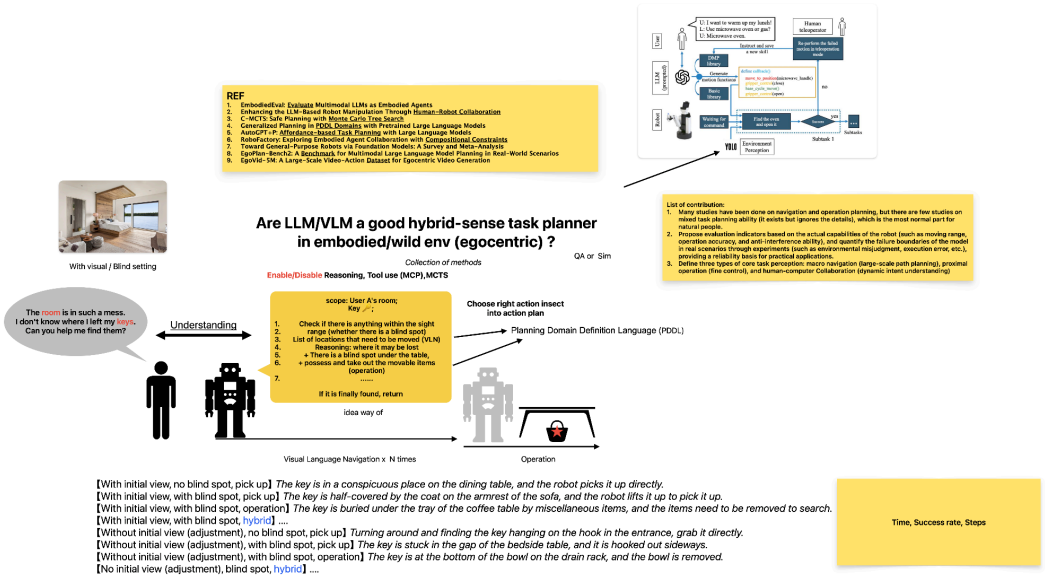
Weekly diary are recorded according to following STAR rules

Situation + **T**ask +**A**ction + R(**R**esult + **R**eflection)

Week 1



Situation:

The initial goal is very broad: to explore how to use large language/visual models (LLMs/VLMs) to build a general embodied agent that can perform complex, multi-step tasks. To make this idea more concrete, I conceived a scenario called "Help me find the keys", in which the agent needs to autonomously plan and execute two-scale actions of navigation and interaction in a cluttered real-world environment through natural language instructions, and finally find the target object.

Task:

1. Conduct literature research on the direction of "Hybrid Task Planning Based on Large Models" to verify feasibility.
2. Identify the key technologies required to implement this scenario (such as planning language, search algorithm, perception module, etc.).
3. Organize the research results and report to the tutor at the Tuesday meeting to determine the scope of feasible and interesting research topics.

Action:

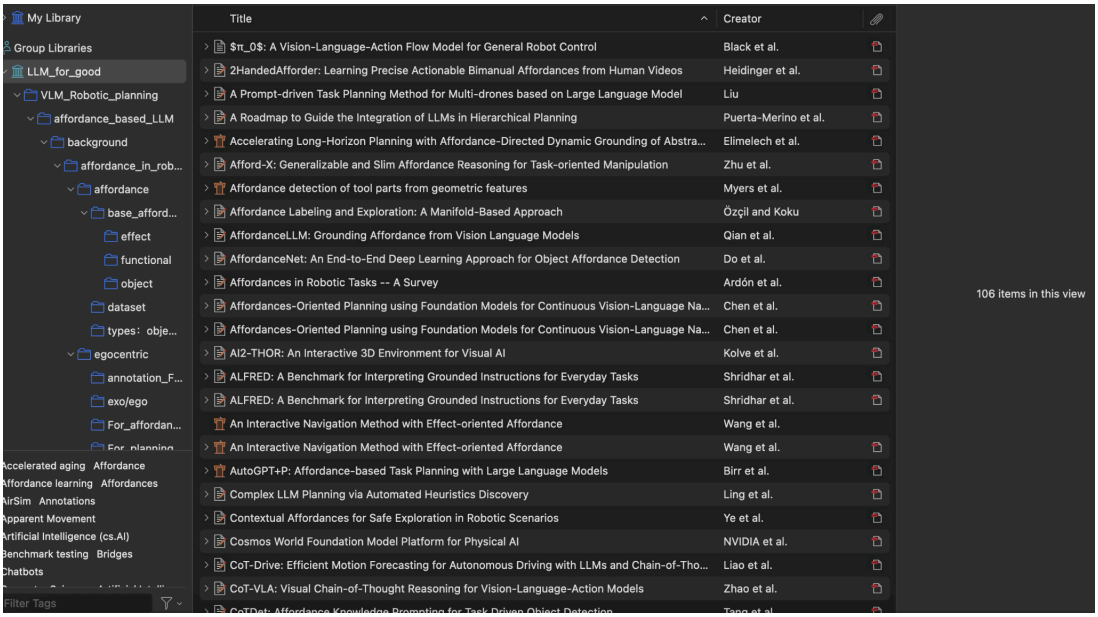
1. Macro-background research: I first read some highly summarized reviews and benchmark papers in the field to understand the capabilities of current general embodied agents.
2. Planning technology exploration: Agents can "think" and "plan", and I investigated the research that combines traditional planning methods with large models.

