Building Embodied EvoAgent: A Brain-inspired Paradigm for Bridging Multimodal Large Models and World Models: Supplementary Materials

Junyu Gao*
State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence
University of Chinese Academy of Sciences (UCAS)
Beijing, China
junyu.gao@nlpr.ia.ac.cn

Xuan Yao*
State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence
University of Chinese Academy of Sciences (UCAS)
Beijing, China
yaoxuan2022@ia.ac.cn

Yong Rui
Lenovo Research, Lenovo Group Ltd.
Beijing, China
yongrui@lenovo.com

Changsheng Xu†
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence
University of Chinese Academy of Sciences (UCAS)
Beijing, China
Peng Cheng Laboratory
Shenzhen, China
csxu@nlpr.ia.ac.cn

1 Full Version of Related Work

Embodied Artificial Intelligence. Embodied AI investigates the interaction between physical bodies and their environments, constituting an interdisciplinary domain at the confluence of artificial intelligence and cognitive science [15, 42, 44, 55, 70, 79]. In recent years, propelled by rapid advancements in multimodal large models for multimodal representation and reasoning, embodied AI has witnessed significant breakthroughs. For example, Vision-Language-Action models (VLAs) [10, 39, 78] empower embodied agents to generate actions based on visual observations and linguistic instructions, thereby facilitating the execution of robotic tasks. PaLM-E [16] employs a decoder-only large language model that autoregressively predicts sequences of text or action decisions. Google's RT series [10, 47] integrates internet-sourced data with information gathered from various simulators to train end-to-end robotic control models. Similarly, Octopus [71] leverages multimodal foundation models to dynamically generate descriptions of observed scenes within ongoing tasks, subsequently feeding these descriptions into large language models to produce follow-up instructions. Additional explorations within the field of embodied AI include the application of multimodal large models or large language models, such as GPT-4V [34, 63], LLaVA [32, 75] and LLaMA [39], for environmental perception or instruction reasoning. Beyond the development of more robust multimodal representations, the comprehensive simulation of the world has emerged as a critical research focus for enhancing the efficacy of embodied AI. For instance, 3D-VLA (3D Vision Language Action) [78] simulates knowledge of the physical world by extracting rich 3D scene information and generates actions by envisioning the task completion process. iVideoGPT [67] serves as an interactive video generation world model, enabling agents to engage in exploration, reasoning, and planning within its framework. Further efforts encompass the development of simulators that more accurately reflect the physical world, such as Habitat [45], alongside the adoption of more effective 3D scene reconstruction techniques, such as 3D Gaussians [56]. At present, a primary limitation in the embodied AI field is the absence of autonomous evolutionary capabilities. Shapiro [55] underscore the pivotal role of continuous interaction and evolution between the body and the physical environment in cognitive processes. Preliminary studies [18, 21] have examined agent evolution grounded in continuous reinforcement learning or test-time adaptation; however, these endeavors remain inadequate for the construction of robust embodied world models and cognitive frameworks.

Multimodal Large Language Model (MLLM). MLLMs are designed to jointly process and comprehend multimodal data, encompassing text, images, videos, and additional modalities. Initial developments in MLLMs predominantly featured visual-language models (VLMs), such as CLIP [52] and ALIGN [37], which undergo pre-training on extensive image-text pairs through contrastive learning for multimodal fusion. Subsequently, Flamingo [2] introduces few-shot learning capabilities, markedly improving the efficacy of VLMs in zero-shot tasks. Leveraging the significant text understanding and generation advancements in large language models (LLMs), such as GPT-3 [11] and the LLaMA series [60],

*Both authors contributed equally to this research.
†Corresponding author.

MLLMs have extended the scope of artificial intelligence (AI) applications in multimodal tasks such as image captioning, visual question answering, video understanding, and cross-modal dialogue. LLaVA [41] achieves end-to-end training by employing visual instruction tuning and datasets generated by GPT-4 [1], attaining performance levels proximate to those of GPT-4V. Over the past two years, MLLMs have expanded into the domains of video and audio processing. Models such as LLaMA-VID [74] and NExT-GPT [69] have demonstrated proficiency in managing long video content and performing multimodal generation. Models like GPT-4o [1], Gemini [58], Claude [8], and Grok exhibiting exceptional multimodal understanding and reasoning capabilities. However, these models remain proprietary, accessible solely through application programming interfaces (APIs). The advent of open-source models, including QwenVL [7], InternLM-Xcomposer [76], CogVLM [64], LLaVA-NeXT [40], and DeepSeek-VL Janus [43], have substantially advanced the field. These developments have also empowered researchers to harness MLLM capabilities with greater flexibility, thereby fostering progress in the domain of embodied AI. Nevertheless, owing to deficiencies in physical interaction and environmental adaptability, the question of how MLLMs can more effectively contribute to embodied intelligence persists as an area of ongoing investigation.

