NeurIPS 2025 Foundation Models Meet Embodied Agents Challenge: Systems, Analysis, and Lessons Learned

EAI Challenge EvalAI OpenReview NeurIPS Docs

Tianwei Bao[1*], Qineng Wang[1*], Kangrui Wang[1], Bryan Yu Zhou[5], Tony Lee[2], Li Erran Li[4], Weiyu Liu[2], Ruohan Zhang[1,2], Yejin Choi[2], Percy Liang[2], Jiayuan Mao[3], Li Fei-Fei[2†], Jiajun Wu[2†], Manling Li[1†]

[1]**Northwestern University** [2]**Stanford University** [3]**UPenn** [4]**Amazon** [5]**UCLA**

{TIANWEIBAO, QINENGW, KANGRUI.WANG}@U.NORTHWESTERN.EDU

MANLING.LI@NORTHWESTERN.EDU, {JIAJUNWU,FEIFEILI}@CS.STANFORD.EDU

[*] **Equal contribution.** [†] **Corresponding authors.**

Editors: Tao Qin, Kun Zhang, Jes Frellsen

Abstract

While Large Language Models (LLMs) have demonstrated impressive high-level reasoning capabilities, their deployment in embodied decision-making tasks remains hindered by a lack of standardized, diagnostic evaluations. We present a retrospective on the NeurIPS 2025 Embodied Agent Interface Challenge, which systematically benchmarked LLMs across four critical reasoning modules: goal interpretation, subgoal decomposition, action sequencing, and transition modeling. Hosted on the BEHAVIOR and VirtualHome simulators, the competition departed from traditional, binary task-success metrics. Instead, it leveraged Linear Temporal Logic (LTL) to evaluate alternative viable trajectories and introduced fine-grained metrics to pinpoint specific failures, such as affordance violations, hallucinations, and missing preconditions. The competition attracted significant engagement globally, with 58 registered teams developing diverse solutions. An analysis of the winning and most innovative submissions reveals a clear trend: while hybrid neuro-symbolic pipelines and test-time compute significantly enhance trajectory feasibility, spatial relationship modeling remains a pervasive bottleneck for current LLMs. In this report, we detail the competition design, summarize the top-performing methodologies, and present key lessons learned to guide the future development of robust, interpretable embodied AI systems.

Keywords: Embodied Decision Making, Embodied Agents, Reasoning and Planning, Physical State Change, Large Language Models (LLMs)

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in zero-shot reasoning, tool use, and high-level task planning (Huang et al., 2022a; Yao et al., 2023; Kojima et al., 2022; Hao et al., 2023; Schick et al., 2023; Shen et al., 2023). Over the past few years, a surge of research has attempted to harness these capabilities for embodied agents, utilizing LLMs to translate human instructions into robot policies (Ahn et al., 2022; Huang et al., 2022b; Brohan et al., 2023; Huang et al., 2023; Singh et al., 2023; Liang et al., 2023), guide open-ended exploration (Wang et al., 2023a,b,c, 2024; Li et al., 2025), and leverage environment feedback for iterative replanning (Huang et al., 2022c; Liu et al., 2023b; Rana et al., 2023; Shinn et al., 2023; Skreta et al., 2024; Mei et al., 2024). However, deploying

