MuEP: A Multimodal Benchmark for Embodied Planning with Foundation Models

Kanxue Li[1,2,6] , **Baosheng Yu**[3] , **Qi Zheng**[4] , **Yibing Zhan**[2,*] , **Yuhui Zhang**[3,2] ,
Tianle Zhang[2] , **Yijun Yang**[5] , **Yue Chen**[2] , **Lei Sun**[2] , **Qiong Cao**[2] ,
Li Shen[2] , **Lusong Li**[2] , **Dapeng Tao**[1,6] and **Xiaodong He**[2]
[1]Yunnan University
[2]JD Explore Academy
[3]University of Sydney
[4]Shenzhen University
[5]University of Technology Sydney
[6]Yunnan Key Laboratory of Media Convergence
likanxue@mail.ynu.edu.cn , zhanyibing@jd.com

Abstract

Foundation models have demonstrated significant emergent abilities, holding great promise for enhancing embodied agents' reasoning and planning capacities. However, the absence of a comprehensive benchmark for evaluating embodied agents with multimodal observations in complex environments remains a notable gap. In this paper, we present MuEP, a comprehensive **Mu**ltimodal benchmark for **E**mbodied **P**lanning. MuEP facilitates the evaluation of multimodal and multi-turn interactions of embodied agents in complex scenes, incorporating fine-grained evaluation metrics that provide insights into the performance of embodied agents throughout each task. Furthermore, we evaluate embodied agents with recent state-of-the-art foundation models, including large language models (LLMs) and large multimodal models (LMMs), on the proposed benchmark. Experimental results show that foundation models based on textual representations of environments usually outperform their visual counterparts, suggesting a gap in embodied planning abilities with multimodal observations. We also find that control language generation is an indispensable ability beyond common-sense knowledge for accurate embodied task completion. We hope the proposed MuEP benchmark can contribute to the advancement of embodied AI with foundation models.

1 Introduction

With the tremendous success of ChatGPT [Wu *et al.*, 2023a], foundation models trained on web-scale data using self-supervision have garnered increasing attention from the community. Foundation models with the great emergent abilities have demonstrated significant performance improvements when adopted for downstream tasks, including fluent interaction [Touvron *et al.*, 2023; OpenAI, 2023], sophisticated literary works creation [Waisberg *et al.*, 2024], image captioning [Alayrac *et al.*, 2022], and code generation [Chen *et al.*, 2021]. These advancements hold great promise for enhancing the reasoning and planning abilities of advanced embodied agents [Yang *et al.*, 2023b; Yang *et al.*, 2023c; Brohan *et al.*, 2022]. Motivated by this, numerous recent studies have utilized foundation models, such as large language models (LLMs) and large multimodal models (LMMs), across a range of tasks as embodied agents, including environmental grounding [Ahn *et al.*, 2022; Driess *et al.*, 2023], vision-language navigation [Brohan *et al.*, 2022; Mu *et al.*, 2023], and task planning [Yao *et al.*, 2023;

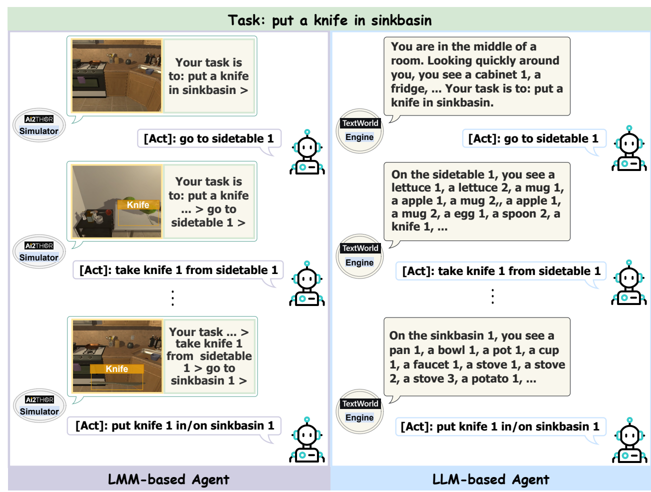

Figure 1: An example of embodied planning driven by different foundation models with vision-language (*left*) and text-only (*right*) observations in the ALFWorld environment.

*Corresponding author. This work was done during the internship at JD Explore. The source code and datasets are available: `https://github.com/kanxueli/MuEP`.

