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

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

Abstract

Embodied artificial intelligence has rapidly developed under the impetus of multimodal learning, robotics, and cognitive science, demonstrating great potential in fields such as navigation and manipulation. However, building embodied agents that can robustly operate in diverse and dynamic environments still faces challenges, such as handling partial observability and environmental adaptability. Multimodal large language models (MLLMs) are vital for embodied intelligence due to their ability to process multimodal information, but they encounter difficulties in understanding spatial environments and performing dynamic decisions and evolution. Inspired by the functional specialization of the left and right hemispheres of the human brain, this paper proposes a brain-inspired learning and evolution paradigm for embodied agents. The method designs an embodied context-augmented MLLM to simulate the language processing and logical analysis capabilities of the left hemisphere, responsible for understanding instructions and visual scenes. At the same time, it constructs a perceptual context-guided world model based on the recurrent state space model to simulate the spatial perception and holistic thinking functions of the right hemisphere, capturing environmental dynamics and predicting future states. By simulating the communication function of the corpus callosum, we propose dynamic communication slots for efficient information exchange between MLLMs and the world model, which also allows the agent to quickly adapt to dynamic environments without requiring extensive computational resources. Experiments show that the proposed paradigm significantly improves the performance of embodied agents in a series of tasks and enhances their generalization ability in zero-shot tasks through embodied exploration experience and online evolution.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Embodied agent, Evolution, Multimodal large language model, World model, Brain-inspired

1 Introduction

Embodied intelligence has flourished in recent years, propelled by advancements in artificial intelligence, robotics, and cognitive science [16, 40, 42, 64]. Embodied intelligence refers to the emergence of intelligent behavior through the interaction between an agent and its spatial environment [51]. This paradigm has demonstrated immense potential across domains such as robotics [9, 22, 63] and autonomous driving [21, 61, 73]. However, developing embodied agents capable of robust operation in diverse and dynamic environments remains a formidable challenge, as it requires addressing partial observability, ensuring effective spatial understanding, and facilitating agent evolution [20, 63, 66].

Currently, multimodal large language models (MLLMs) play a pivotal role in advancing embodied intelligence [30, 32, 37, 59]. Built upon the successes of large language models (LLMs) [1, 45], MLLMs excel in processing and integrating diverse modalities—such as text, images, and audio—demonstrating remarkable performance in cross-modal understanding and reasoning tasks [1, 6, 41, 69]. In the realm of embodied intelligence, MLLMs exhibit significant

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

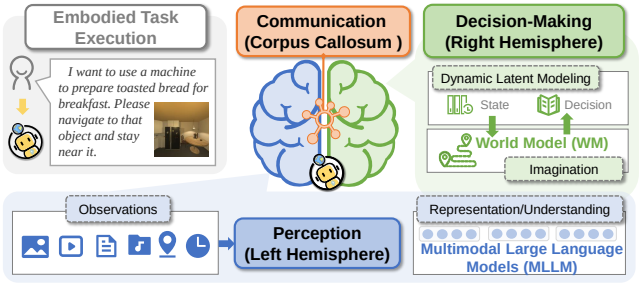

Figure 1: The motivation behind our proposed brain-inspired embodied evolutionary agent stems from the hypothesis that MMLM and WM can functionally emulate the distinct roles of the left and right cerebral hemispheres, while a communication module enables embodied task execution.

promise [14, 15, 18, 19, 44, 65]. Nevertheless, despite their proficiency in multimodal perception and representation, MLLMs face two primary challenges in embodied deployment: **(a)** They lack the capacity for spatial environmental understanding [66] and dynamic behavioral decision-making [14, 40], rendering direct execution of embodied tasks difficult. Although some studies have attempted to design hand-crafted prompts and action spaces, they still fall short of comprehensively modeling the embodied world, struggling to construct dynamic internal representations or predict future states. **(b)** The vast parameter scale of MLLMs hinders dynamic updates during embodied tasks, limiting the autonomous evolution of embodied agents, particularly in resource-constrained settings.

The functional specialization of the human brain's left and right hemispheres [24, 35, 36] offers valuable inspiration for designing embodied agents, as shown in Figure 1. Research indicates [24, 36] that the left hemisphere primarily governs language processing, logical analysis, and reasoning, excelling in detailed tasks such as reading and logical inference. Conversely, the right hemisphere is associated with spatial perception, imagination, and holistic thinking, adept at reconstructing and imagining spatial-visual structures and comprehending the world as a whole. These hemispheres are interconnected via the corpus callosum [36], a bundle of nerve fibers that facilitates interhemispheric communication and integration. Notably, the corpus callosum exhibits dynamic adaptability, incrementally adjusting its activation strength based on experience and learning [36, 46]. In the context of embodied intelligence, MLLMs can be likened to the left hemisphere, providing language comprehension and logical reasoning to interpret embodied operation instructions. To emulate the right hemisphere's capabilities and enhance embodied agents, a promising approach is to construct a world model (WM) [17, 26–28]. Defined as an agent's internal representation of its environment, a world model captures state transitions and predicts future states/actions, offering advantages in decision-making, planning, and environmental adaptation.

Inspired by these insights, this paper proposes the development of a brain-inspired embodied evolutionary agent, introducing a unified paradigm that bridges MLLMs and world models. As illustrated in Figure 1, our approach constructs an embodied context-augmented MLLM to emulate the left hemisphere's language processing and logical analysis, tasked with interpreting instructions and multimodal perception. Concurrently, we design a perceptual context-guided world model, built upon a Recurrent State Space

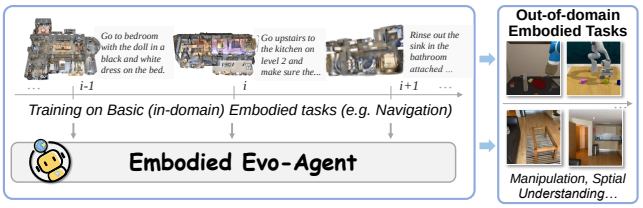

Figure 2: We aim to construct an embodied evo-agent that leverages in-domain task experience to enhance its zero-shot generalization capability on out-of-domain yet related tasks.

