卒　業　研　究　論　文
（2025 年度）

題　　目　：低コスト移動ロボットにおける
VLM を用いたマップレス・ターゲット追従の
性能評価

指導教員　：保坂 忠明

学生番号　：1512221209

氏　　名　：小坂 尚璃

明治大学理工学部 電気電子生命学科

目次

第 1 章

序論

1.1 研究背景

近年，画像と言語を同時に扱う大規模モデル（本稿では Vision-Language Model，VLM と呼ぶ）が普及し，ロボットの目標物認識や移動制御への応用が進んでいる。教育用キットや配送・警備などの現場では，低コストかつ短時間でセットアップできる移動ロボットが求められる。マップレスを「地図を作成・保持・参照せず，単一カメラ画像を直接入力として進行方向を決定する構成」と定義すると，初期セットアップが短時間で済み，環境変更時のリセット負荷が小さい利点がある。一方で，視覚ノイズや照度変化により進行方向が乱れやすく，推論遅延が大きい場合は過去フレーム由来の誤判断が累積する弱点が残る。低コスト移動ロボットではこのシンプルな構成が現実的であり，性能限界を把握することが実装指針として重要である。

1.2 先行研究と課題

SayCan は大規模言語モデルを用いた行動計画を提案し [1]，PaLM-E はマルチモーダル化で計画精度を高めた [2]。RT-2 は画像と言語から直接行動を出力し [3]，VL-Maps は意味情報付き地図で探索性能を向上させた [4]。これらは地図や外部記憶を併用するため高性能だが，準備時間と計算資源を要し，Raspberry Pi のような低価格プラットフォームへの適用は容易でない。マップレスかつ低コストの条件で VLM を評価した報告は少なく，走行安定性に関わるボトルネックは十分に整理されていない。教育・実験向けに普及している小型ロボットでは，推論を外部 PC にオフロードする構成が現実的であり，その場合の遅延と視覚品質の影響を測る必要がある。

1.3 研究目的と評価設定

本研究の目的は，地図や複雑なセンサ統合を用いない安価な構成において VLM 単体制御が達成できる性能を定量化し，失敗要因を明確にすることである。具体的には，OSOYOO Raspberry Pi Car を用いて映像を PC 側で推論し，ロボットは視覚判断のみで「黄色いターゲットボックス（以下，ターゲット）に接近し，前方 20 cm 以内で停止する」タスクを遂行する。マップレス・ターゲット追従に限定し，装置の簡素さを保ったまま実環境での走行指標を測定する。

1.4　貢献

本稿の貢献は二点である。

- ダイレクト制御とハイブリッド制御を同条件で比較し，成功率・到達時間・コマンド適合率・反転回数・推論遅延に基づく限界を示す。
- 失敗事例を分類し，遅延・誤認識・コマンド反転といった要因ごとに改善優先度を整理する。

1.5　論文構成

本論文の構成を示す。第 2 章でハードウェアと制御条件を述べる。第 3 章で実験環境と評価指標を説明し，第 4 章で暫定結果を示す。第 5 章で考察を行い，第 6 章で結論と今後の課題を述べる。付録では補助的な試行結果を提示する。

第 2 章

手法

2.1 ハードウェア構成

対象機体は OSOYOO Raspberry Pi Robot Car（2023 版）である。Raspberry Pi 4B，広角 USB カメラ，モータドライバ（L298N），ステアリング用サーボから構成し，カメラ解像度は 640×480，フレームレートは 5 fps とする。走行は室内の平坦な床面で行う。

2.2 通信とソフトウェア構成

ラズパイ上で ‘raspi_agent.py‘ を 8080 番ポートで起動し，‘GET /frame.jpg‘ と ‘/stream.mjpg‘ によるフレーム配信，‘POST /command‘ による ‘FORWARD/LEFT/RIGHT/BACKWARD/STOP‘ コマンド受信を提供する。PC 側は同一ネットワークから HTTP でアクセスし，VLM の推論結果に基づきコマンドを送信する。自宅用と研究室用の IP を環境変数で切り替え，実験時の誤接続を防ぐ。

2.3 使用モデルとプロンプト

主モデルは GPT-5-mini-20250807（ローカル実行）である。単一フレームを入力し，「ターゲットに到達して停止せよ」と指示した上で，選択肢を ‘LEFT/RIGHT/FORWARD/FORWARD_SLOW/STOP/BACKWARD‘ の一語に限定する。SmolVLM 系モデルも比較のために試行し，結果を付録にまとめる。

2.4 制御条件

- **条件 A（ダイレクト制御）**：VLM が各ループで移動コマンドを直接選び，そのままロボットへ送信する。推論に 5–7 秒要するため，ループ間隔は 10 秒とし，速度は固定とする。
- **条件 B（ハイブリッド制御）**：VLM は「ターゲット位置（LEFT/CENTER/RIGHT）と距離（FAR/MID/NEAR）」のみを JSON で返し，PC 側の単純な規則でコマンドを決定する。規則は「CENTER なら前進，NEAR なら減速または停止，LEFT/RIGHT なら旋回」とする。推論遅延を考慮しループ間隔は 10 秒に揃える。

