NAACL 2025 Tutorial Proposal: Foundation Models meet Embodied Agents

Manling Li[1], **Yunzhu Li**[2], **Jiayuan Mao**[3], **Wenlong Huang**[4]
[1]Northwestern [2]Columbia [3]MIT [4]Stanford
https://foundation-models-meet-embodied-agents.github.io

Keywords: Embodied Agent, Foundation Models, Large Language Models (LLMs), Vision-Language Models (VLMs), Vision-Language-Action Models (VLAs), Embodied Decision Making, Robotics

Duration: Three Hours

Type: Cutting-edge in CL / NLP

Venue Preference: NAACL > ACL/EMNLP, due to visa constraints.

1 Tutorial Content

An embodied agent is a generalist agent that can take natural language instructions from humans and perform a wide range of tasks in diverse environments. Recent years have witnessed the emergence of Large Language Models (LLMs) as powerful tools for building Large Agent Models (LAMs), which have shown remarkable success in supporting embodied agents for different abilities such as goal interpretation (Ding et al., 2023b; Xie et al., 2023; Liu et al., 2023a; Hazra et al., 2023; Joublin et al., 2023; Wu et al., 2023; Wang et al., 2024b; Smirnov et al., 2024), subgoal decomposition (Nottingham et al., 2023; Ahn et al., 2022; Zhu et al., 2023; Chen et al., 2023; Song et al., 2022; Wang et al., 2023a), action sequencing (Silver et al., 2023; Liang et al., 2024; Wang et al., 2023c; Hu et al., 2023; Ni et al., 2023; Zhao et al., 2023; Wang et al., 2024a; Dalal et al., 2024; Liu et al., 2023b; Rana et al., 2023; Liang et al., 2022), and transition modeling (causal transitions from preconditions to post-effects) (Guan et al., 2023; Raman et al., 2022; Smirnov et al., 2024; Singh et al., 2022; Wang et al., 2023b; Li et al., 2024; Wong et al., 2023). However, moving from language models to embodied agents poses significant challenges in understanding lower-level visual details, and long-horizon reasoning for reliable embodied decision-making. We categorize the foundation models into Large Language Models (**LLMs**), Vision-Language Models (**VLMs**), and Vision-Language-Action Models (**VLAs**). We investigate how embodied decision-making abilities differ between these models and how they scale as they get larger.

This tutorial will present a systematic overview of recent advances in foundation models for embodied agents, covering three types of foundation models based on input and output:

- **Large Language Models** (LLMs)
- **Vision-Language Models** (VLMs)
- **Vision-Language-Action Models** (VLAs)

We compare these models and explore their design space to guide future developments, focusing on the following key aspect:

- **Lower-Level Environment Encoding and Interaction**: We are tackling the challenge of helping LLMs truly understand the physical world, especially geometric perception learning. This means teaching it about spatial relationships, how objects are defined and located, and how concepts can be built up from simpler parts, how changes in the world can be modeled as a result of actions, and preconditions and post-effect. In detail, we work on the key challenges:
  - **State/Object Representation**: the ability to interact with its environment, understand intricate visual details, and grasp complex geometric structures;
  - **Action Representation**: the ability to control state transitions from preconditions to post-effects;
  - **Goal Representation**: the ability to interpret goals and ground to the environment;
  - **Trajectory Representation**: the ability to represent a trajectory of action sequence to achieve the goal;

Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 15–24

  - **Reward Representation**: the ability to quantify the progress of goal achievement, interpreting implicit rewards from human feedback or task completion.

- **Longer-Horizon Decision Making**: We are working on enhancing LLM's ability to reason over longer periods. We will formulate the decison making process as Markov Decision Process, including:

  - **Goal Interpretation**: given natural language instructions, output environment-grounded goal states.
  - **Subgoal Decomposition**: given a goal, output a sequence of states to be achieved as subgoals.
  - **Action Sequencing**: given a goal, output a sequence of actions to achieve the goal states.
  - **Transition Modeling**: given an action, predict and control the pre-conditions and post-effects of object states.