3. Exposure to VLA and VLN fields: Think about how agents can perform planned steps in two closely related fields:
    1. Visual language action: Realize physical interactions such as "remove the coat on the sofa" or "open the drawer", and get in touch with the core concept of affordance.
    2. Visual language navigation (VLN): Realize spatial movements such as "walk to the side of the table", and check the basic work in the field of VLN.

Result + **R**eflection:

I sorted out the technology stack required for a hybrid task agent and established a preliminary understanding of the two fields of VLA and VLN. In the discussion with my mentor, I learned that the scope of my initial idea of "building a general hybrid task planning agent" was too divergent for a 10-week project. Although the initial direction was rejected, my mentor pointed out that the differences and connections between VLA and VLN that I found in the research were a starting point.

Week 2 and 3



Situation:

The new directive was to conduct a deep dive into the concept of affordance. The goal for this two-week period was to move from this general concept to a specific, feasible, and motivated research question for review.

Task:

1. Conduct a literature survey on the topic of "affordance," starting from its theoretical origins and extending to its applications in robotics and AI.
2. Identify the main sub-fields, key research paradigms.
3. Develop initial ideas for a potential contribution, inspired by trends in the literature.

Action:

I spent the majority of the two weeks performing an reading of the literature

1. Foundational Reading: I started by reading the foundational work of James J. Gibson on his ecological theory of perception and its later exploration by scholars like Greeno. I reviewed surveys to understand application in robotics.
2. Perception and Learning Methods: I explored the evolution of affordance detection, from early methods based on geometric features to deep learning models. I also looked into learning from diverse data sources.
3. LLM/VLM Integration: I investigated the rapidly advancing frontier of integrating affordance understanding with large models.
4. Robotic Applications: I reviewed a wide array of VLA papers focused on manipulation.
5. Focusing on Navigation: I paid special attention to the more nuanced application of affordance in navigation.

Result + **R**eflection:

After this extensive reading, I noticed a critical pattern: while the VLA/manipulation field had a rich discussion on affordance, its application in VLN was less direct and often implicit. This seemed like a research gap and a focus for my review. Inspired by the trend of papers titled "Can LLM do something?", I began to think about the final, original contribution of my project. Instead of just reviewing what others have done, I could propose a framework to quantify what VLMs truly understand about affordances in a VLN context.

Writing Sample

> DeepDive in Affordance
>
> The concept of "affordances," signifying the action possibilities an environment or object offers an agent, was introduced by psychologist James J. Gibson, with its theoretical underpinnings further explored by Greeno [9]. This foundational idea has influenced robotics and artificial intelligence, spawning a vast body of research aimed at enabling machines to perceive and interact with the world in a human-like, action-oriented manner. Comprehensive overviews of affordances in robotic tasks are provided by Ardón et al. [2], while Yang et al. [27] offer a review of deep robotic affordance learning from a reinforcement learning perspective.
>
> A significant stream of research investigates methods for **learning and perceiving affordances** from diverse sensory inputs. Early approaches focused on detecting

affordances from geometric features [21]. More recent efforts leverage deep learning, with end-to-end models like AffordanceNet [7] for object affordance detection, and methods for learning to segment affordances from RGB images [19]. Affordance understanding is also being pursued from various viewpoints and data modalities, including learning from exocentric images [20], utilizing manifold-based exploration for labeling [22], capturing fine-grained affordances and 3D human-object interaction (HOI) regions from egocentric videos (Yu et al. [31], EgoChoir by Yang et al. [28]), and enabling one-shot affordance grounding for deformable objects [13]. The challenge of inferring scene affordances is even being addressed through innovative techniques like affordance-aware human insertion into scenes [15].

