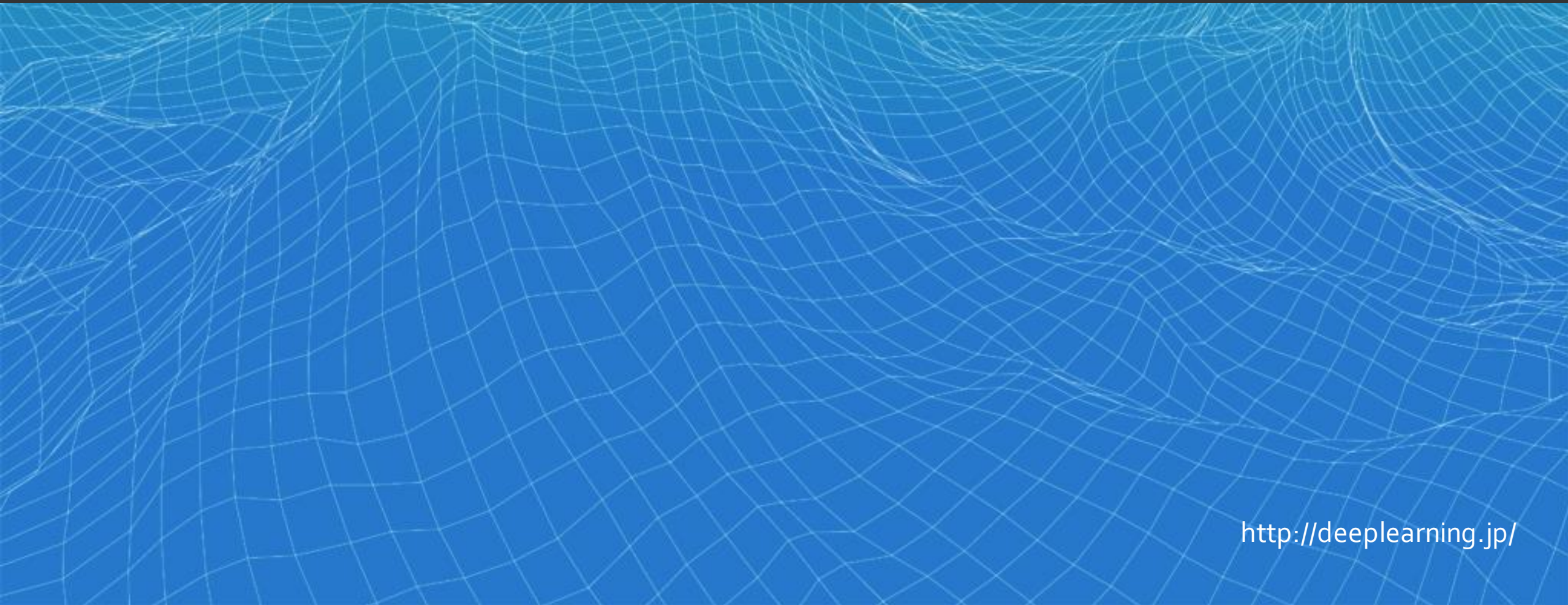
DEEP LEARNING JP
[DL Papers]

PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators (CoRL 2024)

2024.11.14 Ryosuke Takanami, D1, Matsuo-Iwasawa Lab



書誌情報

- **PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators**
  - Project Page: https://poliformer.allen.ai/
  - arXiv: https://arxiv.org/abs/2406.20083
  - Github: https://github.com/allenai/poliformer

- 著者：Kuo-Hao Zeng, Zichen "Charles" Zhang, Kiana Ehsani, Rose Hendrix, Jordi Salvador, Alvaro Herrasti, Ross Girshick, Aniruddha Kembhavi, Luca Weihs

  (PRIOR @ Allen Institute for AI)

- **概要：**
  - **CoRL2024 採択論文 (Outstanding Paper Award)**
  - RGB画像だけを使ったナビゲーションモデルの構築
  - シミュレーション内で大規模にRLすることで解決を目指している

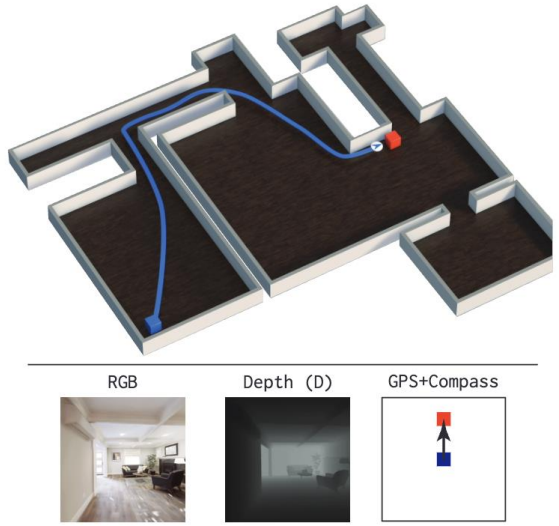
  - 特に断りのない限り，図表等の出典は本論文、本プロジェクトページからの引用

