ICCV 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
manling.li@northwestern.edu, yunzhu.li@columbia.edu, jiayuanm@mit.edu, wenlongh@stanford.edu
https://embodied-foundation-model.github.io/

Abstract

*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 foundation models, which have shown remarkable success in supporting embodied agents for different abilities such as goal interpretation, subgoal decomposition, action sequencing, and transition modeling (causal transitions from preconditions to post-effects). We categorize the foundation models into Large Language Models (**LLMs**), Vision-Language Models (**VLMs**), and Vision-Language-Action Models (**VLAs**). 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.*

Primary Subject Area: *Vision, language, and reasoning*
Subject Area: *Computer vision for robotics, Embodied vision, Scene analysis and understanding, Embodied decision making, Vision-Langauge-Action models*
Duration: *Half-day*
Format: *Mixed in-person attendance, with virtual attendees supported on Zoom.*
Size: *Medium 100-300 participants*

1. Course Description

This tutorial will present a systematic overview of recent advances in foundation models for embodied agents. 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 pre-conditions 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;
  - **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.

2. Tutorial Content

2.1. Motivation and Overview [10min, Jiayuan]

We will define the main research problem and motivate the topic by defining generalist embodied agents. We will categorize the foundation models and outline the road map.

Content	Time
Motivation and Overview	10 mins
Embodied Agent Formulation with MDP	15 mins
Modularized: Virtual Agents	35 mins
- State Estimation	- 10 mins
- Goal Interpretation	- 5 mins
- Policy Learning	- 10 mins
- Reward Modeling	- 10 mins
Modularized: Physical Agents	75 mins
- Physical World Perception	- 25 mins
- High-Level Planning	- 25 mins
- Low-Level Planning	- 25 mins
End-to-End: VLA models	30 mins
Remaining Challenges	15 mins
QA	30 mins

Table 1. Tutorial Outline.

2.2. Embodied Agent Interface based on on Markov Decision Process [15min, Manling]

We will present Embodied Agent Interface [4] to formulate embodied agents based on 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.

2.3. Modularized: Virtual Agents [35min, Manling]

Beginning with language as the cornerstone, we will introduce how LLMs provide the essential reasoning capabilities that enable virtual agents to comprehend instructions and their environment [6]. We then introduce this foundation extended by implementing code as policies [2] which allows agents to formalize decision-making processes as executable code. The RAGEN framework [9] represents a significant advancement in this domain, effectively combining the reasoning capabilities of large language models. We finally point out the future directions such as merging the interpretative power of language models with the procedural clarity of code, these systems represent the next evolution in agent-based AI, opening new possibilities for human-AI collaboration across various domains.

2.4. Modularized: Physical Agents [75min, Jiayuan, Wenlong]

We then will introduce the advances when foundation models meet physical agents.

2.4.1. Physical World Perception [25min, Jiayuan]

Input of States, which corresponds to **State Estimation**, grounding environment to objects and their relations and actions. In this section, we will firstly introduce the fundamental components of physical world perception for embodied agents. We will then layout recent advancements in perception frameworks, beginning with LayoutVLM [3], which integrates visual-language modeling for understanding complex spatial arrangements. We will demonstrate how this approach enables agents to reason about structural organization of physical spaces, facilitating more intuitive navigation capabilities. Next, we will explore keypoint identification [1] and its contributions to precise articulated object understanding. We will illustrate how this granular perception of object structures enables physical agents to perform more accurate manipulations and better predict physical interactions.

2.4.2. High-Level Planning [20min, Wenlong]

We will illustrate how these high-level planning approaches allow physical agents to navigate complex, multi-step tasks by reasoning about action sequences. We will firstly explore how code can serve as policies for these planning systems [2]. We will demonstrate how representing action policies as code provides several advantages, including improved interpretability, better debugging capabilities, and more straightforward integration with existing software systems. We will then present the concept of preconditions and post-effects as fundamental components that enable efficient search and backtracking in planning algorithms [10].

2.4.3. Low-Level Planning [25min, Wenlong]

Throughout this section, we will emphasize the importance of bridging the abstraction gap between high-level planning and physical execution, showing how these methods enable physical agents to translate symbolic plans into effective real-world actions across diverse environments and tasks. We will first present policy learning approaches, with a focus on Inner Monologue [7], which enables agents to develop robust control policies through self-reflection and iterative refinement. We will then explore goal interpretation frameworks that translate high-level objectives into specific motor commands, such as VoxPoser [8]. Next, we will examine ReKep [5] and its contributions to reliable physical execution through keypoint-based representations.

