

# Embodied Arena: A Comprehensive, Unified, and Evolving Evaluation Platform for Embodied AI

<sup>α</sup>Tianjin University, <sup>β</sup>Huawei Noah's Ark Lab, <sup>γ</sup>Shanghai Jiao Tong University,  
<sup>δ</sup>Hong Kong University of Science and Technology (Guangzhou), <sup>λ</sup>Sun Yat-sen University,  
<sup>ρ</sup>PengCheng Laboratory, <sup>ζ</sup>Tongji University, <sup>ε</sup>University College London,  
<sup>μ</sup>Peking University, <sup>ν</sup>Tsinghua University, <sup>σ</sup>Imperial College London, <sup>ξ</sup>King's College London,  
<sup>κ</sup>Institute of Computing Technology, Chinese Academy of Sciences, <sup>η</sup>University of Manchester,  
<sup>θ</sup>Nanjing University, <sup>π</sup>TU Darmstadt

*Full author list in Contributions*

Embodied AI has shown great promise in empowering AI models to perceive, interact with, and ultimately change the physical world. Parallel to the development of large foundation models, Embodied AI is largely falling behind. Located at the center of Embodied AI, three essential challenges emerge and become even more stringent: (1) systematic understanding of the core capabilities needed for Embodied AI is missing in the community, making research lack of clear objectives; (2) despite the proposal of various benchmarks for Embodied AI, there is no unified and standardized evaluation system, leaving the cross-benchmark evaluation and comparison infeasible; (3) different from large language models (LLMs) powered by numerous web-scale data, automated and scalable acquisition methods for embodied data have not been well developed, which poses a critical bottleneck on the scaling of evaluation and training of Embodied AI models. To break the three obstacles, this paper presents **Embodied Arena**, a comprehensive, unified, and evolving evaluation platform and leaderboards for Embodied AI. First, Embodied Arena is established upon a **systematic embodied capability taxonomy** spanning three levels (i.e., perception, reasoning, task execution), seven core embodied capabilities, and 25 fine-grained dimensions. This taxonomy is proposed by absorbing and refining the partial categories in prior works, which allows for unified evaluation and offers systematic objectives for frontier research. Second, Embodied Arena closes the critical gap in standardized evaluation by introducing a **unified embodied evaluation system**. The system is built upon a unified evaluation infrastructure supporting flexible integration of advanced benchmarks and models, which has covered 22 diverse benchmarks across three domains (2D/3D Embodied Q&A, Navigation, and Task Planning) and 30+ advanced models from 20+ worldwide institutes. Third, Embodied Arena is powered by a **novel LLM-driven automated generation pipeline** that ensures the scalability of embodied evaluation data and allows it to keep evolving for diversity and comprehensiveness. Building upon the three major components, Embodied Arena addresses the three essential challenges correspondingly. Moreover, Embodied Arena provides professional support for more advanced models and embodied benchmarks to join, along with frequent maintenance and updates. Through comprehensive evaluation of the growing model population based on evolving evaluation data, Embodied Arena publishes three types of leaderboards (i.e., Embodied Q&A, Embodied Navigation, Embodied Task Planning) with two orthogonal views (i.e., the benchmark view and the capability view), offering a real-time overview of the embodied capabilities of advanced models. Especially, we present **nine findings** summarized from the evaluation results on the leaderboards of Embodied Arena. This helps to establish clear research veins and pinpoint critical research problems, thereby driving forward progress in the field of Embodied AI.

 Website: <https://embodied-arena.com>

## 1. Introduction

On the road towards Artificial General Intelligence (AGI), Embodied AI or Embodied Intelligence, has emerged to be one of the most important research fields in recent years. Complementary to the general understanding, reasoning, tool-use and problem-solving abilities of large foundation models (OpenAI, 2022, Jaech et al., 2024, Guo et al., 2025, Kimi et al., 2025, Yang et al., 2025a), Embodied AI has shown the promise in building various physical agents that are capable of perceiving, interacting with, and ultimately changing the real world (Brohan et al., 2023a,b, Driess et al., 2023, O’Neill et al., 2023, Zhao et al., 2023, Ghosh et al., 2024, Kim et al., 2024, Black et al., 2024, 2025). Notably, OpenVLA (Kim et al., 2024) scaled vision-language-action models to enable generalist robotic manipulation across diverse tasks and  $\pi_0$  (Black et al., 2024) introduced efficient hierarchical planning that bridges high-level reasoning with low-level control, demonstrating the potential of multimodal foundation models in embodied scenarios.

Although notable results have been achieved by the works above, there is still a huge gap between the capabilities of existing embodied agents and complex, diverse real-world application scenarios, preventing the wide-range deployment of Embodied AI techniques. Concurrent with the development of large foundation models, Embodied AI is largely falling behind. Specifically, there are **three critical challenges** that severely limit the advancement of Embodied AI research. First, *what are the core capabilities that a desired Embodied AI model needs?* The answer to this essential question remains unclear. Most ongoing works attempt to push the frontier from concrete aspects such as embodied visual perception, embodied task planning, etc. The lack of anchoring from a systematic view makes it hard to better connect with related works and find an important research purpose. Second, despite the proposal of various benchmarks for Embodied AI, each benchmarks differ a lot in aspects like data formats, evaluation metrics, target embodied capabilities, etc. This makes direct cross-benchmark evaluation and comparison infeasible. Hence, a unified and standardized evaluation system is urgently needed. Third, as the success of large language models (LLMs) stems from scalable training from numerous web-scale data, sufficient and diverse embodied data is crucial to thorough evaluation and training of Embodied AI models. However, most existing embodied data relies heavily on manual scenario construction, task design, and data collection, making scalability impossible. Unfortunately, a scalable, automated acquisition method for embodied data is missing in the field of Embodied AI, which poses a bottleneck on the scaling of evaluation and training of Embodied AI models.

To break the three obstacles and pave the way for the advancement of Embodied AI research, this paper presents **Embodied Arena**, the first comprehensive evaluation platform and leaderboards for Embodied AI. First of all, we propose a **Systematic Embodied Capability Taxonomy** spanning three incremental levels (i.e., perception, reasoning, task execution), seven core embodied capabilities, and 25 fine-grained dimensions. This taxonomy is established by absorbing and refining the partial categories in prior works, which allows for unified evaluation while offering systematic targets for frontier research. Based on the systematic embodied capability taxonomy, Embodied Arena then closes the critical gap in standardized evaluation by introducing a **Unified Embodied Evaluation System**. The system is built upon a unified evaluation infrastructure supporting flexible integration of advanced benchmarks and models, which has covered 22 diverse benchmarks across three domains (2D/3D Embodied Q&A, Navigation, and Task Planning) and 30+ advanced models from 20+ worldwide institutes. Embodied Arena also provides professional support for more advanced models and embodied benchmarks to join, along with frequent maintenance and updates. Moreover, Embodied Arena is powered by a novel **LLM-driven Automated Data Generation Approach** for Embodied AI. By leveraging the general knowledge in LLMs, this approach automates the whole process of scenario construction, task design, and data collection. This automated generation pipeline ensures the scalability of embodied evaluation data and allows it to keep evolving for diversity and comprehensiveness. Building upon these major components, Embodied Arena addresses the three essential challenges correspondingly.

Through comprehensive evaluation among the growing model population based on evolving evaluation data, Embodied Arena publishes three types of leaderboards with two orthogonal views, i.e., *benchmark view* and *capability view*. The benchmark view presents the ranking of models on each benchmark, which is convenient for academic researchers to quote and compare with in their research works; while the capability view instead presents the ranking of models against each embodied capability in the systematic taxonomy, providing an up-to-date overview of the embodied capabilities of advanced models. Through carefully summarizing the comprehensive evaluation results on the leaderboards, especially, we present **nine findings** from a range of important perspectives for useful insights, including the comparison between general multimodal models and embodied models, the limitations of existing benchmarks, the relationship among different embodied capabilities, the scaling law of embodied AI, etc. The ultimate aim of Embodied Arena is to facilitate the establishment of clear research veins and help to identify critical research problems, thereby propelling research progress in the field of Embodied AI.

In the following, we first presents an overview of Embodied Arena and highlight the key features in Section 2. Then we introduce the systematic embodied capability taxonomy as well as a mapping from existing benchmarks to our taxonomy in Section 3. Thereafter, we detail our unified embodied evaluation system in Section 4, followed by the LLM-driven automated generation pipeline for embodied evaluation data in Section 5. Moreover, we summarize nine major findings from our comprehensive evaluation results to provide useful insights that illuminate the current state and future directions of Embodied AI research in Section 6. Finally, the conclusion is summarized in Section 7 and all authors are listed in Section 8.

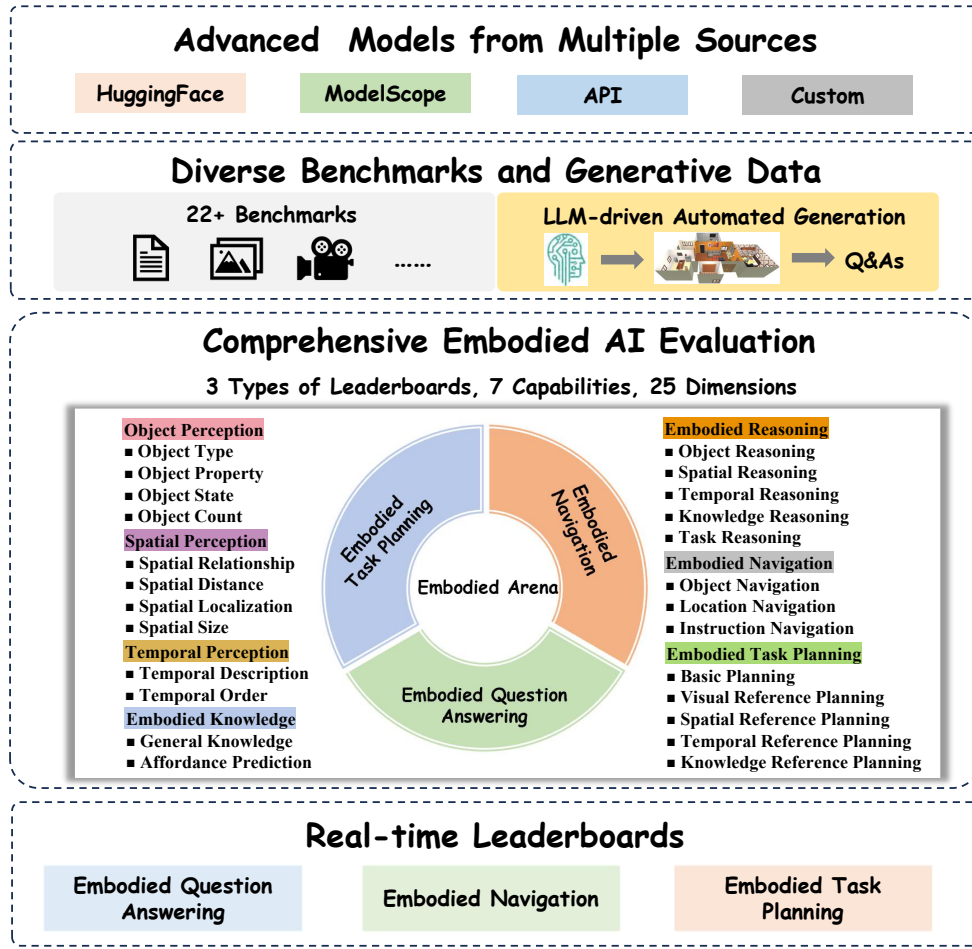
## 2. Overview of Embodied Arena

Embodied Arena is a comprehensive, unified, and evolving evaluation platform and leaderboards for Embodied AI. It features three types of core embodied tasks, a diverse range of high-quality benchmarks, an LLM-driven automated evaluation data generation approach, and a systematic embodied capability taxonomy. A conceptual overview of Embodied Arena is shown in Figure 1.

