Novelty Assessment Report

Paper: Embodied Navigation Foundation Model

PDF URL: https://openreview.net/pdf?id=kkBOIsrCXh

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-27

Abstract

Navigation is a fundamental capability in embodied AI, representing the intelligence required to perceive and interact within physical environments. To achieve such intelligence, recent advanced works leverage Vision-Language Models (VLMs), which demonstrate strong generalizability and possess a well-suited formulation for navigation. However, these approaches remain largely confined to narrow task settings and embodiment-specific architectures. In this work, we introduce a cross-embodiment and cross-task Navigation Foundation Model (NavFoM), trained on eight million navigation samples that encompass quadrupeds, drones, wheeled robots, and vehicles, and spanning diverse tasks such as vision-and-language navigation, object searching, target tracking, and autonomous driving. NavFoM employs a unified architecture that processes multimodal navigation inputs from varying camera configurations and navigation horizons. To accommodate diverse camera setups and temporal horizons, NavFoM incorporates identifier tokens that embed camera view information of embodiments and the temporal context of tasks. Furthermore, to meet the demands of real-world deployment, NavFoM controls all observation tokens using a dynamically adjusted sampling strategy under a limited token length budget. Extensive evaluations on seven public benchmarks demonstrate that our model achieves state-of-the-art or highly competitive performance across different navigation tasks and embodiments without requiring task-specific fine-tuning. Additional real-world experiments further confirm the strong generalizability and practical applicability of our approach.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **cross-embodiment and cross-task embodied navigation**

A total of **50 papers** were analyzed and organized into a taxonomy with **12 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Foundation Models and Generalist Agents for Embodied Navigation**
- **Multi-Agent Navigation and Coordination**
- **Task-Specific Navigation Methods and Representations**
- **Surveys and Overviews of Embodied AI and Navigation**

Complete Taxonomy Tree

- cross-embodiment and cross-task embodied navigation Survey Taxonomy
- Foundation Models and Generalist Agents for Embodied Navigation
  - Cross-Embodiment Foundation Models ★ (6 papers)
  - [0] Embodied Navigation Foundation Model (Anon et al., 2026) View paper
  - [1] Universal actions for enhanced embodied foundation models (Jin-Liang Zheng, 2025) View paper
  - [2] From multimodal llms to generalist embodied agents: Methods and lessons (Andrew Szot, 2025) View paper
  - [5] Robocat: A self-improving generalist agent for robotic manipulation (Bousmalis, 2023) View paper
  - [25] A Bio-Inspired Learning and Control Framework for Cross-Embodiment and Cross-Task Locomotion (Shafiee-Ashtiani, 2025) View paper
  - [44] Pushing the Limits of Cross-Embodiment Learning for Manipulation and Navigation (Yang, 2024) View paper
  - Multi-Task and Generalist Navigation Agents (5 papers)
  - [17] NaviMaster: Learning a Unified Policy for GUI and Embodied Navigation Tasks (Luo Zhihao, 2025) View paper
  - [19] OctoNav: Towards Generalist Embodied Navigation (Gao Chen, 2025) View paper
  - [35] BLM: A Boundless Large Model for Cross-Space, Cross-Task, and Cross-Embodiment Learning (W Tan, 2025) View paper
  - [38] Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks (Jiazhao Zhang, 2024) View paper
  - LLM-Based Robotic Systems and Agentic Reasoning (3 papers)
  - [4] Large language models for multi-robot systems: A survey (Li Peihan, 2025) View paper
  - [14] Agentic LLM-based robotic systems for real-world applications: a review on their agenticness and ethics (Emmanuel K. Raptis, 2025) View paper
  - [24] OmniEAR: Benchmarking Agent Reasoning in Embodied Tasks (Wang Zi-xuan, 2025) View paper
- Multi-Agent Navigation and Coordination
  - Multi-Agent Coordination with Communication and Collaboration (6 papers)
  - [3] Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration (Tan Huajie, 2025) View paper
  - [7] Multi-agent embodied visual semantic navigation with scene prior knowledge (Liu Xinzhu, 2022) View paper
  - [8] Cooperative multi-agent learning for navigation via structured state abstraction (Mohamed K. Abdel-Aziz, 2024) View paper
  - [29] Coordinating multi-agent navigation by learning communication (Dalton Hildreth, 2019) View paper
  - [32] Heterogeneous Embodied Multi-Agent Collaboration (Xin-Zhu Liu, 2024) View paper