2.5. End-to-End: Vision-Language-Action Models [30min, Yunzhu]

In this section, we will introduce end-to-end approaches that integrate vision, language, and action into unified models

for physical agents. We will firstly explore recent architectures that enable direct mapping from observations and instructions to actions. Next, we will examine the benefits and challenges of end-to-end approaches compared to modularized systems. We will illustrate how these integrated models can potentially achieve greater flexibility and generalization across tasks, while discussing the tradeoffs in interpretability, sample efficiency, and robustness. Throughout this section, we will highlight emerging research directions in VLA models, including techniques for improving data efficiency, incorporating physical priors, and enhancing zero-shot generalization to novel tasks and environments.

2.6. Remaining Challenges [15min, Yunzhu]

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.

2.7. Panel and QA [30min]

We will answer questions from audience.

3. Target Audience

We expect audience from robotics community, computer vision (CV) community, natural language processing (NLP) 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 technologies, as well as pre-trained language models and vision-language models. **Planned Materials and Resources:** We will provide learning resources and tools in https://embodied-foundation-model.github.io/ for participants to obtain ready-to-use models and benchmarks.

4. Related Tutorials

The presented topic has not been covered by previous ICCV, CVPR, and ECCV tutorials within the last three years tutorials. 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 ICML 2024 (July 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.

5. Presenter Talks

Please find previous talks at https://youtu.be/OVutTCSeOKs and https://youtu.be/akDSG9FsoCk. **Special Notice:** Four presenters are featured to cover comprehensive expertise: Manling specializes in language-grounded planning, Jiayuan in symbolic planning, Yunzhu in robot learning, and Wenlong in foundation models for embodied agents.

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

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

References

[1] Xiaolin Fang, Bo-Ruei Huang, **Jiayuan Mao**, Jasmine Shone, Joshua B Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Keypoint abstraction using large models for object-relative imitation learning. *ICRA*, 2025. 2

[2] Jacky Liang, **Wenlong Huang**, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2022. 2

[3] Fan-Yun Sun, Weiyu Liu, Siyi Gu, Dylan Lim, Goutam Bhat, Federico Tombari, **Manling Li**, Nick Haber, and Jiajun Wu. Layoutvlm: Differentiable optimization of 3d layout via vision-language models. *CVPR*, 2025. 2

[4] **Manling Li**, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *NeurIPS*, 37: 100428–100534, 2024. 2

[5] **Wenlong Huang**, Chen Wang, **Yunzhu Li**, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. 2

[6] **Wenlong Huang**, P. Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *ArXiv*, abs/2201.07207, 2022. 2

[7] **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. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2022. 2

[8] **Wenlong Huang**, Chen Wang, Ruohan Zhang, **Yunzhu Li**, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pages 540–562. PMLR, 2023. 2

[9] Zihan Wang*, Kangrui Wang*, Qineng Wang*, Pingyue Zhang*, Linjie Li*, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and **Manling Li**. Training agents by reinforcing reasoning, 2025. 2

[10] Lionel Wong, **Jiayuan Mao**, Pratyusha Sharma, Zachary S Siegel, Jiahai Feng, Noa Korneev, Joshua B Tenenbaum, and Jacob Andreas. Learning adaptive planning representations with natural language guidance. *arXiv preprint arXiv:2312.08566*, 2023. 2

A. Biographies

Manling Li (manling.li@northwestern.edu, Primary Contact) 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 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.

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.

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.

Wenlong Huang (wenlongh@stanford.edu) is a Ph.D. student at Stanford, advised by Professor Fei-Fei Li. His research is at the intersection of robotics, machine learning, and foundation models. He received the Stanford School of Engineering Fellowship and ICRA Outstanding Robot Learning Paper Award. He received his B.A. from UC Berkeley in computer science, advised by Professor Deepak Pathak, Dr. Igor Mordatch, and Professor Pieter Abbeel. He has served as a reviewer for NeurIPS, ICML, ICLR, RSS, CoRL, ICRA, IROS, IJRR, RA-L, and Nature Communications. Additional information is available at https://wenlong.page.