The integration of **affordance understanding into advanced AI systems, with Large Language Models (LLMs) and Vision-Language Models (VLMs)**, marks a rapidly advancing frontier. LLMs are employed for high-level task planning and reasoning that incorporates affordance knowledge. This includes grounding natural language instructions in robotic affordances (Ahn et al. [1]), developing affordance-based task planning frameworks (AutoGPT+P by Birr et al. [3]; PLATO by Car et al. [4]), enabling heuristic planning with learnable, affordance-grounded domain knowledge (SayCanPay by Hazra et al. [10]), translating natural language instructions into feasible, affordance-aware motion plans (Text2Motion by Lin et al. [18]), and accelerating long-horizon planning through affordance-directed grounding (Elimelech et al. [8]). VLMs are pivotal for grounding these affordances in perception, with models like AffordanceLLM [24] designed to extract affordance information directly from vision and language inputs. Task-driven object detection is enhanced by prompting affordance knowledge (CoTDet by Tang et al. [25]), and text-driven affordance learning from egocentric vision is also being explored (Yoshida et al. [30], Cheng et al. [6]).

These advancements fuel a wide array of **robotic applications and learning paradigms enhanced by affordance mechanisms**. This includes empowering LLMs for robotic manipulation through dedicated affordance prompting techniques (Cheng et al. [6]), improving VLA models via chain-of-affordance reasoning (Li et al. [17]), and injecting robotic affordance and grounded information into multi-modal LLMs (ManipVQA by Huang et al. [12]). Specific robotic capabilities are being developed, such as learning precise bimanual affordances (2HandedAfforder by Heidinger et al. [11]) and generalizable, slim affordance reasoning for task-oriented manipulation (Afford-X by Zhu et al. [33]). Innovative learning strategies include retrieval-based affordance transfer for generalizable zero-shot manipulation (RAM by Kuang et al. [14]) and one-shot open affordance learning using foundation models (Li et al. [16]). Furthermore, contextual affordances are being explored to ensure safe exploration in robotic scenarios (Ye et al. [29]).

While affordances are studied for manipulation, their application and interpretation in **navigation** present a more varied and nuanced landscape. Researchers are

exploring how robots can learn to move with affordance maps (Qi et al. [23]) and developing benchmarks for affordance-grounded last-mile navigation (MoMa-Kitchen by Zhang et al. [32]). Chen et al. [5] propose "Affordances-Oriented Planning" for continuous vision-language navigation; however, in this context, "affordances" refer to navigable ground identified through segmentation (e.g., using SAM). This interpretation, while valuable for pathfinding, is somewhat more constrained than the richer, interaction-centric understanding of affordances typically seen in manipulation tasks. In contrast, "An Interactive Navigation Method with Effect-oriented Affordance" by Wang et al. [26] offers a more sophisticated view. Their work integrates object affordances (e.g., for interacting with obstacles to clear a path) with "effect affordances" that assess path viability and optimal agent pose for subsequent actions. This definition aligns more closely with the VLA-centric understanding of affordances. Nevertheless, it is noteworthy that both Chen et al. [5] and Wang et al. [26]—much like some other contemporary vision-and-language navigation approaches such as Qiao et al.'s Open-Nav framework [34]—tend to employ composite or modular strategies to achieve their navigation goals, rather than relying on a singular, end-to-end affordance-driven policy.

This evolving landscape of affordance research, with its deep dives into perception, learning, AI integration, and specific applications like navigation, prompts a crucial forward-looking question: Can current or future Vision-Language Models achieve an intrinsic, end-to-end understanding of multifaceted affordances—encompassing object interaction possibilities, environmental layout constraints, and navigational opportunities—that mirrors the holistic, intuitive, and adaptable manner in which humans perceive and act within their surroundings?

Ref:

- Click to unfold
    1. Ahn, Michael, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, et al. “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances.” arXiv, August 16, 2022. https://doi.org/10.48550/arXiv.2204.01691.
    2. Ardón, Paola, Èric Pairet, Katrin S. Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick. “Affordances in Robotic Tasks -- A Survey.” arXiv, April 15, 2020. https://doi.org/10.48550/arXiv.2004.07400.
    3. Birr, Timo, Christoph Pohl, Abdelrahman Younes, and Tamim Asfour. “AutoGPT+P: Affordance-Based Task Planning with Large Language Models.” In Robotics: Science and Systems XX, 2024. https://doi.org/10.15607/RSS.2024.XX.112.
    4. Car, Arvind, Sai Sravan Yarlagadda, Alison Bartsch, Abraham George, and Amir Barati Farimani. “PLATO: Planning with LLMs and