  - [43] Embodied Multi-Agent Task Planning from Ambiguous Instruction (Xin-Zhu Liu, 2022) View paper
  - Decentralized Multi-Agent Navigation and Collision Avoidance (5 papers)
  - [9] Online control barrier functions for decentralized multi-agent navigation (Zhan Gao, 2023) View paper
  - [11] Learning control admissibility models with graph neural networks for multi-agent navigation (Yu, 2023) View paper
  - [21] Safe Multi-Agent Navigation Guided by Goal-Conditioned Safe Reinforcement Learning (Feng Meng, 2025) View paper
  - [26] Decentralized, unlabeled multi-agent navigation in obstacle-rich environments using graph neural networks (Xuebo Ji, 2021) View paper
  - [39] Learning Distributed Safe Multi-Agent Navigation via Infinite-Horizon Optimal Graph Control (Wang Fenglan, 2025) View paper
  - Learning-Based Multi-Agent Navigation Policies (7 papers)
  - [13] Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding (Jaehoon Chung, 2024) View paper
  - [23] ALAN: adaptive learning for multi-agent navigation (Julio Godoy, 2018) View paper
  - [27] Adaptive learning for multi-agent navigation (Julio Godoy, 2015) View paper
  - [40] Differentiable Learning of Scalable Multi-Agent Navigation Policies (Xiaohan Ye, 2023) View paper
  - [45] Learning Graph-Enhanced Commander-Executor for Multi-Agent Navigation (Yang Xin-yi, 2023) View paper
  - [47] MASP: Scalable GNN-based Planning for Multi-Agent Navigation (Yang Xin-yi, 2023) View paper
  - [50] Time-aware MADDPG with LSTM for multi-agent obstacle avoidance: a comparative study (Enyu Zhao, 2024) View paper
  - Environment Co-Optimization for Multi-Agent Navigation (4 papers)
  - [12] Co-Optimization of Environment and Policies for Decentralized Multi-Agent Navigation (Zhan Gao, 2024) View paper
  - [16] Constrained environment optimization for prioritized multi-agent navigation (Zhan Gao, 2023) View paper
  - [20] Environment optimization for multi-agent navigation (Gao, 2022) View paper
  - [36] Co-Optimizing Reconfigurable Environments and Policies for Decentralized Multi-Agent Navigation (Zhan Gao, 2025) View paper
- Task-Specific Navigation Methods and Representations
  - Representation Learning and Perception for Navigation (4 papers)
  - [10] Entl: Embodied navigation trajectory learner (Klemen Kotar, 2023) View paper
  - [30] Analyzing Visual Representations in Embodied Navigation Tasks (Wijmans, 2022) View paper
  - [37] Offline Visual Representation Learning for Embodied Navigation (Yadav, 2022) View paper
  - [48] Learning to Align Multimodal Data for Static and Dynamic Tasks (Paul, 2022) View paper
  - Navigation Policy Learning and Task Decomposition (3 papers)
  - [31] Deep Learning-Powered Embodied Navigation in Simulated Environments (Zhu, 2024) View paper
  - [33] Learning Embodied AI Agents with Task Decomposition (Jia, 2023) View paper
  - [34] Hierarchical Auto-Organizing System for Open-Ended Multi-Agent Navigation (Zhao, 2024) View paper
  - Navigation Benchmarks and Evaluation Frameworks (3 papers)
  - [6] Embodied navigation with multi-modal information: A survey from tasks to methodology (Y. Z. Wu, 2024) View paper
  - [15] Unifying Modern AI with Robotics: Survey on MDPs with Diffusion and Foundation Models (Zhaofan Zhang, 2025) View paper
  - [28] NavSpace: How Navigation Agents Follow Spatial Intelligence Instructions (Yang HaoLin, 2025) View paper
  - Mapping and Spatial Perception Systems (1 papers)
  - [42] A Robust, Task-Agnostic and Fully-Scalable Voxel Mapping System for Large Scale Environments (Jinche La, 2024) View paper
- Surveys and Overviews of Embodied AI and Navigation (4 papers)
  - [18] Multi-agent Embodied AI: Advances and Future Directions (Feng Zhaohan, 2025) View paper
  - [22] Embodied Multi-Agent Systems: Perception, Action, and Learning (H Liu, 2025) View paper
  - [46] Guest Editorial: Special Issue on Embodied AI in Indoor Robotics: Bridging Perception, Interaction, and Autonomy (Yaran Chen, 2025) View paper
  - [49] AIRSHIP: Empowering General-Purpose Intelligent Robots through Open-Source Embodied AI (HC Chou, 2025) View paper