概要

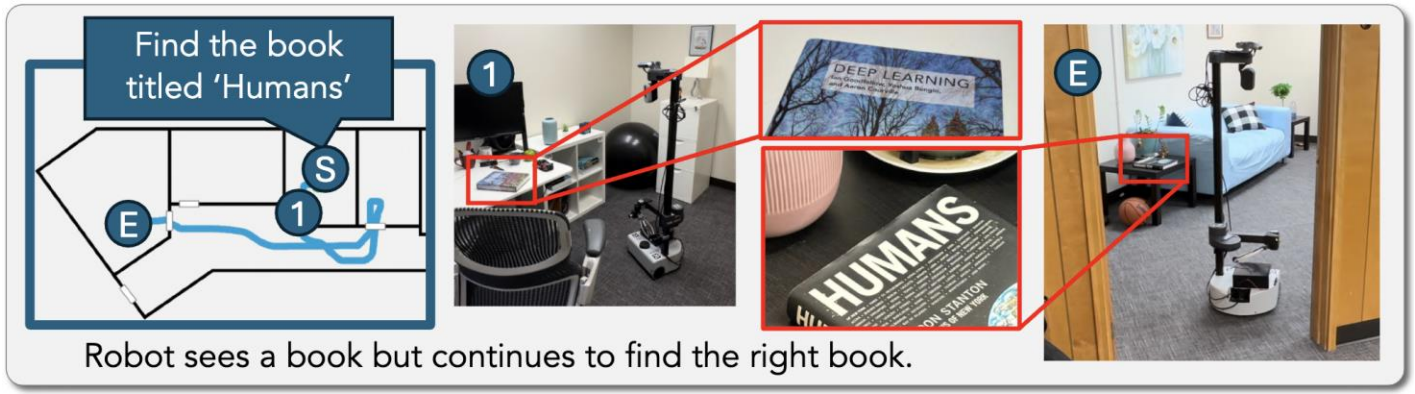
- RLによるナビゲーションモデルの構築は特定の座標をゴールとして移動するPointNavタスクにおいては成功していたが、特定の物体を探し、その物体まで移動するObjectNavタスクにおいては、学習の不安定性からスケールが難しく困難であった
- PoliFormerでは、visual foundation modelを用いたエンコーダと、長期的なメモリを可能にするcasual transformerデコーダを備えた新しいアーキテクチャを提案し、ObjectNavタスクにおいてSoTAを達成した

RLによるナビゲーション

- 代表的なナビゲーションタスク
  - PointNav：エージェントがGPSの座標情報を用いて指定されたポイントに移動するタスク
  - ObjectNav：特定の物体の種類（例：本）を識別し、視覚情報をもとにその物体を探してナビゲートするタスク
- 従来のRLでは、PointNavのような比較的単純なナビゲーションには高い精度で成功してたが、ObjectNavのような複雑なタスクでは、環境探索や記憶力の要求が高く、従来の浅いLSTMベースのネットワークでは限界があり、層の深いTransformerの学習は不安定で困難であった



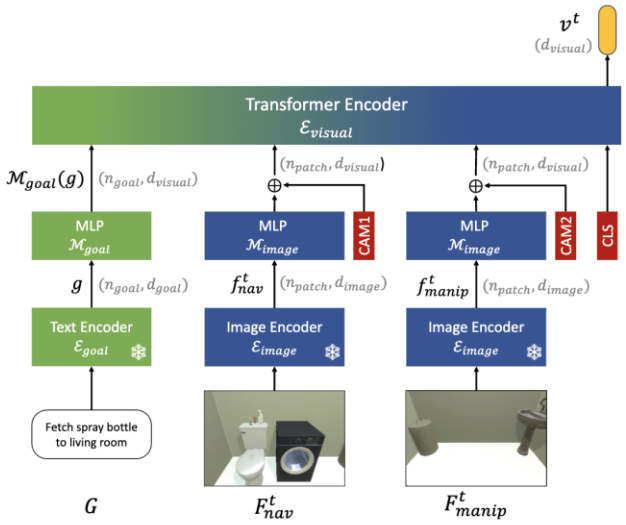
PointNavタスク [Wijmans 24]