Model (RSSM) [28, 43], to simulate the right hemisphere's spatial perception and holistic reasoning, predicting future observations and actions. To effectively connect the MLLM and WM, we draw inspiration from the corpus callosum and introduce a module named dynamic communication slots. This module facilitates bidirectional information exchange: the MLLM transmits high-level instructions and contextual information to the world model, which, in turn, provides spatial structural insights and decision-state feedback to the MLLM. Furthermore, the dynamic communication slots employ an efficient forward representation update mechanism, enabling dynamic updates and reshaping of the embodied agent without explicit gradient backpropagation. This design empowers the recursive state transition process of the world model while simultaneously refining the intermediate representations of the MLLM, enhancing the agent's flexibility and adaptability in dynamic environments.

The proposed agent construction and learning approach significantly improves the performance across a range of embodied tasks. Moreover, as depicted in Figure 2, without supervised training on specific out-of-domain tasks, the agent leverages embodied exploration experience and online evolutionary capabilities from in-domain tasks to achieve enhanced zero-shot embodied execution and spatial comprehension across diverse tasks.

In summary, the contributions of this paper are threefold:

- A brain-inspired paradigm is designed for learning and evolving embodied agents, drawing on the functional lateralization of brain, and integrate embodied context-enhanced MLLM with perceptual context-guided WM within a unified framework.
- We introduce the module of dynamic communication slots, emulating the mechanism of the corpus callosum, to facilitate effective information exchange between the MLLM and the world model. A slot-based message update mechanism is developed, enabling rapid model adaptation.
- Extensive experimental results validate the effectiveness of our approach across a series of embodied tasks. Beyond improving performance on in-domain tasks, the model's accumulated experience and dynamic update capabilities enhance its generalization to zero-shot out-of-domain tasks.

2 Related Work

Embodied Artificial Intelligence. Embodied AI explores the interaction between physical agents and their environments, integrating insights from AI and cognitive science [14, 40, 42, 51, 64, 72]. Recent progress in multimodal large models has driven advances in perception and reasoning, notably through Vision-Language-Action models (VLAs) [9, 15, 37, 44, 65, 71], enabling agents to follow visual-linguistic instructions for task execution. Additionally, models like GPT-4V [34, 59], LLaVA [32, 68], and LLaMA [37]

support embodied tasks by enhancing perception and instruction understanding. Beyond representation, simulating the world has become central to improving generalization in embodied AI [62, 71].

Multimodal Large Language Model (MLLM). Leveraging the significant text understanding and generation advancements in large language models (LLMs), such as GPT-3 [10] and the LLaMA series [56], MLLMs have extended the scope of AI applications in multimodal tasks. Models like GPT-4o [1], Gemini [54], Claude [7], and Grok exhibit strong multimodal competence but remain API-restricted. Open-source alternatives—QwenVL [6], CogVLM [60], InternLM-Xcomposer [69], LLaVA-NeXT [39], and DeepSeek-VL Janus [41], have democratized MLLM access. 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 focus on encoding the evolutionary patterns of world states, environmental responses to agent behaviors, and their intrinsic connections with perceptual inputs, thus constructing internal representation mechanisms for agents [17, 26, 27, 53]. While early work emphasized abstract representations [27], recent efforts advocate for predictive capabilities [23]. Existing architectures span Recurrent State Space Models (RSSMs) [28, 29, 43, 63], Joint-Embedding Predictive Architectures (JEPAs) [4, 8, 23], and Transformer-based models [11, 49, 70]. These models show promise in domains like robotics [50, 57, 58] and autonomous driving [33, 61, 73, 74]. Recent trends integrate MLLMs into world models, but often as monolithic components, lacking modular cooperation and cognitively inspired interfaces.

3 Our Approach

Framework Overview. Embodied agents operating in complex environments face the fundamental challenge of integrating multimodal perception, dynamic decision-making, and model adaptation. To tackle this, we propose a brain-inspired embodied evo-agent framework. As illustrated in Figure 3, our framework emulates this architecture through three loosely coupled yet functionally synergistic components: (1) Embodied Context-enhanced Multimodal Large Language Model (EC-MLLM), which mimics the left hemisphere by interpreting language commands, grounding them in visual observations, and generating embedded task representations; (2) Perceptual Context-guided World Model (PC-WM), which reflects the right hemisphere's capability for spatial and temporal abstraction, by maintaining a latent representation of environment dynamics and enabling future state anticipation across temporally extended interactions; and (3) Dynamic Communication Slots (DCS), which function as an artificial corpus callosum, bridging the two modules via bidirectional message passing and efficient representation alignment. These slots are dynamically updated throughout both training and inference via ongoing agent–environment interaction, without relying on explicit loss supervision or gradient-based updates. This modular, neuro-inspired design facilitates dynamic environment modeling and model evolution.