Target Audience: We expect audience from natural language processing (NLP) community, robotics community, computer vision (CV) community, and machine learning (ML) communities. **Prerequisite Knowledge:** While no specific background knowledge is assumed of the audience, it would be best for the attendees to know about basic deep learning and foundation model technologies, such as pre-trained language models and vision-language models.

Relevance to the CL / NLP Community: CL/NLP audience will learn of recent trends and emerging challenges in leveraging language modeling for embodied agents, as well as learning resources and tools for participants to obtain ready-to-use models and benchmarks, prompting thorough discussions regarding the impact of foundation models on embodied intelligence. In this tutorial, we will comprehensively review existing paradigms for foundations for embodied agents, and focus on their different formulations based on the fundamental mathematical framework of robot learning, Markov Decision Process (MDP), and design a structured view to investigate the robot's decision-making process.

Content	Time
Motivation and Overview	15 mins
Foundation Models meet Virtual Agents:	45 mins
Environment Overview	5 mins
State Estimation	5 mins
MDP Policy Learning	10 mins
Reward Modeling	5 mins
Transition Modeling	5 mins
Large World Model	10 mins
Evaluation	5 mins
Foundation Models meet Physical Agents:	75 mins
MDP Formulation Overview	25 mins
Physical World Perception	
High-level Planning	50 mins
Low-level Planning	
Break	30 mins
Robotic Foundation Models (VLAs)	30 mins
Remaining Challenges	15 mins
QA	30 mins

Table 1: Conference Schedule.

2 Tutorial Outline

2.1 Motivation and Overview [15min]

We will define the main research problem and motivate the topic by presenting embodied decision making of multiple representation levels and ability modules. We will categorize the foundation models and outline the road map.

2.2 Foundation Models meet Virtual Agents based on Markov Decision Process Formulation [45min]

We will ground the abilities of foundation model to the fundamental modules of MDP: The embodied agent receives a natural language goal specification, translates it to the environment objects and their states, relations, and actions as a goal specification, and aims to achieve it through a sequence of state transitions. To abstract the embodied environment, we design the representation to contain *Object*, *State*, *Action*, and, based on that, *Goal* (as final states) and *Trajectory* (as temporally dependent sequences of actions/states). Existing works in embodied task and motion planning (TAMP) have used LLMs to perform varying tasks, serving different abilities. Here in Table 2 we provide a extended list of such works with detailed categorization.

2.2.1 Physical World Perception [10min]

Input of States, which corresponds to **State Estimation**, grounding environment to objects and their relations and actions.

2.2.2 Goal Interpretation [10min]

Input of Goals, which corresponds to **Goal Interpretation**, translating the natural language goal to environment objects and their relations and actions.

2.2.3 High-Level Planning: Subgoal Decomposition and Action Sequencing [10min]

Output of Trajectories, where the output can be a sequence of actions or a sequence of states, which can be regarded as **Action Sequencing** and **Subgoal Decomposition**. The Subgoal Decomposition and Action Sequencing modules are similar in that they both involve trajectory output and evaluate the ordering of decision making. However, the fundamental distinction between them lies in the nature of their outputs. Action sequencing produces imperative actions, while subgoal decomposition generates declarative states.

2.2.4 Low-level Planning: Transition modeling [10min]

Transition modeling can be considered as the low-level controller that governs the state transitions when executing an action. The bottleneck for transition modeling is the ability to search a path to navigate from initial predicates to goal predicates using existing actions. We will define preconditions and post effects for each action enables this search and backtracking.

2.3 Foundation Models meet Embodied Agents [75min]

2.3.1 MDP Formulation Overview and Physical World Perception [25min]

We will introduce the foundation models interacting with physical world: PaLM-E (Huang and Others, 2023) and ConceptGraphs (Garcia and Others, 2023) emphasize high-level planning, while (Chao et al., 2021) and (Yadav and Others, 2024) focus on low-level actions. CoPa (Lee and Others, 2024) bridges this gap with sophisticated VLM and GraspNet optimization. Environmental representations evolve from simple images to complex state graphs (Garcia and Others, 2023). Optimization functions range from imitation learning (Chao et al., 2021) to constrained optimization (Yadav and Others, 2024) and multi-step processes (Lee and Others, 2024). Advanced models introduce spatio-temporal reasoning (Yadav and Others, 2024) and open-vocabulary 3D scene graphs (Garcia and Others, 2023), demonstrating trends toward sophisticated environmental understanding and AI integration in robotic manipulation.