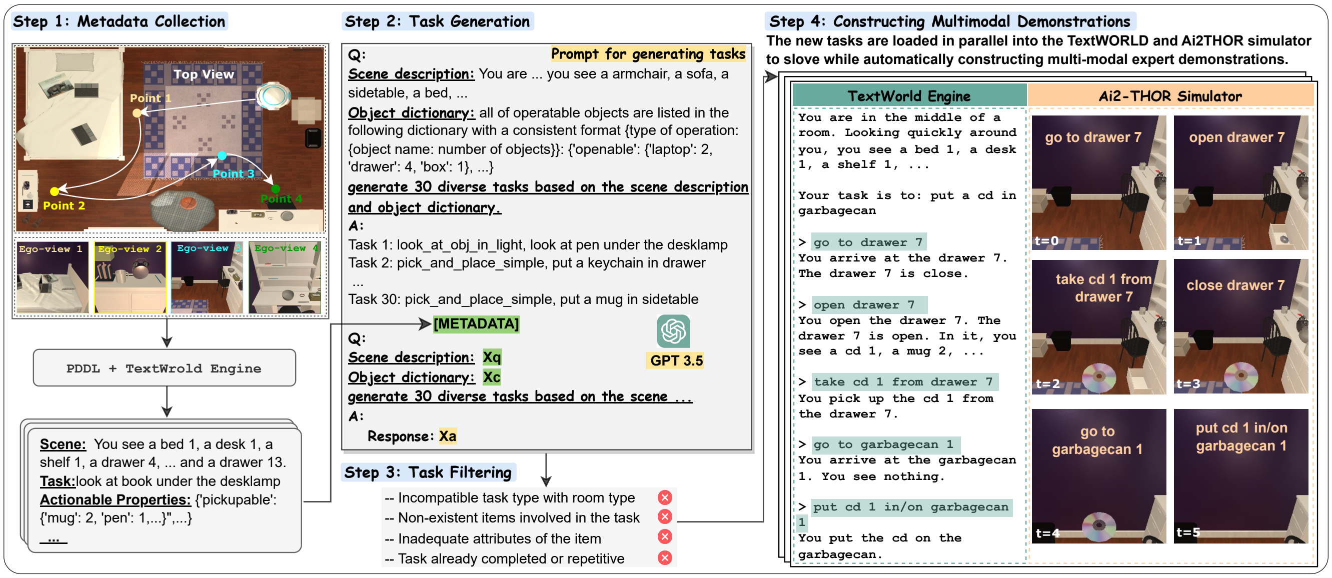

Figure 2: **An automatic pipeline for constructing the proposed MuEP benchmark.** It consists of the following steps: 1) metadata collection, 2) task generation, 3) task filtering, and 4) constructing multimodal demonstrations.

Shinn *et al.*, 2023]. Figure 1 provides an example of embodied planning driven by different foundation models with vision-language and text-only observations.

Despite the increasing number of recent studies on embodied AI with foundation models, a significant challenge persists due to the absence of a comprehensive benchmark for evaluating embodied planning abilities in complex and multimodal scenarios. Many existing benchmarks, including [Wijmans *et al.*, 2019; Ebert *et al.*, 2021; Brohan *et al.*, 2022; Wu *et al.*, 2023b], suffer from limitations such as limited diversity, single modality, and coarse metrics. These drawbacks make them insufficient for a thorough and nuanced evaluation of the capabilities of embodied agents. Moreover, collecting large-scale datasets with multi-round interactions between embodied agents and real-world environments is expensive and time-consuming. As a result, existing embodied datasets are typically gathered on a small scale, as exemplified by the TEACh project, which collected 4,365 crowdsourced data samples at a total cost of $105K [Padmakumar *et al.*, 2022]. To address this challenge, a cheap and scalable solution is automatic data generation. For exampl e, several studies have undertaken data generation and annotation using LLMs and expanded existing datasets [Wang *et al.*, 2022; Xu *et al.*, 2023]. Inspired by this, there are also initial attempts in the field of embodied intelligence such as TaPA [Wu *et al.*, 2023b] and RoboGen [Wang *et al.*, 2023]. However, TaPA may suffer from the issue of illusion in fine-tuned models due to the mismatch between detected objects and ground truth, while RoboGen utilizes popular generative models (e.g., Midjourney and Zero-1-to-3) to synthesize diverse tasks and simulated environments but struggles for a stable generation of complex and contextually consistent training scenarios.

In this paper, we present MuEP, a comprehensive multimodal benchmark for embodied planning. The main dataset construction pipeline is shown in Figure 2, which includes metadata collection, task generation, task filtering, and multimodal demonstration. Specifically, we first collect a small subset of metadata, including scene/task descriptions from the ALFworld [Shridhar *et al.*, 2021] simulator, where the Planning Domain Definition Language (PDDL) [Aeronautiques *et al.*, 1998] and TextWorld [Côté *et al.*, 2019] are utilized to align the text with the agent's egocentric visual observations during exploration. We then employ LLMs to generate different tasks via the in-context learning strategy, i.e., the collected examples as the context are provided to be part of the prompt. By doing this, we can easily generate a large number of different tasks, with those imperfect samples removed according to several heuristic filtering criteria. Lastly, the rule-based planner [Hoffmann and Nebel, 2001] is utilized to generate multimodal data at each demonstration episode. Overall, the MuEP benchmark currently encompasses 14,927 expert demonstration episodes, spreading across 108 varied household scenes and corresponding to 176,593 image-text pairs. Each scene includes diverse objects with variations in shapes, textures, and colors, enriching scene settings and task complexity.

We adopt five distinct metrics to thoroughly evaluate popular foundation models on the proposed MuEP. These include three commonly used metrics: Success Rate (SR), Interaction Step (IS), and Goal-Condition Success (GCS), as well as two additional carefully designed metrics, namely Language Compliance (LC) and Reasoning Disorientation Index (RDI). Notably, the newly introduced metrics specifically target the agent's capabilities in generating compliant structured actions and effectively replanning while avoiding cognitive pitfalls, respectively.