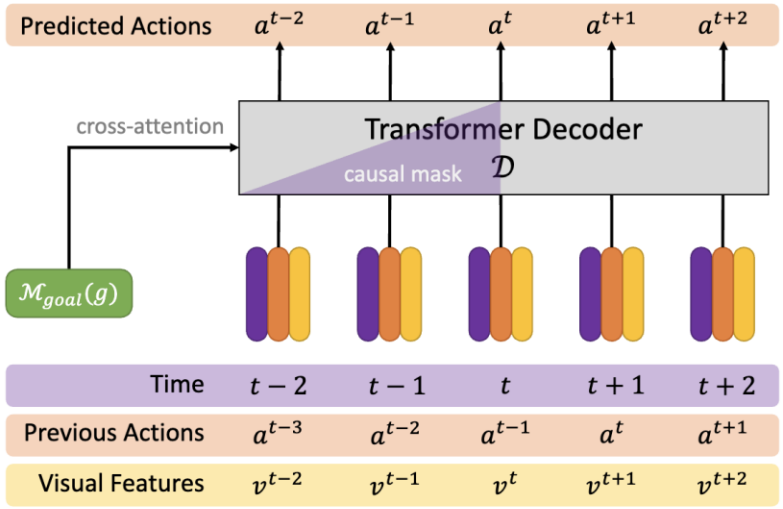


ObjectNavタスク

模倣学習によるナビゲーション

- 近年、模倣学習 (IL) によるTransformerを使用したナビゲーションモデルも台頭
- 層の深いTransformerモデルを大規模データで学習することでObjectNavタスクが解けるようになった
- しかし、ILは探索空間が限られており、エラーカバレッジ（ミスを修正する能力）が不十分であるため、成功率も57%程度で停滞





模倣学習ベースのナビゲーションモデル: SPOC [Ehsani 24]