Embodied Arena evaluates both general large models and Embodied AI models, including leading commercial models and advanced academic models. Embodied Arena is also eagerly calling for more open-source, closed-source models from multiple sources to join, with professional and user-friendly support by different means. The evaluation data of Embodied Arena consists of (1) a diverse range of existing embodied benchmarks, which are carefully integrated and aligned by us, and (2) generative data powered by our LLM-driven automated generation pipeline. Similarly, we also provide support for more benchmarks to join. Hence, Embodied Arena keeps evolving the embodied evaluation data by integrating more benchmarks and generating new data. With the evolving evaluation data, Embodied Arena conducts a comprehensive evaluation for each model based on the unified embodied evaluation system. The evaluation results span three types of embodied tasks (i.e., Embodied Q&A, Embodied Navigation, Embodied Task Planning), against the systematic embodied capability taxonomy (includes seven embodied capabilities with 25 fine-grained dimensions). Finally, three types of leaderboards are summarized and presented for convenient and useful reference to both academia and industry. Embodied Arena reacts in real-time to requests for evaluation and participation and updates the leaderboards and the evaluation system regularly.

Embodied Arena is designed with six core features that distinguish it as the comprehensive evaluation platform for Embodied AI models. These features address the fundamental challenges in Embodied AI evaluation while providing comprehensive support for the research community. We highlight the six key features below:

- **Comprehensive Embodied Capability Taxonomy:** Embodied Arena introduces a systematic cat-



**Figure 1: A conceptual overview of Embodied Arena.** Embodied Arena provides a comprehensive evaluation for advanced models from multiple sources, based on diverse embodied benchmarks and LLM-driven generative data. The evaluation results span three types of embodied tasks (i.e., Embodied Q&A, Embodied Navigation, Embodied Task Planning), against seven embodied capabilities with 25 fine-grained dimensions. Three types of leaderboards are summarized and presented for convenient and useful reference to both academia and industry.

egorization spanning 7 core embodied capabilities decomposed into 25+ fine-grained dimensions, carefully refined from diverse embodied tasks and benchmarks to enable researchers to identify specific capability gaps and track progress across different aspects of Embodied AI.

- **Rich Model Support:** Our platform supports 30+ advanced models from 20+ leading research institutes worldwide, including general multimodal LLMs, specialized embodied models, and both open-source and commercial models through various access methods including API-based evaluation, parameter-based integration, and custom interfaces.
- **Modular Benchmark Integration:** Embodied Arena integrates 22+ evaluation benchmarks across three core domains with flexible extensibility through modular design that enables easy onboarding while maintaining consistent evaluation protocols as the platform evolves with field advancement.
- **Unified Evaluation Infrastructure:** The platform provides a standardized evaluation framework with uniform input/output formats, professional experiment management, and real-time leaderboard systems for transparent result presentation while ensuring consistent protocols and monthly updates.
- **High-quality Evaluation Datasets:** Embodied Arena maintains curated datasets continuously evolved through our LLM-driven automated generation pipeline, ensuring the scalability and diversity of

embodied evaluation data while breaking manual construction bottlenecks.

- **Diverse Evaluation Methodologies:** Our platform employs complementary evaluation paradigms including accuracy-based QA assessment and interactive simulation-based testing, providing thorough assessment with flexibility across different benchmark characteristics for comprehensive embodied capability evaluation.

### 3. Systematic Embodied Capability Taxonomy

Embodied Arena uses a systematic taxonomy of capabilities potentially required for Embodied AI, drawing from cognitive psychology, human experience, and diverse existing tasks and benchmarks in the field (Cheng et al., 2024c, Yang et al., 2024). Specifically, Embodied Arena considers seven core capabilities from low to high level: *Object Perception*, *Spatial Perception*, *Temporal Perception*, *Embodied Knowledge*, *Embodied Reasoning*, *Embodied Navigation*, and *Embodied Task Planning*. More specifically, *Object Perception*, *Spatial Perception*, *Temporal Perception*, and *Embodied Knowledge* are viewed as the *fundamental* embodied capabilities, responsible for multifaceted perception and low-level understanding. *Embodied Reasoning* is treated as an *advanced* embodied capability, as it is built upon the fundamental capabilities for understanding and conceiving solutions to complex questions and tasks. It then moves on to *downstream task-related* embodied capabilities, i.e., *Embodied Navigation* and *Embodied Task Planning*, at the high level of the taxonomy hierarchy. Each core capability consists of multiple fine-grained capability dimensions. Figure 2 illustrates the typical task instance for each capability dimension.

**Object Perception** Recognizing objects via visual inputs is a fundamental capability for embodied models. Here we further divide the object perception into four fine-grained dimensions: *Object Type*, recognizing specific categories of objects; *Object Property*, determining physical properties of objects, e.g., color, shape, material, size, etc; *Object State*, judging states of objects, e.g., open, closed, stationary, etc; *Object Count*, recognizing the number of objects.

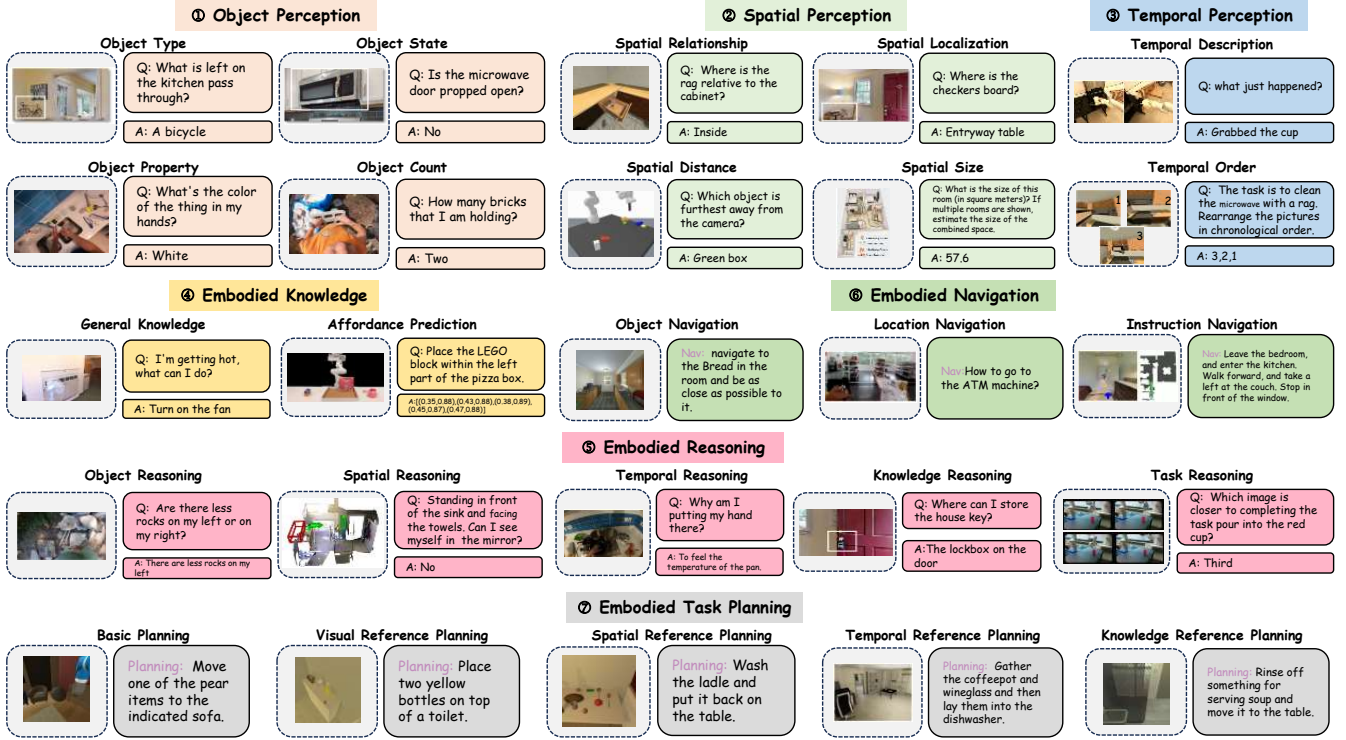
**Spatial Perception** Spatial perception capability is a vital core capability of embodied models. Accurate spatial perception is crucial for embodied agents to successfully perform tasks. Specifically, we further divide the spatial perception into four fine-grained dimensions: *Spatial Relationship*, judging relative relationships (e.g., next to) of objects; *Spatial Distance*, judging relative or absolute distances; *Spatial Localization*, detecting the positions of objects; *Spatial Size*, estimating the size of spaces, e.g., the size of rooms, etc.

**Temporal Perception** Different from perceiving static semantics (such as object types and spatial relationships), temporal perception focuses on semantic content that changes over time. Here we investigate the temporal perception capability from two aspects, temporal description and temporal order. *Temporal Description* recognizes the visual input contents related to the temporal dimension. *Temporal order* judges the timestamp and sequential order of events based on visual inputs.

**Embodied Knowledge** Embodied knowledge refers to the basic cognitive capability of embodied models for the real world. Here we mainly focus on general knowledge and affordance prediction. *General knowledge* requires the embodied models to make judgments on general knowledge based on visual inputs. For example, refrigerators can keep food fresh. *Affordance prediction* requires embodied models to infer possible object manipulations from visual inputs.

**Embodied Reasoning** Reasoning capability plays a crucial role in the decision-making process of complex embodied tasks. Based on basic perception and cognitive capabilities, we further divide the embodied reasoning into five fine-grained dimensions: *Object Reasoning*, reasoning feasible actions on objects and comparing object properties based on object perception results; *Spatial Reasoning*, reasoning about object





**Figure 2: The Systematic Embodied Capability Taxonomy and Exemplary Descriptions.** Embodied Arena encapsulates seven core capabilities: *object perception*, *spatial perception*, *temporal perception*, *embodied knowledge*, *embodied reasoning*, *embodied navigation*, and *embodied task planning*, which contain 25 fine-grained capability dimensions in total.

accessibility, spatial inclusiveness, and spatial imagination, among other aspects, based on spatial perception results; *Temporal Reasoning*, reasoning the causes and consequences of events based on temporal perception results; *Knowledge Reasoning*, reasoning physical dynamics based on prior knowledge and visual inputs; *Task Reasoning*, reasoning the type and location of task-related objects, task progress, among other aspects, based on visual inputs and task instructions.

**Embodied Navigation** Navigation is a core embodied task in Embodied AI. Here we investigate the embodied navigation capability from three aspects: object navigation, location navigation, and instruction navigation. *Object Navigation* refers to the capability to navigate to a goal object from a start position. *Location Navigation* refers to the capability to navigate to a goal location from a start position. *Instruction Navigation* refers to the capability to follow a specified navigation instruction from a start position.

**Embodied Task Planning** Embodied Task Planning refers to decomposing the complex task into a reasonable sequence of sub-steps based on task instructions and visual inputs. Here we investigate the embodied planning capability from five aspects: basic planning, visual reference planning, spatial reference planning, temporal reference planning, and knowledge reference planning. *Basic Planning* refers to the capability to decompose tasks where the instruction specifies object types. *Visual Reference Planning* refers to the capability to decompose tasks where the instruction refers to objects using object properties, states, etc. *Spatial Reference Planning* refers to the capability to decompose tasks with spatial constraints. *Temporal Reference Planning* refers to the capability to decompose tasks with temporal constraints. *Knowledge Reference Planning* refers to the capability to decompose tasks where the instruction refers to objects using object-related knowledge.

