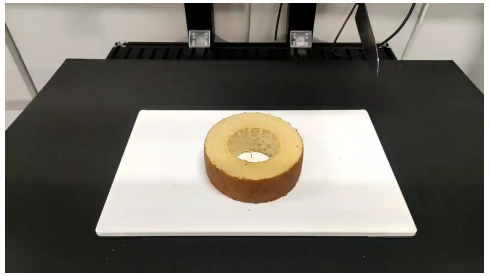
“I want to eat this chiffon cake for each day this week.”



“I want to make tsukemono.”

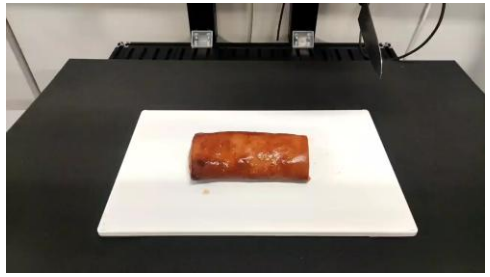


“I want to make wide chips out of this.”

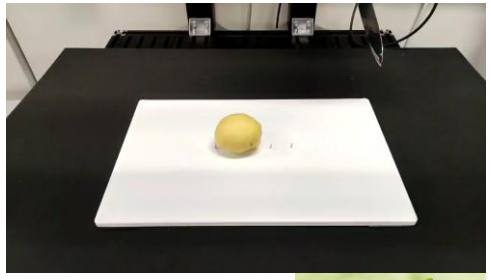


“Split it into 4.”

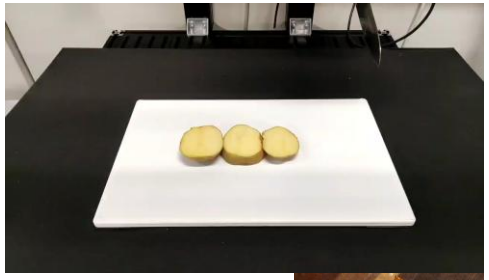
“This is a 490g block of meat (length 21cm, width 8cm), the nutrition facts mentioned that every 100g there’s 150kcal. Cut me several number of slices in a certain length (3cm) just enough if I want to go for a 3km walk after this.”



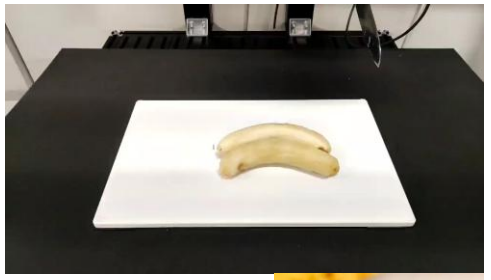
“This is a block of meat (length 16cm, width 8cm). Cut me 4 2cm slices.”



“Prepare it for fondant potato, this potato is quite small.”



“Cut it into French fries.”



“For banana pancake.”

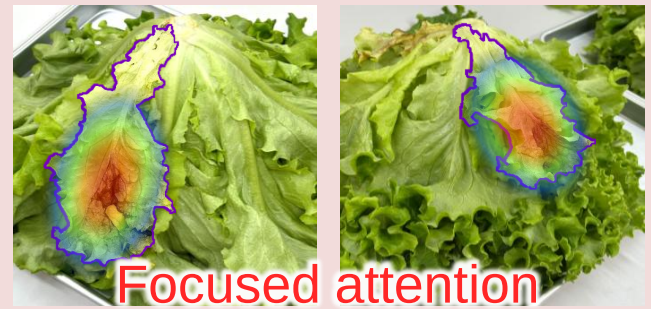
SAMを用いた農作物作業自動化のための模倣学習

農作物は見た目の多様性が大きいため，visuomotor policy は
タスクに無関係な視覚特徴に依存し，未知環境への汎化性能が低下する

レタスの不良葉除去タスク

Visual observation

Fragile visuomotor policy



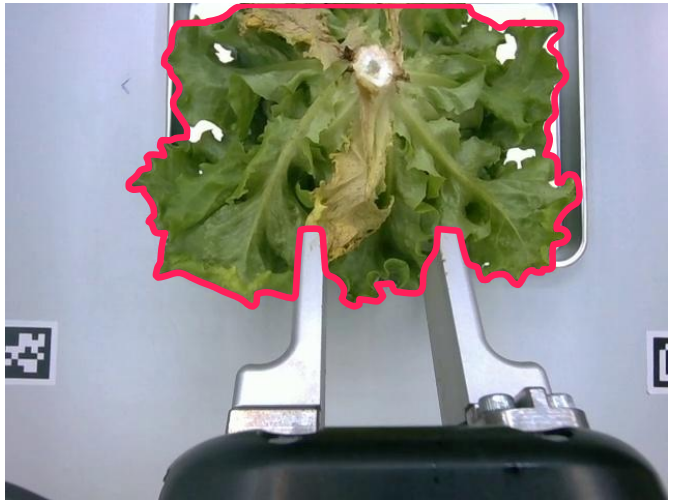
Robust visuomotor policy

画像中の適切な領域に着目することで，
visuomotor policy をロバストに学習できるのでは？

DRAIL：Dual-Region Augmentation for Imitation Learning

[Hattori+, arXiv2026]

農作物マニピュレーションにおける視覚ベース模倣学習のための
タスク関連特徴に注意を誘導する学習フレームワークとして，
Dual-Region Augmentation for Imitation Learning（DRAIL）を提案する。