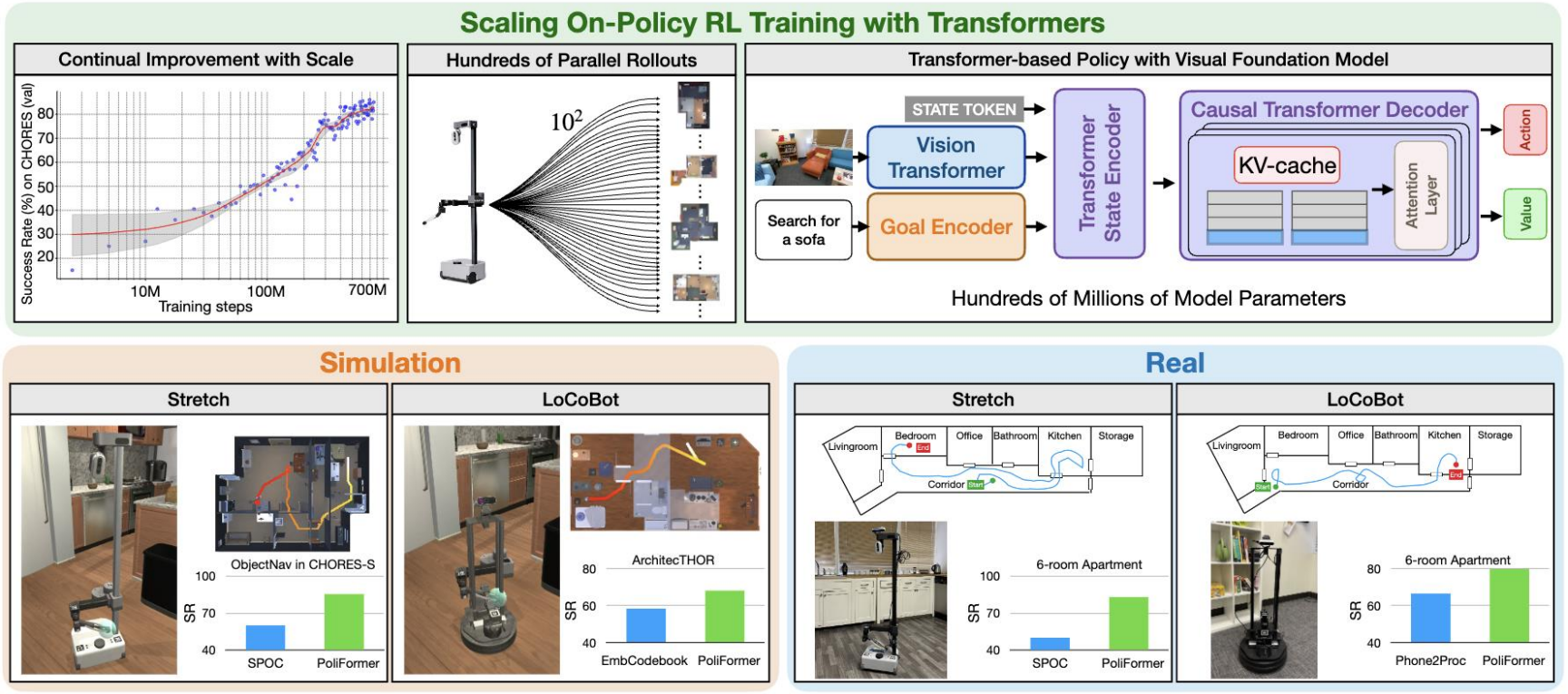
RLとTransformerの統合の必要性

- 長期的な依存関係の理解を必要とするObjectNavタスクにおいては層の深い強力なTransformerを活用したいがRLでは不安定になりがち
- ILでTransformerを安定的に学習できるが、推論するときにdistrubution shiftの影響を受けやすい

ObjectNavタスクにおいてRLでTransformerをうまく学習できるようにしたい

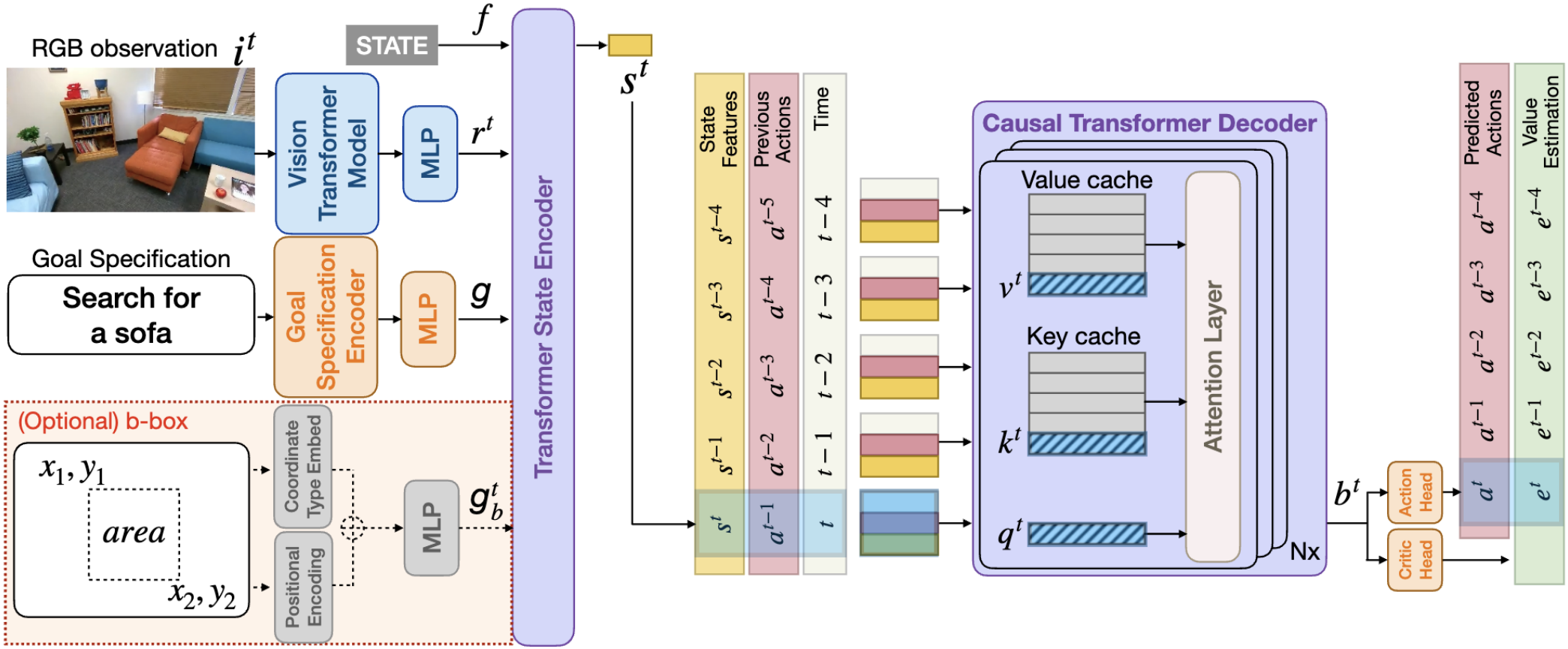
提案手法：PoliFormer

- **PoliFormer**
  - **Visual foundation model**としてDINOv2を使用し、RGB画像の特徴抽出を強化
  - **Casual transformerデコーダとKVキャッシュの利用**により、過去の情報を効率的に保持し、長期的な計画を可能に
  - **大規模な並列ロールアウトと自動生成環境**（PROCTHOR）で多様なインタラクションを学習し、実世界へのゼロショット転送が可能な汎用ナビゲーターの実現