Embodied Tasks	Embodied Benchmarks	Embodied Capabilities
Embodied Q&A	RoBoVQA (Sermanet et al., 2024)	Temporal Perception: Temporal Description Embodied Knowledge: Affordance Prediction Embodied Reasoning: Task Reasoning Embodied Task Planning: Basic Planning
Embodied Q&A	VSI-Bench (Yang et al., 2024)	Object Perception: Object Property, Object Count Spatial Perception: Spatial Relationship, Spatial Distance, Spatial Size Temporal Perception: Temporal Order Embodied Task Planning: Basic Planning
Embodied Q&A	OpenEQA (Majumdar et al., 2024)	Object Perception: Object Type, Object Property, Object State Spatial Perception: Spatial Localization Embodied Knowledge: General Knowledge Embodied Reasoning: Spatial Reasoning, Knowledge Reasoning
Embodied Q&A	Where2Place (Yuan et al., 2024b)	Embodied Knowledge: Affordance Prediction
Embodied Q&A	ERQA (Abeyruwan et al., 2025)	Object Perception: Object Type, Object State Spatial Perception: Spatial Relationship Embodied Reasoning: Spatial Reasoning, Temporal Reasoning, Task Reasoning
Embodied Q&A	UniEQA (Zhang et al., 2025b)	Object Perception: Object Type, Object Property, Object State Spatial Perception: Spatial Relationship Temporal Perception: Temporal Description, Temporal Order, Embodied Knowledge: General Knowledge, Affordance Prediction Embodied Reasoning: Object Reasoning, Task Reasoning Embodied Navigation: Location Navigation
Embodied Q&A	VABench-Point (Yuan et al., 2025a)	Embodied Knowledge: Affordance Prediction
Embodied Q&A	PhyBlock (Ma et al., 2025)	Object Perception: Object Type, Object Property, Object Count Spatial Perception: Spatial Relationship, Spatial Distance Temporal Perception: Temporal Order Embodied Knowledge: Affordance Prediction Embodied Reasoning: Spatial Reasoning, Knowledge Reasoning, Task Reasoning Embodied Task Planning: Spatial Reference Planning
Embodied Q&A	MineAnyBuild (Wei et al., 2025b)	Embodied Reasoning: Spatial Reasoning, Knowledge Reasoning
Embodied Q&A	ScanRefer (Chen et al., 2020)	Spatial Perception: Spatial Localization
Embodied Q&A	Scan2Cap (Chen et al., 2021)	Spatial Perception: Spatial Relationship
Embodied Q&A	ScanQA (Azuma et al., 2022)	Spatial Perception: Spatial Localization
Embodied Q&A	SQA3D (Ma et al., 2023)	Embodied Reasoning: Spatial Reasoning
Embodied Q&A	Multi3DRefer (Zhang et al., 2023)	Spatial Perception: Spatial Localization
Embodied Navigation	MP3D (Chang et al., 2017)	Embodied Navigation: Object Navigation
Embodied Navigation	HM3D (Ramakrishnan et al., 2021)	Embodied Navigation: Object Navigation
Embodied Navigation	EB-Navigation (Yang et al., 2025c)	Embodied Navigation: Object Navigation
Embodied Navigation	R2R-CE (Yang et al., 2025c)	Embodied Navigation: Instruction Navigation
Embodied Navigation	RxR-CE (Yang et al., 2025c, Zhang et al., 2024b)	Embodied Navigation: Instruction Navigation
Embodied Task Planning	ET-Plan-Bench (Zhang et al., 2024c)	Embodied Task Planning: Spatial Reference Planning, Temporal Reference Planning
Embodied Task Planning	EB-ALFRED (Yang et al., 2025c)	Embodied Task Planning: Basic Planning, Visual Reference Planning, Spatial Reference Planning, Knowledge Reference Planning
Embodied Task Planning	EB-Habitat (Yang et al., 2025c)	Embodied Task Planning: Basic Planning, Visual Reference Planning, Spatial Reference Planning, Knowledge Reference Planning

**Table 1: An overview of the mapping from existing embodied benchmarks to the three types of embodied tasks and the systematic embodied capability taxonomy in Embodied Arena.**

Based on our systematic taxonomy of embodied capabilities, we present a mapping in Table 1, from existing embodied benchmarks to the three types of embodied tasks and the systematic embodied capability taxonomy in Embodied Arena. We can observe that all the benchmarks focus on a single type of embodied tasks, and cover different parts of the capability dimensions in our systematic taxonomy. This also indicates the unique value of Embodied Arena in providing a comprehensive evaluation for embodied models against complete embodied capability dimensions.

## 4. Unified Embodied Evaluation System

In this section, we introduce the unified embodied evaluation system in Embodied Arena. This unified evaluation system aims to align the differences among existing embodied benchmarks and provide a standardized evaluation pipeline, thus closing the critical gap in cross-benchmark evaluation and comparison. We detail the components of the system one by one in a logical order in the following.

#### 4.1. Tasks, Benchmarks, and Data

In order to provide an in-depth and comprehensive evaluation of Embodied AI models, Embodied Arena currently covers three core types of embodied tasks: Embodied Question Answering, Embodied Navigation, and Embodied Task Planning. For each task, we carefully select high-quality evaluation benchmarks, which generally have broad academic influence and cover comprehensive and complementary capability dimensions.

Specifically, for Embodied Question Answering, we consider two types of benchmarks: 2D question answering and 3D question answering. The 2D question answering benchmarks include OpenEQA (Majumdar et al., 2024), VSI-Bench (Yang et al., 2024), and ERQA (Abeyruwan et al., 2025), etc., and the 3D question answering benchmarks include the representative ScanQA (Azuma et al., 2022), Scan2Cap (Chen et al., 2021), and SQA3D (Ma et al., 2023), etc. For Embodied Navigation, we select the classic object navigation benchmarks MP3D (Chang et al., 2017), HM3D (Ramakrishnan et al., 2021), EB-Navigation (Yang et al., 2025c), and the instruction navigation benchmarks R2R-CE (Krantz et al., 2020) and RxR-CE (Krantz et al., 2020, Zhang et al., 2024b). For Embodied Task Planning, we consider EB-ALFRED, EB-Habitat (Yang et al., 2025c), and ET-Plan-Bench (Zhang et al., 2024c), which are more diverse in task types.

These benchmarks contain a total of more than 64k task instances. Among them, there are more than 48k embodied question-answer pairs, which are designed to comprehensively evaluate the model’s performance in multiple embodied core capabilities. For embodied navigation tasks, Embodied Arena has accumulated more than 7k tasks, covering diverse navigation challenges of varying difficulty, aiming to comprehensively evaluate the embodied navigation capabilities of the model. In terms of embodied task planning, Embodied Arena provides more than 8k carefully designed tasks to examine the model’s capability in the decomposition and execution of complex embodied tasks. In addition, we provide a **Embodied Wiki** in the Embodied Arena platform, for convenient look-up and reference of the details of each benchmark.

The current platform primarily focuses on perception, spatial reasoning, and high-level navigation and planning capabilities. While manipulation-related reasoning is included through QA and task planning Leaderboards, direct simulation-based manipulation tasks represent an important direction for future platform development. As the field evolves toward more sophisticated embodied agents, future extensions of the platform will incorporate more comprehensive manipulation tasks and closed-loop evaluation capabilities spanning the full perception-decision-action cycle.

Although Embodied Arena collects and integrates existing representative embodied benchmarks as mentioned above, the evaluation data in these benchmarks are static and finite. To this end, Embodied Arena features a novel LLM-driven automated generation framework of embodied evaluation data. We defer the detailed introduction of it to Section 5.

#### 4.2. Models

Embodied Arena evaluates a comprehensive spectrum of AI models, ranging from general-purpose multimodal large language models to specialized Embodied AI models. Our platform encompasses both influential commercial models from leading technology companies and cutting-edge research models that represent the latest advances in Embodied AI. This diverse model ecosystem enables comprehensive cross-model comparison and provides valuable insights into the current landscape of embodied capabilities.



#### 4.2.1. General Multimodal Large Models

General multimodal large language models represent foundation models with robust vision-language understanding capabilities, making them particularly well-suited for embodied question answering and high-level reasoning tasks. These models demonstrate exceptional language comprehension, sophisticated reasoning abilities, and excellent vision-language integration, delivering robust performance across diverse embodied scenarios.

- **OpenAI:** GPT-4o (OpenAI, 2024), GPT-4o mini, o3, o4-mini (OpenAI, 2025)
- **Google DeepMind:** Gemini-1.5-Pro, Gemini-1.5-flash (Team et al., 2024), Gemini-2.5-Pro, Gemini-2.5-flash (Google DeepMind, 2024)
- **Anthropic:** Claude-3.5-Sonnet, Claude-3.7-Sonnet (Anthropic, 2024)
- **Alibaba Group:** Qwen-VL-Max, Qwen2-VL-7B-Instruct, Qwen2-VL-7B, Qwen2-VL-72B-Instruct (Wang et al., 2024), Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-7B, Qwen2.5-VL-72B-Instruct (Bai et al., 2025), mPLUG-Owl3 (Ye et al., 2024)
- **ByteDance:** LongVA-7B (Zhang et al., 2024d), LLaVA-OneVision (Li et al., 2024a), LLaVA-NeXT-Video (Zhang et al., 2025c), pLLaVA-7b (Xu et al., 2024a)
- **Meta AI:** Llama-3.2-11B-Vision-Instruct, Llama-3.2-90B-Vision-Instruct
- **Shanghai AI Lab:** InternVL3 (Zhu et al., 2025), InterVL2.5 (Chen et al., 2024c), InternVL2,
- **NVIDIA:** VILA-1.5 (Lin et al., 2024)
- **Microsoft:** Phi-3-vision-128k-instruct (Abdin et al., 2024)
- **ModelBest:** MiniCPM-V, MiniCPM-V 2.6 (Yao et al., 2024)

These models excel in their strong linguistic understanding and reasoning capabilities, sophisticated vision-language integration, and particular suitability for complex question answering and high-level task reasoning scenarios. Their broad knowledge base and general-purpose design make them effective across multiple Embodied AI domains. However, while these general models provide strong fundamental capabilities, they often lack the specialized design and domain-specific optimizations required for complex embodied AI tasks, motivating the development of more targeted embodied AI models.

#### 4.2.2. Embodied AI Models

Embodied AI models are specifically designed and optimized for embodied intelligence tasks, featuring enhanced spatial understanding, navigation capabilities, and physical interaction reasoning. Unlike general-purpose multimodal models, these models are comprehensively tailored for embodied scenarios through multiple dimensions: architecturally, they incorporate specialized components for spatial-temporal perception, affordance recognition, and action-oriented reasoning; in terms of training data, they leverage embodied-specific datasets including robotic trajectories, 3D scene interactions, and physical manipulation sequences; regarding training paradigms, they often employ supervised finetuning or reinforced post-training approaches adapted for embodied tasks. To better address the diverse requirements of embodied intelligence evaluation, these models are categorized into 2D and 3D embodied models based on their primary application domains and the characteristics of their target environments.

**2D Embodied Models** These models are specifically engineered for 2D visual reasoning benchmarks and embodied question answering tasks that operate within 2D visual representations (Fu et al., 2024, Chen et al., 2024a). They excel at processing egocentric viewpoints, understanding spatial relationships in 2D projections (Liu et al., 2023, Li et al., 2024b), and reasoning about object interactions within constrained visual fields (Cheng et al., 2024b, Liao et al., 2024). These models are primarily applied to VSI-Bench, ERQA, Where2Place, RoboVQA (Sermanet et al., 2024), and other 2D QA benchmarks (Majumdar et al., 2024,

Azuma et al., 2022), where they demonstrate superior performance in tasks requiring fine-grained spatial reasoning (Cai et al., 2024, Ray et al., 2024), temporal understanding from video sequences (Bharadhwaj et al., 2024, Xu et al., 2024b), and affordance prediction in 2D visual contexts (Yuan et al., 2024b, Nasiriany et al., 2024).

