深層状態空間モデルおよびtransformerの並列機構による物体操作前アラインメント予測

Pre-Manipulation Alignment Prediction with Parallel Deep State-Space and Transformer Models

神原 元就　　杉浦 孔明
Motonari Kambara　Komei Sugiura

慶應義塾大学
Keio University

In this work, we address the problem of predicting the future success of open-vocabulary object manipulation tasks. Conventional approaches typically determine success or failure after the action has been carried out. To overcome these challenges, we propose a model, which predicts the alignment between a pre-manipulation egocentric image with the planned trajectory and a given natural language instruction. We introduce a Multi-Level Trajectory Fusion module, which employs a state-of-the-art deep state-space model and a transformer encoder in parallel to capture multi-level time-series self-correlation within the end effector trajectory. Our experimental results indicate that the proposed method outperformed existing methods, including foundation models.

1. はじめに

物体操作は，生活支援 [Wu 23] や農業 [Lehnert 17] など，さまざまな分野で不可欠である．これらの操作タスクにおいて，成否を判定することは，操作の失敗による危険を防ぎ失敗後の冗長な作業を回避することに繋がり，効率性と安全性を向上させる．

本研究では，指示文，一人称視点画像，および事前に生成された軌道より，その軌道が指示文に対して整合しているか否かを予測することに焦点を当てる．このタスクでは，計画軌道がグリッパの開閉状態に関する情報を含んでおり，全体の複雑性が大幅に増大する．したがって，本タスクは，対象物体とエンドエフェクタの将来的なインタラクションを適切に予測する必要があるため，特に困難である．

このタスクは，物体操作全体の効率を向上させる可能性があるため，重要である．[Goko 24] を含む多くの既存手法では，操作が完了した後に成否を予測する．これらのアプローチは，本質的に操作完了後に結果を評価することに依存しており，操作中のリスクを軽減することが困難である．さらに，実行の失敗に基づきタスクの再計画を行う方法では，タスク全体の効率的な実行が難しい．より望ましいアプローチとしては，画像と自然言語指示の両方に基づき，与えられた軌道の整合性を事前に評価し検証することが挙げられる．この事前検証は，コストの高い実行エラーを防ぎ，タスク完了の全体的な効率を向上しうる．

Vision-Language-Action (VLA) モデル [Brohan 22, Kim 24] は，自然言語指示文及びカメラ画像を入力として用いている一方で，実際には十分な整合性を考慮できないことが多い．例えば，優れた VLA モデルの一つである OpenVLA [Kim 24] は，ゼロショット操作タスクにおいてほぼ 0%の成功率であった [Black 24].

本研究では，自然言語指示，一人称視点画像，および事前に生成された軌道の間の整合性を予測するモデルを提案する．また，軌道からマルチレベルの時系列情報を取得するための Multi-Level Trajectory Fusion (MTF) モジュールを導入する．この MTF モジュールは，比較的短い系列間の相関を捉えるのに適した transformer と，treansformer よりも長い系列を効果的に処理できると報告されている最新の状態空間モデル（例：[Gu 22]）を並列に使用する．このアーキテクチャにより，対象物とエンドエフェクタの間のインタラクションを画像および軌道から適切に判断できる．その結果，多くの既存手法とは異なり，操作の実行前に自然言語指示，画像，および計画軌道の整合性を予測することが可能となる．本研究の主な貢献は以下のとおりである．

- 自然言語指示文，軌道，および一人称視点画像の整合性を実行前に検証するモデルを提案する．
- 深層状態空間モデルと transformer エンコーダを並列に使用し，エンドエフェクタ軌道内のマルチレベル時系列自己相関を捉えるための MTF モジュールを導入する．
- 自然言語指示，カメラ画像，およびエンドエフェクタ軌道を効果的に整合させるための Trimodal Cross-Integration モジュールを導入する．