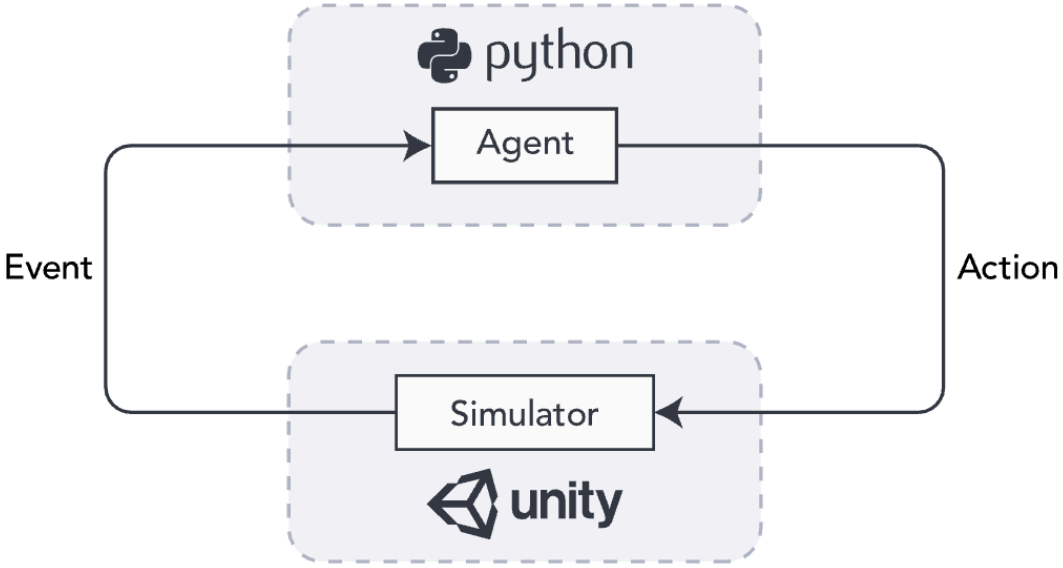
Affordances for Tool Manipulation." arXiv, September 17, 2024. https://doi.org/10.48550/arXiv.2409.11580.
5. Chen, Jiaqi, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K. Wong. "Affordances-Oriented Planning Using Foundation Models for Continuous Vision-Language Navigation." arXiv, August 20, 2024. https://doi.org/10.48550/arXiv.2407.05890.
6. Cheng, Guangran, Chuheng Zhang, Wenzhe Cai, Li Zhao, Changyin Sun, and Jiang Bian. "Empowering Large Language Models on Robotic Manipulation with Affordance Prompting." arXiv, April 17, 2024. https://doi.org/10.48550/arXiv.2404.11027.
7. Do, Thanh-Toan, Anh Nguyen, and Ian Reid. "AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection." arXiv, March 4, 2018. https://doi.org/10.48550/arXiv.1709.07326.
8. Elimelech, Khen, Zachary Kingston, Wil Thomason, Moshe Y. Vardi, and Lydia E. Kavraki. "Accelerating Long-Horizon Planning with Affordance-Directed Dynamic Grounding of Abstract Strategies." In 2024 IEEE International Conference on Robotics and Automation (ICRA), 12688–95, 2024. https://doi.org/10.1109/ICRA57147.2024.10610486.
9. Greeno, James G. "Gibson's Affordances." Psychological Review 101, no. 2 (1994): 336–42. https://doi.org/10.1037/0033-295X.101.2.336.
10. Hazra, Rishi, Pedro Zuidberg Dos Martires, and Luc De Raedt. "SayCanPay: Heuristic Planning with Large Language Models Using Learnable Domain Knowledge." arXiv, January 1, 2024. https://doi.org/10.48550/arXiv.2308.12682.
11. Heidinger, Marvin, Snehal Jauhri, Vignesh Prasad, and Georgia Chalvatzaki. "2HandedAfforder: Learning Precise Actionable Bimanual Affordances from Human Videos." arXiv, March 13, 2025. https://doi.org/10.48550/arXiv.2503.09320.
12. Huang, Siyuan, Iaroslav Ponomarenko, Zhengkai Jiang, Xiaoqi Li, Xiaobin Hu, Peng Gao, Hongsheng Li, and Hao Dong. "ManipVQA: Injecting Robotic Affordance and Physically Grounded Information into Multi-Modal Large Language Models." arXiv, August 22, 2024. https://doi.org/10.48550/arXiv.2403.11289.
13. Jia, Wanjun, Fan Yang, Mengfei Duan, Xianchi Chen, Yinxi Wang, Yiming Jiang, Wenrui Chen, Kailun Yang, and Zhiyong Li. "One-Shot Affordance Grounding of Deformable Objects in Egocentric Organizing Scenes." arXiv, March 3, 2025. https://doi.org/10.48550/arXiv.2503.01092.
14. Kuang, Yuxuan, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. "RAM: Retrieval-Based Affordance Transfer for Generalizable Zero-Shot