Visual observation

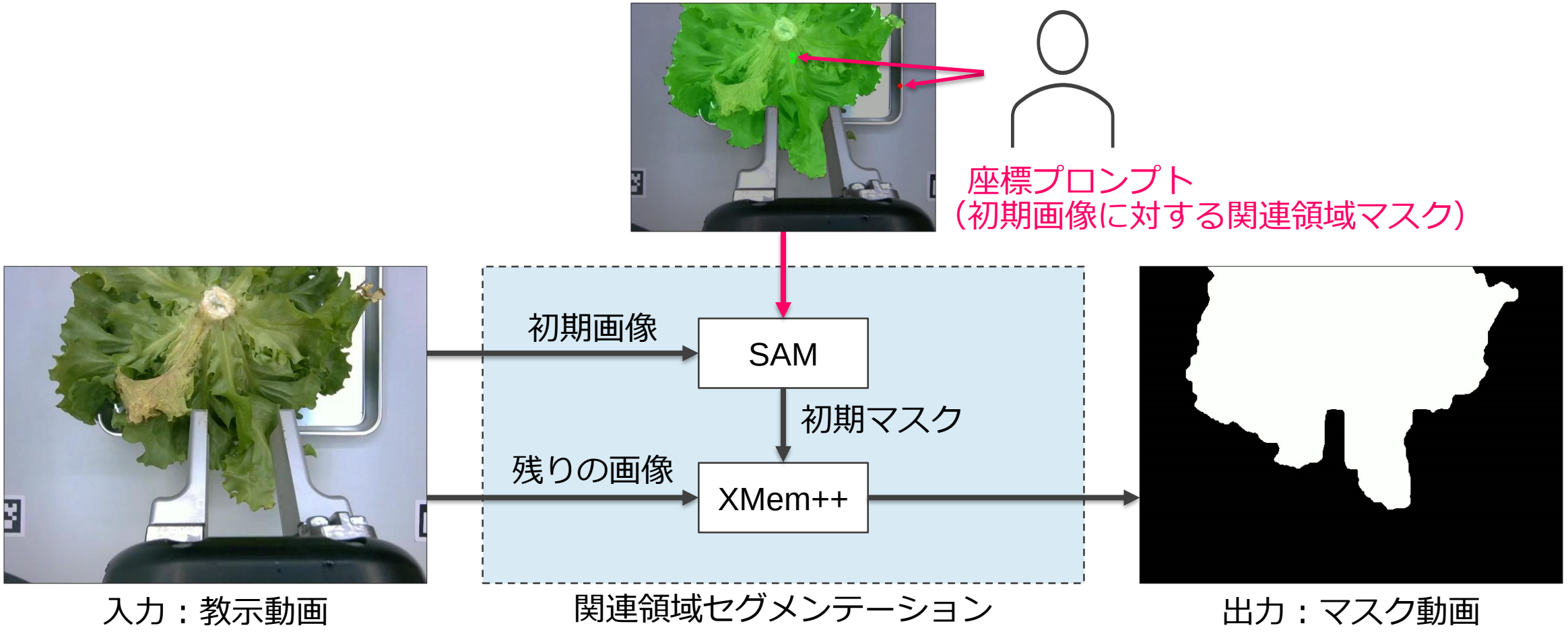
Task-relevant aug.

Task-irrelevant aug.

Augmented visual obs.

SAMを用いた関連/非関連領域抽出

[Hattori+, arXiv2026]



事前学習済みの基盤モデルへのプロンプト入力により，
セグメンテーションマスクを自動生成

実験評価：タスク成功率

[Hattori+, arXiv2026]

We evaluate leaf-selection and positional-alignment success on the same lettuce

Target leaf (largest defect)

	DRAIL	DRAIL w/o task-irr.	DRAIL w/o task-rel.	DRAIL w/o dual
Leaf selection	**80%**	67%	53%	47%
Positional alignment	**73%**	47%	40%	40%

DRAIL achieves the highest success rates

DRAIL (Proposed)

w/o task-irrelevant aug.

w/o task-relevant aug.

w/o dual aug.
(Diffusion Policy)

実験評価：視覚注意の分析

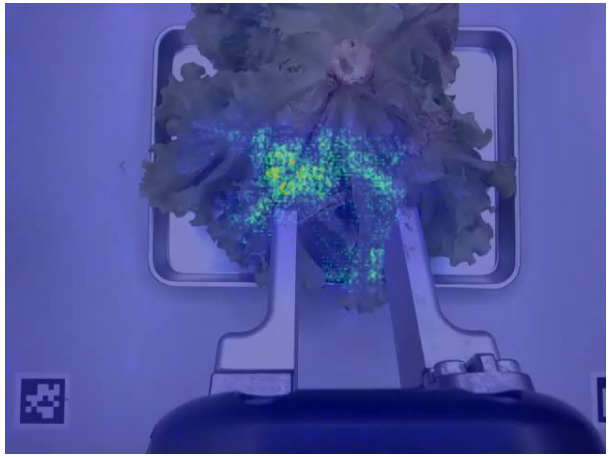
[Hattori+, arXiv2026]

We visualize attention maps to verify that actions are inferred from defect locations

Target leaf (largest defect)