2.3.2 Low-Level and High-Level Planning [50min]

The key difference between LLMs and LAMs lies in their ability to make decisions. While LLMs align instructions with language output, LAMs must align goals with decision-making trajectories. Traditional causal reasoning requires disentangling all elements, but foundation models often entangle them. This means foundation model training does not truly teach reasoning but rather learns input-output distribution mapping. Agent models focus on decision-making ability, aligning goals with decision-making trajectories rather than just instructions with language output. LLMs have been widely used for different abilities such as goal interpretation (Ding et al., 2023b; Xie et al., 2023; Huang et al., 2022b; Lin et al., 2023; Ahn et al., 2022; Liu et al., 2023a; Hazra et al., 2023; Joublin et al., 2023; Zha et al., 2023; Huang et al., 2023; Zhu et al., 2023; Guan et al., 2023; Wake et al., 2023; Wu et al., 2023; Wang et al., 2024b; Smirnov et al., 2024), subgoal decomposition (Nottingham et al., 2023; Ahn et al., 2022; Zhu et al., 2023; Chen et al., 2023; Song et al., 2022; Wang et al., 2023a), action sequencing (Silver et al., 2023; Hao et al., 2023; Liang et al., 2024; Chen et al., 2024; Xu et al., 2023; Wang et al., 2023c; Hu et al., 2023; Liu et al., 2024; Ni et al., 2023; Chalvatzaki et al., 2023; Wu et al., 2024; Mavrogiannis et al., 2023; Zhao et al., 2023; Li et al., 2023; Wang et al., 2024a; Parakh et al., 2023; Dalal et al., 2024; Liu et al., 2023b; Rana et al., 2023; Liang et al., 2022; Wang et al., 2023a; Wong et al., 2023; Huang et al., 2022a), and transition modeling (causal transitions from preconditions to post-effects) (Guan et al., 2023; Raman et al., 2022; Smirnov et al., 2024; Ding et al., 2023a; Singh et al., 2022; Wang et al., 2023b; Li et al., 2024; Wong et al., 2023).

2.4 Robot Foundation Models (Vision-Language-Action Models) [30min]

VLAs combine vision, language, and robotic control to enable robots to understand scenes and act

(Smith and Doe, 2023). Using pretrained models, they aim to improve generalization across tasks and environments (Jones and Brown, 2022; Brown and Green, 2024), potentially enabling complex multi-step tasks from high-level instructions (Kim et al., 2024). VLAs could bridge human-robot communication (Lee and Taylor, 2023) and enhance robotic flexibility across domains (Kim et al., 2024). We will then layout remaining challenges in reliability, safety, and adaptation (Wilson and Martinez, 2024).

2.5 Remaining Challenges [15min]

We will conclude the tutorial by discussing outstanding challenges and promising research directions in three key areas: low-level visual perception, long-horizon decision making, and trustworthy decision making for verifying decision correctness.

2.6 Panel and QA [30min]

We will discuss a dozen potential PhD dissertation topics focused on the new frontiers of model-driven and data-driven improvements to foundation models for embodied agents. It will be done in a more interactive panel format. We will also answer questions from audience.

3 Logistics

3.1 Reading list

Please find reading list in Table 2. We agree to allow the publication of the tutorial materials and presentation in the ACL Anthology. All the materials are openly available at https://embodied-foundation-model.github.io/.

3.2 Tutorial Size / Prior Tutorials

Based on the level of interest in this topic, we expect around 100-150 participants interested in planning and interactions with physical world. No special requirements for technical equipment are needed.

This tutorial has not been presented elsewhere. The presented topic has not been covered by previous AAAI/IJCAI/NeurIPS/CVPR/ACL tutorials in the past five years. There are tutorials on vision-language pretraining at CVPR 2024 (Jun 2024, around 300 audience)* but without much involvement of embodied AI. Another related tutorial is LLMs for Planning tutorial at AAAI 2024 (Feb 2024, around 200 audience)† without discussing the advancement of VLMs and VLAs. In contrast, we focus on the a systematic analysis of foundation models for embodied intelligence, including LLMs, VLMs and VLAs.