Robotic Manipulation." arXiv, July 5, 2024. https://doi.org/10.48550/arXiv.2407.04689.
15. Kulal, Sumith, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, and Krishna Kumar Singh. "Putting People in Their Place: Affordance-Aware Human Insertion into Scenes." arXiv, April 27, 2023. https://doi.org/10.48550/arXiv.2304.14406.
16. Li, Gen, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. "One-Shot Open Affordance Learning with Foundation Models." In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3086–96, 2024. https://doi.org/10.1109/CVPR52733.2024.00298.
17. Li, Jinming, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. "Improving Vision-Language-Action Models via Chain-of-Affordance." arXiv, December 29, 2024. https://doi.org/10.48550/arXiv.2412.20451.
18. Lin, Kevin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. "Text2Motion: From Natural Language Instructions to Feasible Plans." Autonomous Robots 47, no. 8 (December 2023): 1345–65. https://doi.org/10.1007/s10514-023-10131-7.
19. Luddecke, Timo, and Florentin Worgotter. "Learning to Segment Affordances." In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 769–76. Venice, Italy: IEEE, 2017. https://doi.org/10.1109/ICCVW.2017.96.
20. Luo, Hongchen, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. "Learning Affordance Grounding from Exocentric Images." arXiv, March 18, 2022. https://doi.org/10.48550/arXiv.2203.09905.
21. Myers, Austin, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. "Affordance Detection of Tool Parts from Geometric Features." In 2015 IEEE International Conference on Robotics and Automation (ICRA), 1374–81, 2015. https://doi.org/10.1109/ICRA.2015.7139369.
22. Özçil, İsmail, and A. Buğra Koku. "Affordance Labeling and Exploration: A Manifold-Based Approach." arXiv, July 22, 2024. https://doi.org/10.48550/arXiv.2407.15479.
23. Qi, William, Ravi Teja Mullapudi, Saurabh Gupta, and Deva Ramanan. "Learning to Move with Affordance Maps." arXiv, February 14, 2020. https://doi.org/10.48550/arXiv.2001.02364.
24. Qian, Shengyi, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. "AffordanceLLM: Grounding Affordance from Vision Language Models." arXiv, April 17, 2024. https://doi.org/10.48550/arXiv.2401.06341.

25. Tang, Jiajin, Ge Zheng, Jingyi Yu, and Sibei Yang. “CoTDet: Affordance Knowledge Prompting for Task Driven Object Detection.” arXiv, September 3, 2023. https://doi.org/10.48550/arXiv.2309.01093.
26. Wang, Xiaohan, Yuehu Liu, Xinhang Song, Yuyi Liu, Sixian Zhang, and Shuqiang Jiang. “An Interactive Navigation Method with Effect-Oriented Affordance,” 2024. [https://openreview.net/forum?id=1HMjXIgLIV&referrer=the profile of Sixian Zhang](https://openreview.net/forum?id=1HMjXIgLIV&referrer=[the profile of Sixian Zhang](%2Fprofile%3Fid%3D~Sixian_Zhang1)).
27. Yang, Xintong, Ze Ji, Jing Wu, and Yu-Kun Lai. “Recent Advances of Deep Robotic Affordance Learning: A Reinforcement Learning Perspective.” IEEE Transactions on Cognitive and Developmental Systems 15, no. 3 (September 2023): 1139–49. https://doi.org/10.1109/TCDS.2023.3277288.
28. Yang, Yuhang, Wei Zhai, Chengfeng Wang, Chengjun Yu, Yang Cao, and Zheng-Jun Zha. “EgoChoir: Capturing 3D Human-Object Interaction Regions from Egocentric Views.” arXiv, October 13, 2024. https://doi.org/10.48550/arXiv.2405.13659.
29. Ye, William Z., Eduardo B. Sandoval, Pamela Carreno-Medrano, and Francisco Cru. “Contextual Affordances for Safe Exploration in Robotic Scenarios.” arXiv, May 10, 2024. https://doi.org/10.48550/arXiv.2405.06422.
30. Yoshida, Tomoya, Shuhei Kurita, Taichi Nishimura, and Shinsuke Mori. “Text-Driven Affordance Learning from Egocentric Vision.” arXiv, April 3, 2024. https://doi.org/10.48550/arXiv.2404.02523.
31. Yu, Zecheng, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. “Fine-Grained Affordance Annotation for Egocentric Hand-Object Interaction Videos.” arXiv, February 10, 2023. https://doi.org/10.48550/arXiv.2302.03292.
32. Zhang, Pingrui, Xianqiang Gao, Yuhan Wu, Kehui Liu, Dong Wang, Zhigang Wang, Bin Zhao, Yan Ding, and Xuelong Li. “MoMa-Kitchen: A 100K+ Benchmark for Affordance-Grounded Last-Mile Navigation in Mobile Manipulation.” arXiv, March 14, 2025. https://doi.org/10.48550/arXiv.2503.11081.
33. Zhu, Xiaomeng, Yuyang Li, Leiyao Cui, Pengfei Li, Huan-ang Gao, Yixin Zhu, and Hao Zhao. “Afford-X: Generalizable and Slim Affordance Reasoning for Task-Oriented Manipulation.” arXiv, March 5, 2025. https://doi.org/10.48550/arXiv.2503.03556.
34. Qiao, Yanyuan, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. “Open-Nav: Exploring Zero-Shot Vision-and-Language Navigation in Continuous Environment with

Open-Source LLMs.” arXiv, February 11, 2025.
https://doi.org/10.48550/arXiv.2409.18794.