World Model (WM). World models endeavor to emulate the cognitive mechanisms of the human brain, focusing on encoding the evolutionary patterns of world states, the environmental responses to agent behaviors, and their intrinsic connections with perceptual inputs, thereby constructing internal representation mechanisms for agents [19, 24, 57]. The concept of world models was initially proposed by David Ha and Jürgen Schmidhuber [25], who delineated a foundational architecture comprising a visual model (V), a memory model (M), and a controller (C). On the one hand, early investigations [25] concentrates on abstract representations of the external world, aiming to elucidate its fundamental mechanisms and principles. On the other hand, scholars such as LeCun [22] emphasize that world models should not only perceive and model the world but also possess the capability to predict future states. To achieve the above objectives, contemporary world model architectures are predominantly classified into three categories: (1) Recurrent State Space Models (RSSMs) [26, 27, 46, 68] are designed to facilitate predictions within latent spaces by learning dynamic models of the environment from pixel observations and selecting actions through planning in the encoded latent space. (2) Joint-Embedding Predictive Architectures (JEPAs) [5, 9, 22] aim to learn mappings from input data to predicted outputs within a higher-level semantic representation space. (3) Transformer-based Models [12, 53, 77] leverage their robust sequence modeling capabilities to process complex multimodal information and long-term environmental changes. At present, world models have demonstrated considerable application potential in fields such as robotic control [54, 61, 62] and autonomous driving [33, 66, 80, 82]. To enhance the representational capabilities of world models, certain approaches have resorted to harnessing the formidable representational abilities of multimodal large models. However, these methods treat multimodal large models as integral components of the world model, lacking designs for collaborative interaction and cognitively inspired architectures between the two.

2 Detailed Experimental Settings

2.1 In-domain Datasets

Following the evaluation protocol established in [13], we assess the proposed embodied agent on five widely adopted datasets covering core embodied AI tasks such as vision-and-language navigation (VLN) and embodied question answering (EQA). These benchmarks span a broad range of challenges, including multi-turn dialog comprehension, goal-directed object localization, and open-ended QA in realistic 3D environments.

- **CVDN** [59] presents a multi-turn dialog-based navigation task, requiring agents to reach target locations by interpreting a natural conversation between two humans. The dataset contains 2,050 human-human dialogs comprising over 7,400 navigation trajectories, interleaved with question–answer exchanges, and spans 83 real-world houses. It emphasizes the agent's ability to ground dialog history in spatial understanding and perform long-horizon planning in complex indoor environments.
- **SOON** [83] evaluates an agent's ability to locate specific objects based on rich natural language descriptions in large-scale 3D scenes. It provides 3,848 instructions describing absolute object locations, along with 6,326 bounding boxes for 3,923 objects distributed across 90 Matterport scenes. Although the task imposes no constraint on the agent's starting position, the dataset includes over 30,000 long-distance trajectories for evaluating navigation effectiveness. Each instruction integrates attributes, relationships, and region descriptions to help uniquely identify the target object within visually complex environments.
- **R2R** [4] is a foundational VLN benchmark in which agents follow natural language instructions to navigate through photorealistic indoor environments built from Matterport3D. The dataset includes 7,189 unique navigation paths and 21,567 human-written instructions, with each path averaging 9.9 meters in length. The benchmark assesses agents' abilities in instruction following, visual-language grounding, and generalization to unseen environments.
- **REVERIE** [49] focuses on long-range referring expression grounding in realistic 3D spaces. It comprises 10,567 panoramic images and 21,702 high-level goal-oriented instructions spanning 90 buildings. The task requires agents to interpret abstract referring expressions, perform long-horizon navigation, and ground language in visual observations to accurately identify remote target objects.
- **ScanQA** [6] is a large-scale benchmark for embodied question answering over RGB-D reconstructed indoor scenes. It contains 41,363 questions and 58,191 answers, including 32,337 unique questions and 16,999 unique answers. In addition to question–answer pairs, the dataset also provides 3D object localization annotations, supporting fine-grained semantic reasoning and spatial understanding. Questions were collected through a combination of automated generation and human refinement, resulting in broad linguistic diversity across various spatial and functional query types.

Table 1: Overall comparison with state-of-the-art methods on in-domain tasks. * indicates experimental results that we have reproduced.

	CVDN	SOON		R2R		REVERIE		ScanQA	
	GP	SPL	SR	SPL	SR	SPL	SR	Val	ROUGE-L
Separate Model For Each Task									
PREVALENT [29]	3.15	-	-	53	58	-	-	-	-
HOP [50]	4.41	-	-	57	64	26.11	31.78	-	-
HAMT [13]	5.13	-	-	61	66	-	-	-	-
VLN-BERT [30]	-	-	-	57	63	-	-	-	-
GBE [83]	-	13.34	19.52	-	-	-	-	-	-
DUET [14]	-	22.58	36.28	60	69	33.73	46.98	-	-
Meta-Explore [36]	-	**34.84**	**44.69**	62	72	34.03	-	-	-
AZHP [20]	-	-	-	61	72	36.63	48.31	-	-
VLN-SIG [38]	5.52	-	-	62	72	-	-	-	-
VLN-PETL [51]	5.69	-	-	60	65	27.67	31.81	-	-
BEV-BERT [3]	-	-	-	**64**	**75**	36.37	**51.78**	-	-
3D-LLM [31]	-	-	-	-	-	-	-	20.5	35.7
Unified Model For All Tasks									
MT-RCM+Env [65]	4.65	-	-	49	52	-	-	-	-
NaviLLM [79]	6.16	28.09	35.44	58	67	36.63	44.56	**23.3**	38.2
NaviLLM* [79]	5.75	26.19	35.93	54	66	31.01	38.94	22.93	38.2
+BEEA	**6.30**	30.97	38.29	60	69	**37.28**	45.04	23.14	**38.33**

Table 2: Task success rates on 6 subsets of EB-ALFRED and EB-Habitat.