Narrative

Core task: cross-embodiment and cross-task embodied navigation. The field addresses how agents with diverse physical forms and capabilities can navigate and perform tasks across varied environments and objectives. The taxonomy reveals four main branches. Foundation Models and Generalist Agents for Embodied Navigation explores unified architectures that leverage large-scale pretraining and multimodal reasoning to handle multiple robot morphologies and task types, as seen in works like Universal Actions Embodied[1] and Robocat Self-Improving Agent[5]. Multi-Agent Navigation and Coordination focuses on scenarios where multiple agents must navigate shared spaces, often requiring collision avoidance, communication protocols, and cooperative strategies. Task-Specific Navigation Methods and Representations develops specialized techniques for particular problem settings, such as visual representations, hierarchical planning, or instruction following. Finally, Surveys and Overviews of Embodied AI and Navigation provide broad perspectives on the evolving landscape, synthesizing progress across these dimensions.

A central tension emerges between generalist foundation models that aim for broad transferability and specialized methods that optimize for particular embodiments or tasks. Recent efforts like Multimodal LLMs Embodied Agents[2] and Bio-Inspired Cross-Embodiment[25] push toward more flexible policies that can adapt across robot types, while works such as Cross-Embodiment Limits[44] critically examine the boundaries of such transfer. Embodied Navigation Foundation[0] sits squarely within the Cross-Embodiment Foundation Models cluster, emphasizing scalable pretraining and policy adaptation mechanisms that bridge different morphologies and task specifications. Compared to Robocat Self-Improving Agent[5], which focuses on self-improvement through iterative data collection, and Bio-Inspired Cross-Embodiment[25], which draws on biological principles for morphology-agnostic control, Embodied Navigation Foundation[0] appears to prioritize unified representations that facilitate zero-shot or few-shot generalization across diverse navigation scenarios. This positioning reflects ongoing debates about whether cross-embodiment success depends more on architectural universality or on richer inductive biases tailored to embodied reasoning.

Related Works in Same Category

The following **5 sibling papers** share the same taxonomy leaf node with the original paper:

1. Universal actions for enhanced embodied foundation models

Authors: Jin-Liang Zheng, Jianxiong Li, Dongxiu Liu, Yi-nan Zheng, Zhihao Wang, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ dling cross-task, cross-environment, and cross-embodiment â¦ of behaviors for cross-embodiment control, making our 0.5B â¦ as those in manipulation and navigation, and from single-arm â¦

Relationship Analysis

Both papers belong to the Cross-Embodiment Foundation Models category, focusing on training models across diverse robot morphologies (arms, quadrupeds, drones, wheeled robots) to enable shared representations and control policies. They overlap in addressing cross-embodiment navigation and control using vision-language models trained on large-scale heterogeneous data from multiple robot platforms. The key difference is that the original paper (NavFoM) focuses specifically on navigation tasks across embodiments using temporal-viewpoint indicator tokens and trajectory prediction, while the candidate paper (UniAct) addresses general embodied control by learning a universal action space through vector quantization to handle action heterogeneity across different control interfaces and robot types.