Problem Formulation. We consider embodied evolutionary agents as general-purpose task solvers that perform sequential decision-making based on visual perception and natural language instructions across diverse embodied scenarios. This problem can be modeled as a generalized Partially Observable Markov Decision Process (POMDP) [28, 52], formulated as a tuple $\mathcal{M}_e = \{\mathcal{S}_e, \mathcal{A}_e, \mathcal{O}, \mathcal{T}_e, L, T\}$, where each task instance $e \sim \mathcal{E}$ is drawn from a task distribution $\mathcal{E}$ and may involve distinct latent state and action spaces. Here, $\mathcal{S}_e$ denotes the latent state space underlying task e, which is learned and maintained by the world model; $\mathcal{A}_e$ is the task-specific action space; $\mathcal{O}$ is the observation space, where each observation $\boldsymbol{o}_t \in \mathcal{O}$ at timestep t consists of visual images. These raw observations are then processed by the MLLM, which integrates them with language instruction L—specifying the desired task goal—to support downstream reasoning and action selection. $\mathcal{T}_e : \mathcal{S}_e \times \mathcal{A}_e \to \mathcal{S}_e$ defines the environment's transition dynamics. $T \in \mathbb{N}$ denotes the episode horizon—the number of decision-making steps per embodied task. To enable generalization across all embodied tasks, we define a unified latent state space $\mathcal{S} = \bigcup_e \mathcal{S}_e$ and adopt a shared action embedding space $\mathcal{A}$, along with task-specific projection heads $f_a : \mathcal{A} \to \mathcal{A}_e$. This enables the learning of a unified policy $\pi : \mathcal{S} \to \mathcal{A}$ applicable to the full task distribution $\mathcal{E}$. At each timestep t, EC-MLLM encodes the current observation $\boldsymbol{o}_t$ and instruction $\boldsymbol{l}_t \in L$ into a multi-modal semantic embedding $\boldsymbol{z}_t$, while also maintaining a contextual history $\boldsymbol{h}_t$ to capture temporal dependencies. Concurrently, PC-WM maintains a latent state $\boldsymbol{s}_t$ that summarizes the agent's internal understanding of environmental dynamics and produces actions via a policy network $\pi(\boldsymbol{a}_t \mid \cdot)$.

3.1 Embodied Context-enhanced MLLM

In embodied environments, agents must interpret natural language instructions, ground them in visual observations, and construct context-aware semantic representations to support downstream decision making. To this end, we introduce the Embodied Context-enhanced Multi-modal Large Language Model (EC-MLLM), which serves as the agent's perceptual-semantic core.

At each timestep t, EC-MLLM receives embodied observations, which are projected into the same embedding space as the language tokens. These modality-specific embeddings are then interleaved with natural language inputs to form a unified multi-modal token sequence. Specifically, EC-MLLM receives: (1) language tokens $\boldsymbol{l}_t \in \mathbb{R}^{N_l \times d}$ encoding the current instruction; (2) visual tokens $\boldsymbol{v}_t = f_v(\boldsymbol{o}_t) \in \mathbb{R}^{N_v \times d}$ extracted from the raw image observation $\boldsymbol{o}_t \in \mathbb{R}^{N_o \times C \times H \times W}$ using a vision encoder (*e.g.,* ViT); and (3) historical tokens $\boldsymbol{h}_t \in \mathbb{R}^{N_h \times d}$ summarizing prior interaction context. To enhance perception with embodied context, we augment the raw visual tokens $\boldsymbol{v}_t$ with retrieved features from WM-to-MLLM communication slot C_{W2L}. These slot entries encapsulate task-relevant embeddings derived from previous agent-environment interactions during latent state learning guided by the world model. By injecting such embodied context, EC-MLLM is guided to interpret language instructions in light of broader environmental insights such as the world state information. The detailed slot-based augmentation procedure is elaborated in Section 3.3.

These tokens are concatenated into a unified sequence and passed into a Transformer-based multi-modal backbone:

$$\boldsymbol{z}_t = f_{\text{MLLM}}([\boldsymbol{l}_t; \boldsymbol{v}_t^*; \boldsymbol{h}_t]), \quad \boldsymbol{v}_t^* \leftarrow \text{SlotAug}(\boldsymbol{v}_t, C_{W2L}), \tag{1}$$

where $\boldsymbol{z}_t \in \mathbb{R}^{N_t \times d}$ represents token-level semantic embeddings and $N_t = N_l + N_v + N_h$. Additionally, $\boldsymbol{z}_t$ is not directly used for text generation, but instead provides a latent semantic representation of the current perceptual-linguistic context. This embedding is passed

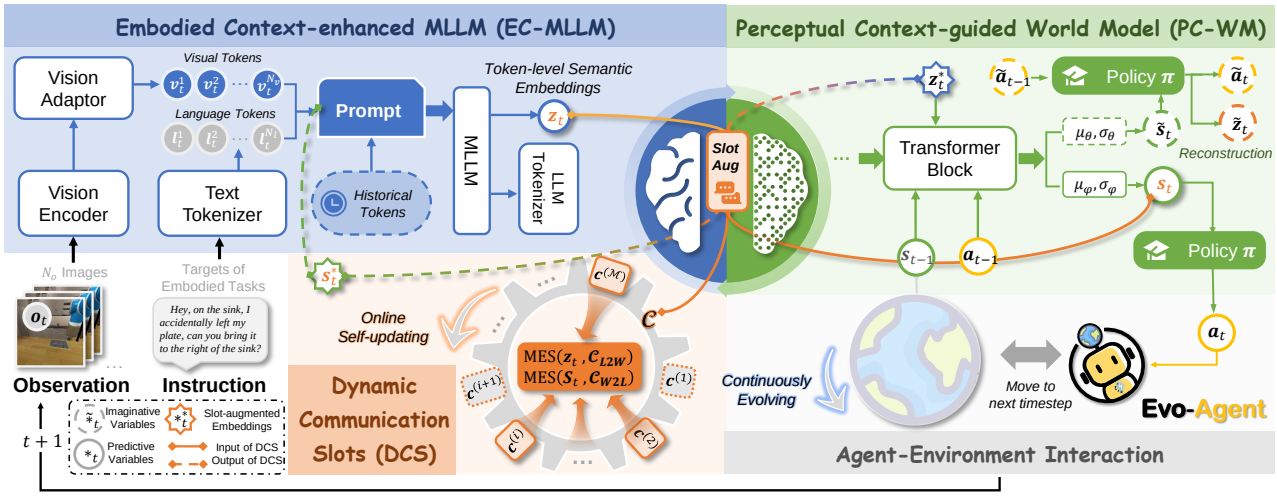

Figure 3: Framework Overview. Our framework comprises three bio-inspired modules: (1) EC-MLLM (left hemisphere) processes language-visual inputs for task understanding; (2) PC-WM (right hemisphere) models environment dynamics through recurrent state space model; (3) DCS (corpus callosum) enables inter-module communication via bidirectional message passing.

to the perceptual context-guided world model (Section 3.2), where it serves as a guiding signal for latent state construction, enabling the agent to align task understanding with environment dynamics.