3.3 Diversity and Inclusion

Topics Promoting Diversity: Our team represents a diverse range of expertise in foundation models for embodied agents, covering LLMs, VLMs and VLAs. Manling is an experts in LLMs for embodied agents from a language modeling background, while Jiayuan focus on LLMs for embodied agents with emphasis on symbolic representations and robotics interfaces. Yunzhu bring strong backgrounds in VLMs supported embodied agents. Collectively, the team has earned multiple best paper awards from top conferences including CVPR, ICCV, ECCV, CoRL, RSS, and NeurIPS. This combination of specializations enables us to provide a comprehensive view of embodied foundation models.

Diversifying representation: We have a diverse organizer team across multiple institutions (Northwestern, Stanford, MIT, Columbia) with varying seniority (ranging from senior PhD students to assistant professors and research scientists), gender (2 out of 3 organizers are female researchers), race, and ethnicity.

Diversifying participation: The call-for-papers aims to bridge researchers from different subcommunities within NLP (such as IE, QA, IR, dialog, etc) and outside NLP (such as representation learning, data science, etc), and will encourage people to create a more comprehensive view of the problem.

4 Tutorial Presenters

Manling Li (manling.li@northwestern.edu) is an assistant professor at Northwestern University and a postdoc at Stanford University. She obtained her PhD in computer science at UIUC in 2023. Her work on multimodal knowledge extraction won the ACL'20 Best Demo Paper and NAACL'21 Best Demo Paper, and LLMs controlling won the ACL'24 Outstanding Paper. She was a recipient of MSR PhD Fellowship, DARPA Riser, EE CS Rising Star, etc. She served on the Organizing Committee of ACL 25 (Virtual Infrastructure

*https://vlp-tutorial.github.io/

†https://yochan-lab.github.io/tutorial/LLMs-Planning/index.html

Co-Chairs), NAACL 25 (Publication Co-Chairs), EMNLP 24 (Demo Co-Chairs), and organized the 1st Knowledgeable LLM workshop at ACL 2024 and AAAI 2025. She has delivered Workshops at multiple conferences including AAAI'21, ACL'21, NAACL'22, AAAI'23, CVPR'23, and IJCAI'24. Additional information is available at https://limanling.github.io. Her previous tutorials include:

- IJCAI'24: Beyond Human Creativity: A Tutorial on Advancements in AI Generated Content (AIGC).
- CVPR'23/AAAI'23: Knowledge-Driven Vision-Language Pretraining.
- NAACL'22: New Frontiers of Information Extraction.
- AAAI'21/ACL'21: Event-Centric Natural Language Understanding.

Jiayuan Mao (jiayuanm@mit.edu) is a Ph.D. student at MIT, advised by Professors Josh Tenenbaum and Leslie Kaelbling. Her research agenda is to build machines that can continually learn concepts (e.g., properties, relations, rules, and skills) from their experiences and apply them for reasoning and planning in the physical world. Her research topics include visual reasoning, robotic manipulation, scene and activity understanding, and language acquisition. Her work is supported by an MIT presidential fellowship. She has co-organized the Workshop on Planning in the Era of LLMs at AAAI 2024, the Workshop on Learning Effective Abstractions for Planning at CoRL 2024, the Workshop on Visual Concepts at ECCV 2024, the workshop on Visually Grounded Interaction and Language (VIGIL) at NAACL 2021, and the Neuro-Symbolic Visual Reasoning and Program Synthesis tutorial at CVPR 2020. She has served as a reviewer for ICML, NeurIPS, ICLR, CVPR, ICCV, ECCV, ACL, CoLM, ICRA, RSS, CoRL, AAAI, IJCAI, ICAPS, T-PAMI, and IJCV. Additional information is available at https://jiayuanm.com. Her previous tutorials include:

- CVPR2020: Neuro-Symbolic Visual Reasoning and Program Synthesis tutorial.