両条件とも VLM 出力と単純規則のみで挙動を決める。

2.5　比較観点と改善優先度

ダイレクト制御とハイブリッド制御を比較し，誤りが生じやすい工程を二段階に分けて評価する。まず，VLM が画像から位置と距離を正しく判断できるかをハイブリッド制御で測定し，誤認識が多い場合は入力前処理や照度制御を優先して改善する。次に，判断は正しいがコマンド生成で逆転や過大旋回が起きる場合は，ダイレクト制御のプロンプト設計やコマンド選択肢の絞り込みを優先する。この結果から，外部記憶の導入や予測モデルの追加が必要となる箇所（判断工程か指令工程か）を明確にし，拡張設計の順序を定める。

第 3 章

実験

3.1　環境とコース

室内直線コース（長さ 2.5 m）と L 字コース（1.5 m ＋ 1.5 m）を作成し，照度は 400–500 lx に保つ。壁や障害物は固定し，背景テクスチャを一定にする。ターゲットは黄色いターゲットボックス（“TARGET”と印字）を終点に配置する。

3.2　手順と試行数

開始位置はターゲットから約 2.5 m 離し，左右位置や向きを少しずつ変えながら複数回試行する。各条件について各コース 10 試行ずつ，計 40 試行を行う。走行中はフレーム画像，VLM 出力，送信コマンド，タイムスタンプを保存し，別カメラで全体動画を記録する。試行には通し番号を付け，コース種別，照度，モデルバージョン，開始・終了時刻をメタ情報として残す。

3.3　評価指標

- 到達時間：開始コマンド送信から停止判定までの時間。
- 成功率：ターゲット前 20 cm 以内で停止した試行の割合。
- コマンド適合率：全送信コマンド中，理想方向・速度に合致すると評価できる割合。
- コマンド反転回数：左右または前後の即時反転の回数。
- 推論遅延：VLM 応答時間の中央値と 95 パーセンタイル。

第 4 章

結果（暫定）

直近のログには試行ごとの到達可否，コマンド系列，推論遅延が保存されている。現在集計中であり，確定値を表 4.1 に示す予定である。表では各条件・各コースの平均到達時間，成功率，コマンド適合率，反転回数中央値，推論遅延中央値／ 95 パーセンタイルを掲載する。コマンド時系列と軌跡オーバーレイを図 4.1，図 4.2 として挿入する計画である。SmolVLM 系モデルは到達率が低く，誤旋回と停止不能が多発したため，詳細を付録で報告する。

表 4.1　各条件の暫定集計（後日更新）

条件	コース	到達時間 [s]	成功率	適合率	反転中央値
ダイレクト	直線	–	–	–	–
ダイレクト	L 字	–	–	–	–
ハイブリッド	直線	–	–	–	–
ハイブリッド	L 字	–	–	–	–

図 4.1　軌跡オーバーレイ（後日実データに置換）

図 4.2　コマンド系列と距離推定の時系列（後日実データに置換）

第 5 章

考察（予定）

予備観察では，ダイレクト制御は推論遅延と誤指示によりコマンド反転が増え，到達時間がばらつく傾向があった。ハイブリッド制御は距離推定の粗さが残るものの，左右誤指示の影響を抑え，成功率が高まると見込む。VLM 単体制御の主要課題は，(1) 応答遅延がループ間隔に近いこと，(2) 単一フレーム依存による瞬間的な誤認識，(3) 照度変動への感度である。これらに対し，フレームスタックによる時系列文脈の付与，簡易外部記憶による状態保持，軽量予測モデルによる補正が有効と考える。

第 6 章

結論

本研究は，低コスト移動体で地図や複雑なセンサ統合を用いない構成において，VLM 単体制御が達成できる性能を評価する枠組みを示した。ダイレクト制御とハイブリッド制御を同条件で比較し，到達時間・成功率・コマンド適合率を中心指標，反転回数と遅延分布を補助指標として計測する計画を立てた。今後，集計したログを反映して結果と考察を更新し，外部記憶や予測モデルを組み込む際の具体的設計指針を示す。

謝辞

本研究の遂行にあたり，実験環境整備に協力いただいた研究室メンバー各位に感謝する。計算資源と機材提供に謝意を表する。

参考文献

[1] Anthony Brohan et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of Robotics: Science and Systems*, 2022.

[2] Danny Driess et al. Palm-e: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[3] Anthony Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, 2023.

[4] Wenlong Huang et al. Vl-maps: Visual language maps for zero-shot object navigation. In *Proceedings of Robotics: Science and Systems*, 2023.

付録 A

SmolVLM 失敗試行の概要

SmolVLM2-mlx を用いて同一タスクを試行したところ，推論時間は平均 0.8 s でループ間隔 0.5 s を上回った。その結果，過去フレームに基づく誤指示が頻発し，10 試行中の成功は 2 回に留まった。誤指示の内訳は左右の即時反転 12 件，停止不能 4 件であった。今後はモデルの軽量化またはフレーム間補間の導入を検討する。