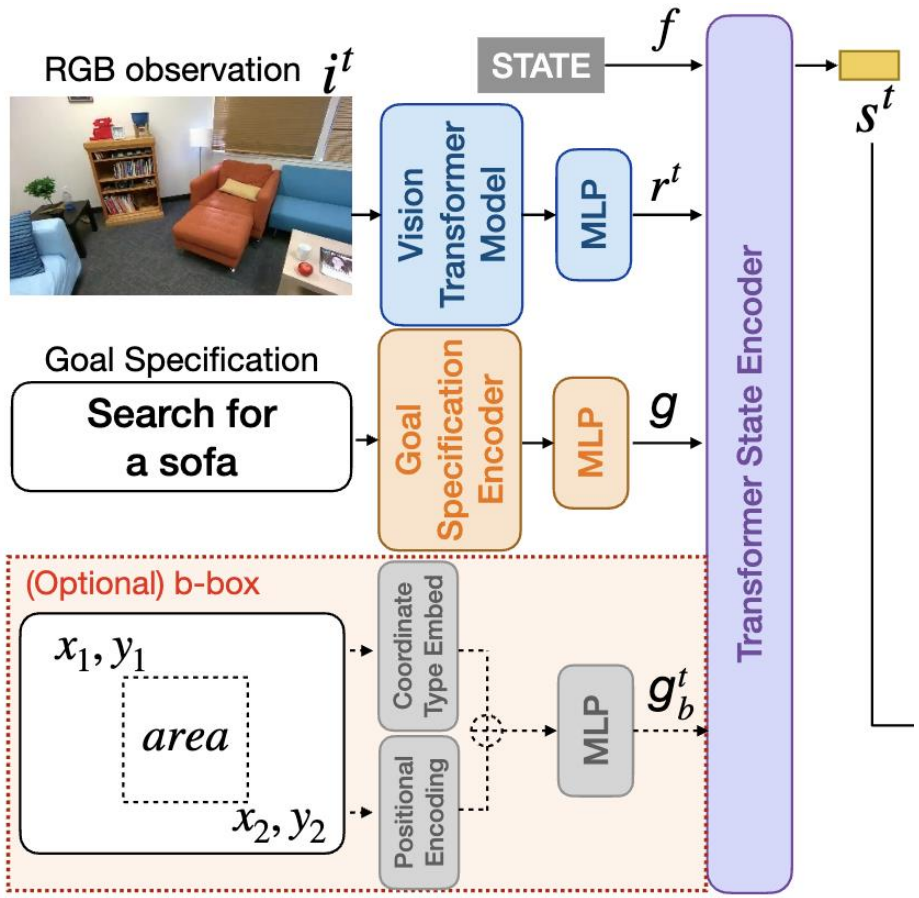


PoliFormerのアーキテクチャ

- Fully transformer based navigation policy
  - 現在のRGB画像とゴール指示文がエンコーダーの入力として与えられ、状態表現を生成
  - 現在と過去の状態表現をcasual transformerに入力することで時系列を考慮したstate beliefを生成
  - 最後にstate beliefを各々線形変換して、actionとvalueを出力