Yunzhu Li (yunzhu.li@columbia.edu) is an Assistant Professor of Computer Science at Columbia University. Before joining Columbia, he was an Assistant Professor at UIUC CS, spent time as a Postdoc at Stanford, and earned his PhD from MIT. His work stands at the intersection of robotics, computer vision, and machine learning, with the goal of helping robots perceive and interact with the physical world as dexterously and effectively as humans do. Yunzhu's work has been recognized through the Best Systems Paper Award and the Finalist for Best Paper Award at CoRL. He is also the recipient of the Sony Faculty Innovation Award, the Adobe Research Fellowship, and was selected as the First Place Recipient of the Ernst A. Guillemin Master's Thesis Award in Artificial Intelligence and Decision Making at MIT. His research has been published in top journals and conferences, including Nature, Science, NeurIPS, CVPR, and RSS, and featured by major media outlets, including CNN, BBC, The Wall Street Journal, Forbes, The Economist, and MIT Technology Review. Additional information is available at https://yunzhuli.github.io. His previous tutorials include:

- CVPR2021: Learning Representations via Graph-structured Networks.
- ICCV2021: Multi-Modality Learning from Videos and Beyond.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Jayant Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego M Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. 2022. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*.

Charlie Brown and David Green. 2024. Generalization capabilities of vla models in diverse robotic tasks. *Robotics and Autonomous Systems*, 120:103567.

Georgia Chalvatzaki, Ali Younes, Daljeet Nandha, An T. Le, Leonardo F. R. Ribeiro, and Iryna Gurevych. 2023. Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning. *Frontiers in Robotics and AI*, 10.

D. Chao, Y. Yang, J. Yang, S. Gu, H. Wu, W. Matusik, and Others. 2021. Cliport: What and where pathways for robotic manipulation. *arXiv preprint arXiv:2109.xxx*.

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. Prompt optimization in multi-step tasks (promst): Integrating human feedback and preference alignment. *ArXiv*, abs/2402.08702.

Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas A. Roy, and Chuchu Fan. 2023. Autotamp: Autoregressive task and motion planning with llms as translators and checkers. *ArXiv*, abs/2306.06531.

Murtaza Dalal, Tarun Chiruvolu, Devendra Singh Chaplot, and Ruslan Salakhutdinov. 2024. Plan-seq-learn: Language model guided rl for solving long horizon robotics tasks. In *ICLR*.

Yan Ding, Xiaohan Zhang, S. Amiri, Nieqing Cao, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. 2023a. Integrating action knowledge and llms for task planning and situation handling in open worlds. *Autonomous Robots*, 47:981 – 997.

Yan Ding, Xiaohan Zhang, Chris Paxton, and Shiqi Zhang. 2023b. Task and motion planning with large language models for object rearrangement. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2086–2092.

M. Garcia and Others. 2023. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. *arXiv preprint arXiv:2309.xxx*.

L. Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. *ArXiv*, abs/2305.14909.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *ArXiv*, abs/2305.14992.

Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. 2023. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. *ArXiv*, abs/2308.12682.

Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *ArXiv*, abs/2311.17842.

Wenlong Huang, P. Abbeel, Deepak Pathak, and Igor Mordatch. 2022a. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *ArXiv*, abs/2201.07207.

Wenlong Huang, F. Xia, Ted Xiao, Harris Chan, Jacky Liang, Peter R. Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022b. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*.

Wenlong Huang, Fei Xia, Dhruv Shah, Danny Driess, Andy Zeng, Yao Lu, Pete Florence, Igor Mordatch, Sergey Levine, Karol Hausman, and Brian Ichter. 2023. Grounded decoding: Guiding text generation with grounded models for embodied agents. In *Neural Information Processing Systems*.

Y. Huang and Others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.xxx*.

Alice Jones and Bob Brown. 2022. Combining vision, language, and action for robotic control. In *Proceedings of the International Conference on Robotics and Automation*, pages 1200–1207. IEEE.