2. 問題設定

本研究は，オープンボキャブラリ物体操作における事前の操作整合性予測（Pre-Manipulation Alignment Prediction, PMAP）に焦点を当てる．このタスクでは，自然言語指示文，一人称視点画像，および事前に生成された軌道の整合性を検証することを目的とする．事前操作画像，軌道，指示文が与えられた場合，モデルはそれらが相互に整合しており，成功する可能性が高い操作であるかを判断する必要がある．図 1 に PMAP タスクの典型的な例を示す．

このエピソードでは，‘Pick the apple from the white bowl” という指示が与えられる．この場合，ロボットがリンゴを白いボウルから把持できた場合は ‘Aligned’，白いボウルを移動させた場合は ‘Hallucinated’ と予測することが求められる．

PMAP タスクでは，入力は一人称視点画像，エンドエフェクタの軌道，および自然言語指示文で構成される．期待される出力は，入力が十分に整合しており，成功する可能性が高いことを示す確率 $p(\hat{y}=1)$ である．y および $\hat{y}$ はラベルと予測ラベルをそれぞれ示し，$y=1$ は軌道，一人称視点画像，および指示文が整合していることを意味する．入力画像は操作直前に

連絡先: 神原元就，慶應義塾大学，神奈川県横浜市港北区日吉 3-14-1，motonari.k714@keio.jp

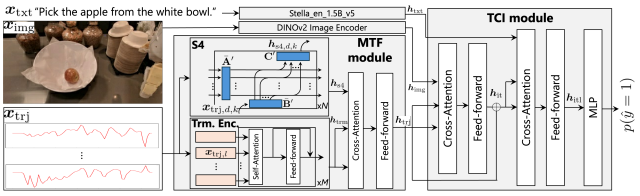

図 1: 提案手法のネットワーク図. 図において, 'MTF module', 'TCI module', N, and M はそれぞれ, Multi-Level Trajectory Fusion module, the Trimodal Cross-Fusion module, S4 ブロックの個数, 及び transformer の層数を示す.

撮影された 1 枚のみである. したがって, モデルはこの単一の事前操作画像と軌道データから最終状態を推測する必要があり, それゆえに本タスクは困難である. 本研究では, 軌道生成は扱わない.

3. 提案手法

図 1 に提案手法の概要を示す. 本手法は, Multi-Level Trajectory Fusion (MTF) モジュールと Trimodal Cross-Integration (TCI) モジュールで構成される.

モデルの入力 $\boldsymbol{x}$ は以下のように定義される.

$$\boldsymbol{x} = \{\boldsymbol{x}_{\text{txt}}, \boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{trj}}\} \tag{1}$$

ここで, $\boldsymbol{x}_{\text{txt}}$, $\boldsymbol{x}_{\text{img}}$, および $\boldsymbol{x}_{\text{trj}} \in \mathbb{R}^{D\times T}$ は, それぞれ自然言語指示, カメラ画像, およびエンドエフェクタ軌道を表す. また, D はマニピュレータの自由度, T は時系列の長さを示す.

3.1 Multi-Level Trajectory Fusion モジュール

Multi-Level Trajectory Fusion (MTF) モジュールは, 学習可能な重みを用いてエンドエフェクタ軌道をフィルタリングする. このフィルタリングにより, 時間的な軌道圧縮を行い, より効率的な表現と処理を可能にする.

このモジュールの入力は $\boldsymbol{x}_{\text{trj}}$, 出力は $\boldsymbol{h}_{\text{trj}}$ である. 本モジュールは transformer エンコーダ, Structured State Space Sequence Model (S4) モジュール [Gu 22], およびクロスアテンション機構で構成される. ロボット分野では, transformer に基づく手法が軌道の時系列依存関係を捉えるために頻繁に用いられている (例：[Brohan 23]). 一方で, 長い時系列信号に対しては, 深層状態空間モデル (deep SSMs) が時系列依存関係の把握において transformer を上回る性能を示すことが知られている [Gu 22]. エンドエフェクタ軌道の長さは制御周波数と全体の実行時間の積で決定される. より複雑な物体操作では, 完了までに時間を要するため, 軌道の長さが増大する. そのため, 図 1 に示すように, 本モジュールでは, transformer を各サブ軌道に適用しつつ, 深層 SSM を軌道全体に適用することで, 異なる階層の時系列依存関係を抽出する.