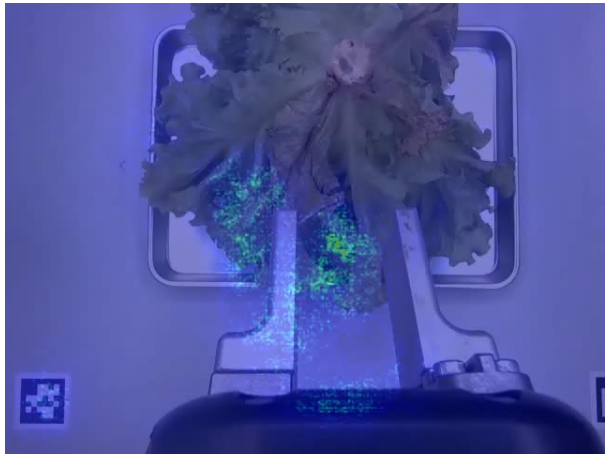
The policy focuses on the defective leaf

The policy attends broadly across the scene

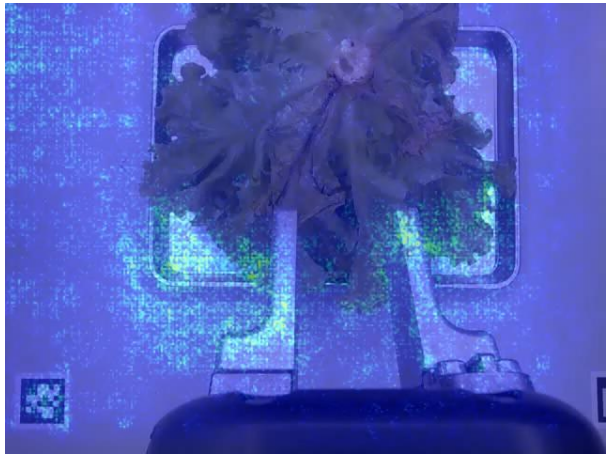


DRAIL (Proposed)

w/o task-irrelevant aug.



w/o task-relevant aug.



w/o dual aug.
(Diffusion Policy)

レタス不良葉除去デモンストレーション

[Hattori+, arXiv2026]

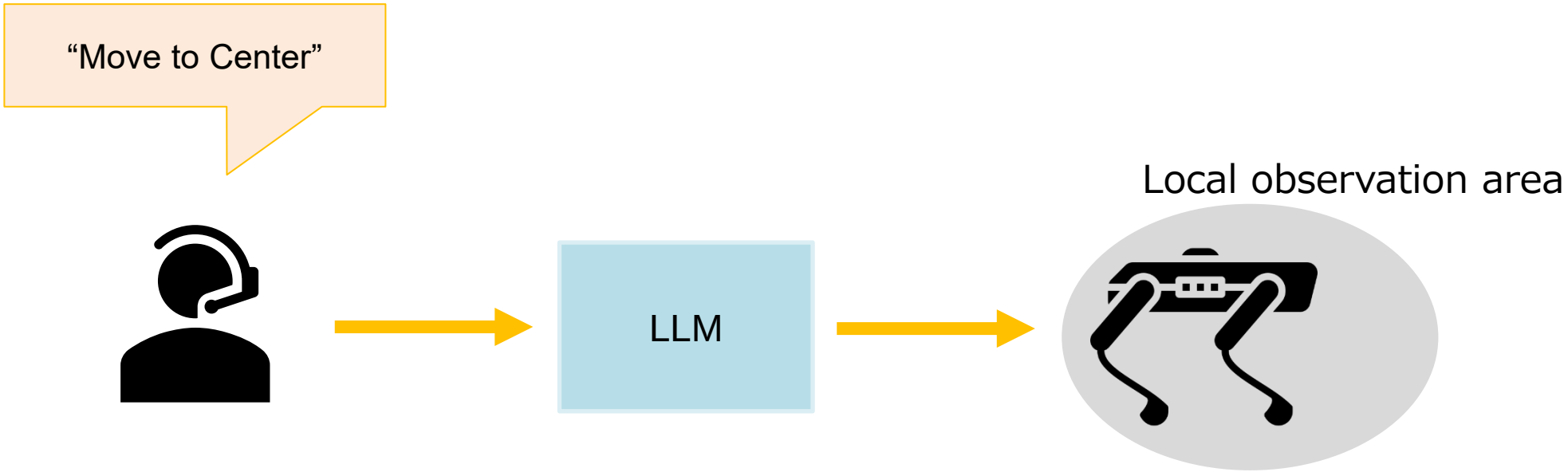
Task-Relevant and Irrelevant Region-Aware Augmentation for Generalizable Vision-Based Imitation Learning in Agricultural Manipulation

Shun Hattori[1], Hikaru Sasaki[1], Takumi Hachimine[1],
Yusuke Mizutani[2], Takamitsu Matsubara[1]

[1] Nara Institute of Science and Technology (NAIST)
[2] TSUBAKIMOTO CHAIN CO.