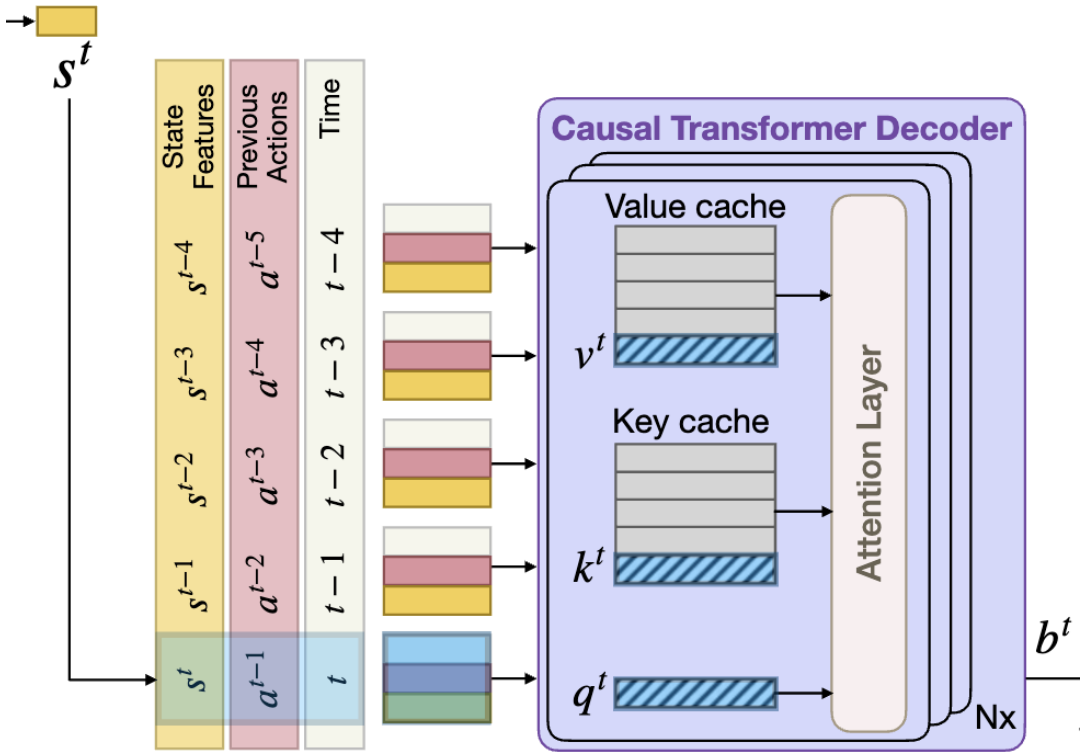
PoliFormerのアーキテクチャ



- エンコーダー側
  - Visual transformer model
    - DINOv2を使用、Sim2Realギャップに強いモデル
    - 学習時は重みを固定
  - Goal encoder
    - 物体カテゴリーを指定する場合はone hot embedding
    - 自然言語で指定する場合はFLAN-T5を使ってembedding
    - Bounding boxで指定する場合はboxの座標と面積をembedding
  - Transformer state encoder
    - Non casual transformerを使用して、goal conditionedな状態表現を生成

PoliFormerのアーキテクチャ

- デコーダー側
  - 現在と過去の状態表現列を入力とすることで長期的な依存関係を抽出
  - 通常のcasual transformerは、各タイムステップで過去すべての情報を再度計算するため、タイムステップ数が増えると計算コストが二乗的に増加
  - そこでPoliFormerではKVキャッシュという過去のタイムステップで得られたKeyとValueの計算結果のキャッシュを用いて、次のタイムステップの計算を新しい状態に対する追加の計算だけに抑えた
  - これにより、計算時間が線形に抑えられ、トレーニングや推論の速度が大幅に向上



シミュレーションでの大規模な強化学習

- 並列環境における分散学習の使用
  - 32台のA6000 GPUを使用してDD-PPOアルゴリズムによって分散学習を実行 (4.5日、シングルノードだと15.3日かかる計算)
  - 具体的な設定：LoCoBotで384並列、Stretch RE-1で192並列、合計700M step学習
- 多様なシミュレーションシーンの生成
  - PROCTHORによる自動生成環境を使用して150,000シーン以上の環境を用意
  - Objaverseも使用してシーン中の物体も多様化

実験

- POLIFORMERの有効性と汎化性能を、従来のナビゲーションモデルと比較
  - **シミュレーション環境でのナビゲーション性能**：複数のベンチマークでの成功率、精度、効率性を測定
  - **実世界環境でのゼロショットナビゲーション**：シミュレーションのみで学習したモデルが、現実環境でどれだけ高い成功率を維持できるかを検証
  - **モデルの拡張性と適応性**：マルチターゲットナビゲーションや人間追従タスクでの性能

実験：シミュレーション環境での評価

Inputs	Model	Loss	CHORES-S ObjectNav Success (SEL)
RGB+text	SPOC [6]	IL	57.0 (46.2)
	SPOC*	IL	60.0 (30.5)
	EmbSigLIP [6]	RL	36.5 (24.5)
	POLIFORMER	RL	**85.5** (**61.2**)
RGB +text+b-box	SPOC	IL	85.0 (61.4)
	POLIFORMER	RL	**95.5** (**71.4**)
RGB+b-box	POLIFORMER	RL	92.0 (73.9)