3.1.1 Structured State Space Sequence Model (S4)

深層 SSM [Gu 22, Smith 23] は, 長い時系列データを扱うための強力な手法として近年注目されている. これらのモデルは, 従来の SSM 理論 [Funahashi 93, Tallec 18] を基盤としつつ, RNN や CNN に着想を得た手法を取り入れることで, 長い時系列データの依存関係を効率的に捉える.

連続系における入力信号 $x_{\text{in}}(t) \in \mathbb{R}$ は, 隠れ状態 $\boldsymbol{h}(t) \in \mathbb{R}^{H\times 1}$ を介して出力信号 $y_{\text{out}}(t) \in \mathbb{R}$ に変換される.

$$\begin{aligned} \boldsymbol{h}'(t) &= \mathbf{A}\,\boldsymbol{h}(t) + \mathbf{B}\,x_{\text{in}}(t), \\ y_{\text{out}}(t) &= \mathbf{C}\,\boldsymbol{h}(t) + D\,x_{\text{in}}(t), \end{aligned} \tag{2}$$

ここで, $\mathbf{A} \in \mathbb{R}^{H\times H}$, $\mathbf{B} \in \mathbb{R}^{H\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times H}$, および $D \in \mathbb{R}$ は学習可能なパラメータである.

S4 [Gu 22] は, 離散 SSM を畳み込みとして再表現することで, 並列化を可能にする. 具体的には, 構造化されたカーネル $\bar{\mathbf{K}}$ を構築し,

$$\boldsymbol{y}_{\text{out}} = \boldsymbol{x}_{\text{in}} * \bar{\mathbf{K}}, \tag{3}$$

とすることで, 時系列全体にわたる並列処理を実現する.

3.1.2 モジュールアーキテクチャ

S4 モジュールは, N 個の S4 ブロックから構成される. 各 S4 ブロックは Eq. 2 に基づいて動作する. 軌道 $\boldsymbol{x}_{\text{trj}}$ は L ステップごとに分割され, transformer エンコーダへ入力される.

最終的に, $\boldsymbol{h}_{\text{s4}}$ と $\boldsymbol{h}_{\text{trm}}$ のクロスアテンションを計算し, $\boldsymbol{h}_{\text{trj}}$ を出力する.

3.2 Trimodal Cross-Integration モジュール

PMAP タスクでは、自然言語指示、画像、軌道の整合性を予測する必要がある。この整合性を予測するためには、入力画像内の各オブジェクトの位置情報を理解し、計画された軌道が指示された操作と対応しているかを判断する必要がある。

そこで、本研究では Trimodal Cross-Integration (TCI) モジュールを導入する。このモジュールは、自然言語指示、軌道、および画像情報の整合性を計算し、観測環境内のエンドエフェクタ軌道が自然言語指示で指定されたタスクを成功裏に達成できる確率を生成する。

TCI モジュールは、2 つのトランスフォーマーデコーダと MLP レイヤーで構成される。このモジュールの入力は $\boldsymbol{h}_{\text{txt}}, \boldsymbol{h}_{\text{img}}, \boldsymbol{h}_{\text{trj}}$ であり、出力は $\boldsymbol{x}_{\text{txt}}, \boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{trj}}$ の整合性の予測確率 $p(\hat{y})$ である。

ここで、$\boldsymbol{h}_{\text{txt}}$ と $\boldsymbol{h}_{\text{img}}$ は、Stella_en_1.5B_v5 [ste 24] により埋め込まれた言語特徴、および DINOv2 [Oquab 24] により抽出された画像特徴である。