2pt
ccccccccccccccc

Model	EB-ALFRED							EB-Habitat						
	Avg	Base	Common	Complex	Visual	Spatial	Long	Avg	Base	Common	Complex	Visual	Spatial	Long
Proprietary MLLMs														
GPT-4o	56.3	64	54	68	46	52	54	59.0	86	44	56	68	36	**64**
GPT-4o-mini	24.0	34	28	36	24	22	0	32.7	74	22	32	22	32	14
Claude-3.5-Sonnet	**64.0**	**72**	**66**	**76**	**60**	**58**	52	**68.0**	**96**	**68**	**78**	70	**38**	58
Gemini-1.5-Pro	62.3	70	64	72	58	52	**58**	56.3	92	52	48	56	38	52
Gemini-2.0-flash	52.3	62	48	54	46	46	**58**	42.3	82	38	38	36	34	26
Gemini-1.5-flash	39.3	44	40	56	42	26	28	39.3	76	32	48	36	32	12
Qwen-VL-Max	41.3	44	48	44	42	38	32	45.3	74	40	50	42	30	36
GPT-4o (Lang)	58.0	62	64	70	52	46	54	56.0	82	52	58	**74**	34	36
GPT-4o-mini (Lang)	31.3	42	36	46	30	20	14	36.7	82	30	34	30	30	14
Open-Source MLLMs														
InternVL2.5-8B	2.0	4	6	2	0	0	0	11.3	36	4	0	10	16	2
+BEEA	3.7	4	8	6	2	0	2	15.3	44	10	0	16	16	6
Qwen2.5-VL-7B-Ins	4.7	10	8	6	2	0	2	14.3	32	2	26	10	14	2
+BEEA	7.7	12	8	14	6	0	6	21.0	56	6	28	16	14	6
InternVL2.5-78B	37.7	38	34	42	34	36	42	49.0	80	42	56	58	30	28
+BEEA	39.3	38	36	52	28	40	42	51.7	82	38	58	68	32	32

2.2 Out-of-domain Datasets

A proficient embodied agent should be able to generalize beyond its training distribution and exhibit spatial intelligence in novel scenarios. Similar to how humans acquire spatial and behavioral understanding through exploration and interaction [17, 23, 35], an embodied agent is expected to implicitly develop robust generalization capabilities by accumulating multimodal experience in diverse environments (e.g., through vision-and-language navigation). To validate the generalization ability of our proposed framework, we conduct zero-shot evaluations across multiple out-of-domain embodied intelligence and spatial reasoning benchmarks.

- **EmbodiedBench** [73] is a large-scale and systematically designed benchmark for evaluating vision-based embodied agents built upon multimodal large language models (MLLMs). It encompasses 1,128 diverse tasks distributed across four distinct simulated household environments. The tasks span a wide difficulty spectrum, ranging from high-level semantic planning tasks

Table 3: Task success rates on 5 subsets of EB-Navigation and EB-Manipulation

cccccccccccccc

2pt

Model	EB-Navigation						EB-Manipulation					
	Avg	Base	Common	Complex	Visual	Long	Avg	Base	Common	Complex	Visual	Spatial
Proprietary MLLMs												
GPT-4o	**57.7**	55.0	60.0	**58.3**	**60.0**	**55.0**	**28.9**	**39.6**	**29.2**	**29.2**	**19.4**	25.0
GPT-4o-mini	32.8	31.7	33.3	35.0	28.3	33.3	4.8	4.2	6.3	2.1	0.0	10.4
Claude-3.5-Sonnet	44.7	**66.7**	51.7	41.7	36.7	26.7	25.4	37.5	16.7	**29.2**	**19.4**	22.9
Gemini-1.5-Pro	24.3	23.3	25.0	25.0	28.3	20.0	21.1	14.6	14.6	22.9	16.7	**35.4**
Gemini-2.0-flash	48.7	63.3	**65.0**	50.0	51.7	13.3	16.7	14.6	8.3	14.6	13.9	31.3
Gemini-1.5-flash	41.7	56.7	50.0	46.7	50.0	5.0	9.6	14.6	10.4	4.2	8.3	10.4
Qwen-VL-Max	39.7	50.0	46.7	41.7	35.0	25.0	18.0	25.0	10.4	18.8	2.8	29.2
GPT-4o (Lang)	17.4	21.7	21.7	26.7	16.7	0.0	16.2	16.7	16.7	14.6	**19.4**	14.6
GPT-4o-mini (Lang)	8.3	3.3	13.3	10.0	15.0	0.0	6.6	12.5	0.0	2.1	2.8	14.6
Open-Source MLLMs												
InternVL2_5-8B	21.3	35.0	23.3	21.7	26.7	0.0	7.0	8.3	2.1	6.3	8.3	10.4
+BEEA	26.7	41.6	25.0	31.6	26.7	8.3	9.2	8.3	6.3	8.3	10.4	12.5
Qwen2.5-VL-7B-Ins	25.7	28.3	30.0	41.7	20.0	8.3	9.6	8.3	8.3	8.3	5.6	16.7
+BEEA	29.7	31.7	38.3	38.3	26.7	13.3	12.1	10.4	12.5	8.3	10.4	18.8
InternVL2_5-78B	30.7	36.7	38.3	33.3	21.7	23.3	18.0	16.7	16.7	14.6	22.2	20.8
+BEEA	33.0	43.3	33.3	35.0	26.7	26.7	23.3	20.8	16.7	18.8	25.0	35.4

Table 4: Results on VSI-Bench. † denotes results on VSI-Bench (tiny) set.