We consider eight representative and open-sourced foun-

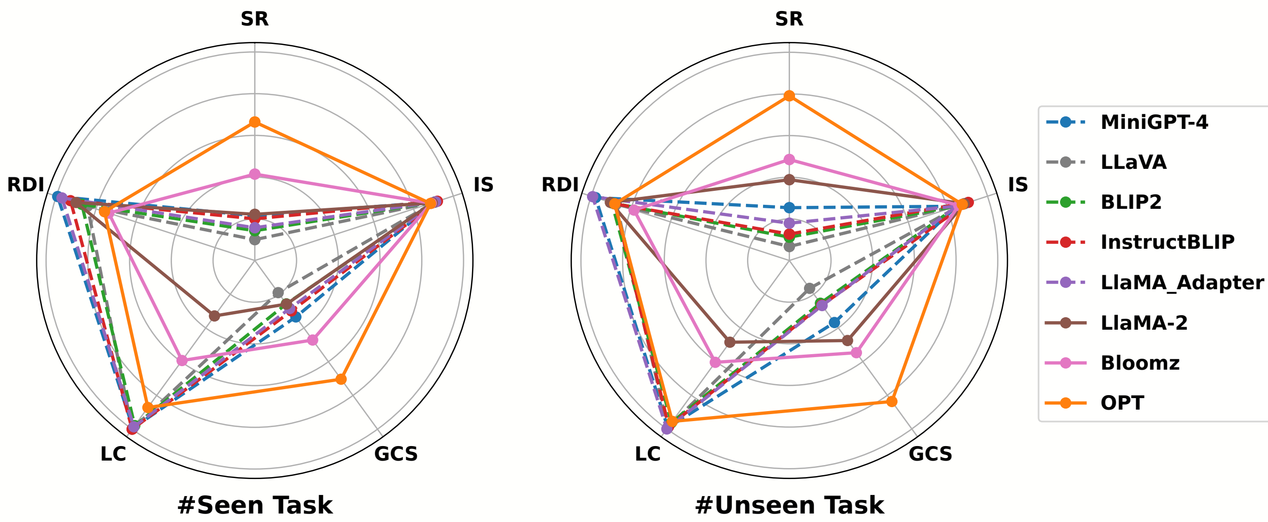

Figure 3: **Performance of LLMs and LMMs in MuEP Benchmark**. Solid lines represent LLMs, and dashed lines represent LMMs.

dation models for embodied planning, employing the parameter-efficient fine-tuning (PEFT) strategies [Mangrulkar *et al.*, 2022; Dettmers *et al.*, 2023]. As shown in Figure 3, these models include LLMs (i.e., LLaMA-2 [Touvron *et al.*, 2023], OPT [Zhang *et al.*, 2022], and Bloomz [Muennighoff *et al.*, 2022]) and LMMs (i.e., BLIP2 [Li *et al.*, 2023], InstructBLIP [Dai *et al.*, 2023], LLaMA-Adapter-v2 [Gao *et al.*, 2023], MiniGPT-4[Zhu *et al.*, 2023], and LLaVA [Liu *et al.*, 2023]). Experimental results show that: 1) Foundation models based on textual representations of environments usually outperform their visual counterparts, suggesting a gap in embodied planning abilities with multimodal observations; 2) Structured control language is crucial for successful embodied task execution; 3) Agents necessitate increased interaction steps to learn and adapt to new dynamic scenes effectively.

In summary, our contribution is twofold: 1) We propose a multimodal benchmark for embodied planning. 2) We evaluate popular LLMs and LMMs on this benchmark. We expect the research presented in this paper to benefit the development of the embodied AI community.

2 Related Work

Embodied AI with Foundation Models. Emerging trends in AI research show the extension of foundation models from basic language tasks to embodied decision-making [Vemprala *et al.*, 2023; Yang *et al.*, 2023a]. Recent research utilizes LLMs for grounding in embodied planning tasks within interactive environments [Ahn *et al.*, 2022; Huang *et al.*, 2022; Huang *et al.*, 2023; Lin *et al.*, 2023]. Some approaches like React [Yao *et al.*, 2023] and Reflexion [Shinn *et al.*, 2023] integrate chain-of-thought [Wei *et al.*, 2022] into embodied agents, enabling them to formulate autonomous problem-solving procedures. Many researchers also propose to refine agent capabilities of reasoning and decision-making in embodied environments through fine-tuning with pre-collected data [Xiang *et al.*, 2023; Mu *et al.*, 2023]. Meanwhile, the advancement in LMMs has been pivotal in more integrated and sophisticated systems, including BLIP2 [Li *et al.*, 2023], InstructBLIP [Dai *et al.*, 2023], LLaMA_Adapter_v2 [Gao *et al.*, 2023], MiniGPT-4 [Zhu *et al.*, 2023], and LLaVA [Liu *et al.*, 2023]. However, there is still an urgent need for research to test foundation-model-based agents in embodied planning. This paper focuses on testing the ability of open-sourced foundation models on embodied multimodal tasks.

Embodied Planning Benchmark. Over the past several years, the emergence of embodied interactive environments has played a pivotal role in the evaluation of Embodied AI [Kolve *et al.*, 2017; Gan *et al.*, 2020], paving the way for groundbreaking approaches in visual navigation [Ramakrishnan *et al.*, 2020; Gan *et al.*, 2021], visual-language tasks [Anderson *et al.*, 2018a; Deitke *et al.*, 2022], and embodied question answering [Das *et al.*, 2018a; Zhou *et al.*, 2023]. ALFWorld [Shridhar *et al.*, 2021] combines the interactive text-based game engine TextWorld [Côté *et al.*, 2019] with ALFRED [Shridhar *et al.*, 2020], a dataset for vision-language tasks in embodied environments, creating a platform for developing AI agents capable of understanding and interacting in complex scenarios using both text and visuals. In this paper, we utilized ALFWorld scenes for testing and its annotated tasks to generate new ones with LLM. While there is an ongoing trend towards the enlargement and enhancement of these datasets, the substantial cost and resource investment required to adapt them for embodied tasks continues to present a significant obstacle. For instance, RT-1 [Brohan *et al.*, 2022] utilized 13 robots over 17 months to collect approximately 130K episodes of data. Recent studies aim to bridge this gap by leveraging LLMs [Bubeck *et al.*, 2023] to generate training data [Wang *et al.*, 2022; Xu *et al.*, 2023], where the research most closely related to ours is TaPA [Wu *et al.*, 2023b]. In contrast to our approach of directly acquiring metadata from the simulator, TaPA utilizes an object detector to generate lists of objects from images captured within the simulator, a process prone to inaccuracies. Moreover, TaPA relies on the evaluation of action plans generated by task