LLMを用いた言語指示に基づくマルチロボット制御

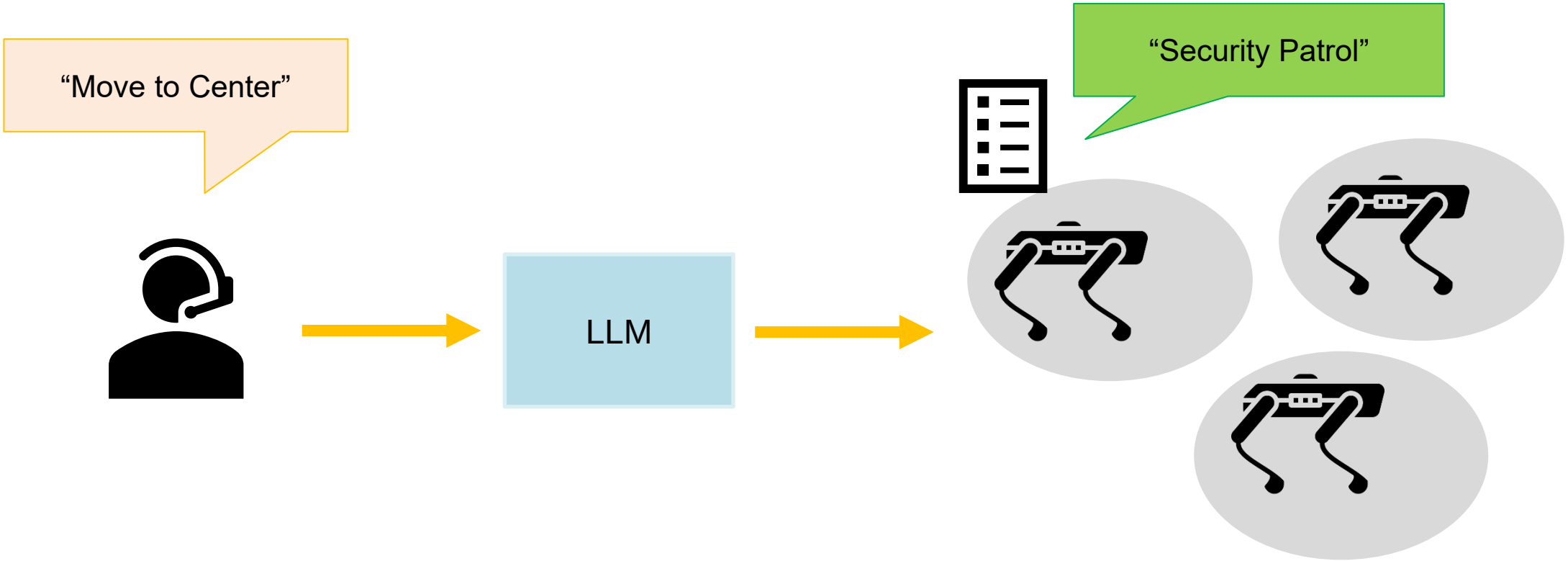
- LLMの発展により自然言語指示によるロボット制御が可能に



本研究の動機：マルチロボットシステムへの拡張

[Yano+, IROS2025]

- 自律分散型：各ロボットは自己の局所センサに基づく意思決定
- タスク保持：言語指示とは別要件の協調タスクを持つ

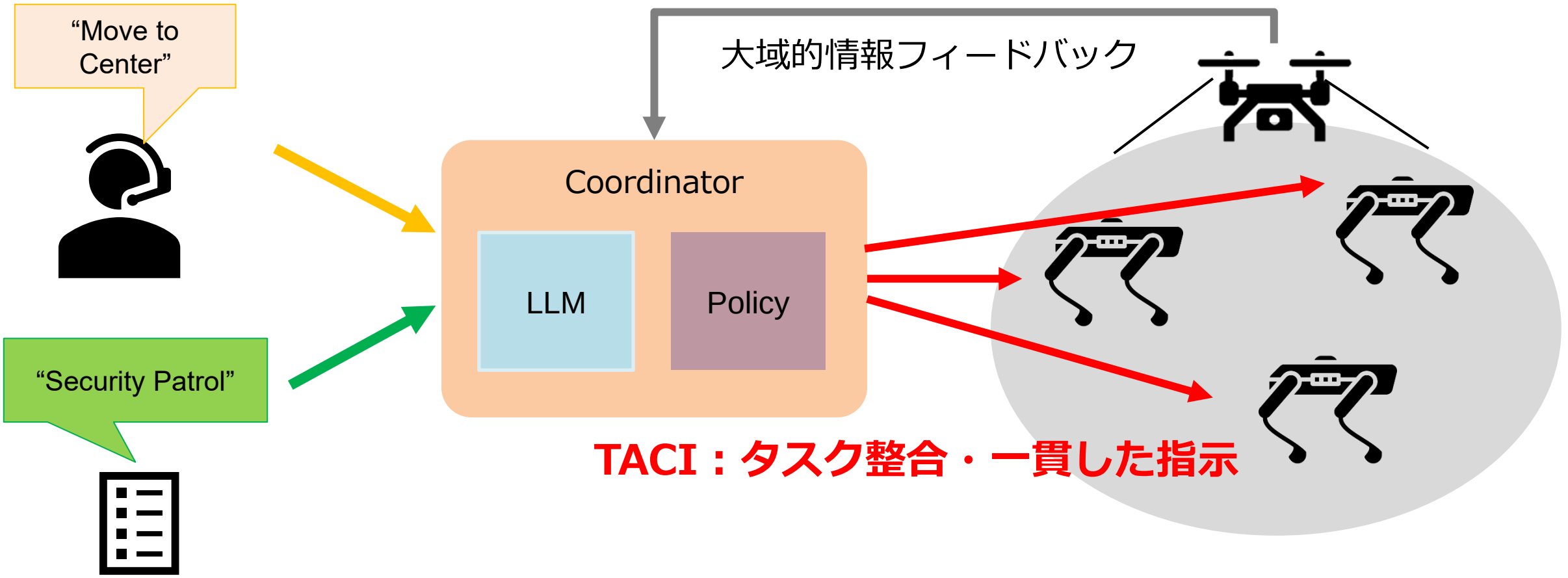


言語指示遵守とタスク整合を両立するマルチロボット制御をどう実現するか？

ICCO: Instruction-Conditioned Coordinator

[Yano+, IROS2025]