We implement f_{MLLM} using state-of-the-art multimodal foundation models (*e.g.*, InternVL [69], QWen-VL [6]), selectively augmented with adapter modules for parameter-efficient adaptation. To validate out-of-domain generalization ability of our proposed framework, EC-MLLM undergoes fine-tuning only on basic embodied instruction-following datasets, such as vision-and-language navigation tasks [3, 47, 55, 75], using schema-based instruction representations and a unified cross-entropy objective $\mathcal{L}_{\text{MLLM}}$ [68, 72].

3.2 Perceptual Context-guided World Model

While EC-MLLM encodes rich perceptual-linguistic representations, such instantaneous context alone is insufficient for embodied agents operating under partial observability and long-horizon goals. Effective decision-making in such settings requires the agent to reason about unobserved dynamics, anticipate future changes, and evaluate the consequences of potential actions. We therefore introduce Perceptual Context-guided World Model (PC-WM)—a latent dynamics module that constructs compact, evolving representations of the environment in response to the agent's actions.

We formulate PC-WM as a latent simulator that approximates the dynamics of the underlying POMDP. Given that the true environment state $\boldsymbol{s}_t \in \mathcal{S}$ is not directly observable, we conceptualize it as a stochastic latent variable and model its temporal evolution through a Transformer-based Recurrent State Space Model (RSSM) [28, 29]. As RSSM sustains a compact latent representation integrating action history with instruction-grounded observations, it supports probabilistic state inference and latent transition forecasting.

Instead of directly processing raw environment images or sensory data, PC-WM takes as input the instruction-aware embeddings $\boldsymbol{z}_t$ produced by EC-MLLM. This semantic abstraction allows PC-WM to model the dynamics of latent world states $\boldsymbol{s}_t$ without requiring explicit pixel-level reconstruction, thereby achieving more efficient state representation learning. At each timestep t, PC-WM captures the environment's latent dynamics by inferring a hidden state $\boldsymbol{s}_t$ from prior interactions and current semantic observations. Given the partial observability and temporally sparse perceptual inputs inherent in embodied tasks, modeling these latent dynamics necessitates integrating not only current observations but also historically grounded visual-semantic cues. Therefore, we enrich the semantic embedding $\boldsymbol{z}_t$—produced by EC-MLLM—with perceptual context from the MLLM-to-WM communication slot C_{L2W}, fostering perception-aware modeling and ensuring long-horizon consistency in latent dynamics. We define a posterior distribution over $\boldsymbol{s}_t$, conditioned on the previous latent state $\boldsymbol{s}_{t-1}$, the action $\boldsymbol{a}_{t-1}$, and the current semantic embedding $\boldsymbol{z}_t$ from EC-MLLM:

$$q_\phi(\boldsymbol{s}_t \mid \boldsymbol{z}_{\le t}, \boldsymbol{a}_{<t}) \sim \mathcal{N}\left(\mu_\phi(\boldsymbol{z}_t^*, \boldsymbol{a}_{t-1}, \boldsymbol{s}_{t-1}), \sigma_\phi(\boldsymbol{z}_t^*, \boldsymbol{a}_{t-1}, \boldsymbol{s}_{t-1})\boldsymbol{I}\right),$$
$$\boldsymbol{z}_t^* \leftarrow \text{SlotAug}(\boldsymbol{z}_t, C_{L2W}), \tag{2}$$

where the posterior network, parameterized by ϕ, predicts the Gaussian parameters (μ_ϕ, σ_ϕ) through a projection network, yielding a learned posterior state distribution. $\boldsymbol{z}_t^*$ represents a perception-augmented semantic observation guided by contextual priors. Rather than introducing a separate policy head, we capitalize on EC-MLLM's robust multimodal reasoning capabilities for action modeling. Its output tokens, conditioned on both linguistic and visual inputs, naturally encode task objectives, spatial semantics, and interaction affordances. Accordingly, we instantiate $\boldsymbol{a}_t$ as a latent action embedding produced by the language modeling head of EC-MLLM, sampled from a learned policy $\pi(\boldsymbol{a}_t \mid \boldsymbol{z}_{\le t}, \boldsymbol{a}_{<t})$.

Beyond inferring the current latent state from real observations, an essential function of the world model is to simulate future states in the absence of new inputs—enabling the agent to anticipate how the environment may change based purely on internal dynamics—without relying on external observations. To support this imagination process, we introduce a prior transition network that simulates latent state trajectories under hypothetical actions, predicting the next latent state conditioned only on the previously

imagined state $\tilde{s}_{t-1}$ and action $\tilde{a}_{t-1}$:

$$p_\theta(s_t \mid \tilde{s}_{t-1}) \sim \mathcal{N}\big(\mu_\theta(\tilde{s}_{t-1}, \tilde{a}_{t-1}),\ \sigma_\theta(\tilde{s}_{t-1}, \tilde{a}_{t-1})\boldsymbol{I}\big), \quad (3)$$

where θ parameterizes the prior network, and $\tilde{a}_{t-1}$ is sampled from current policy $\pi(\tilde{a}_{t-1} \mid \tilde{s}_{t-1})$. This imagination mechanism allows the agent to simulate hypothetical interaction processes without requiring access to actual observations.