2. From multimodal llms to generalist embodied agents: Methods and lessons

Authors: Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ varied domains through a multi-embodiment action tokenizer. â¦ training policies on large multi-task datasets, illustrating the â¦ We use datasets of simulated robot navigation in Habitat. We â¦

Relationship Analysis

Both papers belong to the Cross-Embodiment Foundation Models category, focusing on training models across diverse robot morphologies to enable shared representations and control policies. They overlap in addressing cross-embodiment navigation and manipulation tasks using vision-language models trained on large-scale multi-embodiment datasets (NavFoM uses 8M navigation samples across quadrupeds, drones, wheeled robots, and vehicles; GEA uses 2.2M trajectories across static manipulators, mobile manipulators, and virtual agents). The key difference is that NavFoM specializes in navigation tasks with trajectory-based waypoint prediction and temporal-viewpoint indicator tokens for multi-camera setups, while GEA is a broader generalist agent that extends beyond navigation to include manipulation, video games, UI control, and planning tasks, using a multi-embodiment action tokenizer and combining supervised learning with online reinforcement learning.

3. Robocat: A self-improving generalist agent for robotic manipulation

Authors: Bousmalis, Konstantinos, Konstantinos Bousmalis, Vezzani, Giulia, et al. (116 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

The ability to leverage heterogeneous robotic experience from different robots and tasks to quickly master novel skills and embodiments has the potential to transform robot learning. Inspired by recent advances in foundation models for vision and language, we propose a multi-embodiment, multi-task generalist agent for robotic manipulation. This agent, named RoboCat, is a visual goal-conditioned decision transformer capable of consuming action-labelled visual experience. This data spans a large r...

Relationship Analysis

Both papers belong to the Cross-Embodiment Foundation Models category, focusing on training models across diverse robot morphologies to enable shared representations and control policies. While NavFoM addresses cross-embodiment navigation tasks (quadrupeds, drones, wheeled robots, vehicles) using vision-language models with trajectory prediction for navigation-specific scenarios, RoboCat focuses on cross-embodiment robotic manipulation tasks (various robot arms with different DoF) using visual goal-conditioning and self-improvement loops for pick-and-place and assembly behaviors. The key distinction is that NavFoM specializes in navigation across different locomotion platforms, whereas RoboCat specializes in manipulation across different arm embodiments.

4. A Bio-Inspired Learning and Control Framework for Cross-Embodiment and Cross-Task Locomotion

Authors: Shafiee-Ashtiani, Milad | **Year/Venue**: 2025 • Infoscience (Ecole Polytechnique FÃ©dÃ©rale de Lausanne) | **URL**: View paper

Abstract

Animals exhibit remarkable locomotion skills despite significant sensorimotor delays and operating in uncertain environments. Moreover, mammals acquire these skills within minutes of birth. From the Cambrian explosion to the present day, vertebrate motor control circuits have remained remarkably similar. This shared architecture is rooted in a modular and adaptable design, reflecting an elegant system that enables the complexity of locomotion. At the same time, we are living in an exciting era f...

Relationship Analysis

Both papers belong to the Cross-Embodiment Foundation Models category, focusing on models trained across diverse robot morphologies. The original paper (NavFoM) presents a vision-language foundation model for cross-embodiment navigation tasks (VLN, object search, tracking, autonomous driving) trained on 8 million samples using temporal-viewpoint indicator tokens and budget-aware sampling. The candidate paper presents a bio-inspired learning framework using reinforcement learning with central pattern generators (CPGs) for cross-embodiment locomotion control, emphasizing gait transitions and parkour skills rather than high-level navigation with language instructions.

5. Pushing the Limits of Cross-Embodiment Learning for Manipulation and Navigation

Authors: Yang, Jonathan, Jonathan Yang, Catherine Glossop, Bhorkar, et al. (22 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Recent years in robotics and imitation learning have shown remarkable progress in training large-scale foundation models by leveraging data across a multitude of embodiments. The success of such policies might lead us to wonder: just how diverse can the robots in the training set be while still facilitating positive transfer? In this work, we study this question in the context of heterogeneous embodiments, examining how even seemingly very different domains, such as robotic navigation and manipu...