(a) Stretch RE-1 on CHORES-S

Inputs	Model	PROCTHOR-10K	ARCHITECTHOR	AI2-ITHOR
		Success (SPL)		
RGB+text	PROCTHOR [11][3]	67.7 (49.0)	55.8 (38.3)	70.0 (57.1)
	SGC [63]	70.8 (48.6)	53.8 (34.8)	71.4 (59.3)
	EmbCodebook [86]	73.7 (48.4)	58.3 (35.6)	78.4 (23.7)
	POLIFORMER	**82.4** (**58.5**)	**68.3** (**45.1**)	**85.3** (**72.7**)
RGB +text+b-box	POLIFORMER	90.4 (66.6)	81.9 (55.6)	94.9 (83.5)
RGB+b-box	POLIFORMER	87.4 (56.2)	85.7 (47.6)	92.1 (78.6)

(b) LoCoBot on ProcTHOR-10k (val), ArchitecTHOR and AI2-iTHOR (test)

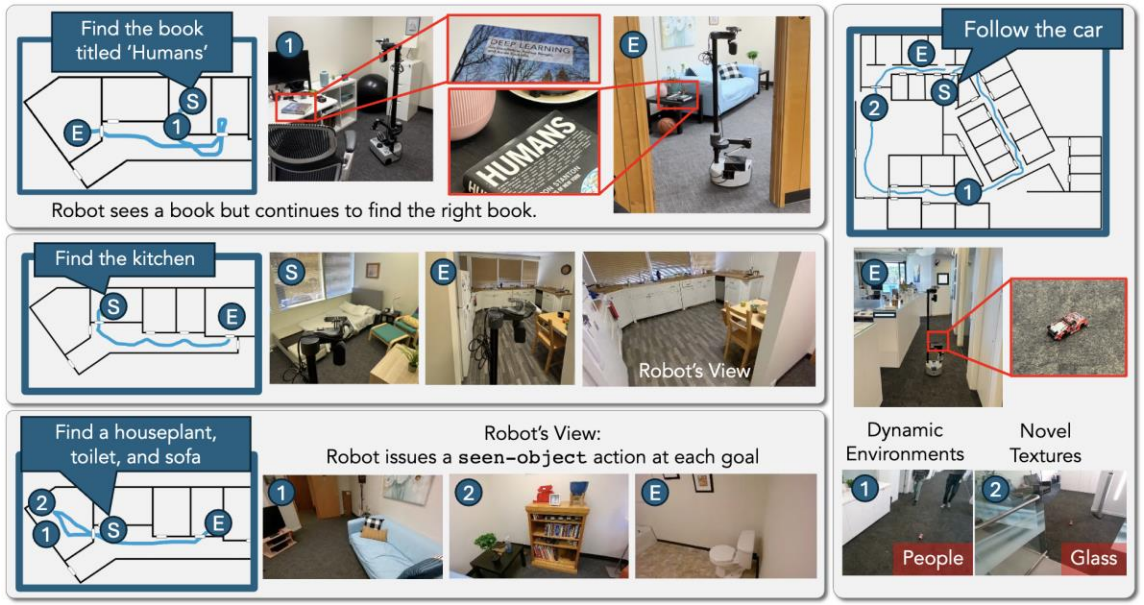
- **実験概要**
  - **ベンチマーク**
    - CHORES-S（ObjectNavタスク）：エージェントが指定された物体を見つけるナビゲーションタスク
    - ProcTHOR-10k：約10,000の自動生成環境でのランダム配置されたオブジェクト間を探索するタスク
    - AI2-iTHOR：多様な部屋やシーン設定で、エージェントが目標物体までナビゲートするタスク
  - **結果**
    - CHORES-Sでは28.5%の精度改善
    - ProcTHOR-10kおよびAI2-iTHORでも、POLIFORMERはSoTAを達成
    - アブレーションでは、エンコーダをスケールアップした場合、成功率が3.2%向上するなどが確認

実験：実世界環境での評価

Model	Stretch RE-1	LoCoBot
ProcTHOR [11]	-	26.7
Phone2Proc [17]	-	66.7
SPOC [6]	50.0	-
POLIFORMER (ours)	**83.3**	**80.0**
SPOC+Detic [6]	83.3	-
POLIFORMER +Detic (ours)	**88.9**	-