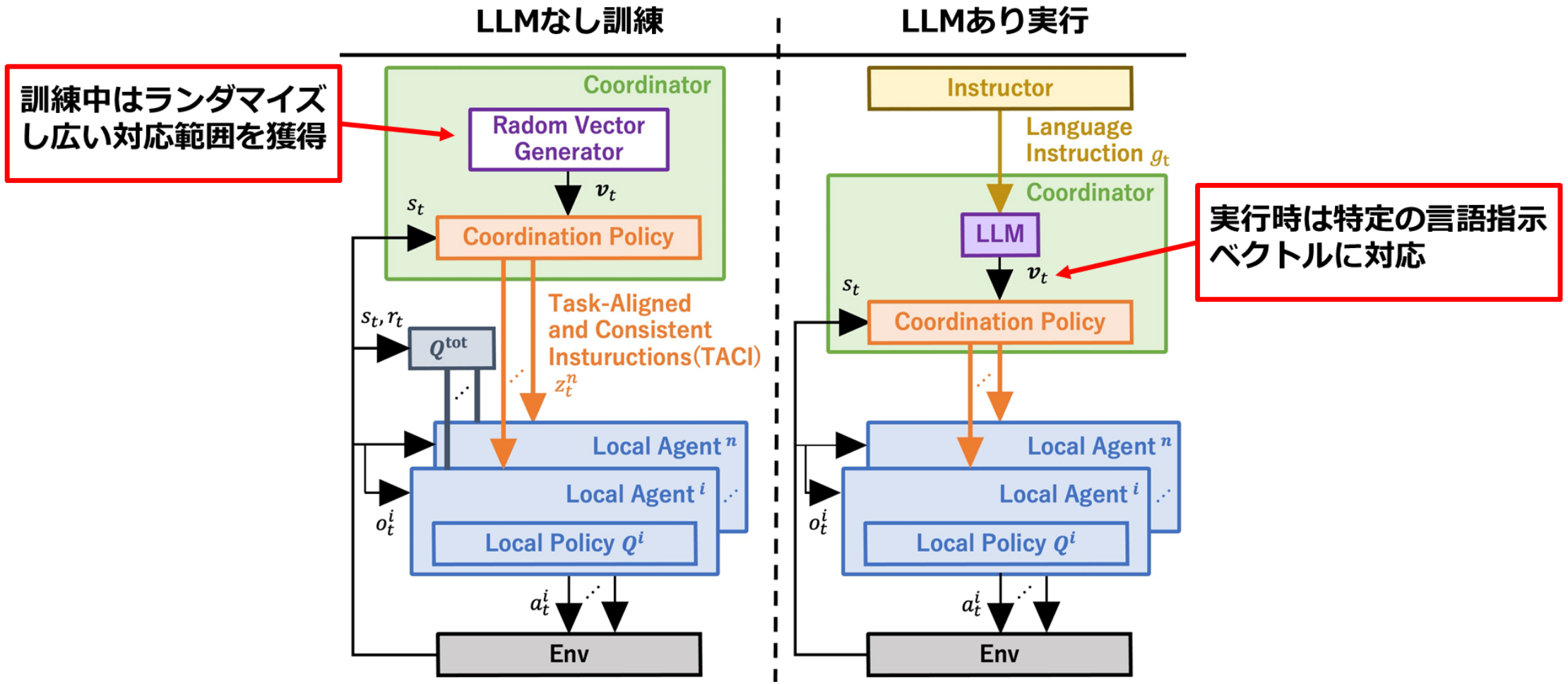
言語指示と協調タスク遂行のバランスを取るMARLフレームワーク



LLMなし学習・LLMあり実行フレームワーク

[Yano+, IROS2025]

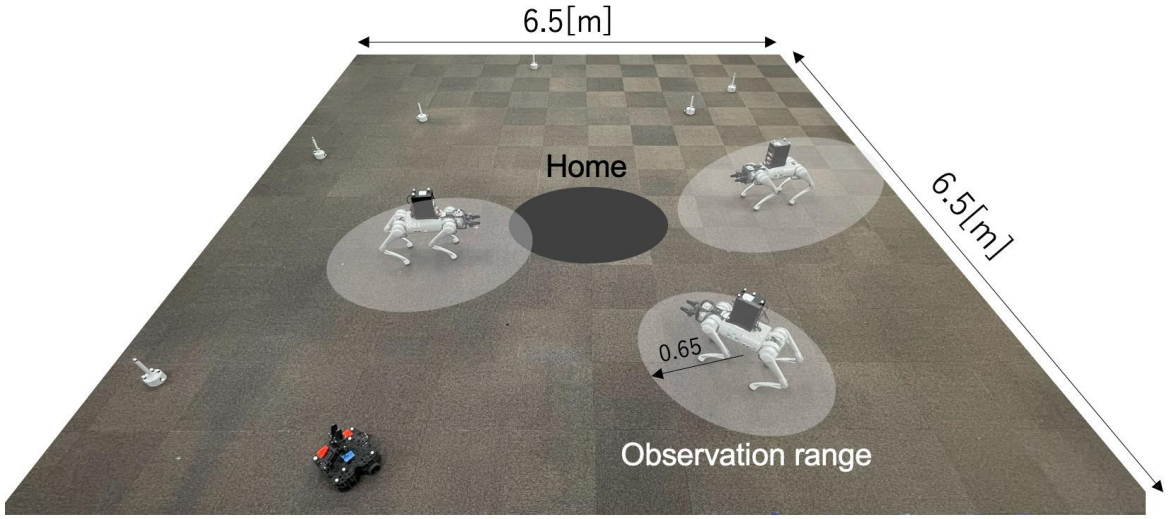
- 学習時はランダム軌道で学習
- 実行時はプロンプトで言語指示→目標軌道へと制限して整合