まず、画像特徴 $\boldsymbol{h}_{\text{img}}$ と軌道特徴 $\boldsymbol{h}_{\text{trj}}$ のクロスアテンションを計算する。

$$\boldsymbol{h}_{\text{it}} = \text{CrossAttn}\big(\boldsymbol{h}_{\text{trj}}, \boldsymbol{h}_{\text{img}}\big), \tag{4}$$

ここで、$\text{CrossAttn}\big(\boldsymbol{X}_a, \boldsymbol{X}_b\big)$ は以下のように定義される。

$$\text{CrossAttn}\big(\boldsymbol{X}_a, \boldsymbol{X}_b\big) = \text{softmax}\left(\frac{\boldsymbol{X}_a\boldsymbol{W}_q(\boldsymbol{X}_b\boldsymbol{W}_k)^\top}{\sqrt{d_k}}\right)\boldsymbol{X}_b\boldsymbol{W}_v, \tag{5}$$

ここで、$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d_{\text{in}}\times d_{\text{out}}}$ は学習可能な重みであり、d_k は $\boldsymbol{X}_b\boldsymbol{W}_k$ の次元数である。結果として得られる $\boldsymbol{h}_{\text{it}}$ は、画像と軌道モダリティ間の整合性を反映する特徴である。

次に、自然言語特徴 $\boldsymbol{h}_{\text{txt}}$ と $\boldsymbol{h}_{\text{it}}$ のクロスアテンションを計算する。

$$\boldsymbol{h}_{\text{itl}} = \text{CrossAttn}\big(\boldsymbol{h}_{\text{txt}}, \boldsymbol{h}_{\text{it}}\big), \tag{6}$$

ここで、$\boldsymbol{h}_{\text{itl}}$ は、$\boldsymbol{h}_{\text{it}}$ と $\boldsymbol{h}_{\text{txt}}$ の整合性を反映する特徴であり、物体操作に関連する自然言語情報が、軌道を条件とした環境情報と整合しているかを表す。

最後に、全結合層を用いて $\boldsymbol{h}_{itl}$ から $p(\hat{y})$ を得る。

$$p(\hat{y}) = \text{softmax}\big(\boldsymbol{W}_o\boldsymbol{h}_{\text{itl}} + \boldsymbol{b}_o\big), \tag{7}$$

ここで、$\boldsymbol{W}_o$ および $\boldsymbol{b}_o$ は学習可能なパラメータである。

4. 実験設定

4.1 データセット

SP-RT-1 データセット [Goko 24] を拡張し, SP-RT-1-Traj データセットを構築した. SP-RT-1 データセットは, RT-1 デー

表 1: SP-RT-1-Traj データセットにおける定量的結果．太字は最良の精度を示す．

Method	Accuracy [%]
Qwen2-VL [Wang 24]	52.1 ± 0
GPT-4o [OpenAI 24]	69.4 ± 0.64
Contrastive λ-Repformer [Goko 24]	77.7 ± 0.65
Ours	**80.1** ± 0.84

タセット [Brohan 22] に基づいて構築され，物体操作前後の画像，物体操作指示文，各エピソードの成功/失敗ラベルを含む．一方，PMAP タスクでは，物体操作前の一人称視点画像，エンドエフェクタ軌道，自然言語指示，およびタスク結果ラベル（成功/失敗）が必要となる．そのため，PMAP タスクに必要な情報を含む SP-RT-1 データセットを基に SP-RT-1-Traj データセットを構築した．SP-RT-1 データセットには各エピソードのエンドエフェクタ軌道が含まれていなかったため，PMAP タスクのために RT-1 データセットから軌道を収集した．データセットの分割は [Goko 24] に従った．