© 2026 T. Bao *et al.*

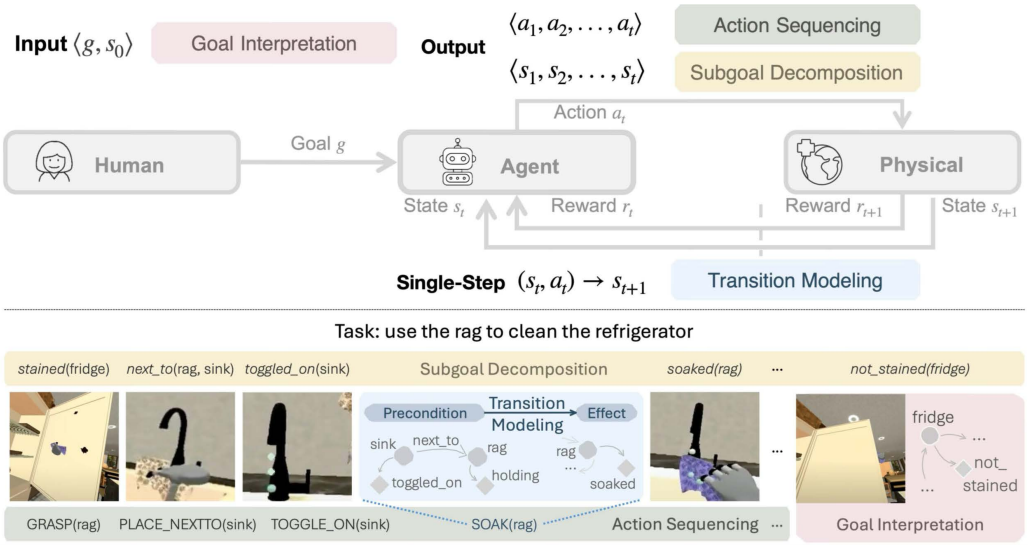


Figure 1: We formalize and standardize the evaluation of LLMs for embodied decision-making within a Markov Decision Process (MDP) framework.

these models as embodied agents in physical or simulated environments remains profoundly challenging. While LLMs excel at generating plausible, human-like text, they frequently struggle to ground their reasoning in the physical constraints of the real world (Bisk et al., 2020; Cherian et al., 2024; Ghaffari and Krishnaswamy, 2024). For instance, an LLM might generate a flawless textual recipe for cooking a meal but fail to recognize that a robot cannot grasp an object inside a closed refrigerator without first opening the door, or it may struggle with spatial relationships and geometric reasoning (Liu et al., 2023a).

Despite the growing intersection of natural language processing and robotics, progress is hampered by a lack of diagnostic evaluation frameworks. Existing embodied AI benchmarks (Anderson et al., 2018; Savva et al., 2019; Shridhar et al., 2020; Deitke et al., 2020; Padmakumar et al., 2022) typically evaluate systems end-to-end, relying heavily on a binary task success rate. Even more recent benchmarks aimed specifically at LLM planners (Deng et al., 2023; Valmeekam et al., 2023; Xie et al., 2024; Liu et al., 2024; Zhou et al., 2024; Qin et al., 2024; Choi et al., 2024) often focus solely on the final action sequence generation without diagnosing the underlying planning logic. When an agent fails to complete a long-horizon task, a single success metric obscures the root cause of the failure: Did the model misinterpret the human instruction? Did it fail to break the goal down into viable subgoals? Did it hallucinate an object that does not exist in the scene? Or did it violate basic physical affordances? Without granular feedback, researchers are left in the dark about how to selectively improve LLM reasoning.

To address these limitations, we organized the NeurIPS 2025 Embodied Agent Interface Challenge. The primary scientific objective of this competition was to answer fundamental questions about LLM capabilities: *How do LLMs reason under embodiment constraints, where exactly do they fail, and how can we systematically evaluate and improve their modular capabilities?*

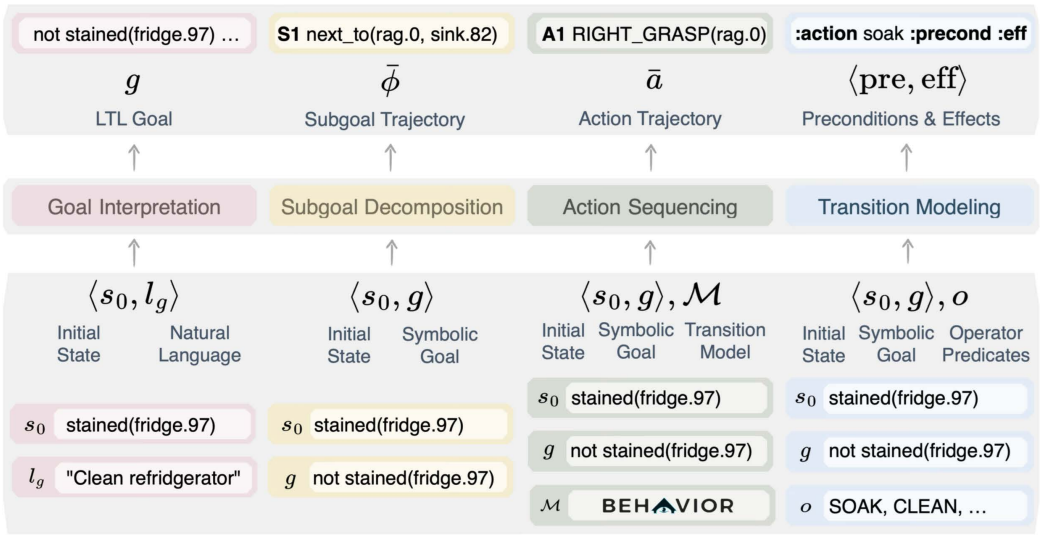


Figure 2: Four evaluated ability modules in the EAI Challenge.