実験環境・評価タスク

[Yano+, IROS2025]

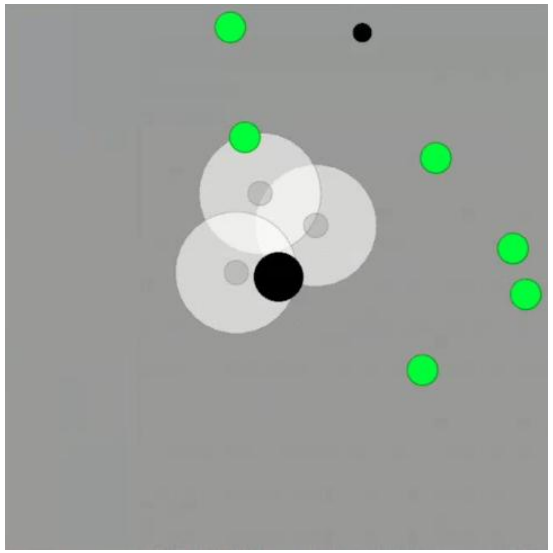
資源収集タスク[B. Liu+, 2021]
資源収集と侵略者防衛を同時に行う必要のあるタスク
Resourcesに当たる(+5), Resourcesをhomeに持ち帰る(+1),
AgentがInvaderに当たる(+4), InvaderがHomeに到達する(-4)




Invader (TurtleBot)


Agent (Go1 with Muzzle)


Resource



● Home ● Resource
● Agent • Invader

実機による資源収集タスクの様子

[Yano+, IROS2025]

ICCO: Learning an Instruction-conditioned Coordinator for Language-guided Task-aligned Multi-robot Control

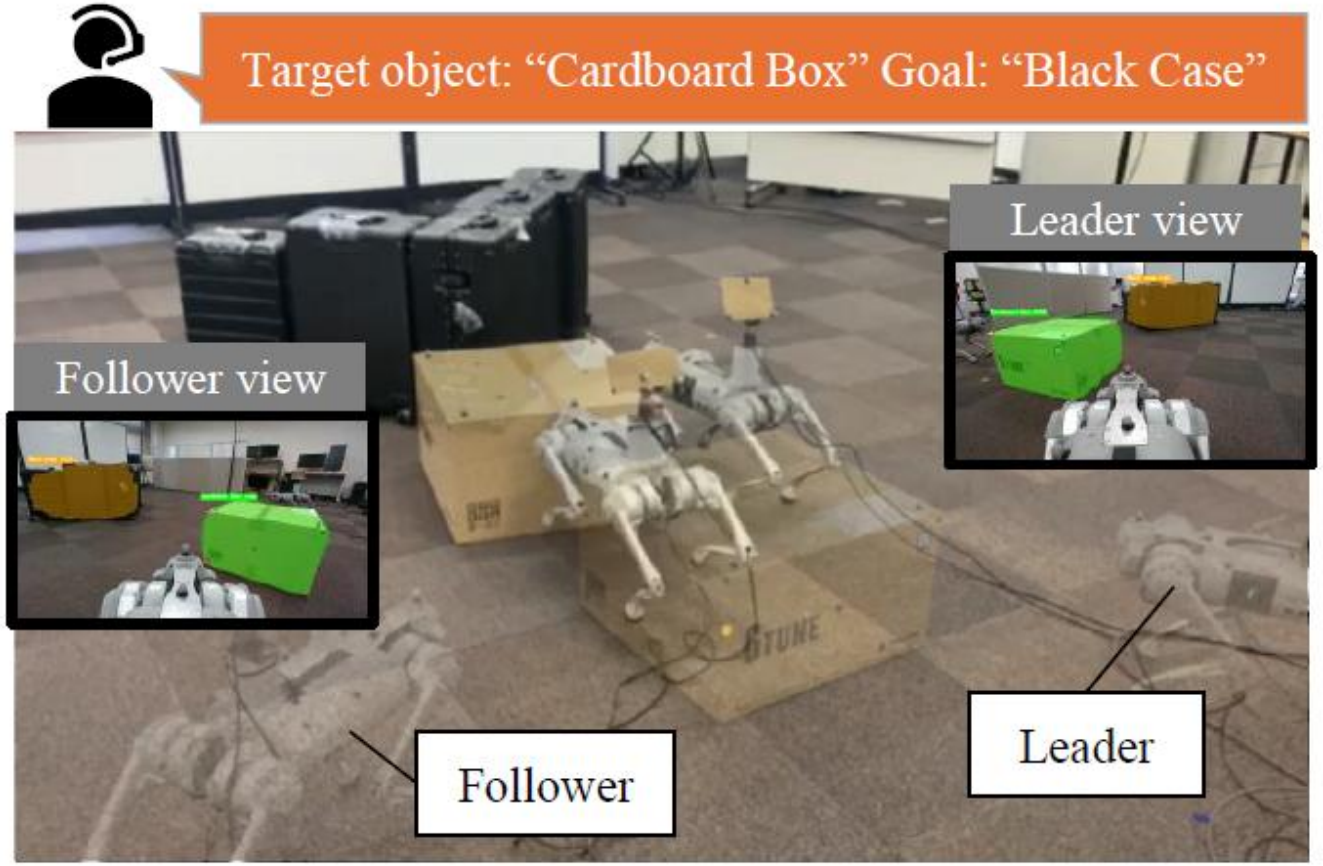
Yoshiki Yano[1], Kazuki Shibata[1], Maarten Kokshoorn[1,2] and Takamitsu Matubara[1]