構築したデータセットには 13,915 サンプルが含まれ，訓練，検証，及びテストセットはそれぞれ 11,915，1,000 サンプル，及び 1,000 サンプルであった．データ収集に使用したマニピュレータの動作空間は 7 次元であり [Brohan 22]，さらにエンドエフェクタの開閉を表す 1 次元を追加した．そのため，各タイムステップの軌道は 8 次元となる．

4.2 手法に関する設定

提案手法において，$M = 2$，$N = 4$，$d' = 512$，とした．最適化関数には Adam を用い，学習率は $1e-5$ とした．また，バッチサイズは 64 とした．実験には NVIDIA GeForce RTX 4090（VRAM 24GB），RAM 64GB，Intel Core i9–13900KF を使用した．提案手法の学習時間は約 0.6 時間であり，特徴抽出を除く推論時間はサンプルあたり 0.34ms であった．

4.3 ベースライン手法

本研究では，ベースライン手法として GPT-4o [OpenAI 24]，Qwen2-VL [Wang 24]，Contrastive λ-Repformer [Goko 24] を使用した．

GPT-4o は，大規模データセットで事前学習された代表的なマルチモーダル大規模言語モデルであり，VQA を含む多様なタスクで顕著な性能を示している．また，Qwen2-VL は最先端の VQA モデルである．Contrastive λ-Repformer は，PMAP タスクと関連性の高い Success Prediction for Open-vocabulary Manipulation (SPOM) タスク [Goko 24] において高い性能を示しているため，選定した．

Contrastive λ-Repformer は SP-RT-1-Traj データセットで学習したが，他のベースラインモデルはゼロショットまたは few-shot 設定で実験を行った．具体的には，GPT-4o へのプロンプトには 2 つのサンプルを例として含めることで，GPT-4o は few-shot 設定で使用した．また，Qwen2-VL は複数画像を適切に扱うことが困難であるため，プロンプトにはサンプル例を含めなかった．また，Contrastive λ-Repformer には，物体操作結果の漏洩を防ぐために，操作後の画像を提供しなかった．これらのベースライン手法は軌道を直接入力として扱うことができないため，カメラ画像に軌道をオーバーレイした画像を入力として使用した．

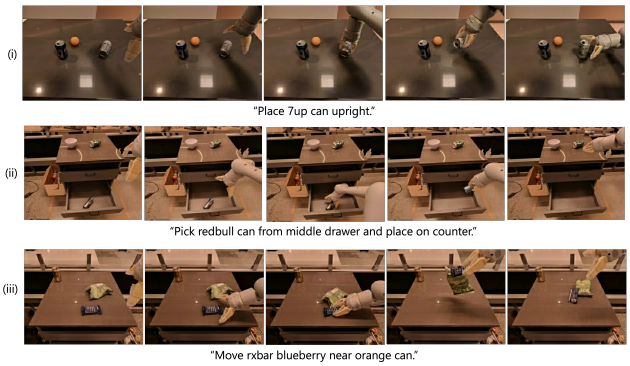

図 2: 定性的結果．パネル (i)，(ii)，及び (iii) はそれぞれ True Positive，True Negative，及び False Positive の例を表す。各パネルにおける最も左の画像は，操作前のシーンを示している。

5. 実験結果

5.1 定量的結果

表 1 に，SP-RT-1-Traj データセット上でのベースライン手法と提案手法の比較結果を示す．各実験は 5 回実施し，各スコアは平均と標準偏差を表す．評価尺度には精度を使用した．

ベースライン手法の精度は，Qwen2-VL が 52.1%，GPT-4o が 69.4%，Contrastive λ-Repformer が 77.7%であった．一方，提案手法は 80.1%の精度を達成し，最良のベースライン手法である Contrastive λ-Repformer を 2.4 ポイント上回った．さらに，この差は統計的に有意であった（$p < 0.01$）．このことから，提案手法が軌道情報を活用して将来の状況を他手法より効果的に予測できることが示唆された．