- **BAAI:** Navid (Zhang et al., 2024b), UniNavid (Zhang et al., 2024a), RoboBrain1.0-7B (Ji et al., 2025), RoboBrain2.0-7B, RoboBrain2.0-32B (Team et al., 2025), MapNav (Zhang and Heidari, 2025)
- **Shanghai AI Lab:** VeBrain (Luo et al., 2025), VLN-R1 (Qi et al., 2025b), StreamVLN (Wei et al., 2025a)
- **Tianjin University:** HuLE-Nav (Han et al., 2024), Embodied-R1 Yuan et al. (2025b)
- **University of Washington:** RoboPoint (Yuan et al., 2024b)
- **Shanghai Jiao Tong University:** SpatialBot (Cai et al., 2024)
- **The University of Hong Kong:** EmbodiedGPT (Mu et al., 2023)
- **Google DeepMind:** LFG (Shah et al., 2023), SpatialVLM (Chen et al., 2024a)
- **Meta AI:** OVRL (Yadav et al., 2022), OVRL-v2 (Yadav et al., 2023a)
- **Peking University:** Space-R (Ouyang et al., 2025), VoroNav (Wu et al., 2024), InstructNav (Long et al., 2024)
- **Huawei Noah’s Ark Lab:** Noah(UniE-VLM), OmniEVA
- **NVIDIA:** Cosmos-reason (Liu et al., 2025), NaVILA (Cheng et al., 2024a)
- **Beihang University:** Robo-Refer (Zhou et al., 2025)
- **Boston University:** SAT (Ray et al., 2024)
- **University of California:** ESC (Zhou et al., 2023)
- **University of California Berkeley:** VLMNav (Goetting et al., 2024)
- **University of Groningen:** L3MVN (Yu et al., 2023)
- **NYUAD Center for Artificial Intelligence and Robotics:** GAMap (Yuan et al., 2024a)
- **Microsoft:** Magma (Yang et al., 2025b)

**3D Embodied Models** These advanced models are architecturally designed for comprehensive 3D scene understanding and complex spatial reasoning tasks that require full volumetric scene comprehension (Hong et al., 2023, Chen et al., 2024b). They incorporate sophisticated 3D feature extraction mechanisms, point cloud processing capabilities, and multi-view geometric reasoning to handle the inherent complexity of three-dimensional environments (Zhu et al., 2024, Li et al., 2024c). These models excel at understanding object relationships in 3D space, reasoning about occlusions and spatial arrangements, and generating contextually aware descriptions of complex indoor scenes (Yang et al., 2025d, Qi et al., 2025a). They are primarily applied to ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2023), Scan2Cap (Chen et al., 2021), and other 3D QA benchmarks, where they demonstrate superior performance in tasks requiring dense captioning of 3D scenes, spatial localization within point clouds, and multi-hop reasoning across complex 3D spatial configurations (Zheng et al., 2025).

- **BIGAI:** LEO (Huang et al., 2024)
- **Shanghai AI Lab:** Grounded 3D-LLM (Chen et al., 2024b), GPT4Scene (Qi et al., 2025a)
- **Peking University:** UniNavid (Zhang et al., 2024a), Navid (Zhang et al., 2024b)
- **The University of Hong Kong:** Video-3D LLM (Zheng et al., 2025), LLaVA-3D (Zhu et al., 2024), GPT4Scene (Qi et al., 2025a)
- **The Chinese University of Hong Kong:** Video-3D LLM (Zheng et al., 2025)
- **UMass Amherst:** 3D-Mem (Yang et al., 2025d)

### 4.3. Infrastructure

Our unified evaluation infrastructure forms the backbone of Embodied Arena, ensuring consistent, reliable, and scalable assessment across all benchmarks through carefully designed system architecture and standardized protocols. The infrastructure is built with modularity, extensibility, and reproducibility as core design principles.

**Standardized Evaluation Framework** The platform implements a standardized evaluation framework with uniform input/output formats that enable seamless comparison across diverse models and benchmarks. This framework abstracts away benchmark-specific implementation details while preserving the unique characteristics of each evaluation task. The standardized interface supports various model access methods, including API-based evaluation for commercial models, parameter-based integration for open-source models, and custom interfaces for specialized architectures.

**Flexible Model Integration** Embodied Arena supports comprehensive evaluation of models from different sources (open-source, commercial) through various access methods (model parameters, API endpoints, custom implementations). This flexibility ensures broad accessibility and participation while maintaining evaluation consistency and fairness across different model types and deployment scenarios.

**Professional Management:** The infrastructure includes comprehensive experiment tracking and management capabilities that provide detailed performance analysis and ensure reproducible evaluation results. Each evaluation run is meticulously logged with complete metadata including model configurations, benchmark parameters, execution environment details, and performance metrics.

### 4.4. Evaluation Methods

Embodied Arena extensively supports the comprehensive evaluation of models from different sources (open-source, commercial) by different means (model parameters, API), offering flexibility and convenience for users to join. The platform leverages a diverse range of well-curated Embodied AI benchmarks, ensuring high alignment with canonical evaluation methods and the best completeness compared to prior works.

#### 4.4.1. Evaluation Metrics

During the evaluation phase, we select the corresponding evaluation metric based on the characteristics of the benchmark itself, which generally include the following types:

##### Embodied Question Answering:

- **Exact Matching Accuracy:** Applied to benchmarks requiring precise categorical responses such as VSI-Bench, Where2Place, and ERQA (Du et al., 2024). This metric evaluates the model’s ability to provide accurate factual answers and correct spatial reasoning outputs.
- **Fuzzy Matching Accuracy:** Employed for benchmarks involving natural language generation and open-ended responses:
  - *Rule-based Metrics* (CIDEr, BLEU, ROUGE, MRA): Applied to benchmarks like RoboVQA (Sermanet et al., 2024), Scan2Cap (Chen et al., 2021), and ScanQA (Azuma et al., 2022) for evaluating generated descriptions and spatial explanations
  - *LLM-based Evaluation:* Utilized for benchmarks such as OpenEQA (Majumdar et al., 2024) and UniEQA (Zhang et al., 2025b), leveraging large language models to assess semantic correctness and reasoning quality in generated responses (Zheng et al., 2023)

### Embodied Navigation Evaluation:

- **Success Rate:** Primary metric for navigation benchmarks including EB-Navigation, R2R-CE (Krantz et al., 2020), and RxR-CE, measuring the percentage of successfully completed navigation episodes
- **Path Length Weighted Success Rate (SPL):** Evaluates navigation efficiency by considering both success and path optimality

### Embodied Task Planning Evaluation:

- **Task Completion Success Rate:** Applied to benchmarks such as EB-ALFRED, EB-Habitat, and ET-Plan-Bench (Zhang et al., 2024c), measuring the percentage of successfully completed task sequences

#### 4.4.2. Scoring Rules

The scoring rules for the embodied capability leaderboards and the embodied task leaderboards are as follows:

**Embodied Task Leaderboards:** Given  $N$  benchmarks, let there be a benchmark  $B^n$  ( $n = 1, 2, \dots, N$ ) consisting of  $M$  fine-grained original capability dimensions. For each capability dimension  $m$  ( $m = 1, 2, \dots, M$ ),  $k_m^n$  denotes the total number of questions in the  $m$ -th capability dimension, and  $c_m^n$  is the number of questions answered correctly in the  $m$ -th capability dimension. Each question has a score within  $[0, 1]$ .

- **Score Calculation for a Single Benchmark**

- *Capability Dimension Score*  $S_m^n$ :  $S_m^n = \frac{c_m^n}{k_m^n} \times 100$ , where  $c_m^n \in [0, k_m^n]$  and  $S_m^n \in [0, 100]$ .
- *Total Score of All Capability Dimensions*  $A_{total}^n$ :  $A_{total}^n = \frac{1}{M} \sum_{m=1}^M S_m^n$ .

- **Total Score across  $N$  Benchmarks:**  $B_{total} = \frac{1}{N} \sum_{n=1}^N A_{total}^n$ .

**Embodied Capability Leaderboards:** Given  $N$  benchmarks, let there be a benchmark  $B^n$  ( $n = 1, 2, \dots, N$ ) with  $M^n$  original capability dimensions. Our taxonomy defines  $D = 25$  fine-grained capability dimensions and  $P = 7$  core capabilities. Let  $\phi : (n, i) \rightarrow j$  be the mapping function from the  $i$ -th original dimension of benchmark  $B^n$  to the  $j$ -th taxonomy dimension. For each taxonomy dimension  $j$  ( $j = 1, 2, \dots, D$ ), let  $k_j^n$  and  $c_j^n$  denote the total number of questions and correctly answered questions respectively, aggregated from all original dimensions of benchmark  $B^n$  that map to dimension  $j$ .

- **Score Calculation for a Single Benchmark**

- *Fine-grained Capability Dimension Score*  $S_j^n$ :  $S_j^n = \frac{c_j^n}{k_j^n} \times 100$  if  $k_j^n > 0$ , otherwise undefined.
- *Core Capability Dimension Score*  $C_p^n$ :  $C_p^n = \frac{1}{|I_p|} \sum_{j \in I_p} S_j^n$ , where  $I_p$  is the set of fine-grained dimensions belonging to core capability  $p$ .
- *Total Score of All Capability Dimensions*  $A_{total}^n$ :  $A_{total}^n = \frac{1}{|J^n|} \sum_{j \in J^n} S_j^n$ , where  $J^n$  is the set of taxonomy dimensions covered by benchmark  $B^n$ .

- **Total Score of  $N$  Benchmarks on Fine-grained Capability Dimension  $j$ :**  $T_{total}^j = \frac{\sum_{n=1}^N c_j^n}{\sum_{n=1}^N k_j^n} \times 100$  (only for benchmarks where  $k_j^n > 0$ ).

- **Total Score of  $N$  Benchmarks for All Capability Dimensions**  $B_{total}$ :  $B_{total} = \frac{1}{D} \sum_{j=1}^D T_{total}^j$  (only including dimensions with valid scores).

## 4.5. Leaderboards

Embodied Arena features a comprehensive leaderboard system designed to provide clear, actionable insights into model performance across different perspectives and granularities. Through comprehensive evaluation among the growing model population based on evolving evaluation data, Embodied Arena publishes three types of leaderboards, i.e., Embodied Q&A, Embodied Navigation, Embodied Task Planning, with two orthogonal views, i.e., *benchmark view* and *capability view*. The benchmark view presents the ranking of models on each benchmark, which is convenient for academic researchers to quote and compare with in their research works; while the capability view instead presents the ranking of models against each embodied capability in the systematic taxonomy, providing an up-to-date overview of embodied capabilities of advanced models. Moreover, to ensure evaluation integrity and community engagement, our leaderboard system implements structured monthly updates with transparent submission policies and real-time performance tracking. The ultimate aim of Embodied Arena is to facilitate the establishment of clear research veins and help to identify critical research problems, thereby propelling research progress in the field of Embodied AI.