Methods	Numerical Answer				Multiple-Choice Answer			
	Obj. Count	*Abs. Dist.*	*Obj. Size*	*Room Size*	*Rel. Dist.*	*Rel. Dir.*	*Route Plan*	*Appr. Order*
†Human Level	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100.0
GPT-4o	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Pro	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
LLaVA-NeXT-Video-72B	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
InternVL2.5-8B	7.0	33.4	42.4	41.3	38.0	39.7	25.7	36.0
+BEEA	**7.4**	**34.4**	**43.0**	**43.7**	**38.9**	**42.0**	**27.3**	**37.1**
Qwen2.5-VL-7B	**23.0**	16.5	48.0	23.3	37.5	39.7	28.9	29.5
+BEEA	22.9	**18.1**	**49.7**	**25.2**	**38.4**	**40.9**	**30.0**	**30.5**
InternVL2.5-78B	46.7	30.3	55.4	45.4	46.6	37.6	28.9	30.9
+BEEA	**47.2**	**34.8**	**56.9**	**46.7**	**48.0**	**41.2**	**32.1**	**35.4**

(e.g., organizing a room, preparing breakfast) to low-level sensorimotor actions (e.g., pick-and-place, navigation, and object manipulation). Each task is paired with multimodal instructions, combining natural language and visual context. The benchmark comprehensively assesses six core embodied intelligence capabilities: spatial reasoning, commonsense reasoning, long-horizon planning, visual perception, action grounding, and language understanding. Its diversity and task richness make it a valuable resource for testing the generalization, compositionality, and robustness of vision-language agents under out-of-distribution conditions.

- **VSI-Bench** [72] is a diagnostic benchmark targeting the evaluation of spatial perception and reasoning abilities in multimodal models, particularly MLLMs. The dataset comprises over 5,000 multiple-choice question–answer pairs derived from 288 annotated videos of real-world indoor scenes. Each video captures dynamic scenes from a first-person or third-person perspective, mimicking real-world embodied exploration. The benchmark is organized into eight spatial task categories, including object counting, object search, distance estimation, spatial relationship identification, collision prediction, navigation route planning, and goal-driven scene understanding. Each task is designed to evaluate a specific aspect of visual-spatial intelligence in complex and cluttered environments. By focusing on spatial inference grounded in visual sequences, VSI-Bench complements static scene-based benchmarks and provides critical insights into the temporal and geometric reasoning abilities of modern embodied agents.

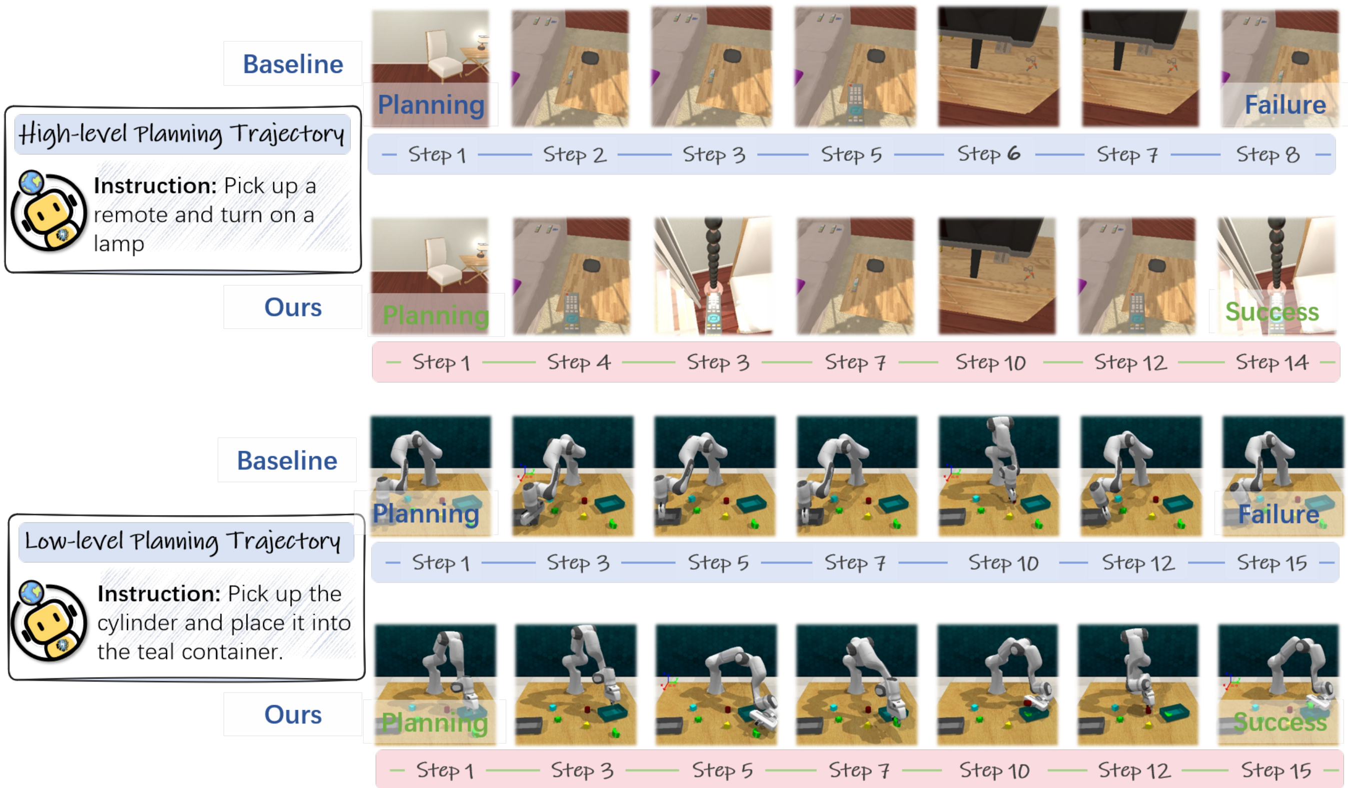

Figure 1: Embodied execution comparison in EB-ALFRED and EB-Manipulation using our BEEA and InternVL2.5-78B.