Relationship Analysis

Both papers belong to the Cross-Embodiment Foundation Models category, focusing on training unified models across diverse robot morphologies for navigation and manipulation tasks. They overlap in addressing cross-embodiment generalization using foundation model approaches with shared representations across quadrupeds, drones, wheeled robots, and manipulators. The key difference is that the original paper (NavFoM) focuses exclusively on navigation tasks across embodiments using vision-language models with 8 million navigation samples, while the candidate paper (Pushing the Limits) jointly trains on both navigation and manipulation data to study heterogeneous cross-embodiment transfer, demonstrating that navigation data can improve manipulation performance and vice versa.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Cross-embodiment and cross-task Navigation Foundation Model (NavFoM)

Description: The authors propose NavFoM, a unified navigation foundation model trained on 8 million samples covering multiple embodiments (quadrupeds, drones, wheeled robots, vehicles) and diverse navigation tasks (vision-and-language navigation, object searching, target tracking, autonomous driving). The model uses a unified architecture that processes multimodal navigation inputs from varying camera configurations and navigation horizons without requiring task-specific fine-tuning.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Autonomous visual navigation for mobile robots: A systematic literature review

URL: View paper

Brief Assessment

Autonomous Visual Navigation[67] is a systematic literature review that surveys existing navigation approaches. It does not present a unified foundation model trained on 8 million samples across multiple embodiments and tasks as proposed in the original paper.

2. Design of AI based Autonomous Navigation System Using Swarm Intelligence Techniques for Agriculture Application

URL: View paper

Brief Assessment

Swarm Intelligence Agriculture[71] focuses on multi-robot coordination using swarm intelligence (PSO-ACO) for agricultural field navigation, not on building a unified foundation model across diverse embodiments and navigation tasks as proposed in the original paper.

3. X-mobility: End-to-end generalizable navigation via world modeling

URL: View paper

Brief Assessment

X-Mobility World Modeling[66] focuses on world modeling architecture for navigation with cross-embodiment deployment, but does not claim to be a unified foundation model trained on diverse navigation tasks (VLN, object search, tracking, autonomous driving) across multiple embodiments as NavFoM does.

4. A Cross-Environment and Cross-Embodiment Path Planning Framework via a Conditional Diffusion Model

URL: View paper

Brief Assessment

Cross-Environment Path Planning[70] focuses on path planning for robotic manipulators using diffusion models for joint-space trajectory generation, not vision-language navigation across diverse embodiments and tasks as in NavFoM.

5. Compass: Cross-embodiment mobility policy via residual rl and skill synthesis

URL: View paper

Brief Assessment

Compass Mobility Policy[65] focuses on cross-embodiment mobility through residual RL and skill synthesis for point-to-point navigation, not on building a unified navigation foundation model across diverse navigation tasks (VLN, object search, tracking, autonomous driving) as proposed in the original paper.

6. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities

URL: View paper

Brief Assessment

Embodied Gap Study[63] focuses on evaluating existing VLN methods across different robot embodiments in physically realistic settings, rather than proposing a unified navigation foundation model trained on diverse tasks and embodiments like NavFoM.

7. Antcar: simple route following task with ants-inspired vision and neural model

URL: View paper

Brief Assessment

Antcar Ants-Inspired Vision[68] focuses on a biologically-inspired visual navigation system for route following using ant-inspired vision and mushroom bodies neural models. This is fundamentally different from NavFoM's cross-embodiment, cross-task foundation model trained on 8 million samples across multiple robot types and diverse navigation tasks.

8. Pushing the Limits of Cross-Embodiment Learning for Manipulation and Navigation

URL: View paper

Brief Assessment

Cross-Embodiment Limits[44] focuses on heterogeneous cross-embodiment learning combining manipulation and navigation tasks, while NavFoM specifically addresses navigation-only tasks across multiple embodiments with a unified architecture for diverse navigation scenarios (VLN, object search, tracking, autonomous driving).

9. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation

URL: View paper

Brief Assessment

Scaling Cross-Embodied Learning[64] focuses on a unified policy for manipulation, navigation, locomotion and aviation using transformer architecture, while the original paper specifically addresses navigation tasks with vision-language models. The candidate does not challenge the novelty of NavFoM's navigation-specific foundation model approach with identifier tokens and temporal sampling strategies.

10. X-Nav: Learning End-to-End Cross-Embodiment Navigation for Mobile Robots

URL: View paper

Brief Assessment

X-Nav Cross-Embodiment[69] focuses on cross-embodiment transfer for low-level control in wheeled and quadrupedal robots using reinforcement learning, not on a unified foundation model spanning diverse navigation tasks (VLN, object search, tracking, autonomous driving) across multiple embodiments (quadrupeds, drones, wheeled robots, vehicles) as proposed in the original paper.

Contribution 2: Temporal-Viewpoint Indicator (TVI) tokens

Description: The authors introduce TVI tokens as a mechanism to organize visual tokens by encoding both viewpoint (camera angle) and temporal information. These tokens enable flexible processing of arbitrary camera arrangements and support unified training across image QA, video QA, and navigation tasks with different camera configurations.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Simultaneous multi-view camera pose estimation and object tracking with squared planar markers

URL: View paper

Brief Assessment

Multi-View Pose Tracking[56] focuses on simultaneous camera pose estimation and object tracking using planar markers in multi-camera setups. It does not address encoding temporal and viewpoint information for organizing visual tokens in vision-language navigation tasks or foundation models.

2. Active SLAM With Dynamic Viewpoint Optimization for Robust Visual Navigation

URL: View paper

Brief Assessment

Active SLAM Viewpoint[57] focuses on SLAM-based viewpoint optimization for robotic navigation using feature maps and keyframes, not on organizing visual tokens with temporal-viewpoint encoding for multi-task vision-language models.

3. NaviFormer: A Spatio-Temporal Context-Aware Transformer for Object Navigation

URL: View paper

Brief Assessment

NaviFormer Transformer Navigation[55] focuses on encoding spatial layouts and temporal pose trajectories for object navigation in static environments, not on organizing multi-view camera configurations with temporal information for diverse navigation tasks as in the original paper's TVI tokens.

4. Real-time vision-aided localization and navigation based on three-view geometry

URL: View paper

Brief Assessment

Vision-Aided Localization[62] focuses on three-view geometry constraints for camera localization using epipolar constraints and translation vectors, not on organizing visual tokens with temporal-viewpoint embeddings for multi-task navigation training as in the original paper.

5. Spatiotemporal Contrastive Learning for Cross-View Video Localization in Unstructured Off-road Terrains

URL: View paper

Brief Assessment

Spatiotemporal Contrastive Learning[59] focuses on cross-view localization between ground and aerial imagery in off-road environments, not on organizing multi-camera visual tokens for embodied navigation tasks. The technical approaches and problem domains are fundamentally different.

6. Henet: Hybrid encoding for end-to-end multi-task 3d perception from multi-view cameras

URL: View paper

Brief Assessment

Henet Hybrid Encoding[54] focuses on multi-task 3D perception from multi-view cameras in autonomous driving, using hybrid image encoders for different temporal frames. While it processes multi-view camera data, it does not introduce viewpoint-temporal indicator tokens for organizing visual tokens across arbitrary camera configurations and navigation tasks as described in the original contribution.

7. Virtual video camera: Image-based viewpoint navigation through space and time

URL: View paper

Brief Assessment

Virtual Video Camera[61] focuses on image-based rendering for view interpolation in captured video footage, not on encoding mechanisms for multi-view visual navigation in embodied AI systems.

8. Learning View-invariant and Novel Spatio-temporal Features Under Uncertainty from Video

URL: View paper

Brief Assessment

View-Invariant Spatiotemporal Features[60] focuses on learning view-invariant representations for action recognition and rPPG health sensing, not on encoding camera viewpoint and temporal information through specialized tokens for multi-view navigation tasks.