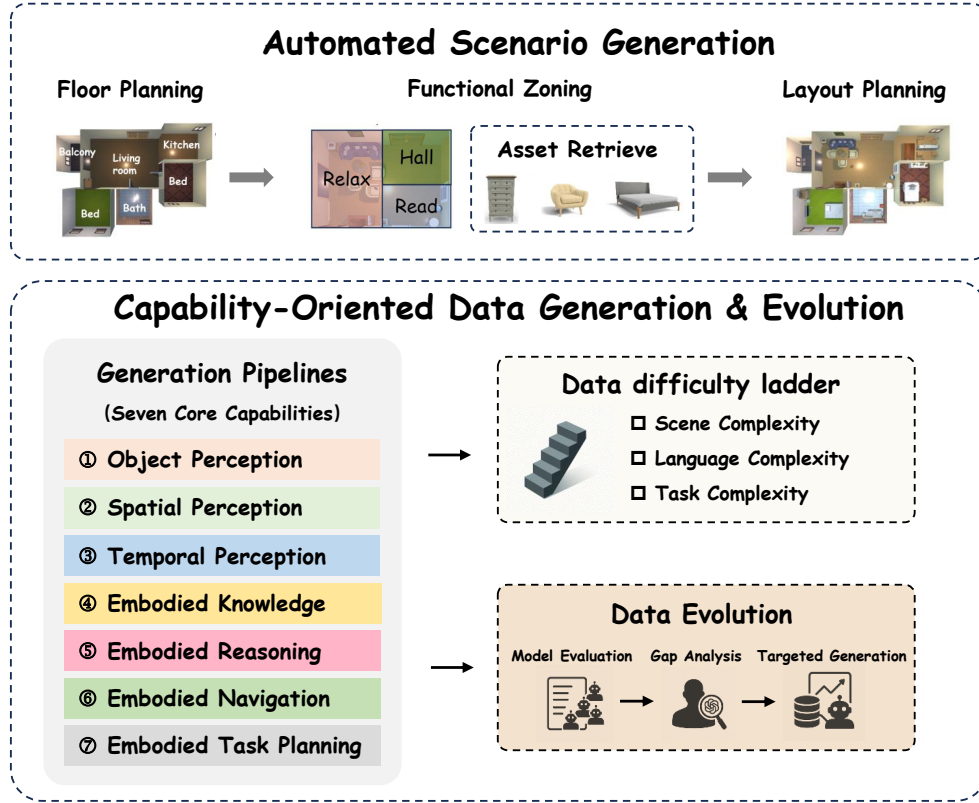
**Three Types of Leaderboards** Embodied Arena establishes comprehensive evaluation through three specialized leaderboards targeting distinct Embodied AI domains. The *Embodied Question Answering* leaderboard evaluates models across visual reasoning tasks, assessing capabilities from 2D visual understanding to 3D spatial comprehension. The *Embodied Navigation* leaderboard focuses on spatial movement and pathfinding abilities, evaluating object navigation, location navigation, and instruction-following capabilities. The *Embodied Task Planning* leaderboard assesses high-level reasoning and decomposition skills, examining models' abilities to break down complex tasks into executable sequences. Together, these leaderboards provide comprehensive coverage of core Embodied AI competencies.

**Two Orthogonal Comparison Views** Embodied Arena provides complementary insights through its dual-view leaderboard system. The *benchmark view* presents model rankings on individual benchmarks, enabling direct comparison and facilitating academic citation. In contrast, the *capability view* aggregates performance across our seven core embodied capabilities (detailed in Section 3), providing strategic insights into model strengths and weaknesses. Together, these orthogonal views deliver both granular benchmark-specific analysis and holistic capability assessment, enabling targeted improvement while maintaining systematic understanding of Embodied AI advancement.

**Update and Submission Policies** Embodied Arena maintains evaluation integrity through structured update protocols and submission policies. The leaderboards are updated monthly with performance snapshots taken on the first working day of each month, ensuring consistent and timely community engagement. To ensure fairness, each organization is limited to one evaluation submission per month, with results typically processed and updated within seven working days of submission. This systematic approach provides dynamic performance tracking while maintaining evaluation reliability and preventing potential gaming of the leaderboard system through excessive submissions.

**Platform Accessibility and Community Engagement** Embodied Arena operates as an open evaluation platform designed to foster community-driven advancement in embodied AI research. The platform provides multiple pathways for researcher participation: *open evaluation access* allows researchers to submit models for assessment through standardized API interfaces, *benchmark contribution* enables community members to propose and integrate new evaluation tasks following our established guidelines, and *transparent methodology* ensures all evaluation protocols and baseline implementations remain publicly accessible. Through comprehensive documentation, integration templates, and testing scripts, Embodied Arena lowers barriers to participation while maintaining evaluation consistency. This open architecture not only democratizes access





**Figure 3: Illustration of the automated data generation pipeline.** The pipeline includes two modules: Automated Scenario Generation and Capability-Oriented Data Generation & Evolution. The former is responsible for generating diverse and realistic high-fidelity scenarios, while the latter builds generation pipelines to ensure the continuous evolution of the evaluation set.

to comprehensive embodied AI evaluation but also enables the platform to evolve continuously with field advancement through community contributions and feedback.

## 5. Automated Data Generation for Embodied AI Evaluation

Current evaluation benchmarks for embodied tasks suffer from fundamental limitations in adaptability, scalability, and task diversity, which restrict their effectiveness. In contrast, **Embodied Arena** is designed as a continuously evolving evaluation platform—one that actively identifies model weaknesses and autonomously generates new, targeted data to maintain the comprehensiveness and cutting-edge nature of the benchmark over time. Specifically, we identify three core limitations in existing benchmarks:

- **Static evaluation:** Conventional benchmarks are typically constructed once and remain fixed, without adapting to model performance. This static nature introduces the overfitting risk, whereby agents achieve high performance on existing data but fail to generalize to out-of-distribution (OOD) data.
- **Limited scalability:** Current benchmarks rely heavily on manual annotation, which is both labor-intensive and time-consuming, rendering it infeasible to collect large-scale evaluation datasets efficiently.
- **Limited diversity:** Most handcrafted data focus on a small set of tasks, making it difficult to evaluate the broad spectrum of embodied capabilities or generalization to novel tasks.

To overcome these challenges, we introduce an LLM-driven evaluation data generation framework built

on high-fidelity simulation. Our framework consists of two key modules: *Automated Scenario Generation*, which constructs realistic and diverse simulation environments, and *Capability-Oriented Data Generation & Evolution*, which establishes data generation pipelines and continuously injects targeted data to adapt the evaluation set based on the model limitations.

### 5.1. Automated Scenario Generation

Since embodied agents operate in complex physical environments, the ability to simulate realistic and diverse scenes is essential for evaluating their embodied capabilities across a broad range of tasks. To address this, the Automated Scenario Generation Module is designed to automatically build multi-room indoor environments through a structured, hierarchical process that mirrors real-world scene building process. The generation pipeline consists of three stages: (1) *Floor Planning*, which defines room types and their spatial relationships to ensure logical connectivity and functional plausibility; (2) *Functional Zoning*, which divides each room into activity-specific zones (e.g., cooking, dining, storage); and (3) *Layout Planning*, which populates each zone with diverse assets and applies layout optimizations.

To ensure the generated scenes are aligned with human common sense and real-world affordances, we leverage large language models (LLMs) and vision–language models (VLMs) throughout the generation process. These models guide decisions such as object spatial relations and scene semantics, enabling the creation of environments that are not only diverse but also semantically coherent. Once the scene is constructed, a high-fidelity rendering pipeline produces rich outputs, including RGB images, depth maps, and a structured object graph. Domain randomization techniques are applied to introduce variability in textures, lighting, and viewpoints, enhancing generalization for downstream tasks. Furthermore, the module offers a *targeted scene* mode, in which users provide high-level descriptors — such as “cluttered kitchen with partially hidden utensils” or “open-concept living room featuring scattered numeric signs”. The system then samples room layouts, places assets, and applies refinements to construct these concrete indoor environments, yielding reproducible scenes that match the specified requirements.

### 5.2. Capability-Oriented Data Generation & Evolution

To generate datasets for the seven core embodied capabilities, we design simulator-driven procedural pipelines. Each pipeline (i) defines a capability-specific task template and specification, (ii) loads scenes and task-relevant assets, (iii) executes scripted procedures, (iv) leverages the simulator’s privileged access to automatically extract object types, positions, attributes, and other ground-truth annotations as the basis for dataset construction, and (v) performs automated filtering and selection to retain only unambiguous, high-quality data before storage.

Building on these pipelines, we introduce the difficulty ladder that generates data along three dimensions:

- Scene Complexity (number of objects, degree of occlusion)
- Language Complexity (instruction length, semantic complexity)
- Task Complexity (Horizon length, temporal dependencies)

At each level, we generate visual-instruction-answer triplets. During evaluation, agents can unlock levels sequentially, enabling granular diagnosis of strengths and weaknesses and providing a built-in curriculum for progressive fine-tuning.

Procedural data generation relies on privileged information obtained from simulation, which ensures correctness but may also introduce ambiguous cases that lead to model hallucinations. For example, certain

2D-Embodied QA Benchmarks										
		Property		Value						
		All		All						
#	Model	↓ Total Score	UniEQA	OpenEQA	VSI	ERQA	RoboVQA	Where2Place	MineAnyBuild	PhyBlock
1	Gemini-2.5-Pro	49.25	60.65	65.75	44.10	55.65	49.00	49.58	53.21	43.60
2	o3	46.07	55.44	61.43	57.80	41.51	52.25	33.46	52.50	40.60
3	Qwen-VL-Max	42.47	50.81	52.19	33.98	39.32	65.75	26.84	42.42	28.90
4	RoboBrain2.0-32B	41.48	36.20	27.14	42.69	45.07	57.75	73.59	36.43	17.32
5	Embodied-R1-3B	39.53	38.14	34.51	26.56	35.24	22.51	69.50	37.07	26.20
6	Gemini-2.0-flash	38.52	51.47	48.36	45.40	40.18	36.75	33.80	42.10	39.30
7	GPT-4o	37.20	48.47	51.06	34.00	34.79	41.50	22.35	48.80	38.80
8	InternVL3-38B	37.19	49.58	45.83	45.31	41.68	45.50	20.14	41.27	26.10
9	o4-mini	35.74	49.71	54.41	26.43	30.72	50.25	25.06	52.76	23.30
10	InternVL3-14B	34.91	46.38	44.16	44.21	38.76	47.25	16.56	44.35	22.50
11	RoboBrain2.0-7B	33.68	33.78	24.15	36.10	33.83	57.50	62.34	34.87	13.56
12	Space-R	33.48	41.36	39.38	41.76	37.76	53.00	15.32	37.43	17.40

**Figure 4: A Screenshot of 2D Embodied QA Leaderboard from Embodied Arena.** The leaderboard shows performance rankings across multiple 2D embodied question answering benchmarks, with large-scale general multimodal models generally achieving higher overall scores than specialized embodied models. This convenient web interface enables researchers to easily analyze model performance patterns and compare capabilities across different approaches.

assets in corners may be recognizable in simulation even when only a small portion is exposed, yet such cases remain challenging for human observers. To ensure data quality, we adapt a sampling-based inspection method, where human evaluation is used to filter the data and remove cases that are difficult to discern for the human eye. Although this process introduces a certain degree of manual overhead, it offers a more reliable safeguard for data accuracy and validity.

To support the long-term effectiveness of Embodied Arena, we introduce a data evolution mechanism driven by model performance analysis. By regularly analyzing model capability, we generate the related data via the aforementioned pipelines. These targeted additions enrich the evaluation set with fresh, challenging data tailored to current model limitations. In this way, Embodied Arena evolves alongside model progress, maintaining comprehensiveness while continuously advancing in difficulty and evaluation value.

## 6. Insights from Embodied Arena Evaluation and Leaderboards

Through comprehensive evaluation across the diverse benchmarks and model ecosystem on our platform, Embodied Arena reveals several major insights that illuminate the current state and future directions of Embodied AI research. These findings emerge from a systematic analysis of performance patterns across 30+ models and 22+ benchmarks, providing empirical evidence for understanding the fundamental capabilities and limitations of contemporary embodied intelligence systems.