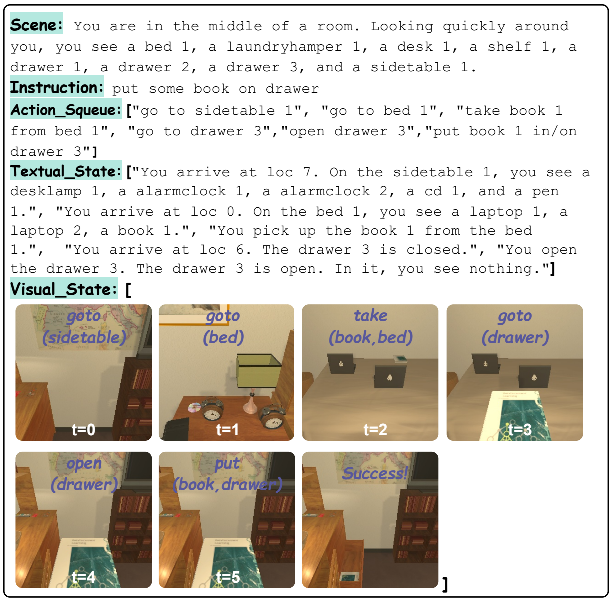

Figure 4: **Illustration of multimodal data in MuEP.** Each data point contains the scene, the instruction, and task sequences with corresponding multimodal states.

planners (LLMs) by 30 volunteers, deviating from an automated and objective evaluation system. Additionally, existing benchmarks mainly use Success Rate (SR) and Interaction Step (IS) as primary evaluation metrics [Shridhar *et al.*, 2021; Deitke *et al.*, 2022], while these coarse metrics fall short in providing a systematic evaluation of emerging embodied agents. To address this limitation, we further introduce two novel fine-grained metrics: Language Compliance (LC) and Reasoning Disorientation Index (RDI).

3 The MuEP Dataset

This section presents the details of MuEP.

3.1 Dataset Construction

The main dataset construction pipeline consists of four stages: metadata collection, task generation, task filtering, and multimodal demonstration, as follows.

Metadata Collection. We first extract metadata from the ALFWorld [Shridhar *et al.*, 2021], including scene information, available tasks, and item attributes and states, as shown in Figure 2-Step 1. We use PDDL, a standard planning language, and the TextWorld engine to programmatically align the agent's egocentric visual observations with accurate textual descriptions during exploration. PDDL encodes the textual representations of each visual scene in our dataset, similar to the approach used in ALFWorld. TextWorld then generates corresponding textual environments from this PDDL encoding, ensuring that textual and visual information are consistently aligned. Furthermore, we also gather information about the inventory, attributes, and states of objects within the scene. This information acts as constraints to reduce hallucinatory outputs and the occurrence of invalid tasks. In contrast to TaPA [Wu *et al.*, 2023b], which relies on object detection/segmentation, our method directly obtains error-free metadata from Ai2-THOR's [Kolve *et al.*, 2017] configuration. This approach not only streamlines the process but also significantly enhances the accuracy and quality of the generated tasks.

Task Generation. For each piece of metadata collected in the previous step, we explore LLM's in-context learning capability for task generation. Considering the impressive performance of GPT, we incorporate the text-davinci-003 model from OpenAI. We adopt an example-based prompt to enhance task generation, which involves querying with specific examples, including *scene information*, *object dictionary*, *the target number of tasks*, and *30 manually crafted task examples along with their types*. The object dictionary includes operational attributes of items (e.g., pickupable, receptacle, and cookable), along with quantity details. These attributes form constraint conditions during task generation, ensuring relevance and feasibility. Specifically, we prompt LLM to generate different tasks that can be executed within that particular scene. To account for different room types linked to distinct task types, we have created four distinct prompt templates. Through iterative queries with diverse scenes and object dictionaries, we can generate an extensive array of tasks that capture relevant object attributes and scene characteristics.

Task Filtering. While prompt templates have proven effective, occasionally, some generated tasks may not meet practical standards. We thus introduce a four-fold criterion to remove imperfect tasks: 1) the task is incompatible with the room scene, e.g., the "Heat & Place" category of tasks will be filtered when meeting in an environment that contains no items used for heating, such as microwaves; 2) the task involves invalid items; 3) the task involves inadequate item attributes, e.g., "put a bread in book" is invalid because, in the Ai2-THOR environment, books are not considered receptacles — they lack the necessary attribute for holding objects; 4) the task is already completed in the scene. For instance, if a "keychain" is already placed on the "shelf" in the current scene, the task "put a keychain on the shelf" will be removed.

Multimodal Demonstration. We utilize a parallel text-visual environment to construct the dataset, which allows the simultaneous acquisition of visual observations from Ai2-THOR and corresponding textual observations in TextWorld, as depicted in Figure 2-Step 4. PDDL serves as a bridge between the visual and the textual environments. We employ a rule-based planner [Hoffmann and Nebel, 2001] to generate action sequences. Inaccessible to the agent during inference, the planner relies on metadata and encodes the environment as fully observable states that contain perfect knowledge of world dynamics. Figure 4 illustrates the data format as $\{$*Scene, Instruction, Action_Sequence, Textual_State, Visual_State*$\}$. "Scene" details the agent's current environment, "Instruction" specifies the task, "Action_Sequence" records the planner's steps, and "Visual_State" and "Textual_State" provide visual and textual representations of the post-action environment, respectively. This design is well-

	Annotations		Virtual Scene			Evaluation			
	Scale	Auto-generation	Quality	Interaction	Obs.	Metric Types	Dynamic Scene	LLM	LMM
R2R [Anderson *et al.*, 2018b]	21k+	✗	Low	✗	Ego	2	✗	✗	✓
Touchdown [Chen *et al.*, 2019]	9.3k+	✗	Low	✗	Ego	2	✗	✗	✓
EQA [Das *et al.*, 2018b]	5k+	✗	Low	✗	Ego	2	✗	✗	✓
ALFRED [Shridhar *et al.*, 2020]	25k+	✗	High	✓	Ego	3	✓	✗	✓
ALFWorld [Shridhar *et al.*, 2021]	25k+	✗	High	✓	Ego	2	✓	✓	✓
TaPA [Wu *et al.*, 2023b]	6k+	✓	High	✗	3rd Person	1	✗	✓	✗
MuEP	170k+	✓	High	✓	Ego	5	✓	✓	✓

Table 1: **Comparison of different embodied planning datasets**. The advantage of MuEP lies in its expansive scale, comprehensive evaluation metrics, dynamic scene evaluation, and capability to test both LLMs and LMMs.

suited for training both LLM and LMM agents because it enables multi-round interactions and extends task horizons in dynamic scenes.