5.2 定性的結果

図 2 に定性的結果を示す．パネル (i)，(ii)，(iii) は，それぞれ True Positive，True Negative，False Positive の例である．各パネルは特定のエピソードから抽出された画像列を示しており，最も左の画像が物体操作前の画像である．この画像のみがモデルの入力として使用された．

図 2 (i) に示すエピソードでは，指示は “Place 7up can upright.” であった．環境には 2 つの缶があり，マニピュレータは正しい缶を選択し，誤りなく直立させた．そのため，入力情報は適切に整合していた．提案手法はこのサンプルを正しく予測したが，全てのベースライン手法は誤って判定した．この結果は，TCI モジュールがテキストと視覚データを効果的に整合させ，意図された対象物を識別し，その状態を適切に認識できることを示している．

図 2 (ii) では，指示は “Pick redbull can from middle drawer and place on counter.” であった．マニピュレータは引き出しの中から缶を取り出し，カウンター上へ移動させた．しかし，把持位置が最適でなかったため，缶が滑って転がり，正しく配置されなかった．すなわち，この軌道は幻覚的（hallucinated）であると見なされた．提案手法はこの結果を正しく予測したが，GPT-4o と Contrastive λ-Repformer は誤った推論を行った．これは，提案手法が軌道情報を利用して，計画された軌道と一人称視点画像が指示文と整合しているかをより正確に予測できることを示している．

図 2 (iii) では，指示は “Move rxbar blueberry near orange can.” であった．マニピュレータはブルーベリー味のプロテインバーを把持し，オレンジ缶に近づけたが，依然として十分に近いとは言えない距離であった．このため，ラベルは ‘Hallucinated’

表 2: MTF モジュールに関する ablation study．表において，‘Trm.’ は transformer エンコーダを示す．

Model	S4 [Gu 22]	Trm.	Accuracy [%]
(i)	✓	✓	**80.1** ± 0.84
(ii)	✓		79.7 ± 1.15
(iii)		✓	79.9 ± 0.78
(iv)			71.9 ± 1.76

表 3: TCI モジュールにおける軌道特徴と画像特徴のクロスアテンション機構の性能に関する ablation study．‘Cross-Attention’ は、式 (4) の適用を示す．

Model	Cross-Attention	Accuracy [%]
(i)	✓	**80.1** ± 0.84
(ii)		75.4 ± 1.02

となった．提案手法はこのサンプルを誤って予測した．このエピソードでは，複数の物体間の空間関係を認識する必要があり，整合性の予測が困難であった．

5.3 Ablation Study

Ablation study として，以下の 2 条件について行った．

MTF モジュールに関する ablation study: MTF モジュールにおける S4 およびトランスフォーマーエンコーダの影響を調査するため，それぞれを削除した場合の性能を比較した．結果は表 2 に示す．モデル (iv) は，軌道を重み付けするために全結合層のみを適用したモデルである．

表 2 によると，モデル (ii)，(iii)，(iv) の精度は，それぞれモデル (i) と比較して 0.4，0.2，8.2 ポイント低下した．この結果から，S4 とトランスフォーマーエンコーダの双方がわずかながら性能向上に寄与していることが分かる．また，両者の性能差は大きくないものの，いずれかを削除すると精度が低下することから，それぞれの重み付け機構が全結合層よりも有効であることが示唆された．

TCI モジュールに関する ablation study: クロスアテンション機構が性能に与える影響を評価するため，この機構を削除したモデルを検証した．表 3 に示す通り，モデル (ii) の精度はモデル (i) と比較して 4.7 ポイント低下した．この結果は，画像特徴を軌道情報と事前に整合させることで，環境と操作の理解が向上することを示唆している．

6. 結論

本研究では，自然言語指示文，物体操作前の一人称視点画像，および事前に生成されたエンドエフェクタ軌道の整合性を事前に検証することに焦点を当てた．本研究の貢献は以下のとおりである．