Week 4



Situation:

I needed to select a suitable experimental environment to build proposed evaluation. A requirement was that the environment must support not just navigation, but also rich object interaction, as experimental tasks involve testing object and effect affordances.

Task:

1. Evaluate popular simulation environments and select one that aligns with the needs of my proposed experiment (i.e., interactivity, and scene diversity).
2. Set up the chosen environment and thoroughly learn its core API.

Action:

Select Environment: My initial investigation led me to Habitat, a popular framework for VLN research due to its use of high-fidelity scans of real-world scenes. However, standard environments, while realistic, do not support the level of dynamic object interaction. I chose AI2-THOR for two main reasons:

1. High Interactivity: It is specifically designed to support rich physical interactions with objects in the environment.

2. Procedural Generation: Its ProcTHOR-10k dataset provides a vast number of procedurally generated scenes, which is ideal for testing model generalization.

Learn the AI2-THOR Framework: Studying the AI2-THOR documentation and dive in core Controller class during this week.

1. Initialization: initialize the controller, setting key parameters like agentMode, visibilityDistance, scene, gridSize, and camera properties to create a consistent experimental setup.
2. Movement: all discrete agent movement commands, including cardinal direction movements (MoveAhead, MoveLeft, etc.), rotations (RotateRight), and direct placement using the Teleport action.
3. Perception: understand how the agent perceives its environment. I learned that after every action, the event.metadata object returns information within the agent field of view and visibilityDistance.
4. Extract key object metadata such as objectId, objectType, distance, and, most importantly, the boolean flags are pickupable and moveable.
5. Navigable Space: Investigated the GetReachablePositions action. This function, which returns a list of all grid points accessible to the agent. It provides a direct, programmatic way to determine the ground truth for navigation affordance (traversability) within any given scene.

Result + **R**eflection:

My primary takeaway this week is the importance of ensuring the chosen experimental environment's capabilities align with the research question. I learned how an abstract concept like "navigation affordance" translates into a concrete API call (GetReachablePositions). Similarly, "object affordance" maps to object properties like pickupable. Being able to access information about object visibility, distance, and properties directly from the simulator will be essential for creating accurate evaluation metrics for my experiments.

Week 5


Situation:

My focus now is to implement the first phase of evaluation: Task 1 - Object Affordance Recognition. The goal is to generate a large-scale dataset of first-person views, each paired with detailed ground-truth metadata.

Task:

1. Design and implement an automated data collection pipeline that interacts with the AI2-THOR environment.
2. The pipeline programmatically navigate to diverse viewpoints within a scene.
3. At each viewpoint, it extract and save a high-resolution visual image and its corresponding, precise scene metadata.
4. The final output be structured image-metadata pairs, ready for use in the downstream object affordance evaluation tasks.

Action:

I designed a "Random Point Sampling" strategy. The logic is that by sampling from all possible reachable locations in a scene, the agent can capture a wide variety of perspectives.

- Environment: set gridSize=0.25 and visibilityDistance=5 to define the agent's movement and perception capabilities. set render resolution to 640x640 to balance image detail with file size.
- Get Reachable Space, the script calls controller.step(action='GetReachablePositions') to obtain a matrix of all coordinates the agent can safely occupy. This forms a 2D sampling space.

- Random Sampling Loop: The script iterates a set number of times. In each loop, it randomly selects a coordinate position from the reachable space and a random rotation from {0, 90, 180, 270}.It uses controller.step(action='Teleport', ...) to move the agent to the new pose.
- Data and Ground Truth: The script parses the event.metadata['objects'] list, filtering for all objects where the visible property is True. It then saves the RGB image from event.frame as a PNG file and stores the agent position along with the list of visible objects' metadata (ID, type, distance) in a JSON file.

Result + **R**eflection:

I moved from theoretical knowledge of the AI2-THOR API to practical application. This process revealed nuances I hadn't fully appreciated before, such as the exact conditions under which an object's visible flag becomes True (i.e., it must be both in the camera's FOV and within the visibilityDistance). My another lesson this week is that a brute-force approach to data collection can be wasteful. Purely random sampling generates a lot of data, much of it is low-value. But the pipeline I built is a functional prototype that will be refined.

Week 6

Situation:

The "Random Point Sampling" proved inefficient due to its "blind sampling" nature, producing a low yield of useful data and generating samples that lacked any contextual or sequential

relationship. As reflection pointed out, discussing how affordance perception changes as an agent approaches an object is impossible with isolated, random data points.