3.2 Dataset Metrics and Statistics

This subsection first explains the evaluation metrics and then summarizes the statistics of MuEP by comparing with previous embodied datasets.

Evaluation Metrics. To form a more nuanced evaluation, MuEP provides five types of metrics, including three commonly used metrics success rate, interaction step, and goal-condition success, and two additional designed metrics, language compliance, and reasoning disorientation index. All metrics are complementary to each other and can comprehensively test the capabilities of foundation models.

Success Rate (SR) refers to the ratio of the number of successful plays and the total plays. Task completion requires two vital conditions: the target location and the target state. For example, the task "placing a heated potato on a table" is deemed successful if the potato is in a heated state and located on the table.

Interaction Step (IS) measures the number of steps an agent interacts with the environment to complete a task. This metric evaluates the efficiency of an agent's interactions in achieving the task objectives. Fewer interaction steps indicate a better capability for reasoning and planning.

Goal-Condition Success (GCS) measures how close an agent is to entirely completing a task by counting the completion of all the sub-goals. For instance, the task "placing a heated potato on a table" includes three sub-goals: finding the potato, heating it, and placing it on the table. If the agent only puts the potato on the table without heating it, the task's GCS score is 2/3 or about 66%, indicating partial success.

Language Compliance (LC) was designed to measure the structuredness and compliance of the action generated by foundation models, providing a crucial tool for assessing their capability in delivering effective and syntactically correct action commands. Foundation models exhibit fluid and free expressive capabilities, enhancing human-machine interaction with naturalness. Nevertheless, their capacity to generate strictly formatted language in embodied tasks remains unclear. To address this issue, we introduce the 'Language Compliance' metric. The metric LC is given as $LC = \hat{a}/a^*$, where $\hat{a}$ is the number of valid actions and a^* is the total number of actions.

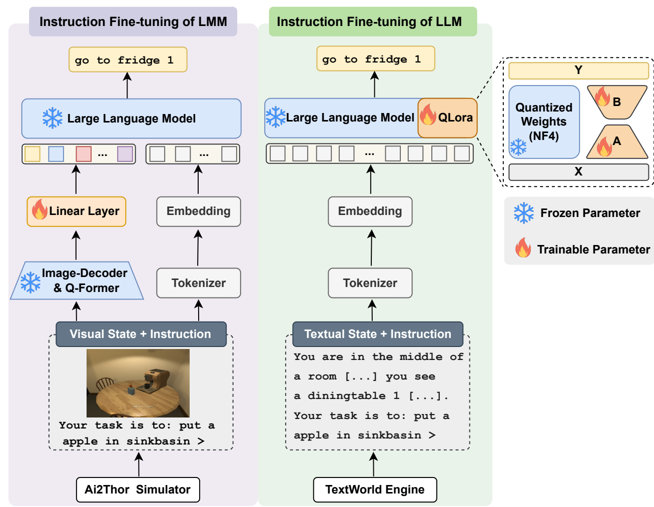

Figure 5: **Architectures for fine-tuning LLMs and LMMs.** The illustration provides more details of PEFT and resource conservation in model optimization.

Reasoning Disorientation Index (RDI) reflects whether an agent persistently takes repetitive actions in response to the same environmental feedback. Specifically, the RDI quantifies the tendency of an agent to resort to repetitive actions or iterative, ineffective solutions when faced with consistent environmental feedback. The RDI is calculated by observing the agent's actions in specific tasks and checking for repetitive or cyclical patterns. The metric RDI is given as $RDI = p_{rdi}/p_{fail}$, where p_{rdi} is the number of planning with RDI occurrences and p_{fail} is the total failed tasks. A higher RDI value indicates a higher likelihood of the agent encountering cognitive traps, necessitating enhanced reasoning strategies.

Dataset Summary. MuEP comprises 15,247 demonstration episodes, corresponding to 176,593 image-text pairs, where Figure 4 illustrates the multimodal data. A comparison between MuEP and other embodied planning datasets is shown in Table 1. MuEP outshines others with its extensive scale of over 170k+ instances, robust generative task support, high-quality virtual scene interactions, and superior metrics for dynamic scene evaluation. Furthermore, MuEP can be used to evaluate embodied agents using both LLMs and LMMs.