Evaluation Metrics. For navigation tasks in in-domain datasets, we employ Success Rate (SR), Success weighted by Path Length (SPL) and Goal Progress (GP) as evaluation metrics. Here, SR is the proportion of successfully executed instructions with the NE less than 3 meters; SPL denotes SR penalized by Path Length, which is calculated as $\frac{1}{E}\sum_{i=1}^{E} S_i \frac{l_i}{max(p_i,l_i)}$, where E is the number of tasks, S_i denotes the success as a binary value, l_i and p_i denote the shortest path and actual path length for the ith task. GP indicates the average agent progress towards the goal location. For 3D-QA tasks, we adopt the Exact Match (EM) accuracy metric. Regarding out-of-domain datasets, we utilize the respective evaluation metrics specified in each benchmark: for instance, execution accuracy for both high-level and low-level embodied tasks in EmbodiedBench, and question-answering accuracy across different spatial intelligence dimensions in VSI-Bench.

Baselines. We construct our embodied evolution agent using representative and popular multimodal large language models, including InternVL2.5-8B/78B [76] and Qwen2.5-VL-7B-Ins [7]. Additionally, since the original in-domain MLLM, NaviLLM [79] were trained based on the Vicuna language model [48], we conduct comparative evaluation and ablation studies on this baseline model to validate the effectiveness of our proposed framework. Furthermore, we comprehensively compare our results with those reported by other methods in respective benchmarks [72, 73, 79].

2.3 Implementation Details

In this study, we employ multiple open-source MLLMs to validate the effectiveness of the BEEA framework, including Vicuna-based NaviLLM [79], InternVL2.5-8B/78B [76], and Qwen2.5-VL-7B-Ins [7]. We conduct multi-task fine-tuning using in-domain datasets. Following [28, 79], we utilize the Adam optimizer with a learning rate of 3e-5 and train for 5000 steps. A training strategy alternating between teacher forcing and student forcing is adopted. Diverse tasks—including embodied navigation and question answering—are reformulated as generation problems via schema-based instruction representations. Please refer to the [79] for dataset details and schema examples. For the Vicuna-based NaviLLM, adhering to [79], we perform full-parameter fine-tuning on in-domain datasets. For the other three representative MLLMs, to better assess the enhancement of zero-shot generalization on out-of-domain tasks, fine-tuning is conducted solely on in-domain navigation datasets. Notably, for these three MLLMs, all parameters of the visual encoders and LLMs are frozen during training. Following [81], we introduce a Q-Former-like structure [41] between the visual encoder and LLM for parameter fine-tuning. For hyperparameters, we set K to 10, and β_{L2W}/β_{W2L} and γ are set to 0.95 and 0.9, respectively. During testing, with respect to the sampling strategy for action generation, we referred to [72, 73, 79] and employ varying temperatures and greedy strategies. It is worth noting that the dynamic communication slot is dynamically updated during testing. All models are trained using 8 Nvidia A100 GPUs.

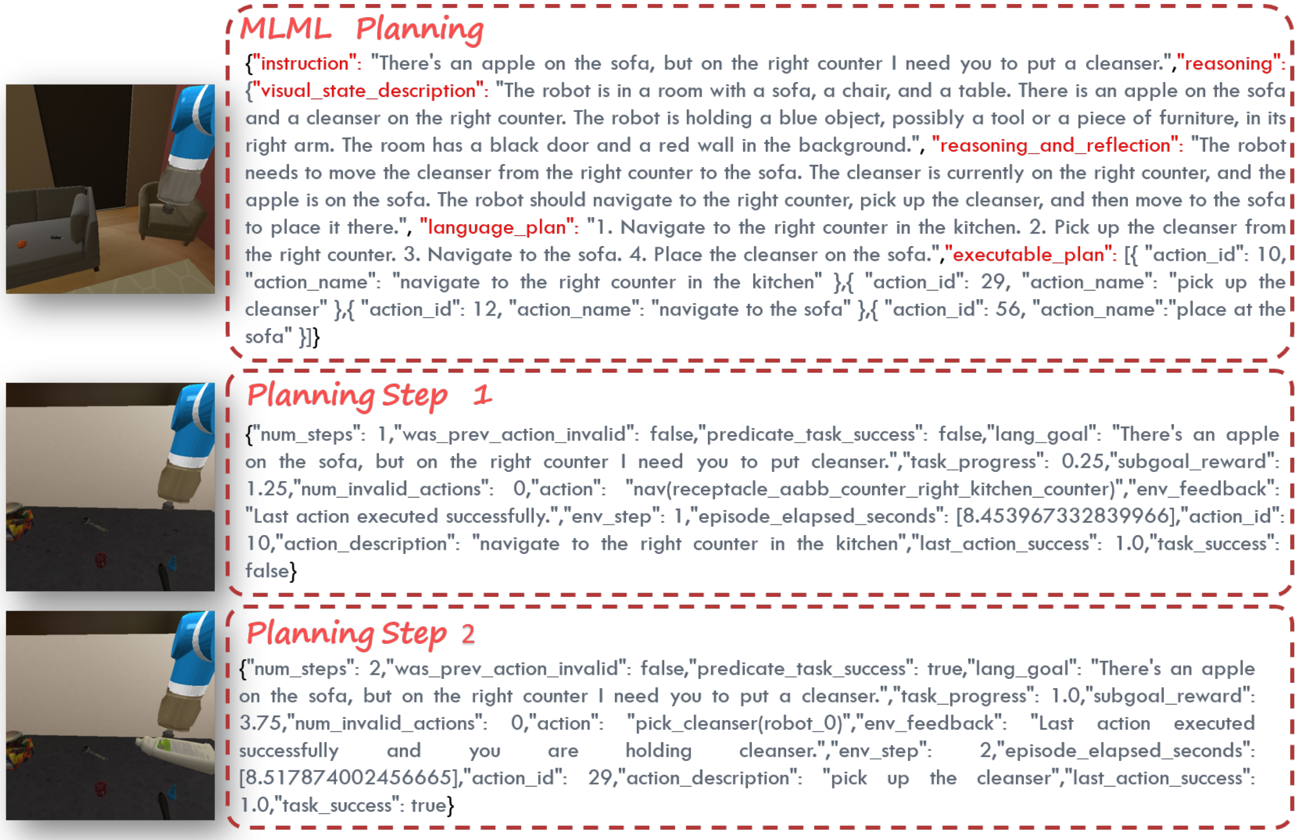

Figure 2: In-context learning example in EB-ALFRED using our BEEA with InternVL2.5-78B.