1 Nara Institute of Science and Technology (NAIST)
2 Delft University of Technology

VLMを用いたリーダー・フォロワー戦略の強化学習

[Despature+, arXiv2026]

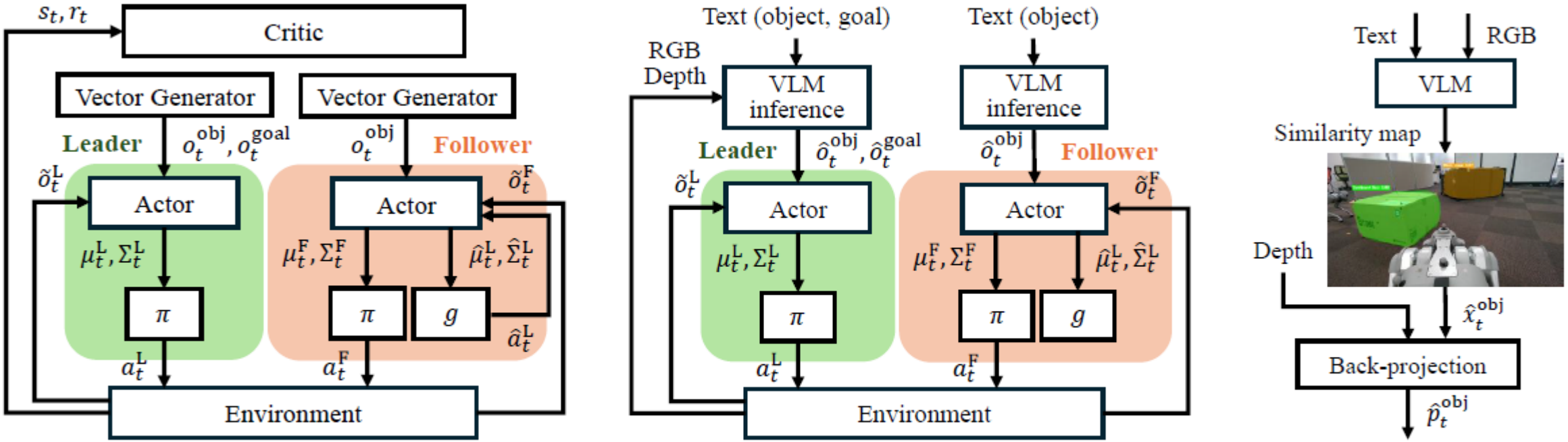
- リーダー：視覚と言語指示（目標）に基づいて行動
- フォロワー：視覚のみに基づいて行動（目標は知らない）
- 分散協調方策を強化学習によって獲得したい



VLMなし訓練・VLMあり実行フレームワーク

[Despature+, arXiv2026]

- VLMからの出力を対象物と目標の位置に設定
- 訓練時はカノニカル物体・目標で学習（多様性排除）



(a) VLM-free training　(b) VLM-based execution　(c) VLM inference

Fig. 2: Overview of the CoLF framework for vision-language-guided multi-robot cooperative transport.

実験結果

[Despature+, arXiv2026]

CoLF: Learning Consistent Leader–Follower Policies for Vision-Language-Guided Multi-Robot Cooperative Transport