Frank Joublin, Antonello Ceravola, Pavel Smirnov, Felix Ocker, Joerg Deigmoeller, Anna Belardinelli, Chao Wang, Stephan Hasler, Daniel Tanneberg, and Michael Gienger. 2023. Copal: Corrective planning of robot actions with large language models. *ArXiv*, abs/2310.07263.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Emma Lee and Frank Taylor. 2023. Bridging human communication and robotic action through vla models. In *Conference on Neural Information Processing Systems*, pages 5500–5512.

J. Lee and Others. 2024. Copa: General robotic manipulation through spatial constraints of parts with foundation models. *arXiv preprint arXiv:2403.xxx*.

Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. 2023. Interactive task planning with language models. *ArXiv*, abs/2310.10645.

Zhaoyi Li, Kelin Yu, Shuo Cheng, and Danfei Xu. 2024. LEAGUE++: EMPOWERING CONTINUAL ROBOT LEARNING THROUGH GUIDED SKILL ACQUISITION WITH LARGE LANGUAGE MODELS. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Jacky Liang, Wenlong Huang, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500.

Jacky Liang, Fei Xia, Wenhao Yu, Andy Zeng, Montse Gonzalez Arenas, Maria Attarian, Maria

Bauza, Matthew Bennice, Alex Bewley, Adil Dostmohamed, Chuyuan Fu, Nimrod Gileadi, Marissa Giustina, Keerthana Gopalakrishnan, Leonard Hasenclever, Jan Humplik, Jasmine Hsu, Nikhil Joshi, Ben Jyenis, Chase Kew, Sean Kirmani, Tsang-Wei Edward Lee, Kuang-Huei Lee, Assaf Hurwitz Michaely, Joss Moore, Kenneth Oslund, Dushyant Rao, Allen Z. Ren, Baruch Tabanpour, Quan Ho Vuong, Ayzaan Wahid, Ted Xiao, Ying Xu, Vincent Zhuang, Peng Xu, Erik Frey, Ken Caluwaerts, Ting-Yu Zhang, Brian Ichter, Jonathan Tompson, Leila Takayama, Vincent Vanhoucke, Izhak Shafran, Maja Mataric, Dorsa Sadigh, Nicolas Manfred Otto Heess, Kanishka Rao, Nik Stewart, Jie Tan, and Carolina Parada. 2024. Learning to learn faster from human feedback with language model predictive control. *ArXiv*, abs/2402.11450.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. 2023. Text2motion: from natural language instructions to feasible plans. *Autonomous Robots*, 47:1345 – 1365.

B. Liu, Yuqian Jiang, Xiaohan Zhang, Qian Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. Llm+p: Empowering large language models with optimal planning proficiency. *ArXiv*, abs/2304.11477.

Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. 2024. Delta: Decomposed efficient long-term robot task planning using large language models. *ArXiv*, abs/2404.03275.

Zeyi Liu, Arpit Bahety, and Shuran Song. 2023b. Reflect: Summarizing robot experiences for failure explanation and correction. *ArXiv*, abs/2306.15724.

A. Mavrogiannis, Christoforos Mavrogiannis, and Yiannis Aloimonos. 2023. Cook2ltl: Translating cooking recipes to ltl formulae using large language models. *ArXiv*, abs/2310.00163.

Zhe Ni, Xiao-Xin Deng, Cong Tai, Xin-Yue Zhu, Xiang Wu, Y. Liu, and Long Zeng. 2023. Grid: Scene-graph-based instruction-driven robotic task planning. *ArXiv*, abs/2309.07726.

Kolby Nottingham, Prithviraj Ammanabrolu, Alane Suhr, Yejin Choi, Hannaneh Hajishirzi, Sameer Singh, and Roy Fox. 2023. Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. In *International Conference on Machine Learning*.

Meenal Parakh, Alisha Fong, Anthony Simeonov, Abhishek Gupta, Tao Chen, and Pulkit Agrawal. 2023. Lifelong robot learning with human assisted language planners. *arXiv:2309.14321*.

S. Sundar Raman, Vanya Cohen, Eric Rosen, Ifrah Idrees, David Paulius, and Stefanie Tellex. 2022. Cape: Corrective actions from precondition errors using large language models. In *ICRA*.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D. Reid, and Niko Sünderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In *Conference on Robot Learning*.