Task:

1. Analyze the defects of the previous strategy and design a new aware data collection pipeline, from "stateless random teleporting" to a "goal-oriented navigation" process.
2. Develop a method for visual masks to highlight target objects in the captured images, which will be visual prompts for VLMs.
3. Finalize the question-answer and evaluation metrics for Task 1-Object Affordance.

Action:

Object-Oriented Path Sampling: The core idea is to simulate an agent moving towards a specific object ("landmark"), collecting data at each step along the path.

Path: It uses the AI2-THOR action controller.step(action='GetShortestPathToObject', ...) to compute the optimal path (a sequence of waypoints) from the starting position to the target object. The agent moves along this generated path, and at each step (every 0.25 meters), the data collection module is triggered to save the image and metadata

Visual Mask Generation: For a given target object, I extract its boolean instance segmentation mask from event.instance_masks. I use contour detection on the mask to find the boundary. The detected boundary is drawn onto this layer with a conspicuous color (e.g., bright green). This method effectively highlights the object while preserving its visual texture for the VLM to analyze.

Task 1 Evaluation:

- Task 1.1 (Sanity Check): Object Identification. Question: "What is the highlighted object?" The answer is compared to the ground-truth objectType from the metadata.
- Task 1.2. Question: "Can you move the highlighted object?" The VLM's "Yes/No" answer is compared to the ground-truth isPickupable.
- Task 1.3: For objects with state, Question: "Is the highlighted object open?" The VLM's "Yes/No" answer is compared to the isOpen boolean.

Result + **R**eflection:

The Importance of Context: The process should mimic the task I want to evaluate. I have designed and implemented a path-based data collection pipeline. The evaluation protocol for Task 1 is with clear question formats.

With a sequence of images showing an agent approaching a chair, I can ask more context-dependent questions, such as "At what distance does the chair become 'movable'?" This is a much richer research question than what could be asked with isolated data points.

Week 7

Situation:

After functional data collection pipeline , my work this week is to establish quantifiable experimental procedure in vlm. The focus was narrowed to the most fundamental aspect: "Binary Affordance Judgment," assessing its ability to determine through different prompt designs.

Task:

1. Design a matrix of prompts with varying complexity.
2. Select a representative set of VLM models and complete the technical setup for local deployment to enable offline inference.

Action:

To ensure objectivity and scalability, I use the existing annotations within the ProcTHOR-10k dataset as our ground truth, rather than relying on manual labeling. In ProcTHOR, the pickupable and moveable properties are defined by an object's physical mass (ref in paper). Also adopting this physics-based standard ensures our results are comparable with other research in the community.

Prompt Engineering Designed: I designed a prompt matrix:

- Category A (Baseline): Includes a Simple Query (v1) and a Simple Role-Play (v2) to establish baseline performance.
- Category B (In-depth Persona): Provides the model with a detailed robot persona, including physical constraints (pv01).
- Category C (Few-Shot Learning): Includes examples of task -> judgment (pv02) and, task -> reasoning -> judgment (pv04). This is designed to diagnose "shortcut learning" by testing if the model learns a true reasoning process or just mimics output formats.
- Category D (Chain-of-Thought - CoT): Employs Zero-shot CoT (pv03) to force the model to output its reasoning steps, and a Knowledge Generation + Judgment prompt (pv05) to test its ability to activate internal knowledge.

VLM Testbed Selection and Deployment: Considering the practical constraints of robotics research (i.e., the need for offline model execution), I selected models with parameters suitable for local deployment.

Models Selected: Google Gemma-3 (1B & 4B), LLaVA-v1.5 (7B)

Deployment: use Ollama due to its ease of use, and the LLaVA model using VLLM for its high-performance inference.

Result + **R**eflection:

This week taught me the difference between testing and evaluation. By systematically varying the prompt structure and context, I can now isolate and test specific VLM capabilities, such as comparison between the two few-shot prompts (pv02 vs. pv04) to probe for "shortcut learning." . And moving from APIs to local deployment was a learning experience. I gained hands-on skills with serving frameworks like Ollama and VLLM.

Week 8

Situation:

The goal of this week is to create visual prompts. These masks will serve as direct visual cues for the VLM, allowing us to test its understanding of navigable spaces (navigation affordance).

Task:

1. Implement the functionality to generate and save two primary types of segmentation masks: a ground mask for walkable areas and a mask for a targeted interactive object.
2. Design and implement a clear data storage structure, including a comprehensive metadata file to link all data elements for a single navigation task.
3. Address practical implementation challenges to ensure the data collection pipeline is robust and the generated data is meaningful for the research questions.