3 Quantitative Comparisons

3.1 Experimental Results on In-domain Tasks

Comparison with Full Parameter-tuning Methods. Following NaviLLM [79], we adopt a multimodal perception framework comprising a Vicuna-based language model and a Vision Transformer (ViT)-based visual encoder, with full parameter multi-task fine-tuning during training. As evidenced by Table 1, our proposed BEEA model outperforms the baseline NaviLLM across all tasks and metrics. Notably, BEEA significantly improves navigation performance over NaviLLM on SOON, R2R, and REVERIE. This enhancement primarily stems from the synergistic collaboration between the MLLM and the WM, facilitated by dynamic communication slots analogous to the corpus callosum in biological systems. In addition, both BEEA and NaviLLM, as unified multi-task models, achieve superior performance over specialized models on CVDN, SOON, and ScanQA, while attaining comparable results to task-specific models on R2R and REVERIE . This demonstrates that multimodal multi-task training frameworks exhibit greater potential in tasks requiring complex language understanding and interaction (e.g., CVDN, SOON, ScanQA).

3.2 Experimental Results on Out-of-domain Datasets

Comparison Results on EmbodiedBench. EmbodiedBench is designed to evaluate the performance of MLLMs in vision-driven embodied agent tasks. This benchmark encompasses four environments: EB-ALFRED and EB-Habitat (high-level tasks) and EB-Navigation and EB-Manipulation (low-level tasks). The results in Table 2 and Table 3 demonstrate that MLLMs exhibit significantly better performance in high-level tasks compared to low-level tasks, with proprietary models like GPT and Claude outperforming smaller-scale open-source MLLMs. Notably, with the integration of the proposed embodied evo-agent framework, various MLLMs show consistent improvements across most metrics. Additionally, long-horizon planning emerges as the most challenging subtask, with model performance generally declining by 20%-30% compared to base tasks, highlighting the current limitations of MLLMs in complex sequential decision-making. The improvement achieved by our approach in this subtask underscores the potential of bridging MLLM and WM to enhance the execution of real-world tasks.

Comparison Results on VSI-Bench. To evaluate whether the proposed method can implicitly enhance an agent's ability to understand and reason space after undergoing only basic embodied exploration pretraining (e.g., vision-and-language navigation), we conducted validation experiments on VSI-Bench, a video-based benchmark specifically designed for assessing the visual-spatial

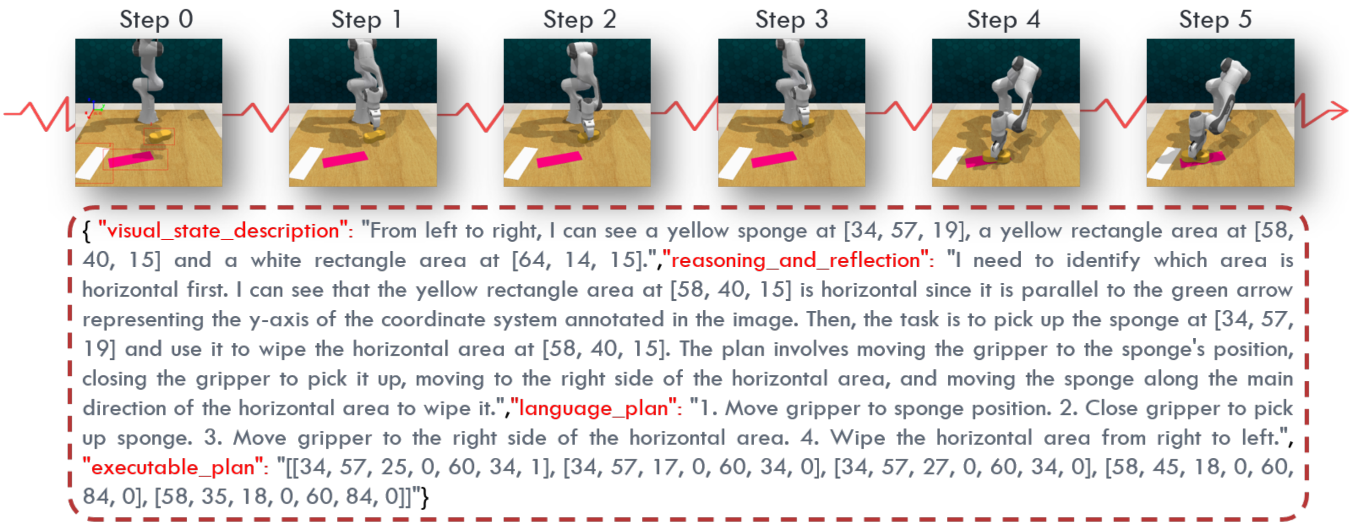

Figure 3: In-context learning example in EB-Manipulation using our BEEA with InternVL2.5-78B.