To ensure that imagined trajectories remain semantically meaningful and behaviorally coherent, we formulate a variational training objective that jointly supervises both generative reconstruction and latent dynamics modeling. Specifically, we optimize a variational lower bound on the data log-likelihood [38], which encourages the model to reconstruct observations and predict actions while regularizing the latent state transitions. The agent first observes inputs over T timesteps and subsequently imagines latent trajectories for an additional T_{im} steps within an embodied task. The training objective is defined as:

$$\mathcal{L}_{\text{WM}} = \sum_{t=1}^{T+T_{im}} \mathbb{E}_q \left[\log p(z_t \mid s_t) + \log p(a_t \mid s_t)\right] - \sum_{t=1}^{T} D_{KL}\big(q(s_t \mid z_{\le t}, a_{<t}) \,\|\, p(s_t \mid s_{t-1})\big), \quad (4)$$

where the first term encourages semantic reconstruction and action prediction, while the second term enforces consistency between the posterior and prior distributions over latent dynamics. For notational simplicity, we unify the latent-action variables across both observation and imagination phases as (s_t, a_t), omitting the hat notation $(\tilde{s}_t, \tilde{a}_t)$. In practice, the reconstruction objective is implemented as a semantic prediction loss, formulated as a normalized L2 distance between the predicted and actual embeddings. The action prediction loss is implemented as a standard cross-entropy objective over token distributions. By optimizing this objective, PC-WM learns to model how latent states evolve under different actions, enabling action-conditioned imagination, language-aligned planning, and robust decision-making in embodied tasks.

3.3 Dynamic Communication Slots

To enable adaptive information exchange between EC-MLLM and PC-WM, we propose a pair of Dynamic Communication Slots (DCS) that facilitate neuro-inspired, attention-based communication. Analogous to the corpus callosum in the human brain, these slots mediate bidirectional representation flow between perceptual and latent execution modules, with evolving contents driven by semantic alignment. We introduce two distinct dynamic communication slots: C_{L2W} and C_{W2L}, corresponding to MLLM-to-WM and WM-to-MLLM communication pathways, respectively. Each communication slot is defined as a set of message vectors $C = \{c^{(i)}\}_{i=1}^{M}$, $c^{(i)} \in \mathbb{R}^d$ encodes a contextual message from past interactions.

MLLM-to-WM Communication. The slot C_{L2W} stores semantic-level multimodal embeddings z_t generated by EC-MLLM, capturing grounded task semantics fused from both language and perceptual modalities. These representations serve as episodic cues that support posterior inference in PC-WM by injecting historically aligned semantic context. We define an attention-based communication operator that retrieves a task-relevant message from the slot $C_{L2W} = \{c_z^{(i)}\}_{i=1}^{M}$, based on the current semantic query z_t:

$$\text{MES}(z_t, C_{L2W}) = \sum_{i=1}^{M} \alpha_i \cdot c_z^{(i)}, \quad \alpha_i = \frac{\exp\left(z_t^\top \cdot \eta(c_z^{(i)})\right)}{\sum_j \exp\left(z_t^\top \cdot \eta(c_z^{(j)})\right)}, \quad (5)$$

where $\eta(\cdot)$ denotes a lightweight residual MLP projection that maps slot entries into an attention-compatible key space.

The retrieved semantic message is then fused with the current semantic embedding using a soft interpolation strategy:

$$\text{SlotAug}(z_t, C_{L2W}) = \beta_{L2W} \cdot z_t + (1 - \beta_{L2W}) \cdot \text{MES}(z_t, C_{L2W}), \quad (6)$$

where $\beta_{L2W} \in [0, 1]$ is a gating hyperparameter.

WM-to-MLLM Communication. The slot $C_{W2L} = \{c_s^{(i)}\}_{i=1}^{M}$ maintains latent states s_t learned by PC-WM. These latent representations encapsulate abstract, temporally coherent information about the environment's dynamics, distilled from agent-environment interactions, enabling embodied guidance for downstream perception. To integrate this distilled knowledge into the agent's current visual understanding, we enable EC-MLLM to access these latent dynamics via messages transmitted from PC-WM. Specifically, the current visual embedding v_t, which encodes egocentric observations at timestep t, acts as a query to retrieve semantically relevant latent priors from the slot C_{W2L}: $\text{SlotAug}(v_t, C_{W2L}) = \beta_{W2L} \cdot v_t + (1 - \beta_{W2L}) \cdot s_t^*$, where $s_t^* = \text{MES}(s_t, C_{W2L})$, $\beta_{W2L} \in [0, 1]$ is a gating hyperparameter. The resulting enhanced visual embedding v_t^* is subsequently fed to EC-MLLM, enabling it to ground the next-step observation o_{t+1} in a context-enriched representation.

Additionally, the communication slots C are initialized randomly and progressively updated during training and online testing via an attention-based momentum mechanism. At each timestep, the current query q_t (i.e., z_t for C_{L2W} and v_t for C_{W2L}) is matched against existing slot entries using cosine similarity. The top-K most similar entries are selected and softly updated as follows:

$$\left\{c^{(k)}\right\} = \arg\text{Top}_K\left(\text{Sim}[q_t, c^{(i)}]\right), c^{(k)} \leftarrow \gamma c^{(k)} + (1 - \gamma) q_t \quad (7)$$

where $\text{Sim}[\cdot, \cdot]$ denotes cosine similarity and $\gamma \in [0, 1]$ is a momentum coefficient controlling the update strength. This dynamic updating process enables the communication pathways to integrate temporally accumulated embodied experience and maintain stable alignment between perception and reasoning modules. Through these neuro-inspired, continuously evolving slots, our framework sustains coherent, bidirectional message exchange between EC-MLLM and PC-WM, enabling efficient fusion of perceptual cues and semantic understanding throughout task execution.