- **Finding 1: Embodied models surpass general models of similar sizes on specialized benchmarks, while top-tier closed-source general models achieve strong overall performance with large model size and massive training data.**

2D-Embodied QA Benchmarks			Navigation Benchmarks - Unified Evaluation Framework				Task Planning Benchmarks			
#	Model	↓ Total Score	#	Model	↑ Avg Rank	Total SR	#	Model	↑ Avg Rank	Total SR
👑	Gemini-2.5-Pro	49.25	👑	StreamVLN	1.0	54.90	👑	o3	1.0	72.17
👑	o3	46.07	👑	NaVILA	2.0	51.65	👑	Claude-3.7-Sonnet	3.0	66.42
👑	Qwen-VL-Max	42.47	👑	UniNavid	3.0	47.85	👑	Gemini-2.5-Pro	4.7	61.05
4	RoboBrain2.0-32B	41.48	4	MapNav	4.0	36.15	4	GPT-4o	5.0	62.34
5	Embodied-R1-3B	39.53	5	Navid	5.0	30.60	5	Qwen-VL-Max	8.3	47.81
6	Gemini-2.0-flash	38.52	6	o3	5.0	34.80	6	Gemini-2.0-flash	10.0	50.04
7	GPT-4o	37.20	7	VLN-R1	6.0	26.25	7	InternVL3-38B	11.7	45.16
8	InternVL3-38B	37.19	8	Gemini-2.0-flash	8.3	21.60	8	RoboBrain2.0-32B	12.7	40.98
9	o4-mini	35.74	9	Claude-3.7-Sonnet	10.3	18.93	9	GPT-4o mini	13.3	42.59
10	InternVL3-14B	34.91	10	Gemini-2.5-Pro	12.3	27.37	10	InternVL3-14B	17.7	29.90
11	RoboBrain2.0-7B	33.68	11	InternVL3-14B	12.7	15.20	11	Embodied-R1-3B	21.7	17.97
12	Space-R	33.48	12	GPT-4o	14.3	21.67	12	InternVL3-8B	22.0	24.04
13	InternVL3-8B	32.42	13	RoboBrain1.0-7B	17.0	10.57	13	RoboBrain2.0-7B	22.0	23.61

**Figure 5: A Screenshot of Cross-Benchmark Performance Variation from Embodied Arena.** The figure displays ranking comparisons across 2D Embodied QA, Navigation, and Task Planning leaderboards, revealing significant performance fluctuations where models excel in specific domains but struggle to maintain consistent rankings across all task types. This comprehensive view from our web platform facilitates easy analysis of model strengths and weaknesses across different embodied capabilities.

**Core Finding:** As illustrated in Figure 4, our evaluation reveals a significant phenomenon that massive general models achieve strong performance through large model size and broad knowledge, while specialized models excel through targeted embodied training. Large commercial general models, e.g., GPT-o3 (OpenAI, 2025), Gemini-2.5-Pro (Google DeepMind, 2024), Claude-3.7 (Anthropic, 2024), leverage their massive scale to achieve overall 10-20% performance advantages across most benchmarks, demonstrating that model size and extensive pre-training data clearly matter. However, when we compare models of similar scales, specialized embodied models consistently outperform the general multimodal models. RoboBrain2.0-32B (BAAI, 2025) achieves 73.59% on Where2Place versus GPT-o3’s 33.46%, while navigation specialists like StreamVLN (Wei et al., 2025a) and UniNavid (Zhang et al., 2024a) reach 30-57% success rates while the general models of similar sizes such as InternVL3 (Zhu et al., 2025) achieve less than 10% success rates.

**In-depth Analysis:** Massive model scale powered by large-scale pre-training results in clear advantages — commercial models leverage massive parameter scales (likely hundreds of billions to trillions) trained on internet-scale datasets, providing both reasoning capacity and extensive world knowledge. However, specialized embodied data also demonstrates remarkable power. Models like RoboBrain2.0 (BAAI, 2025), Embodied-R1, and RoboPoint (Yuan et al., 2024b) show that fine-tuning small-scale open-source general models with high-quality embodied datasets can produce dramatic performance gains. By incorporating specialized training data focused on spatial identification, affordance prediction, and manipulation sequences, these models vastly outperform their original versions and can even match large commercial models on specific embodied benchmarks. This effect is particularly evident in navigation models like StreamVLN (Wei et al., 2025a) and NaVILA (Zheng et al., 2024), which use vision-language-action training paradigms to develop capabilities that general pre-training cannot provide. This demonstrates that targeted incorporation of embodied-specific data represents a viable pathway to achieving competitive performance even with constrained computational resources. This post-training optimization pathway promises to be a key research direction for enabling smaller specialized embodied models to match or surpass large closed-source general models through better data quality and targeted architectural innovations.

- **Finding 2: Individual benchmarks with limited scope are insufficient and biased for embodied eval-**

**uation. Embodied models exhibit more or less overfitting on benchmark-specific data rather than developing comprehensive embodied capabilities.**

**Core Finding:** Models exhibit dramatic performance variations across different benchmarks, revealing fundamental limitations in both current evaluation benchmarks in Figure 5. For instance, RoboBrain-v1-7B (Ji et al., 2025) achieves top performance on RoboVQA (Sermanet et al., 2024) across all metrics but performs poorly on spatial understanding benchmarks like Where2Place (Yuan et al., 2024b) and VSI-Bench (Yang et al., 2024). Similarly, while specialized navigation models dominate VLN tasks with success rates above 50%, they struggle with basic question answering tasks. Only a few large-scale models like Gemini-2.5-Pro (Google DeepMind, 2024), GPT-o3 (OpenAI, 2024), and InternVL3-38B (Zhu et al., 2025) maintain relatively balanced performance across question answering, navigation, and task planning.

**In-depth Analysis:** This performance inconsistency across benchmarks directly validates the core motivation for establishing our comprehensive embodied leaderboard system, which aims to provide holistic model evaluation beyond isolated task performance. Each benchmark evaluates only limited capability dimensions, and no single benchmark currently covers all embodied capabilities comprehensively, causing significant ranking fluctuations as models demonstrate varying strengths across embodied AI capabilities — spatial reasoning benchmarks like Where2Place (Yuan et al., 2024b) favor affordance prediction training while task planning benchmarks like EB-ALFRED (Yang et al., 2025c) advantage instruction following abilities. Moreover, current embodied models exhibit concerning overfitting phenomena where capabilities appear artificially enhanced through injecting benchmark-correlated data for specialized ability improvements — performance on one benchmark can be improved simply by adding specific related training datasets, but this comes at the expense of performance on other benchmarks rather than achieving true comprehensive enhancement of embodied capacity. To this end, one possible solution is developing automated data generation systems that create diverse scenarios and tasks for comprehensive evaluation and training. These systems will enable unified evaluation paradigms that assess models’ ability to integrate perception, temporal reasoning, and planning in realistic scenarios, providing robust assessment that resists benchmark-specific overfitting.

- **Finding 3: The embodied reasoning capabilities of models are strongly dependent on their fundamental embodied capabilities. Among the five fundamental embodied capabilities, object perception and spatial perception turn out to be the major bottlenecks.**

**Core Finding:** The comprehensive evaluation results across multiple benchmarks indicate that the defects in the models’ fundamental embodied capabilities directly restrict their performance in the advanced reasoning capability. Specifically, the models’ fundamental embodied capabilities (object perception, spatial perception, temporal perception, and embodied knowledge) show a significant positive correlation with advanced reasoning capability (Spearman’s rank correlation coefficient  $\rho = 0.80$ ,  $p < 0.0001$ ), and each fundamental embodied capability also exhibits a significant positive correlation with advanced reasoning capability ( $\rho$  ranging from 0.68 to 0.77,  $p < 0.0001$ ). Meanwhile, the models’ performance on advanced reasoning capability (with an average score of 33.64) is generally worse than their overall performance on fundamental embodied capabilities (with an average score of 38.84). Among the fundamental embodied capabilities, the models’ object perception (average score 38.33) and spatial perception (average score 28.62) capabilities are the worst. These results collectively reveal the deep dependence of the model’s advanced reasoning abilities on its fundamental embodied capabilities.

**In-depth Analysis:** The embodied capabilities of the evaluated models exhibit a hierarchical decline, primarily stemming from three core factors: first, there is a structural imbalance in the pre-training data for different embodied capabilities, with labeled data related to reasoning capabilities being particularly scarce; second, the deficiency of specific embodied capabilities easily leads to performance declines of models in other ones



due to the dependencies among embodied capabilities; third, the reasoning capabilities of embodied models are still lacking while the methods for enhancing embodied reasoning capabilities have not been well studied yet.

- **Finding 4: The fundamental and advanced embodied capabilities of models are significantly positively correlated with their performance on downstream embodied tasks. Furthermore, in an end-to-end manner, there is a moderate correlation between the model’s embodied capabilities and downstream task performance. In contrast, in a task-oriented agentic framework manner, there is a strong correlation between the model’s embodied capabilities and downstream task performance.**

**Core Finding:** Based on the comprehensive ranking of the models in embodied capabilities (e.g., object perception, spatial perception, temporal perception, embodied knowledge, and embodied reasoning) and downstream tasks (e.g., embodied navigation, embodied task planning), we find that models’ embodied capabilities are highly positively correlated with downstream task performance (Spearman’s rank correlation coefficient  $\rho = 0.80$ ,  $p < 0.0001$ ). Moreover, each embodied capability shows a significant positive correlation with downstream task performance, with correlation coefficients  $\rho$  ranging from 0.73 to 0.83. Among these, embodied knowledge demonstrates the strongest positive correlation ( $\rho = 0.83$ ,  $p < 0.0001$ ), and the remaining capabilities show stable correlation coefficients around 0.75. All p-values for the above analyses are less than 0.0001, reaching an extremely high level of statistical significance. Notably, the strength of this correlation is significantly influenced by how the model is applied in downstream tasks: in an end-to-end manner, the model’s embodied capabilities and downstream task performance exhibit only a moderate correlation ( $\rho = 0.40$ ,  $p > 0.08$ ), which does not reach statistical significance; However, in a task-oriented agentic framework manner (i.e., integrating the general models into agentic frameworks targeted for downstream embodied tasks), the two exhibit a strong positive correlation ( $\rho = 0.79$ ,  $p < 0.0001$ ). In addition, we find that when general models are applied in an end-to-end manner, their overall success rate in navigation tasks is only 5.80%. In contrast, when using task-oriented agentic frameworks (such as EmbodiedBench (Yang et al., 2025c) and ET-Plan-Bench (Zhang et al., 2024c)), general models’ success rates in navigation and task planning tasks increase to 36.21% and 40.08%, respectively.

**In-depth Analysis:** These evaluation results reveal two key insights about how to convert the embodied capabilities of models into performance on downstream embodied tasks. On one hand, the embodied capabilities of models are the foundation for the performance in downstream embodied tasks. Therefore, further enhancing the embodied capabilities like perception and reasoning is an essential necessity. On the other hand, compared with the end-to-end approach, task-oriented agentic framework provides an effective pathway for better utilizing the fundamental embodied capabilities of general models in downstream embodied tasks, although these agentic frameworks usually require manual design and lack generality. Leveraging learning-based methods like RL to optimize or generate the agentic framework can be promising to facilitate the transformation of embodied capabilities into task performance.

- **Finding 5: The scaling law for embodied tasks has yet to emerge. Larger model size does not consistently lead to stronger embodied capabilities, although a positive correlation exists locally for specific models and capabilities. More embodied data leads to improved task-specific performance, albeit with increased overfitting.**

**Core Finding:** For the same model architecture, increasing the number of parameters can improve performance on specific embodied benchmarks. General multimodal models show more consistent evidence of this, whereas for embodied models, this phenomenon is inconsistent across models and capabilities. For instance, InternVL3-38B > InternVL3-14B > InternVL3-8B (Zhu et al., 2025) across all three fundamental capabilities. Similar trends are observed for RoboBrain2.0 (32B vs. 7B) (BAAI, 2025) and Qwen2.5-VL-Instruct (7B vs. 3B) (Bai et al., 2025) on embodied QA and embodied task planning. Nevertheless, the scaling effect is

not consistent: in embodied navigation, smaller models (RoboBrain2.0-7B and Qwen2.5-VL-3B-Instruct) achieve better performance than their larger counterparts. Moreover, from the perspective of embodied data, increasing the amount of task-specific data can significantly improve models' specific capabilities. For example, StreamVLN (Wei et al., 2025a), NaVILA (Zheng et al., 2024), and UniNavid (Zhang et al., 2024a) outperform GPT-o3 (OpenAI, 2025) in instruction-following navigation. However, the constructed embodied datasets usually fail to deliver consistent performance improvements in all capabilities. For example, Embodied-R1 (Yuan et al., 2025b) and SpaceR (Ouyang et al., 2025), trained on their respective embodied datasets, surpass the base model Qwen-2.5-VL-3B-Instruct in some capabilities. However, they also suffer from performance drops in others.