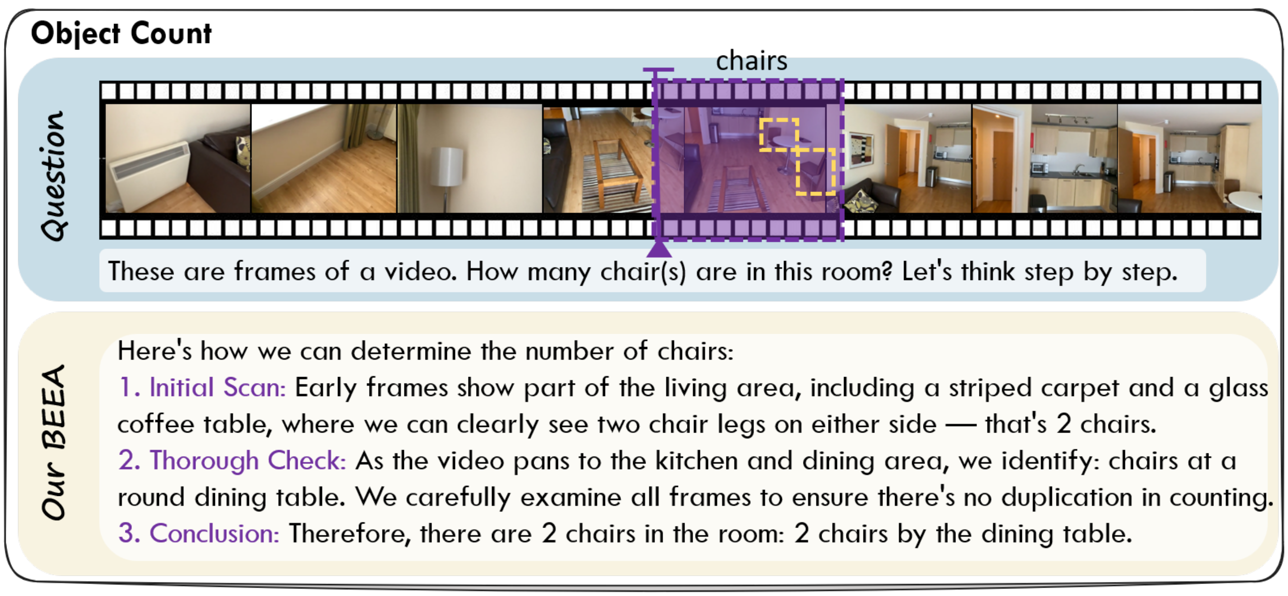

Figure 4: Chain-of-Thought example in VSI-Bench using our BEEA.

intelligence of MLLMs. As shown in Table 4, the proposed BEEA method significantly improves the performance of various baseline MLLMs. The most substantial improvements are observed in this table, which aligns with showing the strength of joint using MLLM and WM in capturing spatial relationships and global topological structures. Meanwhile, the gains in measurement estimation tasks are comparatively smaller, likely because precise numerical estimation relies not only on spatial reasoning but also on finer-grained visual perception capabilities. Despite these advancements, a notable performance gap remains compared to human-level accuracy, particularly in complex spatial reasoning tasks. Future research will focus on enhancing global 3D feature representation modeling and incorporating external knowledge.

4 Qualitative Analysis

For EmbodiedBench, As shown in Figure 1, by visualizing the agent execution process, it is evident that our proposed method significantly outperforms the baseline approach in terms of task execution coherence and success rate. In high-level tasks (e.g., complex instruction execution in EB-ALFRED), the proposed method more accurately decomposes task steps. For instance, in the task of "pick up a remote and turn on a lamp" the baseline method may produce invalid actions or omit steps whereas our method generates a complete action sequence. In low-level tasks (e.g., object grasping in EB-Manipulation), the baseline model frequently exhibits action misalignment or failure due to inadequate spatial awareness. In contrast, the proposed method, leveraging an enhanced world model and communication mechanism with the multimodal large model, markedly improves grasping precision (e.g., successfully picking up the cylinder and placing it into the teal container.). These findings validate the robustness and generalization capability of the proposed method across tasks of varying complexity levels. Following [73], we provide text-based in-context learning (ICL) demonstrations in Figure 2 and Figure 3, showing the proposed method's understanding ability for various tasks.

For the VSI-Bench, we also follow [72] to provide a Chain-of-Thought example, showing the effectiveness of our proposed method for spatial understaning. As shown in Figure 4, our method demonstrates accurate reconstruction of spatial relationships, as evidenced by precise timestamped descriptions and coherent step-by-step reasoning that align closely with ground truth.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

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

[3] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. 2022. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385* (2022).

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*. 3674–3683.

[5] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*.

[6] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. 2022. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*.

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* (2023).

[8] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[9] Adrien Bardes, Jean Ponce, and Yann LeCun. 2023. Mc-jepa: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698* (2023).

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818* (2023).

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *NeurIPS* (2020).

[12] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. 2024. Genie: Generative interactive environments. In *ICML*.

[13] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *NeurIPS* (2021).

[14] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*.

[15] Shoubin Chen, Zehao Wu, Kai Zhang, Chunyu Li, Baiyang Zhang, Fei Ma, Fei Richard Yu, and Qingquan Li. 2025. Exploring Embodied Multimodal Large Models: Development, Datasets, and Future Directions. *arXiv preprint arXiv:2502.15336* (2025).

[16] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *ICML*.

[17] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. 2017. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience* 20, 11 (2017), 1504–1513.

[18] Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yuwei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. 2025. EvoAgent: Agent Autonomous Evolution with Continual World Model for Long-Horizon Tasks. *arXiv preprint arXiv:2502.05907* (2025).

[19] Karl Friston, Rosalyn J Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. 2021. World model learning and inference. *Neural Networks* (2021).

[20] Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. 2023. Adaptive Zone-Aware Hierarchical Planner for Vision-Language Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14911–14920.