Task Type		Large Multimodal Model					Large Language Model		
		BLIP2	**InstructBLIP**	**LlaMA_Adapter_V2**	**LLaVA**	**MiniGPT-4**	**LlaMA-2**	**Bloomz**	**OPT**
Pick & place	SR	25.71	20.00	25.71	11.43	**31.43**	28.57	51.43	***77.14***
	IS	6.89	***4.86***	6.78	11.25	8.09	11.20	8.78	**8.59**
	GCS	25.71	20.00	25.71	11.43	31.43	28.57	51.43	77.14
	LC	100	100	94.29	100	100	28.57	60.00	77.14
	RDI	11.43	8.57	2.86	22.86	0.00	11.43	34.29	17.14
Examine in Light	SR	7.69	**15.38**	0.00	0.00	7.69	***84.62***	53.85	***84.62***
	IS	**9.00**	11.50	-	-	**9.00**	9.82	***7.14***	11.00
	GCS	42.31	34.62	34.62	11.54	34.62	84.62	57.69	88.46
	LC	100	100	100	100	100	84.62	92.31	100
	RDI	53.85	30.77	7.69	15.38	7.69	0.00	38.46	7.69
Clean & place	SR	11.11	**40.74**	18.52	22.22	33.33	3.70	51.85	***74.07***
	IS	12.00	**8.73**	11.80	9.17	10.11	***6.00***	11.43	10.95
	GCS	27.16	49.38	34.57	34.57	45.68	6.17	58.02	74.07
	LC	96.30	100	100	100	100	14.81	70.37	96.30
	RDI	7.41	0.00	0.00	14.81	0.00	18.52	22.22	14.81
Heat & place	SR	**12.50**	**12.50**	6.25	6.25	6.25	31.25	25.00	***37.50***
	IS	7.00	6.50	7.00	7.00	***6.00***	**12.80**	16.75	14.67
	GCS	31.25	31.25	27.08	27.08	29.17	39.58	25.00	37.50
	LC	100	100	100	100	93.75	62.50	43.75	87.50
	RDI	6.25	6.25	0.00	6.25	0.00	0.00	18.75	56.25
Cool & place	SR	12.00	16.00	**24.00**	8.00	20.00	16.00	44.00	***72.00***
	IS	8.67	10.00	9.67	***7.00***	9.80	**9.75**	15.27	12.61
	GCS	26.67	32.00	38.67	24.00	37.33	18.67	45.33	72.00
	LC	92.00	100	100	100	100	40.00	56.00	100
	RDI	8.00	0.00	4.00	12.00	0.00	0.00	12.00	28.00
Pick two & place	SR	8.33	8.33	4.17	4.17	**12.50**	0.00	16.67	***45.83***
	IS	17.50	9.00	11.00	***8.00***	9.67	-	**10.5**	16.55
	GCS	10.42	16.67	12.50	6.25	20.83	10.42	39.58	66.67
	LC	100	100	100	100	100	4.17	41.67	70.83
	RDI	8.33	8.33	4.17	16.67	0	20.83	33.33	29.17

Table 2: **Performance comparison by task category of models on #Seen tasks.** All values are presented as percentages

4 Experiments

This section details and analyzes our experiments.

4.1 Baseline Choices

To assess the performance of mainstream open-source large models in embodied tasks, we fine-tuned and analyzed foundation models on the MuEP dataset. For a fair comparison, all models are at around the 7B magnitude.

LLMs: Our first choice is the LLAMA-2-7B [Touvron *et al.*, 2023] model from the LLAMA family, known for its range of models with 7B to 65B parameters and its high quality. This selection balances model performance with resource efficiency, aiming to reduce training time and resource consumption. Further exploring the LLM domain, we included the OPT-6.7B [Zhang *et al.*, 2022] model, comparable in size to LLAMA-7B, to assess the impacts of using different foundational models. Additionally, the Bloomz-3B [Muennighoff *et al.*, 2022] model was trained to explore the performance of models with varying parameter magnitudes.

LMMs: We selected five representative LMMs, including BLIP2 [Li *et al.*, 2023], InstructBLIP [Dai *et al.*, 2023], LLaMA_Adapter_V2 [Gao *et al.*, 2023], MiniGPT-4 [Zhu *et al.*, 2023], and LLaVa [Liu *et al.*, 2023]. These models integrate pre-trained image encoders with large language models and underwent extensive pre-training on massive image-text pairs. This process aligned text-image characteristics, setting a strong foundation for subsequent fine-tuning. All models were then fine-tuned on diverse, small-scale instruction-following datasets, enhancing their basic instruction-following capabilities. Table 3 includes detailed configurations of all baseline models in our experiments.

4.2 Experimental and Implementation Details

Experimental Settings. We conduct extensive experiments on 134 #Unseen and 140 #Seen test tasks from the original ALFworld to evaluate the performance of foundation models and also to compare with previous works. #Seen tasks entail executing in familiar rooms from training, yet with different instantiations of object locations, quantities, and visual appearances, whereas #Unseen tasks involve new tasks in entirely unknown rooms, characterized by different objects and scene layouts not encountered during training. Each model is evaluated under a constraint of a maximum of 30 steps per task.

Agent		Model Configuration					# Seen					# Unseen				
		LLM	#VE	#ToP	#TuP	#TuM	SR ↑	IS ↓	GCS ↑	LC ↑	RDI ↓	SR ↑	IS ↓	GCS ↑	LC ↑	RDI ↓
LMMs	BLIP2	Vicuna	ViT-g	7B	3.15M	FC Layer	14.29	9.10	25.71	97.86	12.86	11.19	14.80	25.37	96.27	10.45
	InstrutBLIP	Vicuna	ViT-g	7B	3.15M	FC Layer	20.00	***<u>8.00</u>***	29.88	100	7.14	12.69	***<u>9.71</u>***	26.49	98.51	11.94
	LlaMA_Adapter_v2	Vicuna	ViT-g	7B	3.15M	B-Tuning	15.71	8.91	28.45	98.57	2.86	17.91	15.33	26.74	100	0.75
	LLaVA	LlaMA	ViT-l	7B	400M	FC Layer	10.00	9.21	19.05	100	15.71	6.72	12.22	16.54	99.25	12.69
	MiniGPT-4	Vicuna	ViT-l	7B	3.10M	FC Layer	**21.43**	9.10	33.45	99.29	0.71	**25.37**	15.68	36.82	99.25	2.24
LLMs	LlaMA-2	-	-	7B	7995M	Q-Lora	22.14	**10.61**	25.83	32.86	10.00	38.81	**11.71**	47.51	48.51	9.70
	Bloomz	-	-	3B	3932M	Q-Lora	41.43	11.12	47.14	59.29	26.43	48.51	14.51	54.73	60.45	21.64
	OPT	-	-	6.7B	7549M	Q-Lora	***<u>66.43</u>***	11.49	70.36	87.14	24.29	***<u>79.10</u>***	12.75	83.71	95.52	11.94

Table 3: **Results on various downstream tasks and scenarios**. "#VE", "#ToP", "#TuP", and "#TuM" denotes the visual encoder, the total number of parameters, the tuning parameters, and the tuning module, respectively. All values are presented as percentages.