9. Beings: Bayesian embodied image-goal navigation with gaussian splatting

URL: View paper

Brief Assessment

Beings Gaussian Splatting[53] focuses on Bayesian image-goal navigation using 3D Gaussian splatting for scene representation, not on encoding camera viewpoint and temporal information for multi-view visual navigation tasks.

10. Learning multi-view camera relocalization with graph neural networks

URL: View paper

Brief Assessment

Multi-View Camera Relocalization[58] focuses on camera pose estimation using graph neural networks for multi-view sequences, not on organizing visual tokens with temporal-viewpoint identifiers for navigation tasks across different embodiments and camera configurations.

Contribution 3: Budget-Aware Temporal Sampling (BATS) strategy

Description: The authors propose BATS, a token sampling strategy that dynamically samples navigation history tokens based on an exponential forgetting curve while respecting a fixed token budget. This approach balances navigation performance with inference efficiency and adapts to varying numbers of cameras, addressing practical deployment constraints.

This contribution was assessed against **2 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Learning Adaptive and Temporally Causal Video Tokenization in a 1D Latent Space

URL: View paper

Brief Assessment

Adaptive Video Tokenization[51] focuses on video reconstruction and generation tasks with token allocation for visual content compression, not navigation history sampling for embodied AI agents under deployment constraints.

2. Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning

URL: View paper

Brief Assessment

RL of Thoughts[52] focuses on training a lightweight navigator model using reinforcement learning to dynamically select logic blocks for LLM reasoning tasks, not on temporal sampling strategies for navigation history under token budget constraints in embodied navigation.

Appendix: Text Similarity Detection

Textual similarity detection checked 26 papers and found 1 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Embodied Navigation Foundation Model View paper
- [1] Universal actions for enhanced embodied foundation models View paper
- [2] From multimodal llms to generalist embodied agents: Methods and lessons View paper
- [3] Roboos: A hierarchical embodied framework for cross-embodiment and multi-agent collaboration View paper
- [4] Large language models for multi-robot systems: A survey View paper
- [5] Robocat: A self-improving generalist agent for robotic manipulation View paper
- [6] Embodied navigation with multi-modal information: A survey from tasks to methodology View paper
- [7] Multi-agent embodied visual semantic navigation with scene prior knowledge View paper
- [8] Cooperative multi-agent learning for navigation via structured state abstraction View paper
- [9] Online control barrier functions for decentralized multi-agent navigation View paper
- [10] Entl: Embodied navigation trajectory learner View paper
- [11] Learning control admissibility models with graph neural networks for multi-agent navigation View paper
- [12] Co-Optimization of Environment and Policies for Decentralized Multi-Agent Navigation View paper
- [13] Learning team-based navigation: a review of deep reinforcement learning techniques for multi-agent pathfinding View paper
- [14] Agentic LLM-based robotic systems for real-world applications: a review on their agenticness and ethics View paper
- [15] Unifying Modern AI with Robotics: Survey on MDPs with Diffusion and Foundation Models View paper
- [16] Constrained environment optimization for prioritized multi-agent navigation View paper
- [17] NaviMaster: Learning a Unified Policy for GUI and Embodied Navigation Tasks View paper
- [18] Multi-agent Embodied AI: Advances and Future Directions View paper
- [19] OctoNav: Towards Generalist Embodied Navigation View paper
- [20] Environment optimization for multi-agent navigation View paper
- [21] Safe Multi-Agent Navigation Guided by Goal-Conditioned Safe Reinforcement Learning View paper
- [22] Embodied Multi-Agent Systems: Perception, Action, and Learning View paper
- [23] ALAN: adaptive learning for multi-agent navigation View paper
- [24] OmniEAR: Benchmarking Agent Reasoning in Embodied Tasks View paper
- [25] A Bio-Inspired Learning and Control Framework for Cross-Embodiment and Cross-Task Locomotion View paper
- [26] Decentralized, unlabeled multi-agent navigation in obstacle-rich environments using graph neural networks View paper
- [27] Adaptive learning for multi-agent navigation View paper
- [28] NavSpace: How Navigation Agents Follow Spatial Intelligence Instructions View paper
- [29] Coordinating multi-agent navigation by learning communication View paper
- [30] Analyzing Visual Representations in Embodied Navigation Tasks View paper
- [31] Deep Learning-Powered Embodied Navigation in Simulated Environments View paper
- [32] Heterogeneous Embodied Multi-Agent Collaboration View paper
- [33] Learning Embodied AI Agents with Task Decomposition View paper
- [34] Hierarchical Auto-Organizing System for Open-Ended Multi-Agent Navigation View paper
- [35] BLM: A Boundless Large Model for Cross-Space, Cross-Task, and Cross-Embodiment Learning View paper
- [36] Co-Optimizing Reconfigurable Environments and Policies for Decentralized Multi-Agent Navigation View paper
- [37] Offline Visual Representation Learning for Embodied Navigation View paper
- [38] Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks View paper
- [39] Learning Distributed Safe Multi-Agent Navigation via Infinite-Horizon Optimal Graph Control View paper
- [40] Differentiable Learning of Scalable Multi-Agent Navigation Policies View paper