- 指示文に指定されたタスクが適切に実行可能かどうかを，軌道と物体操作前画像の整合性に基づいて予測するモデルを提案した．
- 最先端の状態空間モデルである S4 とトランスフォーマーエンコーダを並列に用いることで，エンドエフェクタ軌道内の多階層の時系列自己相関を捉えるための MTF モジュールを導入した．
- 自然言語指示，カメラ画像，およびエンドエフェクタ軌道の三者を効果的に整合させるための Trimodal Cross-Integration モジュールを導入した．
- 実験結果により，提案手法が複数のベースライン手法よりも高い予測精度を達成することを示した．

今後の展望として，提案手法を VLA モデル [Brohan 22,Kim 24] と組み合わせることが考えられる．Sec. 1. で述べたように，多くの既存の VLA モデルはモダリティ間の整合性が不十分であり，不適切な軌道を生成することがある．さらに，多くのモデルは生成した軌道の妥当性を評価する機構を持たず，誤った軌道の実行を防ぐことが難しい．提案手法をこれらの VLA モデルに適用することで，VLA モデルが生成した複数の候補軌道の中から適切な軌道を選択し，実行することが可能となる．これにより，より効果的な物体操作タスクの実行が実現できると考えられる．

謝辞

本研究の一部は，JSPS 科研費 23H03478，JST ムーンショット，JSPS 特別研究員奨励費 JP23KJ1917 の助成を受けて実施されたものである．

参考文献

[Black 24] Black, K., Brown, N., Driess, D., Esmail, A., et al.: π_0: A Vision-Language-Action Flow Model for General Robot Control, *arXiv preprint arXiv:2410.24164* (2024)

[Brohan 22] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., et al.: RT-1: Robotics Transformer for Real-World Control at Scale, *arXiv preprint arXiv:2212.06817* (2022)

[Brohan 23] Brohan, A., Chebotar, Y., Finn, C., et al.: Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, in *CoRL*, pp. 287–318 (2023)

[Funahashi 93] Funahashi, K.-i. and Nakamura, Y.: Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks, *Neural networks*, Vol. 6, No. 6, pp. 801–806 (1993)

[Goko 24] Goko, M., Kambara, M., Saito, D., Otsuki, S., and Sugiura, K.: Task Success Prediction for Open-Vocabulary Manipulation Based on Multi-Level Aligned Representations, in *CoRL* (2024)

[Gu 22] Gu, A., Goel, K., and Ré, C.: Efficiently Modeling Long Sequences with Structured State Spaces, in *ICLR* (2022)

[Kim 24] Kim, M., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., et al.: OpenVLA: An Open-Source Vision-Language-Action Model, in *CoRL* (2024)

[Lehnert 17] Lehnert, C., English, A., McCool, C., et al.: Autonomous Sweet Pepper Harvesting for Protected Cropping Systems, *IEEE RA-L*, Vol. 2, No. 2, pp. 872–879 (2017)

[OpenAI 24] OpenAI, : GPT-4o (Accessed: Nov. 2024.)

[Oquab 24] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., et al.: DINOv2: Learning Robust Visual Features without Supervision, *TMLR*, pp. 1–31 (2024)

[Smith 23] Smith, J. T. H., Warrington, A., and Linderman, S. W.: Simplified State Space Layers for Sequence Modeling, in *ICLR* (2023)

[ste 24] Stella en 1.5B v5 (2024), Accessed: Dec. 2024

[Tallec 18] Tallec, C. and Ollivier, Y.: Can Recurrent Neural Networks Warp Time?, in *ICLR* (2018)

[Wang 24] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of The World at Any Resolution, *arXiv preprint arXiv:2409.12191* (2024)

[Wu 23] Wu, J., Antonova, R., Kan, A., et al.: Tidybot: Personalized Robot Assistance with Large Language Models, *Autonomous Robots*, Vol. 47, No. 8, pp. 1087–1102 (2023)