[21] Junyu Gao, Xuan Yao, and Changsheng Xu. 2024. Fast-Slow Test-Time Adaptation for Online Vision-and-Language Navigation. In *ICML*.

[22] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. 2024. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504* (2024).

[23] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. 2021. Embodied intelligence via learning and evolution. *Nature communications* 12, 1 (2021), 5721.

[24] David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. *NeurIPS* (2018).

[25] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122* (2018).

[26] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. [n. d.]. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR*.

[27] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104* (2023).

[28] Mingfei Han, Liang Ma, Kamila Zhumakhanova, Ekaterina Radionova, Jingyi Zhang, Xiaojun Chang, Xiaodan Liang, and Ivan Laptev. 2024. RoomTour3D: Geometry-Aware Video-Instruction Tuning for Embodied Navigation. *arXiv preprint arXiv:2412.08591* (2024).

[29] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13137–13146.

[30] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. A Recurrent Vision-and-Language BERT for Navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1643–1653.

[31] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. [n. d.]. 3d-llm: Injecting the 3d world into large language models. *NeurIPS* ([n. d.]).

[32] Yining Hong, Zishuo Zheng, Peihao Chen, Yian Wang, Junyan Li, and Chuang Gan. 2024. Multiply: A multisensory object-centric embodied large language model in 3d world. In *CVPR*.

[33] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080* (2023).

[34] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. 2023. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842* (2023).

[35] Karen A Hunt, Vanisha Mistry, Nicholas A Bockett, Tariq Ahmad, Maria Ban, Jonathan N Barker, Jeffrey C Barrett, Hannah Blackburn, Oliver Brand, Oliver Burren, et al. 2013. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature* 498, 7453 (2013), 232–235.

[36] Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, and Songhwai Oh. 2023. Meta-Explore: Exploratory Hierarchical Vision-and-Language Navigation Using Scene Object Spectrum Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6683–6693.

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

[38] Mohit Bansal Jialu Li. 2023. Improving Vision-and-Language Navigation by Generating Future-View Image Semantics. In *CVPR*.

[39] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In *8th Annual Conference on Robot Learning*.

[40] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS* (2023).

[42] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2024. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *arXiv preprint arXiv:2407.06886* (2024).

[43] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525* (2024).

[44] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. 2024. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093* (2024).

[45] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *ICCV*.

[46] Masashi Okada and Tadahiro Taniguchi. 2022. DreamingV2: Reinforcement learning with discrete world models without reconstruction. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 985–991.

[47] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*.

[48] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277* (2023).

[49] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*. 9982–9991.

[50] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. 2022. Hop: history-and-order aware pre-training for vision-and-language navigation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 15418–15427.

[51] Yanyuan Qiao, Zheng Yu, and Qi Wu. 2023. VLN-PETL: parameter-efficient transfer learning for vision-and-language navigation. In *ICCV*.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[53] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. 2023. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109* (2023).

[54] Dhruv Shah, Ajay Sridhar, Arjun Bhorkar, Noriaki Hirose, and Sergey Levine. 2023. Gnm: A general navigation model to drive any robot. In *ICRA*.

[55] Lawrence Shapiro. 2019. *Embodied cognition.* Routledge.

[56] Ola Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, Timothy Chen, Roya Firoozi, Monroe Kennedy III, and Mac Schwager. 2024. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. *arXiv preprint arXiv:2405.04378* (2024).

[57] Richard S Sutton. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990.* Elsevier, 216–224.

[58] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[59] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning.* 394–406.

[60] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[61] Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024. Repvit: Revisiting mobile cnn from vit perspective. In *CVPR*.

[62] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. 2023. Dreamwalker: Mental planning for continuous vision-language navigation. In *ICCV*.

[63] Jiaqi Wang, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Bao Ge, and Shu Zhang. 2024. Large language models for robotics: Opportunities, challenges, and perspectives. *Journal of Automation and Intelligence* (2024).

[64] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2024. Cogvlm: Visual expert for pretrained language models. *NeurIPS* (2024).

[65] Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020. Environment-agnostic multitask learning for natural language grounded navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16.* Springer, 413–430.

[66] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *CVPR*.

[67] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. 2024. ivideogpt: Interactive videogpts are scalable world models. *NeurIPS* 37 (2024).

[68] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning.*

[69] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *ICML*.

[70] Zhiyuan Xu, Kun Wu, Junjie Wen, Jinming Li, Ning Liu, Zhengping Che, and Jian Tang. 2024. A survey on robotics with foundation models: toward embodied ai. *arXiv preprint arXiv:2402.02385* (2024).

[71] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Haoran Tan, Chencheng Jiang, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. 2024. Octopus: Embodied vision-language programmer from environmental feedback. In *ECCV*.

[72] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171* (2024).

[73] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. 2025. EmbodiedBench: Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents. *arXiv preprint arXiv:2502.09560* (2025).

[74] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* (2023).

[75] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. 2024. Uni-NaVid: A Video-based Vision-Language-Action Model for Unifying Embodied Navigation Tasks. *arXiv preprint arXiv:2412.06224* (2024).

[76] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320* (2024).

[77] Weipu Zhang, Gang Wang, Jian Sun, Yetian Yuan, and Gao Huang. 2023. Storm: Efficient stochastic transformer based world models for reinforcement learning. *NeurIPS* (2023).

[78] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *ICML*.

[79] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards learning a generalist model for embodied navigation. In *CVPR*.

[80] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. 2024. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*.

[81] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. 2024. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *ECCV*.

[82] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. 2023. Query-centric trajectory prediction. In *CVPR*.

[83] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 2021. Soon: Scenario oriented object navigation with graph-based exploration. In *CVPR*. 12689–12699.