4 Experiments

In this section, we evaluate the capabilities of embodied execution, generalization, and evolution of the brain-inspired embodied evolutionary agent (BEEA) proposed in this paper. Firstly, training on basic embodied tasks such as navigation enhances the effectiveness of the foundation model. Subsequently, we conduct zero-shot generalization across diverse embodied tasks, validating the proposed paradigm's contribution to improving embodied execution and spatial intelligence capabilities. For *detailed experimental settings and comparative results*, please refer to the **Supplementary Materials**.

Table 1: Overall comparison with specialized models and the unified baseline model NaviLLM on in-domain tasks. We report GP for CVDN, SPL for SOON, R2R, and REVERIE, and report EM Accuracy for ScanQA. * indicates experimental results that we have reproduced.

Method	CVDN	SOON	R2R	REVERIE	ScanQA
HAMT [12]	5.13	-	61	30.20	-
DUET [13]	-	22.58	60	33.73	-
VLN-PETL [48]	5.69	-	60	27.67	-
BEV-BERT [2]	-	-	**64**	36.37	-
3D-LLM [31]	-	-	-	-	20.5
NaviLLM* [72]	5.75	26.19	54	31.01	22.93
+BEEA	**6.30**	**30.97**	60	**37.28**	**23.14**

4.1 Evaluation Datasets and Setups

In-domain Datasets. Following [30, 72], we utilize unseen validation sets corresponding to the five training datasets to evaluate the proposed embodied agent's ability to handle in-domain tasks (vision-and-language navigation and embodied question-answering).

- **CVDN** [55] tasks agents with target navigation using dialog history, requiring dialog understanding and action translation. It includes 7,415 instances across training and test sets.
- **SOON** [75] requires the agent to locate a target object based on a detailed goal description. The dataset includes 3,848 instruction sets and over 30,000 long-distance trajectories.
- **R2R** [3] offers step-by-step navigation instructions in photorealistic environments, comprising 10,800 panoramic views and 7,189 trajectories.
- **REVERIE** [47] includes 10,567 panoramic images and 21,702 high-level instructions, centered on localizing distant target objects within 90 buildings.
- **ScanQA** [5] challenges models to answer text questions about 3D scenes using RGB-D scan data, featuring 41k question–answer pairs across 800 indoor environments.

Out-of-domain Datasets. A proficient embodied agent should, during environmental exploration (e.g., in vision-and-language navigation), implicitly enhance its generalization capability and spatial awareness for unseen tasks through accumulated experience, much like how humans develop spatial and behavior cognition [16, 25, 35]. To validate that our proposed framework possesses this capability, we conduct zero-shot generalization tests across multiple out-of-domain embodied intelligence/spatial intelligence datasets.

- **EmbodiedBench** [67] is a comprehensive benchmark for evaluating vision-driven embodied agents based on MLLMs. It features 1,128 diverse tasks across four environments, spanning high-level semantic tasks (e.g., household planning) and low-level atomic actions (e.g., navigation, manipulation). The benchmark assesses six core capabilities, including embodied commonsense reasoning, spatial awareness, and long-horizon planning.
- **VSI-Bench** [66] is a video-based visual-spatial intelligence benchmark comprising 5,000+ question-answer pairs derived from 288 real-world indoor scene videos. Designed to evaluate multimodal large language models (MLLMs), it assesses spatial perception and reasoning through eight task categories, including object counting, distance estimation, and route planning, *etc.*

Table 2: Comparative results of parameter-efficient fine-tuning for multiple MMLMs on in-domain datasets.

Method	CVDN	SOON	R2R	REVERIE
InternVL2.5-8B	5.40	21.37	46	28.31
+BEEA	**5.67**	**23.59**	**51**	**31.93**
InternVL2.5-78B	5.74	27.87	59	35.42
+BEEA	**5.98**	**30.14**	**63**	**38.37**
Qwen2.5-VL-7B	5.21	23.22	50	26.96
+BEEA	**5.87**	**24.65**	**55**	**29.45**

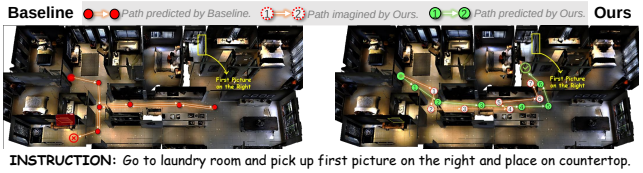

Figure 4: Trajectory visualization comparison with NaviLLM on the REVERIE dataset.

Evaluation Metrics. For navigation tasks in in-domain datasets, we employ Success weighted by Path Length (SPL) and Goal Progress (GP) as evaluation metrics. 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 in EmbodiedBench, and question-answering accuracy in VSI-Bench.

Baselines. We construct our embodied evolution agent using representative and popular multimodal large language models, including InternVL2.5-8B/78B [69] and Qwen2.5-VL-7B-Ins [6]. Additionally, since the original in-domain MLLM, NaviLLM [72] were trained based on the Vicuna language model [45], 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 [66, 67, 72].

Implementation Details. We conduct multi-task fine-tuning using in-domain datasets. Following [30, 72], we utilize the Adam optimizer with a learning rate of 3e-5 and train for 5000 steps. For hyperparameters, we set K to 10, 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 [66, 67, 72] and employ varying temperatures and greedy strategies. During testing, with respect to the sampling strategy for LLM-based action generation, we referred to [66, 67, 72] 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.