- [41] BLM$_1$: A Boundless Large Model for Cross-Space, Cross-Task, and Cross-Embodiment Learning View paper
- [42] A Robust, Task-Agnostic and Fully-Scalable Voxel Mapping System for Large Scale Environments View paper
- [43] Embodied Multi-Agent Task Planning from Ambiguous Instruction View paper
- [44] Pushing the Limits of Cross-Embodiment Learning for Manipulation and Navigation View paper
- [45] Learning Graph-Enhanced Commander-Executor for Multi-Agent Navigation View paper
- [46] Guest Editorial: Special Issue on Embodied AI in Indoor Robotics: Bridging Perception, Interaction, and Autonomy View paper
- [47] MASP: Scalable GNN-based Planning for Multi-Agent Navigation View paper
- [48] Learning to Align Multimodal Data for Static and Dynamic Tasks View paper
- [49] AIRSHIP: Empowering General-Purpose Intelligent Robots through Open-Source Embodied AI View paper
- [50] Time-aware MADDPG with LSTM for multi-agent obstacle avoidance: a comparative study View paper
- [51] Learning Adaptive and Temporally Causal Video Tokenization in a 1D Latent Space View paper
- [52] Rl of thoughts: Navigating llm reasoning with inference-time reinforcement learning View paper
- [53] Beings: Bayesian embodied image-goal navigation with gaussian splatting View paper
- [54] Henet: Hybrid encoding for end-to-end multi-task 3d perception from multi-view cameras View paper
- [55] NaviFormer: A Spatio-Temporal Context-Aware Transformer for Object Navigation View paper
- [56] Simultaneous multi-view camera pose estimation and object tracking with squared planar markers View paper
- [57] Active SLAM With Dynamic Viewpoint Optimization for Robust Visual Navigation View paper
- [58] Learning multi-view camera relocalization with graph neural networks View paper
- [59] Spatiotemporal Contrastive Learning for Cross-View Video Localization in Unstructured Off-road Terrains View paper
- [60] Learning View-invariant and Novel Spatio-temporal Features Under Uncertainty from Video View paper
- [61] Virtual video camera: Image-based viewpoint navigation through space and time View paper
- [62] Real-time vision-aided localization and navigation based on three-view geometry View paper
- [63] Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities View paper
- [64] Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation View paper
- [65] Compass: Cross-embodiment mobility policy via residual rl and skill synthesis View paper
- [66] X-mobility: End-to-end generalizable navigation via world modeling View paper
- [67] Autonomous visual navigation for mobile robots: A systematic literature review View paper
- [68] Antcar: simple route following task with ants-inspired vision and neural model View paper
- [69] X-Nav: Learning End-to-End Cross-Embodiment Navigation for Mobile Robots View paper
- [70] A Cross-Environment and Cross-Embodiment Path Planning Framework via a Conditional Diffusion Model View paper
- [71] Design of AI based Autonomous Navigation System Using Swarm Intelligence Techniques for Agriculture Application View paper