**In-depth Analysis:** The scaling phenomena regarding model size and data amount for embodied models have not emerged generally across embodied benchmarks and capabilities. Different from LLMs and general multimodal models, which often share the base model architecture among several canonical choices, embodied models vary considerably in how they are constructed. Moreover, embodied models are often released in only one or a narrow range of sizes. Therefore, a thorough investigation of the scaling law regarding model size for embodied models requires more effort in building consistent model architectures and providing multiple model sizes. For the scaling regarding embodied data, the task-specific performance improvement accompanied by overfitting is likely to stem from insufficient diversity, scope, and scale. Hence, a scalable approach for data construction or generation is essential for advancing research in this regard.

- **Finding 6: Reasoning models exhibit strong overall performance on multiple benchmarks by reinforced finetuning (RFT). However, whether RFT can yield stronger out-of-distribution generalization than SFT remains a key open question.**

**Core Finding:** Reasoning models fine-tuned with RFT have demonstrated significant and consistent performance improvements across multiple benchmarks. We observe that the latest records on most benchmarks have been set by these reasoning models. For instance, GPT-o3 (OpenAI, 2025) achieves remarkable and stable performance on task planning benchmarks such as EB-ALFRED, EB-Habitat, and EB-Navigation, exhibiting no obvious capability shortcomings. Space-R (Ouyang et al., 2025) establishes a new SOTA for embodied models on OpenEQA (Majumdar et al., 2024) (37.70 points), while also maintaining stable performance on other embodied QA tasks. Furthermore, Embodied-R1 (Yuan et al., 2025b) achieves breakthrough results on affordance prediction tasks like VABench-Point (66 points). In the context of VLN-R1 (Qi et al., 2025b), fine-tuning the Qwen2-VL-7B model (Wang et al., 2024) with VLN data via supervised learning significantly enhances its navigation capabilities. Building upon this, the application of GRPO for RFT further boosts performance, increasing the success rate from 24.9 to 30.2. Nevertheless, although the majority of recently emerged models are reasoning-based (e.g., Gemini-2.5-Pro, o3, and RoboBrain2.0), whether RFT can yield superior generalization compared to SFT and enhance capabilities beyond the training data distribution remains a question that requires and merits further exploration.

**In-depth Analysis:** Lagging behind the development of LLMs and general multimodal models, the ability of slow thinking or reasoning has not been universally empowered for existing embodied models. Pivotal to the ability of slow thinking, RL training has demonstrated its effects in enabling models to activate and combine fundamental perceptual abilities into complex reasoning skills. It allows the models to fully utilize fundamental embodied capabilities to address complex tasks, rather than merely pattern matching from training examples. This RL-finetuning paradigm is particularly suited for embodied tasks involving multi-step reasoning, sequential decision-making, and precise manipulation, and offers promising directions for future training strategies in embodied AI.

- **Finding 7: 3D representations are essential for embodied understanding but face challenges in the alignment with language modality. Strategic integration with 2D-3D representation can effectively**

### leverage pre-trained language alignment to unlock superior spatial understanding.

**Core Finding:** The evaluation results on 3D Embodied QA benchmarks show that models that effectively integrate 2D visual features with 3D spatial priors significantly outperform those relying solely on naive 3D data processing. The top-performing models — GPT4Scene-HDM (Qi et al., 2025a) (71.00), LL3DA (Chen et al., 2024b) (62.11), 3DRS (65.77), OmniEVA (64.66), and Video-3D LLM (Zheng et al., 2025) (64.92) — all adopt visual-spatial integration strategies, while the models that adopt native 3D processing methods like LEO (Huang et al., 2023) (48.48) consistently underperform with a 15-25% performance gap.

**In-depth Analysis:** The evaluation reveals a fundamental architectural principle: 3D geometric representations provide important spatial awareness that 2D understanding cannot deliver, yet they encounter challenges in achieving sufficient language modality alignment. Traditional approaches that directly process point clouds or voxels through 3D encoders show consistently poor performance across 3D embodied benchmarks. This indicates that naive 3D representation methods face inherent challenges in language modality alignment and cannot effectively leverage pre-training model capacity. In contrast, current leading models consistently employ strategic integration methods that combine rich 2D visual features with explicit 3D spatial information through various encoding mechanisms through position encoding, multi-view synthesis, coordinate injection, etc. The above represents a compromise technical solution constrained by the current lack of native 3D foundation language models. However, from a long-term perspective, exploring how to achieve in-depth alignment between native 3D information and language through multi-stage training mechanisms or innovative architectural designs remains a more critical research direction in the field of embodiment.

- **Finding 8: Embodied Navigation methods can be derived by harnessing models with either end-to-end or agentic framework:** E2E frameworks show VLN-specialized models outperforming general models through enhanced embodied capabilities, while agentic frameworks achieve better performance via structured pipeline design to integrate extensible modular architecture and external knowledge especially for long-horizon tasks.

**Core Finding:** E2E frameworks show that VLN-specialized models demonstrate substantial performance advantages over general multimodal models through targeted architectural innovations and domain-specific training data. The top-performing VLN models — StreamVLN (Wei et al., 2025a) (54.90%), NaVILA (Zheng et al., 2024) (51.65%), and UniNavid (Zhang et al., 2024a) (47.85%) — achieve dramatically higher success rates compared to leading general models like Claude-3.7-Sonnet (Anthropic, 2024) (20.17%), Gemini-2.5-Pro (Google DeepMind, 2024) (27.37%), and GPT-4o (OpenAI, 2024) (21.67%). Notably, specialized VLN models dominate the entire top-5 rankings, with even mid-tier VLN models like MapNav (Zhang et al., 2025a) (36.15%) and Navid (Zhang et al., 2024b) (30.60%) outperforming most general foundation models. This performance gap is particularly pronounced in navigation-specific metrics, where StreamVLN (Wei et al., 2025a) achieves 56.90% on VLN-CE R2R compared to general models that struggle to exceed 25% success rates. And agentic frameworks demonstrate the potential to alleviate these built-in model limitations: specialized agentic frameworks like OmniEVA achieve top performance (59.10% on MP3D, 74.20% on HM3D) and OVRL (Yadav et al., 2023b) also achieve competitive results (62.00% on HM3D).

**In-depth Analysis:** This performance gap between VLN-specialized and general models stems from architectural and training differences addressing embodied navigation challenges. VLN-specific architectures demonstrate superior performance by leveraging historical frames rather than relying solely on current frames, e.g., models like NaVid (Zhang et al., 2024b) and UniNavid (Zhang et al., 2024a) allocate more tokens to current frames for improved decision accuracy. The combination of RxR and R2R datasets with Habitat simulation enables large-scale vision-language-action data construction for VLN, allowing effective supervised fine-tuning and strong validation performance on unseen splits. Building upon these capabilities,

designing efficient hierarchical agentic frameworks represents a promising direction for leveraging VLN capabilities. Such frameworks could decompose complex navigation tasks into subtasks, integrate multi-modal reasoning with spatial planning, and provide error recovery mechanisms. Agentic approaches activate and amplify existing foundation model capabilities through external reasoning pipelines rather than requiring costly model retraining or architectural modifications. This enables consistent performance enhancement through pipeline optimization, where new knowledge sources, memory architectures, and reasoning strategies can be systematically integrated without modifying the foundation model itself.

- **Finding 9: Embodied pointing is critical to both enhancing fundamental embodied capabilities and improving downstream embodied task performance. Supervised/reinforced fine-tuning for pointing not only significantly enhances performance on pointing tasks, but can also lead to improvements of fundamental embodied abilities. Pointing tasks under complex instructions remain a major challenge for most models.**

**Core Findings:** We evaluate the pointing capabilities of models on Where2Place (Yuan et al., 2024b) and VABench-P (Yuan et al., 2025a). The data shows that training on dedicated pointing data significantly boosts pointing performance. On the Where2Place benchmark, the top three performers—RoboBrain2.0 (BAAI, 2025), Embodied-R1 (Yuan et al., 2025b), and RoboRefer (Zhou et al., 2025) are all embodied models fine-tuned with pointing data. For instance, RoboBrain2.0-7B improved by 45.65% after incorporating data from the Spatial Referring Dataset (Zhou et al., 2025). However, performance diverges sharply on the more challenging VABench-P benchmark. While most top-tier models score above 60 on Where2Place, they fail to surpass 40 on VABench-P, with only Embodied-R1 (66) and Qwen-VL-Max (42) as notable exceptions. Moreover, models specifically optimized for pointing, such as RoboPoint (19.09) and Roborefer (4.62), also underperform significantly on this benchmark. More importantly, enhancing pointing capability appears to promote the model’s generalization on other tasks. Taking Embodied-R1 as an example, after RFT on a dataset containing partial spatial reasoning and pointing data, it achieved stable performance improvements across several OOD benchmarks, including OpenEQA (26.19  $\rightarrow$  34.51), ERQA (32.61  $\rightarrow$  35.24), and UniEQA (33.62  $\rightarrow$  38.14).

**In-depth Analysis:** Why do models optimized for pointing perform well on Where2Place but exhibit uncertainty on VABench-P? We posit that this discrepancy is primarily attributed to the inherent complexity of VABench-P. Compared to Where2Place, its tasks feature more intricate instructions and diverse scenes, requiring models to seamlessly integrate instruction understanding, spatial reasoning, and multimodal pointing capabilities. This integrated challenge reveals a critical trade-off: some embodied models, despite being fine-tuned on pointing data, appear to overfit. This specialization on specific pointing tasks may weaken their broader understanding and reasoning abilities, causing poor performance when faced with varied instructions or novel tasks. Fundamentally, the enhancement of general capabilities through pointing training is rooted in its role as a critical *grounding mechanism* for embodied AI. It compels the model to anchor abstract language to precise spatial coordinates, thereby serializing and integrating sub-tasks like perception, reasoning, and planning onto points. This anchoring process strengthens the model’s cognitive integration and boosts its generalization capabilities, as evidenced by Embodied-R1’s strong performance on several OOD benchmarks. Therefore, pointing tasks in complex environments are not merely about localization; they are intuitive, expressive, and provide the precise anchor points required for subsequent manipulation, making them an effective metric for evaluating multimodal understanding and reasoning (Cheng et al., 2025). In summary, mastering embodied pointing remains a crucial core capability that advanced embodied models must develop.

For complete and detailed discussions of the evaluation results on each benchmark, please refer to the

leaderboard page of Embodied Arena website<sup>1</sup>.

## 7. Conclusion

We introduce Embodied Arena, a comprehensive, unified, evolving evaluation platform and leaderboards for embodied AI models. It features three types of core embodied tasks, a diverse range of high-quality benchmarks, an LLM-driven automated evaluation data generation framework, and a systematic embodied capability taxonomy. Moreover, Embodied Arena offers professional support for advanced models and new benchmarks to join. With three types of real-time leaderboards and two evaluation views, Embodied Arena presents a multifaceted overview of embodied capabilities of advanced models. This offers a convenient way for researchers in both academia and industry to obtain useful insights and helps pinpoint critical research directions, thereby propelling the research progress in the field of Embodied AI. As the field evolves toward more sophisticated embodied agents, future extensions of the platform will incorporate more comprehensive manipulation tasks and closed-loop evaluation capabilities.

