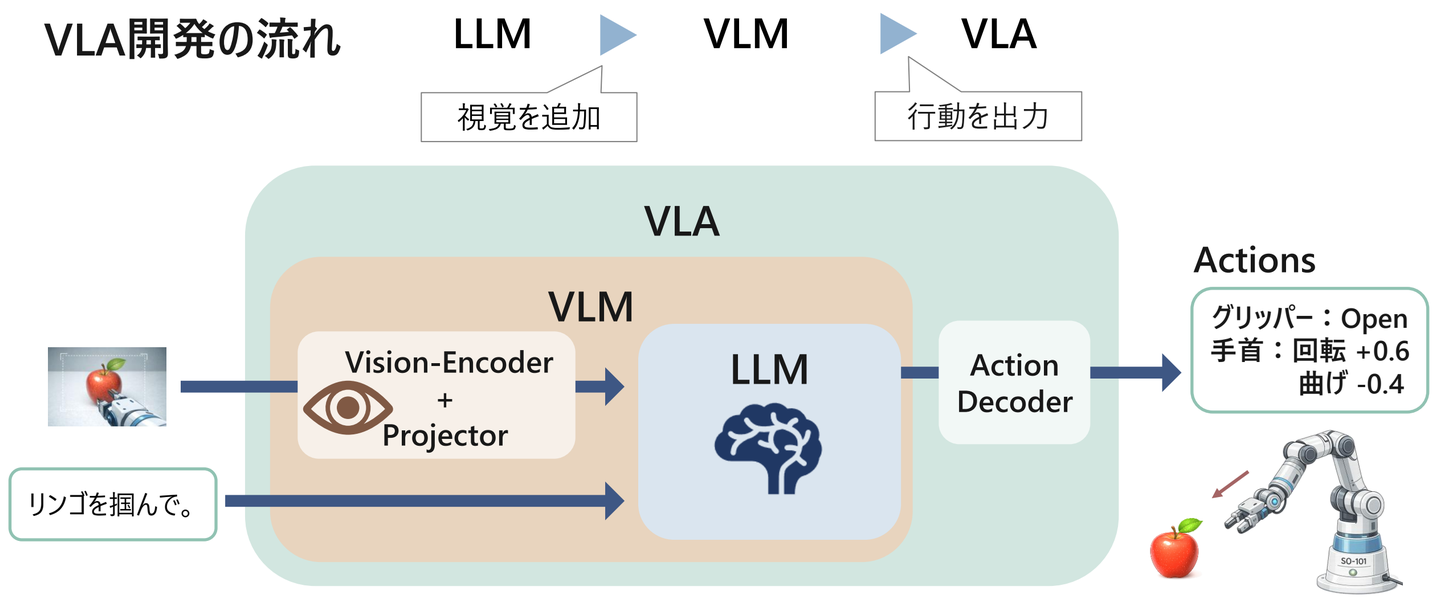
FT-LLM 2026 自由型タスク

日本語ロボット基盤モデルの開発

RAMEN （Team035）

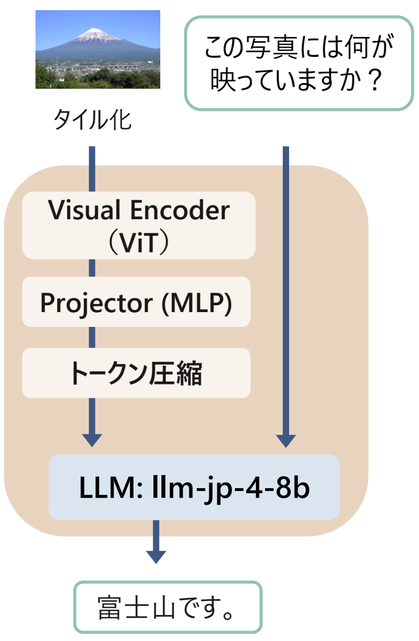
背景

- **Vision-Language-Action（VLA）モデル**：画像と言語指示を入力とし、ロボットの行動を生成するモデル
- 国内におけるVLA開発の知見はまだ少ない
  →・先進的な取り組みとしてVLA開発に挑戦
  ・知見を共有し、国内におけるVLA開発の促進へとつなげる
- 8BはVLAとしてはモデルが大きいため、llm-jp-3.1-1.8b-instruct4についても学習



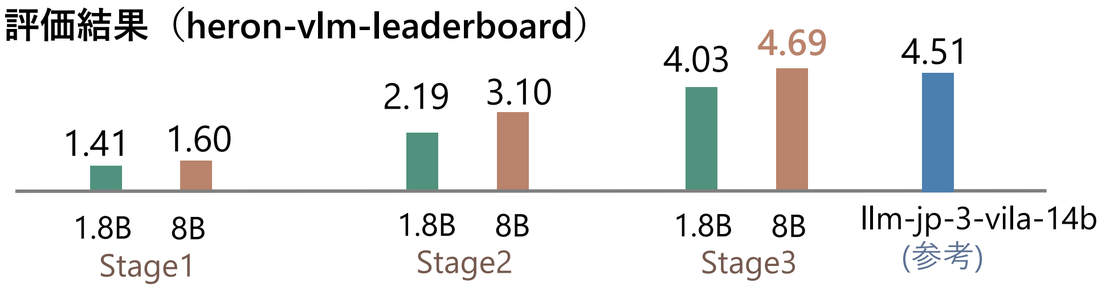
VLM（Vision-Language-Model）開発

NVILAアーキテクチャを採用し、３ステージで学習



	Stage1: Alignment	Stage2: Pretrain	Stage3: SFT
学習対象モジュール	Projector	Projector LLM	Visual Encoder Projector LLM
データ	Ja image-text pairs LLaVA-Pretrain 約1.1M 件	Ja image text pairs Ja interleaved data mmc4-core 約12M 件	llava-instruct-ja ja-photos-conversation ja-vg-vqa-conversation synthdog-ja, LLaVA Instruct 約1M 件