Task Type		Large Multimodal Model					Large Language Model		
		BLIP2	InstructBLIP	LlaMA_Adapter_V2	LLaVA	MiniGPT-4	LlaMA-2	Bloomz	OPT
Pick & place	SR	8.33	4.17	16.67	**25.00**	4.17	62.50	75.00	***<u>83.33</u>***
	IS	***<u>4.00</u>***	8.00	16.75	14.33	9.00	**10.27**	10.39	10.65
	GCS	8.33	4.17	16.67	25.00	4.17	62.50	75.00	83.33
	LC	95.83	100	100	100	100	66.67	87.50	100
	RDI	12.50	0.00	0.00	4.17	0.00	0.00	25.00	16.67
Examine in Light	SR	22.22	44.44	11.11	0.00	**61.11**	44.00	44.00	***<u>88.90</u>***
	IS	12.75	***<u>10.50</u>***	11.50	-	15.00	16.50	19.25	**14.50**
	GCS	61.11	61.11	16.67	5.56	80.56	66.67	55.56	91.67
	LC	100	100	100	100	100	50.00	72.22	94.44
	RDI	22.22	33.33	0.00	33.33	5.56	16.67	33.33	11.11
Clean & place	SR	9.68	9.68	9.68	3.23	**22.58**	6.45	67.74	***<u>74.19</u>***
	IS	25.00	10.67	19.33	***<u>6.00</u>***	18.71	***<u>6.00</u>***	14.43	13.61
	GCS	26.88	25.81	25.81	19.35	38.71	9.68	72.04	79.57
	LC	93.55	96.77	100	100	96.77	16.13	70.97	90.32
	RDI	3.23	12.90	3.23	16.13	0.00	12.90	12.90	9.68
Heat & place	SR	19.23	19.23	**38.46**	7.69	26.92	69.23	7.69	***<u>88.46</u>***
	IS	11.60	***<u>8.20</u>***	14.90	9.00	12.71	12.39	26.00	**11.57**
	GCS	32.05	34.62	51.28	23.08	38.46	75.64	12.82	88.46
	LC	96.15	96.15	100	100	100	76.92	11.54	92.31
	RDI	15.38	11.54	0.00	0.00	3.85	3.85	26.92	7.69
Cool & place	SR	5.56	0.00	22.22	0.00	**27.78**	38.89	38.89	***<u>88.89</u>***
	IS	30.00	-	**11.00**	-	17.00	***<u>8.71</u>***	16.14	13.56
	GCS	18.52	16.67	33.33	14.81	40.74	38.89	42.59	88.89
	LC	100	100	100	100	100	66.67	50.00	100
	RDI	0.00	11.11	0.00	22.22	5.56	0.00	16.67	5.56
Pick two & place	SR	0.00	0.00	5.88	0.00	**17.65**	11.76	***<u>52.94</u>***	47.06
	IS	-	-	27.00	-	**18.00**	***<u>13.50</u>***	14.89	13.75
	GCS	5.88	20.59	8.82	2.94	26.47	41.18	70.59	70.59
	LC	94.12	100	100	100	100	17.65	76.47	100
	RDI	11.76	5.88	0.00	5.88	0.00	29.41	17.65	23.53

Table 4: **Performance comparison by task category of models on #Unseen tasks.** All values are presented as percentages.

Implementation Details. The overall model fine-tuning framework is illustrated in Figure 5. For LMMs, core components like the ViT, Q-Former, and LLM itself were frozen to maintain stability, with the projection layer fine-tuned. Notably, the LlaMa_adapter_v2 model only underwent fine-tuning on the bias of adapter layers [Gao *et al.*, 2023]. LMMs incorporate two main types of inputs: text instructions (i.e., task directives and historical actions) and image observations. Specifically, the historical actions provide crucial contextual information for understanding the extent of task completion and aids in the agent's subsequent reasoning process. Concurrently, the visual inputs offer a first-person perspective of the agent's external environments. If not otherwise stated, we always adhere to the prompt templates used in the original works during fine-tuning. As illustrated on the right side of Figure 5, we employed Q-LoRA [Dettmers *et al.*, 2023] to fine-tune all LLMs efficiently. LLM's input consisted exclusively of text-based instructions and environmental observations. Despite representing different modalities, visual and textual observations were treated as equivalent in our dataset. All experiments were accelerated by four Tesla V100 GPUs.

4.3 Results and Discussions

As shown in the main results Table 3 and Figure 3, for LLM agents, OPT achieved the best performance in terms of SR (i.e., 79.10%). However, OPT and Bloomz models showed higher tendencies toward reasoning disorientation, as reflected by their RDI scores. Conversely, LlaMA-2-7B demonstrated the best performance in IS and RDI, but its lower SR suggested limited stability in task completion even compared with Bloomz-3B. Regarding LMM agents, MiniGPT-4 outperformed others in both #Seen and #Unseen scenarios in terms of the SR. However, it required more interaction steps to complete tasks, suggesting a higher frequency of task completion but with increased interactions. In contrast, InstructBLIP, while slightly underperforming in SR, had the most outstanding IS metric, indicating higher efficiency in task execution. Overall, although LLMs demonstrate significantly superior performance in terms of SR, there is notable potential for improvement in terms of LC.

To offer insights into the capacities of the model for each task category, a more comprehensive evaluation per task type is presented in Tables 2 and 4. This comprehensive comparison not only pinpointed specific weaknesses in the model's capabilities but also shed light on the difficulty levels associated with all tasks. For instance, we find that the 'Pick two & place' task was particularly challenging for all models, with the highest success rate of 52.94% achieved. Additionally, the success rate of LlaMA-2-7B's in the task of 'Pick two & place' was 0% (#Seen tasks), primarily due to its extremely low LC score of 4.17%, highlighting the importance of improving LlaMA-2-7B's in terms of the LC ability.