## 8. Contributions

This work represents a collaborative effort from researchers across multiple institutions worldwide. The authors are listed below:

**Contributors:** Fei Ni, Min Zhang, Pengyi Li, Yifu Yuan, Lingfeng Zhang, Yuecheng Liu, Peilong Han, Longxin Kou, Shaojin Ma, Jinbin Qiao, David Gamaliel Arcos Bravo, Yuening Wang, Xiao Hu, Zhanguang Zhang, Xianze Yao, Yutong Li, Zhao Zhang, Ying Wen, Ying-Cong Chen, Xiaodan Liang, Liang Lin, Bin He, Haitham Bou-Ammar, He Wang, Huazhe Xu, Jiankang Deng, Shan Luo, Shuqiang Jiang, Wei Pan, Yang Gao, Stefanos Zafeiriou, Jan Peters, Yuzheng Zhuang, Yingxue Zhang, Yan Zheng, Hongyao Tang, Jianye Hao.

The development of Embodied Arena involved contributions across multiple areas including benchmark integration, model evaluation infrastructure, automated data generation pipelines, capability taxonomy design, and comprehensive analysis. Each contributor brought expertise from their respective institutions to create this unified evaluation platform for Embodied AI.

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norrick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy,

<sup>1</sup>Please refer to the detailed discussions below the leaderboard tables in the website <https://embodied-arena.com>.



Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL <https://arxiv.org/abs/2404.14219>.

Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D’Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

Anthropic. Claude 3.7 Sonnet, 2024. URL <https://www.anthropic.com/news/claude-3-7-sonnet>.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

BAAI. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. *arXiv preprint arXiv:2405.01527*, 2024.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al.  $\pi_0$ : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

- Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura M. Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky.  $\pi_0.5$ : a vision-language-action model with open-world generalization. [arXiv preprint arXiv:2504.16054](#), 2025.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. [arXiv preprint arXiv:2307.15818](#), 2023a.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In *Robotics: Science and Systems*, 2023b.
- Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. [arXiv preprint arXiv:2406.13642](#), 2024.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.
- Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024a.
- Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.
- Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021.
- Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. [arXiv preprint arXiv:2405.10370](#), 2024b.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024c.

- An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Biyik, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453, 2024a.
- An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. arXiv preprint arXiv:2406.01584, 2024b.
- Long Cheng, Jiafei Duan, Yi Ru Wang, Haoquan Fang, Boyang Li, Yushan Huang, Elvis Wang, Ainaz Eftekhari, Jason Lee, Wentao Yuan, et al. Pointarena: Probing multimodal grounding through language-guided pointing. arXiv preprint arXiv:2505.09990, 2025.
- Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models, 2024c. URL <https://arxiv.org/abs/2311.15596>.
- Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In International Conference on Machine Learning, volume 202, pages 8469–8488, 2023.
- Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. arXiv preprint arXiv:2406.05756, 2024.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In European Conference on Computer Vision, pages 148–166. Springer, 2024.
- Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In Robotics: Science and Systems, 2024.
- Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. arXiv preprint arXiv:2411.05755, 2024. URL <https://jirl-upenn.github.io/VLMnav/>.
- Google DeepMind. Gemini 2.0 Flash, 2024. URL <https://deepmind.google/technologies/gemini/>.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.
- Peilong Han, Min Zhang, Hongyao Tang, YAN ZHENG, et al. Hule-nav: Human-like exploration for zero-shot object navigation via vision-language models. In NeurIPS 2024 Workshop on Behavioral Machine Learning, 2024.
- Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems, 36:20482–20494, 2023.

- Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. [arXiv preprint arXiv:2311.12871](#), 2023.
- Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world, 2024. URL <https://arxiv.org/abs/2311.12871>.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. [arXiv preprint arXiv:2412.16720](#), 2024.
- Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, Xinda Xue, Qinghang Su, Huaihai Lyu, Xiaolong Zheng, Jiaming Liu, Zhongyuan Wang, and Shanghang Zhang. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. [arXiv preprint arXiv:2502.21257](#), 2025.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. [arXiv preprint arXiv:2406.09246](#), 2024.
- Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. [arXiv preprint arXiv:2501.12599](#), 2025.
- Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In [European Conference on Computer Vision](#), 2020.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. [arXiv preprint arXiv:2408.03326](#), 2024a.
- Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. [arXiv preprint arXiv:2406.02537](#), 2024b.
- Zeju Li, Chao Zhang, Xiaoyan Wang, Ruilong Ren, Yifan Xu, Ruifei Ma, Xiangde Liu, and Rong Wei. 3dmit: 3d multi-modal instruction tuning for scene understanding. In [IEEE International Conference on Multimedia and Expo Workshops](#), pages 1–5, 2024c.
- Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. [arXiv preprint arXiv:2409.09788](#), 2024.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In [IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 26689–26699, 2024.
- Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. [Transactions of the Association for Computational Linguistics](#), 11:635–651, 2023.

- Jiajun Liu, Hao Ma, Ankur Jain, Xiang Li, Fan Chen, Ming-Chang Li, Tzu-Ming Lin, Liang Sun, Dong Yu, et al. Cosmos-Reason 1: From Physical AI Common Sense to Embodied Decisions. NVIDIA Research, 2025. URL [https://research.nvidia.com/publication/2025-03\\_cosmos-reason-1-physical-ai-common-sense-embodied-decisions](https://research.nvidia.com/publication/2025-03_cosmos-reason-1-physical-ai-common-sense-embodied-decisions).
- Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024.
- Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025.
- Liang Ma, Jiajun Wen, Min Lin, Rongtao Xu, Xiwen Liang, Bingqian Lin, Jun Ma, Yongxin Wang, Ziming Wei, Haokun Lin, Mingfei Han, Meng Cao, Bokui Chen, Ivan Laptev, and Xiaodan Liang. Phylblock: A progressive benchmark for physical understanding and planning via 3d block assembly. *arXiv preprint arXiv:2506.08708*, 2025.
- Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023.
- Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36:25081–25094, 2023.
- Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024.
- Jake O’Neill, Abraham Arthurs, Fábio Avila Belbute-Peres, Julian Balaguer, Sarah Bechtle, Gemma Bidoia, Kyle Burden, Erwin Chang, Sheila Chen, Todor Davchev, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.
- OpenAI. Gpt-3.5. Technical report, OpenAI, 2022. URL <https://platform.openai.com/docs/models/gpt-3-5>.
- OpenAI. GPT-4o Technical Report, 2024. URL <https://openai.com/research/gpt-4o>.
- OpenAI. o3 and o4min, 2025. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- OpenAI. Introducing openai o3 and o4-mini, 2025. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.



- Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. arXiv preprint arXiv:2501.01428, 2025a.
- Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. arXiv preprint arXiv:2506.17221, 2025b.
- Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. URL <https://arxiv.org/abs/2109.08238>.
- Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. arXiv preprint arXiv:2412.07755, 2024.
- Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In IEEE International Conference on Robotics and Automation, pages 645–652, 2024.
- Dhruv Shah, Michael Robert Equi, Błażej Osiniński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In Conference on Robot Learning, pages 2683–2699. PMLR, 2023. URL <https://sites.google.com/view/lfg-nav/>.
- BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoyu Lyu, Yijie Xu, Jiayu Shi, Mengfei Du, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Junkai Zhao, Xiaojie Zhang, Shanyu Rong, Huaihai Lyu, Zhengliang Cai, Yankai Fu, Ning Chen, Bolun Zhang, Lingfeng Zhang, Shuyi Zhang, Dong Liu, Xi Feng, Songjing Wang, Xiaodan Liu, Yance Jiao, Mengsi Lyu, Zhuo Chen, Chenrui He, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobrain 2.0 technical report, 2025. URL <https://arxiv.org/abs/2507.02029>.
- Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.
- Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. arXiv preprint arXiv:2507.05240, 2025a.
- Ziming Wei, Bingqian Lin, Zijian Jiao, Yunshuang Nie, Liang Ma, Yuecheng Liu, Yuzheng Zhuang, and Xiaodan Liang. Mineanybuild: Benchmarking spatial planning for open-world ai agents. arXiv preprint arXiv:2505.20148, 2025b.

- Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. [arXiv preprint arXiv:2401.02695](#), 2024.
- Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning, 2024a. URL <https://arxiv.org/abs/2404.16994>.
- Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. [arXiv preprint arXiv:2407.15208](#), 2024b.
- Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation, 2022. URL <https://arxiv.org/abs/2204.13226>.
- Karmesh Yadav, Arjun Majumdar, Ram Ramrakhya, Naoki Yokoyama, Alexei Baevski, Zsolt Kira, Oleksandr Maksymets, and Dhruv Batra. Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav, 2023a. URL <https://arxiv.org/abs/2303.07798>.
- Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. In [Workshop on Reincarnating Reinforcement Learning at ICLR 2023](#), 2023b.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. [arXiv preprint arXiv:2505.09388](#), 2025a.
- Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. [arXiv preprint arXiv:2502.13130](#), 2025b.
- Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. [arXiv preprint arXiv:2412.14171](#), 2024.
- Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multimodal large language models for vision-driven embodied agents. [arXiv preprint arXiv:2502.09560](#), 2025c.
- Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning. In [Computer Vision and Pattern Recognition Conference](#), pages 17294–17303, 2025d.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. [arXiv preprint arXiv:2408.01800](#), 2024.

- Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models, 2024. URL <https://arxiv.org/abs/2408.04840>.
- Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE, 2023. URL <http://dx.doi.org/10.1109/IROS55552.2023.10342512>.
- Shuaihang Yuan, Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, and Yi Fang. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance, 2024a. URL <https://arxiv.org/abs/2410.23978>.
- Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024b.
- Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*, 2025a.
- Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation, 2025b. URL <https://arxiv.org/abs/2508.13998>.
- Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024a.
- Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024b.
- Lingfeng Zhang, Yuening Wang, Hongjian Gu, Atia Hamidizadeh, Zhanguang Zhang, Yuecheng Liu, Yutong Wang, David Gamaliel Arcos Bravo, Junyi Dong, Shunbo Zhou, Tongtong Cao, Xingyue Quan, Yuzheng Zhuang, Yingxue Zhang, and Jianye Hao. Et-plan-bench: Embodied task-level planning benchmark towards spatial-temporal cognition with foundation models. *arXiv preprint arXiv:2410.14682*, 2024c.
- Lingfeng Zhang, Xiaoshuai Hao, Qinwen Xu, Qiang Zhang, Xinyao Zhang, Pengwei Wang, Jing Zhang, Zhongyuan Wang, Shanghang Zhang, and Renjing Xu. Mapnav: A novel memory representation via annotated semantic maps for vision-and-language navigation. *arXiv preprint arXiv:2502.13451*, 2025a. URL <https://arxiv.org/abs/2502.13451>.
- Min Zhang, Xian Fu, Jianye Hao, Hongyao Tang, Yan Zheng, and Peilong Han. Unieqa & unieval: A unified benchmark and evaluation platform for multimodal foundation models in embodied question answering, 2025b. URL <https://huggingface.co/datasets/TJURL-Lab/UniEQA>.
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024d.
- Qin Zhang and Mohsen Heidari. Quantum data sketches, 2025. URL <https://arxiv.org/abs/2502.13451>.

- Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15225–15236, October 2023.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. Transactions on Machine Learning Research, 2025c.
- Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In Robotics: Science and Systems, 2023.
- Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13624–13634, 2024.
- Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In Computer Vision and Pattern Recognition Conference, pages 8995–9006, 2025.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595–46623, 2023.
- Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308, 2025.
- Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation, 2023. URL <https://arxiv.org/abs/2301.13166>.
- Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. arXiv preprint arXiv:2409.18125, 2024.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.