Tom Silver, Soham Dan, Kavitha Srinivas, Joshua B. Tenenbaum, Leslie Pack Kaelbling, and Michael Katz. 2023. Generalized planning in pddl domains with pretrained large language models. In *AAAI Conference on Artificial Intelligence*.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. Progprompt: Generating situated robot task plans using large language models. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530.

Pavel Smirnov, Frank Joublin, Antonello Ceravola, and Michael Gienger. 2024. Generating consistent pddl domains with large language models. *ArXiv*, abs/2404.07751.

John Smith and Jane Doe. 2023. Vision-language-action models: A comprehensive survey. *Journal of Artificial Intelligence*, 15(3):300–325.

Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2986–2997.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Neural Information Processing Systems*.

Naoki Wake, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi. 2023. Chatgpt empowered long-step robot control in various environments: A case application. *IEEE Access*, 11:95060–95078.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi (Jim) Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *ArXiv*, abs/2305.16291.

Huaxiaoyue Wang, Gonzalo Gonzalez-Pumariega, Yash Sharma, and Sanjiban Choudhury. 2023b. Demo2code: From summarizing demonstrations to synthesizing code via extended chain-of-thought. *ArXiv*, abs/2305.16744.

Huaxiaoyue Wang, K. Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, Gonzalo Gonzalez-Pumariega, Aditya Kompella, Maximus Adrian Pace, Yash Sharma, Xiangwan Sun, Neha Sunkara, and Sanjiban Choudhury. 2024a. Mosaic: A modular system for assistive and interactive cooking. *ArXiv*, abs/2402.18796.

J. Wang, Jiaming Tong, Kai Liang Tan, Yevgeniy Vorobeychik, and Yiannis Kantaros. 2023c. Conformal temporal logic planning using large language models: Knowing when to do what and when to ask for help. *ArXiv*, abs/2309.10092.

Shu Wang, Muzhi Han, Ziyuan Jiao, Zeyu Zhang, Yingnian Wu, Song-Chun Zhu, and Hangxin Liu. 2024b. Llm3: Large language model-based task and motion planning with motion failure reasoning. *ArXiv*, abs/2403.11552.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023d. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *ArXiv*, abs/2302.01560.

Grace Wilson and Hugo Martinez. 2024. Challenges in deploying vla models for real-world robotic applications. Technical Report SAIL-TR-2024-001, Stanford AI Lab.

Li Siang Wong, Jiayuan Mao, Pratyusha Sharma, Zachary S. Siegel, Jiahai Feng, Noa Korneev, Joshua B Tenenbaum, and Jacob Andreas. 2023. Learning adaptive planning representations with natural language guidance. *ArXiv*, abs/2312.08566.

Yike Wu, Jiatao Zhang, Nan Hu, LanLing Tang, Guilin Qi, Jun Shao, Jie Ren, and Wei Song. 2024. Mldt: Multi-level decomposition for complex long-horizon robotic task planning with open-source large language model. *ArXiv*, abs/2403.18760.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. 2023. Embodied task planning with large language models. *ArXiv*, abs/2307.01848.

Yaqi Xie, Chenyao Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh. 2023. Translating natural language to planning goals with large-language models. *ArXiv*, abs/2302.05128.

Mengdi Xu, Peide Huang, Wenhao Yu, Shiqi Liu, Xilun Zhang, Yaru Niu, Tingnan Zhang, Fei Xia, Jie Tan, and Ding Zhao. 2023. Creative robot tool use with large language models. *ArXiv*, abs/2310.13065.

R. Yadav and Others. 2024. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *Private Manuscript*.

Lihan Zha, Yuchen Cui, Li-Heng Lin, Minae Kwon, Montse Gonzalez Arenas, Andy Zeng, Fei Xia, and Dorsa Sadigh. 2023. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. *ArXiv*, abs/2311.10678.

Mandi Zhao, Shreeya Jain, and Shuran Song. 2023. Roco: Dialectic multi-robot collaboration with large language models. *ArXiv*, abs/2307.04738.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Y. Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *ArXiv*, abs/2305.17144.