Action:

Concept Grounding for "Visual Prompts": I treat "Visual Prompts" as a direct, pixel-level method to inject scene info to a model, contrasting it with language prompts. My implementation focused on two types:

> State-Change Prompt: Generating a mask for a specific object (e.g., a chair) allows us to ask the model to reason about its presence, to test its understanding for occlusion.
>
> Navigation Affordance Prompt: Implementation in house_collect.py: enabled renderSemanticSegmentation and renderInstanceSegmentation in the Controller. It triggered at each step of a path. And I developed a method to first identify all possible floor colors in a scene using event.color_to_object_type. For each frame, the script iterates through the semantic_segmentation_frame and creates a binary mask where pixels matching the floor colors are marked as the walkable area.

Object Mask & Target Tracking: At the start of a path, the _select_for_masking function identifies the nearest visible chair and "tracks" its unique objectId. This simulates an agent focus on an object during a navigation task with dynamic perspective.

Data Structure and API Design: Each navigation task is saved in separate directory. I created a navigation_task_summary file for each task, including path nodes, the ID of any modified door, and the objectId of the object for which a mask was generated. This allows downstream training or evaluation scripts to easily parse the data.

Result + **R**eflection:

It ensures that the visual prompts generated are relevant to the agent's ongoing task, simulating a more realistic cognitive process of focus, creating a mask of a chair is equivalent to asking the VLM a direct visual question: "Pay attention to this object." This reframes the data collection process as an integral part of the experimental design itself and shows a robust, well-documented data structure ensuring easy use for downstream modules.

Week 9-10

Situation:

With the object affordance data pipeline established and visual prompt created, the project's focus these two weeks start the more dynamic and complex Navigation Affordance and Effect Affordance evaluation design. These tasks require the generation of ground-truth navigation paths, which involves translating a high-level goal (e.g., "go to the door") into a precise sequence of low-level, first-person agent poses (position + orientation). And also we need to synthesize all these elements into a coherent review that not only summarizes existing work but also presents experimental framework as contribution.

Task:

1. Implement an automated path-planning module, able to construct a navigation graph of the environment which accurately identifies the optimal target coordinate for a given destination object.
2. The final output will be a ground-truth path, represented as a sequence of agent position and pose, to be used in evaluating Tasks 2 and 3.
3. Synthesize the detailed experimental designs for Object, Navigation, Effect Affordance tasks into a clear, unified protocol.
4. Structure and write the review manuscript for clarity, consistency, and academic rigor.

Action:

1. World Map Construction: Created a PathPlanner class that initializes a networkx Graph from all reachable positions provided by AI2-THOR. The continuous-space target coordinate is then mapped to its closest representative node on the discrete navigation graph.
2. Precise Target location: Implemented a get_target_xz function with a logic to determine the most suitable interaction point for a target object, instead of just its object center.
3. Agent Position Calculate: For each waypoint i in the path, we computes the required yaw angle for the agent to be facing the next waypoint i+1, thus converting an abstract path of coordinates into a sequence of concrete, first-person agent poses (position + rotation).
4. Task 2 (Navigation Affordance): The agent is placed at the start or midpoint of a pre-computed optimal path. Its task is to output the correct next move. This task directly tests the ability to perceive and act upon navigational affordances under various input conditions (with/without a ground mask, with/without depth infomation).
5. Task 3 (Effect Affordance): The agent follows a ground-truth plan until a physical obstacle appears. The task will measure if the agent can identify the blockage and choose the correct subsequent action (re-plan a path for an immovable obstacle, or attempt to move a movable one if we give permission). This tests the model's understanding of effect affordance.
6. Writing the Paper:
    a. Chapters 2.1-2.3: I revised and polished the literature review chapters, ensuring the narrative smoothly transitions from Gibson's theory to the specific challenges and applications of affordance in VLN.
    b. Chapter 4: I wrote out the complete experimental framework, detailing the environment, agent, the three core tasks, the VLM models to be evaluated, the different input modalities, and the comprehensive set of evaluation metrics.

Result + **R**eflection:

The system can now automatically produce ground-truth paths as sequences of executable agent poses. And the Importance of a "Good" Target mean simply navigating to an object center point might place the agent in a suboptimal or even incorrect position for interaction. This output is now ready to be used as the supervision signal for the Navigation Affordance task (Task 2) and as the initial plan that gets disrupted in the Effect Affordance/Re-planning task (Task 3). The experimental framework has been deepened by the implementation of the "Retest" logic, allowing us to move from testing data-level perception to physics-level interaction and re-planning.

Over these 10 weeks, I put a lot of effort, start a research project in miniature. It started with a broad, ill-defined idea, which was refined through critical feedback into a focused topic. It then progressed through a deep literature review, practical implementation and prototyping, and finally culminated in the synthesis of existing knowledge and the creation of a proposal. I learned that a research contribution does not always have to be a new model or algorithm. Designing a well-motivated question itself, it identifies key questions, defines how to measure them, and provides a benchmark for the community to build upon, which is a critical part of scientific progress.