評価結果（heron-vlm-leaderboard）



8Bモデルにおいて、llm-jp-3-vila-14bと比較して高スコアを確認

出力例

この映像の中で明らかに人間ではないのはどれでしょうか？

↓

VLM: llm-jp-4-8b-nvila

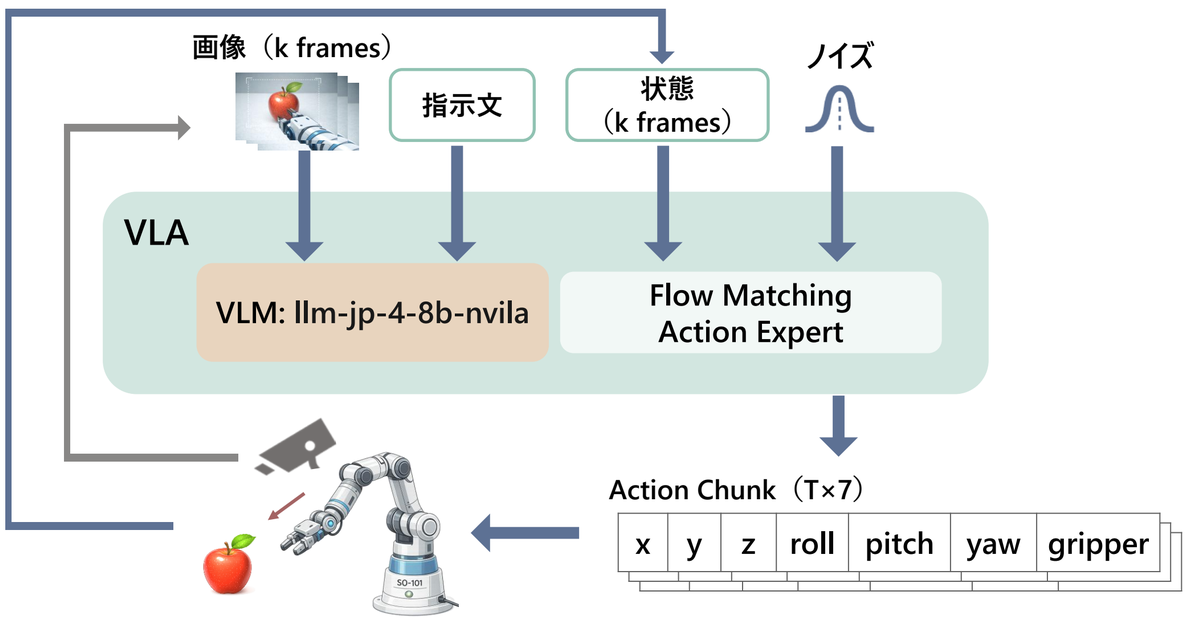
↓

映像の中で明らかに人間ではないのは、白い顔の影だけが見え、口や鼻がなく、手も腕も無い存在です。これは、アニメ「千と千尋の神隠し」に登場するキャラクター「カオナシ」に似た特徴を持っています。

VLA（Vision-Language-Action）開発

VILAFlowPolicy

VILA系VLMベースのVLAポリシーをLeRobot上に実装



学習

学習対象

- LLM (LoRA)
- Flow Matching Action Expert

以下のデータを日本語に翻訳して使用

- **rail-berkeley/bridge_data_v2**:
  Berkeley/Stanford が収集した実機ロボットデータ
- **jxu124/OpenX-Embodiment**:
  Google の RT-1 学習用データセット

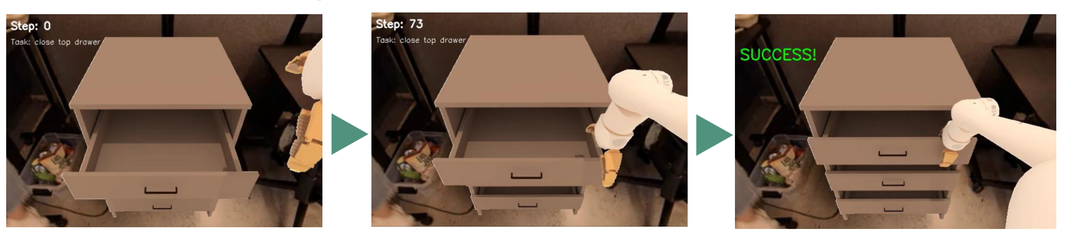
評価

SimplerEnv によるシミュレータ評価（10ep）
指示文を日本語に翻訳し、日本語VLAとしての挙動を確認

タスク	成功率（1.8B）	成功率（8B）
google_robot_close_top_drawer	60%	0%
google_robot_open_top_drawer	10%	0%
google_robot_pick_coke_can	20%	0%
google_robot_move_near	20%	0%
widowx_spoon_on_towel	10%	0%
widowx_carrot_on_plate	0%	0%

8Bは一般的なVLAモデルに比べてサイズが大きいため、
十分な学習量を確保できなかった可能性がある

成功例: close top drawer



まとめ

- LLM→VLM→VLAまで一連の開発パイプラインを構築
- llm-jp-4-8b及びllm-jp-3.1-1.8b-instruct4を基盤とした日本語VLAを開発し、1.8Bモデルで複数のタスクに成功
- 今後の展望：性能向上、実機での検証