A Appendix

Table 2: Readling List, focusing on existing Embodied Agent planning works usage of Large Language Models: Each "FMs" refers to the usage of FMs to perform an ability module. For example, Ada (Wong et al., 2023) uses FMs for Action Sequencing and Transition Modeling, while LLM+P (Liu et al., 2023a) uses FMs for Goal Interpretation. (Part 1)

Existing Work	Ref.	Goal Interpretation	Action Sequencing	Subgoal Decomposition	Transition Modeling
SayCan	(Ahn et al., 2022)	FMs	FMs		
Ada	(Wong et al., 2023)	FMs			FMs
LLP+P	(Liu et al., 2023a)	FMs			
AutoTAMP	(Chen et al., 2023)		FMs		FMs
Code as Policies	(Liang et al., 2022)	FMs	FMs	FMs	
Voyager	(Wang et al., 2023a)	FMs	FMs		
Demo2Code	(Wang et al., 2023b)	FMs		FMs	FMs
LM as ZeroShot Planner	(Huang et al., 2022a)		FMs	FMs	
SayPlan	(Rana et al., 2023)	FMs	FMs		FMs
Text2Motion	(Lin et al., 2023)		FMs		
LLMGROP	(Ding et al., 2023b)	FMs	FMs		
REFLECT	(Liu et al., 2023b)	FMs	FMs		
Generating Consistent PDDL Domains with FMs	(Smirnov et al., 2024)	FMs			FMs
PlanSeqLearn	(Dalal et al., 2024)		FMs		
COWP	(Ding et al., 2023a)	FMs	FMs		FMs
HumanAssisted Continual Robot Learning	(Parakh et al., 2023)		FMs		
DECKARD	(Nottingham et al., 2023)	FMs			FMs
MOSAIC	(Wang et al., 2024a)		FMs		
Interactive Task Planning with Language Models	(Li et al., 2023)		FMs	FMs	
RoCo	(Zhao et al., 2023)		FMs		
Cook2LTL	(Mavrogiannis et al., 2023)	FMs			
InnerMonologue	(Huang et al., 2022b)		FMs		
MLDT	(Wu et al., 2024)		FMs		
Learning to Reason over Scene Graphs	(Chalvatzaki et al., 2023)	FMs	FMs		FMs
GRID	(Ni et al., 2023)	FMs	FMs		

Table 3: Categorization of Existing Embodied Agent Planning Works' Usage of Large Language Models (Part 2)

Existing Work	Ref.	Goal Interpretation	Action Sequencing	Subgoal Decomposition	Transition Modeling
LLMplanner	(Song et al., 2022)	FMs	FMs		
DELTA	(Liu et al., 2024)		FMs		
Look Before You Leap	(Hu et al., 2023)	FMs	FMs		
CAPE	(Raman et al., 2022)	FMs	FMs		
HERACLEs	(Wang et al., 2023c)		FMs		
RoboTool	(Xu et al., 2023)		FMs		FMs
PROMST	(Chen et al., 2024)		FMs		
LLM3	(Wang et al., 2024b)	FMs	FMs		
Ghost in the Minecraft	(Zhu et al., 2023)		FMs		
PlanBench	(Valmeekam et al., 2022)	FMs	FMs		
TaPA	(Wu et al., 2023)	FMs	FMs	FMs	
ChatGPT Robot Control	(Wake et al., 2023)		FMs		
LLM World Models for Planning	(Guan et al., 2023)	FMs	FMs		
DEPS	(Wang et al., 2023d)	FMs	FMs		
Grounded Decoding	(Huang et al., 2023)		FMs		
ProgPrompt	(Singh et al., 2022)	FMs	FMs		
DROC	(Zha et al., 2023)		FMs		FMs
LMPC	(Liang et al., 2024)	FMs	FMs		
GPTPDDL	(Xie et al., 2023)		FMs		
RAP	(Hao et al., 2023)		FMs		
LEAGUE++	(Li et al., 2024)		FMs		FMs
CoPAL	(Joublin et al., 2023)	FMs	FMs		
SayCanPay	(Hazra et al., 2023)	FMs	FMs		
LLMGenPlan	(Silver et al., 2023)		FMs		