Several key observations regarding the performance disparities between LLMs and LMMs are as follows: **1)** Foundation models based on textual representations often tend to outperform their visual counterparts. This is mainly because embodied agents based on LLMs directly leverage PDDL, while there are still gaps for current LMMs to capture comparable information from visual representations. **2)** Generating formatted control language is of great importance for embodied task completion. The LC metric highlights the significance of the formatted control language. Our results suggest that higher LC values in LLMs correspond to increased success rates (SR) and Goal-Condition Success (GCS). **3)** New environments present heightened challenges for task planning. The IS metric reflects the number of interaction steps. Our results suggest that agents operating in "Unseen" scenarios (unfamiliar environments) typically require significantly more steps to complete tasks compared to those in "Seen" scenarios. **4)** According to Table 3, LLM agents demonstrated significantly better performance in #Unseen scenarios, whereas LMM agents exhibited consistent performance across both scenarios. We owe this observation to variations, such as the difference in object location and appearance. They would not be reflected in the current textual representation of scene information, yet they can be noticed in visual information. Therefore, the LLM-based agents, when meeting tasks in seen scenes, tend to output actions they have already learned, even though the generated control languages are ineffective after multiple action failures, leading to worse RDI and LC metrics and further affecting the SR metric.

5 Conclusions and Future Work

We introduce MuEP, a comprehensive multimodal benchmark for embodied planning that tackles the challenges posed by diversity, modality limitations, and the coarse metrics in existing benchmarks. Our evaluation of recent foundation models highlights the superior performance of text-based models over their visual or multimodal counterparts in tasks related to embodied planning. Moreover, our newly proposed metrics, LC and RDI, offer valuable insights into agents' action generation and cognitive planning capabilities. These insights serve as crucial guidance for future research in multimodal embodied agent planning. We envision MuEP as a pivotal benchmark, contributing to the advancement of embodied agents in higher-level reasoning and planning abilities. In the future, we plan to expand the benchmark in terms of scale, diversity, and complexity. Additionally, we aim to investigate the disparity between textual and visual modalities in embodied planning tasks.

Acknowledgments

This work was partially supported by the Major Science and Technology Innovation 2030 "New Generation Artificial Intelligence" key project (No. 2021ZD0111700), National Natural Science Foundation of China under Grant 62172354, Yunling Scholar Talent Program of Yunnan Province under Grant K264202230207, and Yunnan Provincial Key Laboratory of Media Convergence Open Project Funding Project Number: 220235203.

References

[Aeronautiques *et al.*, 1998] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. Pddl: the planning domain definition language. *Technical Report, Tech. Rep.*, 1998.

[Ahn *et al.*, 2022] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

[Anderson *et al.*, 2018a] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

[Anderson *et al.*, 2018b] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[Brohan *et al.*, 2022] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana

Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[Bubeck *et al.*, 2023] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[Chen *et al.*, 2019] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, 2019.

[Chen *et al.*, 2021] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[Côté *et al.*, 2019] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games Workshop, IJCAI*, 2019.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[Das *et al.*, 2018a] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, 2018.

[Das *et al.*, 2018b] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[Deitke *et al.*, 2022] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. *NeurIPS*, 35:5982–5994, 2022.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[Ebert *et al.*, 2021] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021.

[Gan *et al.*, 2020] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020.

[Gan *et al.*, 2021] Chuang Gan, Siyuan Zhou, Jeremy Schwartz, Seth Alter, Abhishek Bhandwaldar, Dan Gutfreund, Daniel L. K. Yamins, James J. DiCarlo, Josh H. McDermott, Antonio Torralba, and Joshua B. Tenenbaum. The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied AI. *CoRR*, 2021.

[Gao *et al.*, 2023] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[Hoffmann and Nebel, 2001] Jörg Hoffmann and Bernhard Nebel. The ff planning system: Fast plan generation through heuristic search. *JAIR*, 2001.

[Huang *et al.*, 2022] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[Huang *et al.*, 2023] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model. *arXiv preprint arXiv:2305.11176*, 2023.

[Kolve *et al.*, 2017] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[Lin *et al.*, 2023] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *arXiv preprint arXiv:2303.12153*, 2023.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[Mangrulkar *et al.*, 2022] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods, 2022.

[Mu *et al.*, 2023] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *arXiv preprint arXiv:2305.15021*, 2023.

[Muennighoff *et al.*, 2022] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

[OpenAI, 2023] OpenAI. Gpt-4 technical report, 2023.

[Padmakumar *et al.*, 2022] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI*, 2022.

[Ramakrishnan *et al.*, 2020] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *ECCV*. Springer, 2020.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

[Shridhar *et al.*, 2020] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. 2020.

[Shridhar *et al.*, 2021] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vemprala *et al.*, 2023] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. Chatgpt for robotics: Design principles and model abilities. *Microsoft Auton. Syst. Robot. Res*, 2023.

[Waisberg *et al.*, 2024] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Google's ai chatbot "bard": a side-by-side comparison with chatgpt and its utilization in ophthalmology. *Eye*, 2024.

[Wang *et al.*, 2022] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[Wang *et al.*, 2023] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation. *arXiv preprint arXiv:2311.01455*, 2023.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

[Wijmans *et al.*, 2019] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *CVPR*, 2019.

[Wu *et al.*, 2023a] Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 2023.

[Wu *et al.*, 2023b] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*, 2023.

[Xiang *et al.*, 2023] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language models meet world models: Embodied experiences enhance language models. *arXiv preprint arXiv:2305.10626*, 2023.

[Xu *et al.*, 2023] Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*, 2023.

[Yang *et al.*, 2023a] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.

[Yang *et al.*, 2023b] Yijun Yang, Tianyi Zhou, Jing Jiang, Guodong Long, and Yuhui Shi. Continual task allocation in meta-policy network via sparse prompting. In *International Conference on Machine Learning*, pages 39623–39638. PMLR, 2023.

[Yang *et al.*, 2023c] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. *arXiv preprint arXiv:2311.16714*, 2023.

[Yao *et al.*, 2023] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

[Zhang *et al.*, 2022] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[Zhou *et al.*, 2023] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *CoRR*, 2023.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.