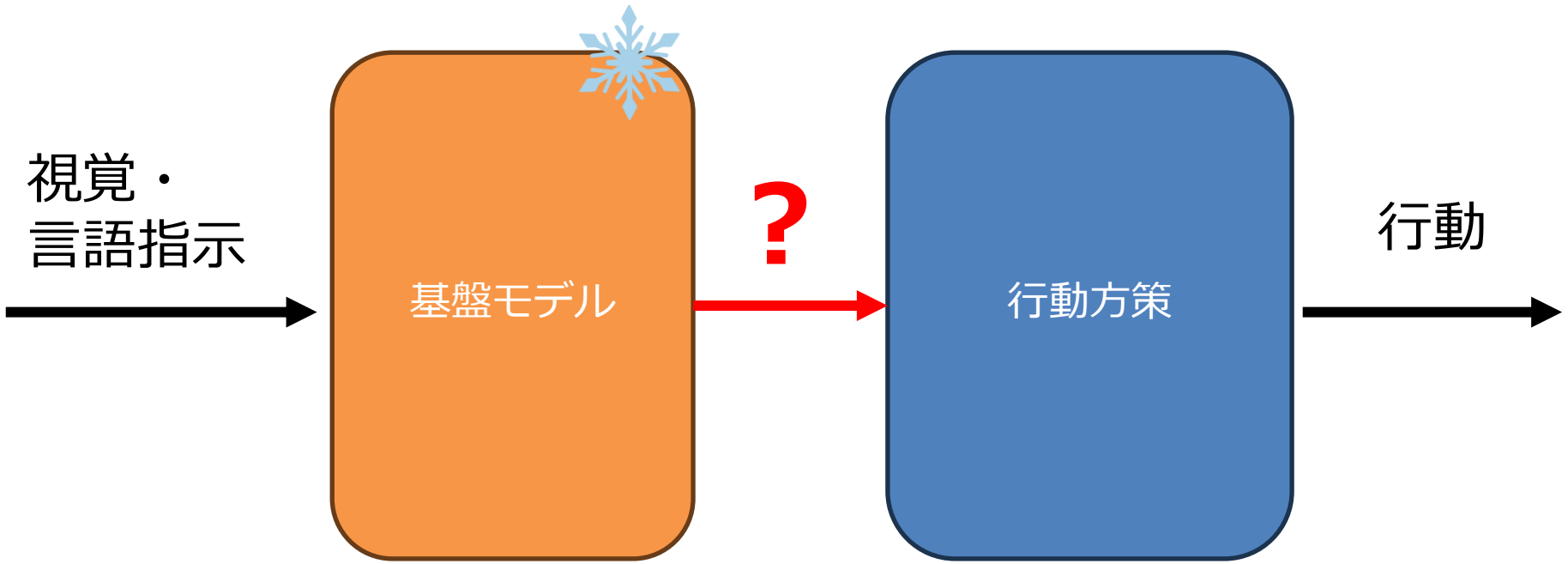
Anonymous Authors

1

VLMにより物体認識を分離することでカノニカル物体で学習した方策を多様な物体に展開できる

まとめ: 基盤モデルによるロボット学習の高度化

- ロボット学習のボトルネックはデータ
- 基盤モデルは「行動方策」としては成立しない
- 「認識モジュール」として分離して使うべき
  → 分離設計がスケーラブルなロボット学習の鍵



謝辞

Nara Institute of Science and Technology NAIST

NAIST Robot Learning Laboratory

松原崇充の自己紹介

山口生まれ，兵庫育ち
神戸高専（電子）→　大阪府立大編入（パワエレ）→　NAIST（情報）

本務

- 奈良先端科学技術大学院大学先端科学技術研究科 教授 (2022.4-現在)

兼務

- ATR脳情報研究所 客員研究員（2008.4-）
- 福島国際研究教育機構 客員研究員（2025.7-）

学会活動

- 日本ロボット学会: 2023-2024 理事（国際），2025-代議員
- IEEE ICRA: 2023-2025 Editor (Robot Learning)
- IJRR 2023- Associate Editor
- Neural Networks 2023- Associate Editor

松原の興味：フィジカルAIと実世界応用

様々な企業様とのコラボレーションを通じて、強化学習を実社会へ応用

化学プラントの自動操業

https://www.yokogawa.co.jp/news/press-releases/2022/2022-03-22-ja/

AIが設計した制御ルールにより35日間連続運転

Zhu+CEP2020, Zhu+CCE2022

小型船舶の自動航行

現場で数十分の試行錯誤によりナビゲーション・定点保持に成功

Cui+JFR2020

ゴミクレーンの自動運転

現場での数十回の試行錯誤によりゴミのばらまき性能改善

Sasaki+RAL2020, Kwon+Access2023

最近はSim2realのアプローチに注力

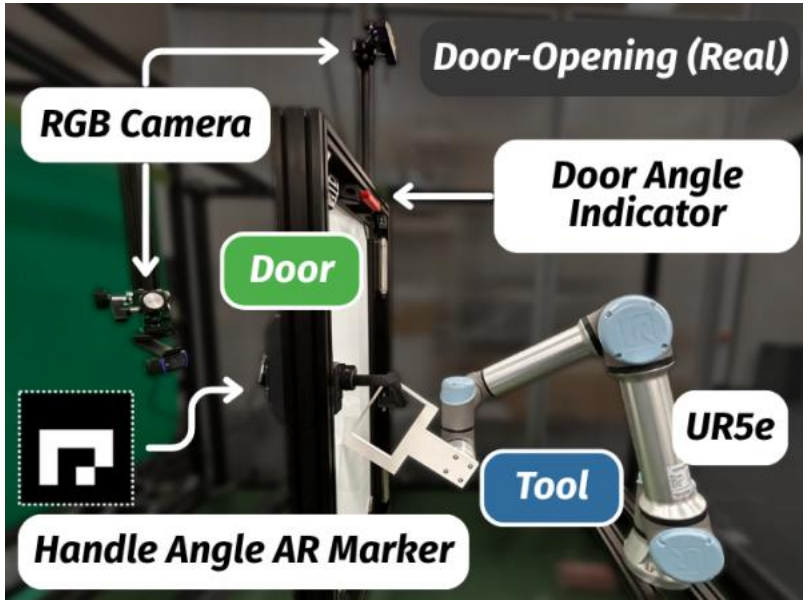
1. 油圧ショベル掘削作業

Kadokawa+RAS23
Kadokawa+T-ASE25

2. 粗研削作業

Hachimine+RAL-ICRA24
Hachimine+RAL-IROS25

3. 使用寿命を考慮した道具作業



Wu+IEEE-Access26