- **実験概要**
  - **目的**
    - シミュレーションのみでトレーニングされたPoliFormerを、実際のロボット（LoCoBotとStretch RE-1）でテストし、ゼロショットでのナビゲーション性能を検証
  - **実験結果**
    - LoCoBotでは13.3%の成功率向上、Stretch RE-1では33.3%の成功率向上
    - シミュレーション環境で学習した知識が、現実環境で高い汎化能力を発揮し、追加の適応トレーニングなしでシミュレーションとほぼ同等の性能を実現

実験：モデルの拡張性と適応性



- **実験概要**
  - **手法の拡張**
    - **PoliFormer-BOXNAV**という拡張版を実装
    - バウンディングボックス入力によるマルチターゲットナビゲーションや、人間追従タスクなど多様なシナリオに対応可能な手法
  - **結果**
    - **マルチターゲットナビゲーション**：複数のオブジェクトを同時に認識し、それぞれに対して効率的に移動することが可能に
    - **人間追従**：目標を人間に設定し、動的に移動する対象を追従する実験でも高い精度を発揮

まとめ

- **まとめ**
  - PoliFormerは、TransformerとRLを組み合わせたナビゲーションモデルで、シミュレーションと実世界の両方で高い性能を発揮
  - DINOv2(Visual foundation model)とKVキャッシュ付きのcasual transformerデコーダーを用い、視覚情報とゴール情報を統合しつつ、長期的な記憶を効率的に保持
  - その結果、従来のRLモデルを上回る成功率を達成し、シミュレーションから実世界へのゼロショット転送も高精度で実現

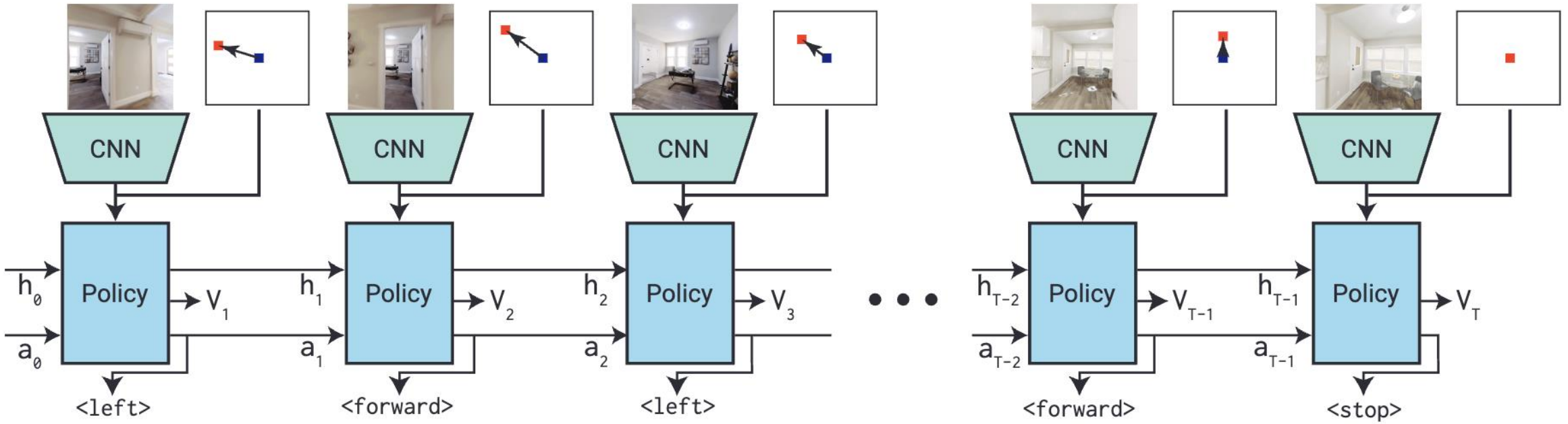
- **感想**
  - 今回ロボットごとにモデルを学習していてロボットに対する汎化性があるのかが気になった
  - かなりのGPUを使用して、分散しているとはいえ、かなりの時間をかけているのでまだサンプル効率の問題があるような気がする

Reference

- Zeng, K. et al. PoliFormer: Scaling On-Policy RL with Transformers Results in Masterful Navigators. CoRL. 2024.
- Wijmans, E. et al. DD-PPO: LEARNING NEAR-PERFECT POINTGOAL NAVIGATORS FROM 2.5 BILLION FRAMES. ICLR. 2020.
- Ehsani, K. et al. SPOC: Imitating Shortest Paths in Simulation Enables Effective Navigation and Manipulation in the Real World. CVPR. 2024.

RLによるナビゲーション

- PointNavにおけるSoTAアーキテクチャ
  - LSTMベースのモデル



LSTMベースのナビゲーションモデル [Wijmans 24]