4.2 Experimental Results on In-domain Tasks

Comparison with Full Parameter-tuning Methods. Following NaviLLM [72], 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. 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

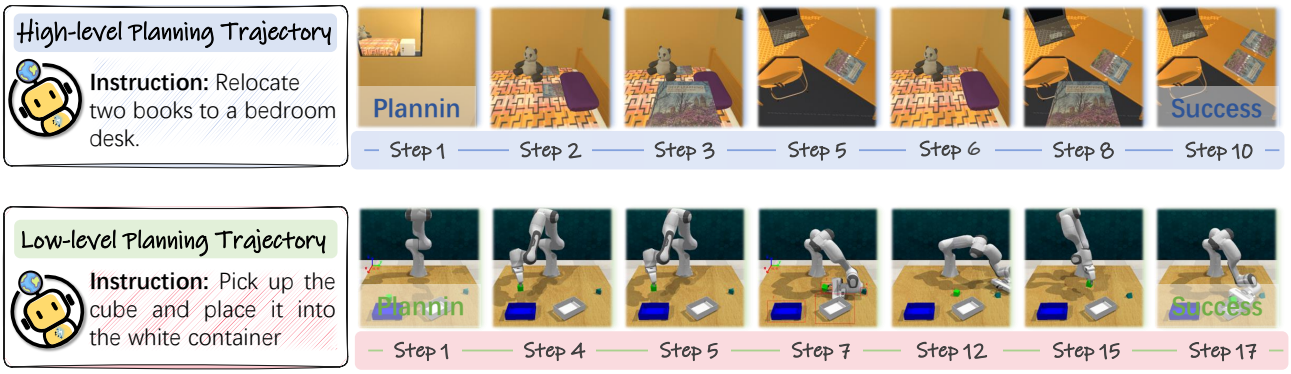

Figure 5: Embodied execution examples in EB-ALFRED and EB-Manipulation using our BEEA with InternVL2.5-78B.

Table 3: Task success rates on 3 subsets of EB-ALFRED, EB-Habitat,EB-Navigation, and EB-Manipulation of EmbodiedBench. GPT-4o and Claude-3.5 are SoTA proprietary MLLMs for reference. Superior results compared to baselines are shown in bold.

Model	EB-ALFRED			EB-Habitat			EB-Navigation			EB-Manipulation		
	Base	Complex	Long	Base	Complex	Long	Base	Complex	Long	Base	Complex	Spatial
GPT-4o	64	68	54	86	56	64	55.0	58.3	55.0	39.6	29.2	25.0
Claude-3.5-Sonnet	72	76	52	96	78	58	66.7	41.7	26.7	37.5	29.2	22.9
InternVL2.5-8B	**4**	2	0	36	0	2	35.0	21.7	0	**8.3**	6.3	10.4
+BEEA	**4**	**6**	**2**	**44**	**10**	**6**	**41.6**	**31.6**	**8.3**	**8.3**	**8.3**	**12.5**
Qwen2.5-VL-7B	10	6	2	32	26	2	28.3	**41.7**	8.3	8.3	**8.3**	16.7
+BEEA	**12**	**14**	**6**	**56**	**28**	**6**	**31.7**	38.3	**13.3**	**10.4**	**8.3**	**18.8**
InternVL2.5-78B	**38**	42	**42**	80	56	28	36.7	33.3	23.3	16.7	14.6	20.8
+BEEA	**38**	**52**	**42**	**82**	**58**	**32**	**43.3**	**35.0**	**26.7**	**20.8**	**18.8**	**35.4**

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).

Conducting Parameter-Efficient Fine-Tuning. In addition to conducting full-parameter fine-tuning, we place particular emphasis on exploring parameter-efficient fine-tuning, with the following considerations: (1) preserving the original capabilities of MLLMs (by freezing parameters) can maintain their ability to handle out-of-domain reasoning and generalization tasks; (2) achieving efficiency in fine-tuning; and (3) facilitating a clearer comparison with the original MLLMs to validate our effectiveness on out-of-domain tasks. Consequently, we fine-tune only the dynamic communication slots and the additional parameters introduced by MLLM/WM. To better evaluate out-of-domain generalization capabilities in subsequent experiments, we fine-tune the model using only the basic navigation tasks, without incorporating the ScanQA dataset. As demonstrated in Table 2, experimental results across three models of varying sizes and architectures—InternVL2.5-8B/78B and Qwen2.5-VL-7B-Ins—consistently confirm the effectiveness of the proposed approach. Furthermore, we observe that: (1) compared to full-parameter fine-tuning, fine-tuning a limited number of parameters constrains the model's performance on in-domain tasks; (2) employing MLLMs with greater capacity significantly enhances embodied exploration capabilities.

Does our proposed embodied evolutionary agent genuinely enhance the planning capabilities of MLLMs? To investigate whether the introduction of the WM augments the MLLM's performance, we visualize the agent's execution process using the NaviLLM baseline as a reference. Given that our world model encodes high-level features rather than raw images, direct visualization of latent states remains challenging. Consequently, we adopt a trajectory-based evaluation approach—implicitly reflecting the model's dynamic planning capabilities through predicted and executed navigation sequences. Figure 4 presents the trajectory visualization results, where the consistency between the predicted paths of our method and the actual executed paths substantiates the model's ability to model environmental dynamics and perform forward planning. In contrast, the baseline method, lacking such capabilities, resulted in navigation failures.

4.3 Experimental Results on EmbodiedBench

Comparison Results. 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 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, highlighting the current limitations of MLLMs

Table 4: Evaluation on VSI-Bench. † indicates 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
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**

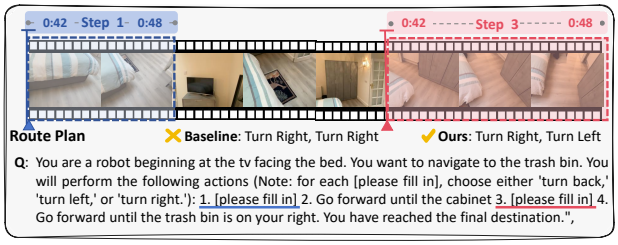

Figure 6: Qualitative comparison on the VSI-Bench dataset using the InternVL2.5-78B model.

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.

Qualitative Analysis. Figure 5 illustrates the successful execution of embodied tasks by our proposed BEEA agent, showcasing its capability in both high-level planning and low-level manipulation. In the high-level planning trajectory, the agent demonstrates its ability to perform complex household tasks, such as relocating objects. In the low-level planning trajectory, the agent showcases its fine-grained manipulation skills in a controlled setting.

4.4 Experimental Results on VSI-Bench

Comparison Results. 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 intelligence of MLLMs. As shown in Table 4, the proposed BEEA method significantly improves the performance of various baseline MLLMs. showing the strength of joint using MLLM and WM in capturing spatial relationships and global topological structures. Future research will focus on enhancing global 3D feature representation modeling and incorporating external knowledge.

Qualitative Analysis. The visualization of question-answering results for spatial understanding on the VSI-Bench, as shown in Figure 6, provides compelling qualitative evidence of the superiority of our proposed method over baseline approaches. For more examples and discussions, please refer to **Supplementary Materials**.

Table 5: Ablation study of the proposed BEEA.

Method	CVDN	SOON	R2R	REVERIE	ScanQA
w/o MLLM	3.56	16.24	41	27.65	13.1
w/o WM	5.71	29.80	58	34.55	22.93
w/o DCS	5.84	28.42	57	32.79	22.98
w/o Evolution	6.05	28.88	58	34.32	22.57
Ours	**6.30**	**30.97**	**60**	**37.28**	**23.14**

4.5 Ablation Study

To validate the effectiveness of each module in the proposed framework, we conduct ablation experiments on the in-domain datasets.

MLLM is of paramount importance. Robust language perception and logical reasoning capabilities serve as the cornerstone of embodied intelligence. It can be readily inferred that the absence of a pretrained MLLM (primarily an LLM) would lead to a significant decline in the embodied execution capabilities of the agent. As in Table 5, the results of baseline variant *w/o MLLM* substantiate this view: when the LLM and visual encoding modules are randomly initialized, the agent fails to effectively perform embodied tasks.

WM plays a critical role in embodied agents. When the world model is excluded from modeling the dynamic changes in the environment, the framework degrades to relying solely on the MLLM augmented with a set of learnable parameters, and the dynamic communication slots are reduced to a mechanism for enhancing internal feature communication within the multimodal model. A comparison between *w/o WM* and *Ours* reveals that the world model significantly contributes to improving the agent's performance.

Is a simple combination of MLLM and WM sufficient? In this paper, we draw an analogy to the function of the corpus callosum in the human brain and design dynamic communication slots to connect the MLLM and WM. To assess the efficacy of this collaborative mechanism, we remove the dynamic communication slots, directly passing the perceptual features from the MLLM to the WM. Consequently, this variant model loses its capacity for efficient dynamic evolution. The results of *w/o DCS* demonstrate that the corpus callosum-inspired design is effective in unlocking the collaborative potential between the WM and MLLM.

Does the agent's capability truly evolve? Owing to the efficient design of the DCS, our method enable unsupervised evolution during online testing through direct slot feature updates. To verify this deign, we disable online updates while the agent perform embodied tasks. The comparison between the last two rows of Table 5 highlights the effectiveness of these updates, providing motivation for further development agent evolution algorithms.

5 Conclusion

This paper proposes an innovative brain-inspired framework that integrates MLLMs with world models, emulating the functional specialization of the human brain's left and right hemispheres. By incorporating dynamic communication slots, the framework achieves efficient information exchange and online evolution, significantly enhancing the performance of embodied agents in dynamic environments and zero-shot tasks. In future work, to further enhance the model's plasticity, strategies such as modular self-growing learning with assemblable components could be introduced.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2023YFC3310700, in part by the National Natural Science Foundation of China under Grants 62036012, 62236008, U21B2044, 62472422, and U2333215, and in part by Beijing Natural Science Foundation under Grant 4242051.

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] 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).

[3] 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.

[4] 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*.

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

[6] 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).

[7] 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).

[8] 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).

[9] 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).

[10] 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).

[11] 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*.

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

[13] 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*.

[14] 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).

[15] 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*.

[16] 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.

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

[18] Junyu Gao, Mengyuan Chen, and Changsheng Xu. 2023. Vectorized evidential learning for weakly-supervised temporal action localization. *IEEE transactions on pattern analysis and machine intelligence* 45, 12 (2023), 15949–15963.

[19] Junyu Gao, Mengyuan Chen, and Changsheng Xu. 2025. Learning Probabilistic Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

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

[21] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2019. Graph convolutional tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4649–4659.

[22] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2021. Learning to Model Relationships for Zero-Shot Video Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2021), 3476–3491.

[23] 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).

[24] Onur Güntürkün, Felix Ströckens, and Sebastian Ocklenburg. 2020. Brain lateralization: a comparative perspective. *Physiological reviews* (2020).

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

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

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

[28] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR*.

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

[30] 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).

[31] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. *NeurIPS* (2023).

[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] Vyacheslav R Karolis, Maurizio Corbetta, and Michel Thiebaut de Schotten. 2019. The architecture of functional lateralisation and its relationship to callosal connectivity in the human brain. *Nature communications* 10, 1 (2019), 1417.

[37] 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*.

[38] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

[39] 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).

[40] 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).

[41] 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).

[42] 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).

[43] 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.

[44] 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*.

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

[46] Marta Pietrasanta, Laura Restani, and Matteo Caleo. 2012. The corpus callosum and the visual cortex: plasticity is a game for two. *Neural plasticity* 2012, 1 (2012), 838672.

[47] 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.

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

[49] 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).

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

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

[52] Gautam Singh, Skand Peri, Junghyun Kim, Hyunseok Kim, and Sungjin Ahn. 2021. Structured world belief for reinforcement learning in pomdp. In *ICML*.

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

[54] 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).

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

[56] 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).

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

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

[59] 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).

[60] 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).

[61] 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*.

[62] 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).

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

[64] 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).

[65] 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*.

[66] 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).

[67] 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).

[68] 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).

[69] 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).

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

[71] 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*.

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

[73] 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*.

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

[75] 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.