Building upon the framework proposed in the Embodied Agent Interface (Li et al., 2024), we standardized the evaluation of embodied decision-making using a Markov Decision Process (MDP) formulation as shown in Figure 1. The challenge decomposed the planning problem into four critical, testable modules: **Goal Interpretation**, **Subgoal Decomposition**, **Action Sequencing**, and **Transition Modeling**.

To ensure rigorous and scalable evaluation, the competition was hosted across two diverse simulation environments: **VirtualHome** (Puig et al., 2018), which emphasizes high-level symbolic planning in household scenarios, and **BEHAVIOR** (Srivastava et al., 2022), which features complex, temporally extended goals grounded in realistic physics. A core novelty of this competition was the use of Linear Temporal Logic (LTL) (Pnueli, 1977) to formalize both state-based and temporally dependent goals. Unlike rigid sequence matching, LTL allows the evaluation pipeline to recognize and reward alternative, valid paths to task completion. Furthermore, we introduced fine-grained evaluation metrics that automatically categorize reasoning errors, providing an unprecedented level of diagnostic feedback to participants.

In this retrospective report, we summarize the outcomes of the NeurIPS 2025 Embodied Agent Interface Challenge, highlight key lessons learned, and outline promising directions for future research in the embodied AI community.

2. Competition Design: Data, Tasks, and Baselines

The core design philosophy of the Embodied Agent Interface Challenge was to move away from monolithic, end-to-end evaluations. Instead, we formulated embodied decision-making as a standardized, modular Markov Decision Process (MDP). This allowed us to isolate, evaluate, and diagnose specific reasoning abilities of LLMs using fine-grained metrics.

2.1. Task Formulation and Interface

We defined an embodied decision-making problem as a tuple $\langle \mathcal{U}, \mathcal{S}, \mathcal{A}, g, \phi, \bar{a} \rangle$, representing objects, states, actions, goals, subgoals, and action trajectories, respectively. Within this

framework, participants were tasked with developing or fine-tuning models to solve one or more of the following four ability modules shown in Figure 2:

- **Goal Interpretation ($\mathcal{G}$):** Given an initial state s_0 and a natural language instruction l_g, the model must output a formal goal specification g in Linear Temporal Logic (LTL) (Pnueli, 1977).
- **Subgoal Decomposition (Φ):** Given s_0 and an LTL goal g, the model generates a temporally ordered sequence of intermediate LTL subgoals $\bar{\phi} = \{\phi_i\}_{i=1}^{k}$.
- **Action Sequencing ($\mathcal{Q}$):** Given s_0, goal g, and a transition model $\mathcal{M}$, the LLM acts as a planner to generate an executable action sequence $\bar{a} = \{a_i\}_{i=1}^{n}$.
- **Transition Modeling ($\mathcal{T}$):** Given s_0, g, and an action operator o, the model predicts the preconditions and post-effects $\langle pre, eff \rangle$ in PDDL (Planning Domain Definition Language) format (McDermott et al., 1998).

2.2. Datasets and Simulators

The competition utilized two complementary simulation environments to ensure diverse and robust evaluation:

- **VirtualHome (Puig et al., 2018):** Emphasizes high-level symbolic planning in diverse household scenarios. Tasks in VirtualHome have an average action sequence length of 8.76 steps, requiring broad semantic understanding of everyday objects.
- **BEHAVIOR (Srivastava et al., 2022):** Features low-level manipulation tasks grounded in realistic physics. BEHAVIOR tasks are significantly longer (averaging 14.6 steps) and contain complex, quantified goal conditions expressed in the BEHAVIOR Domain Definition Language (BDDL).

To prevent data contamination, we split the dataset into a publicly released validation set used during the Development Phase and a hidden test set used for the Final Evaluation Phase. The hidden test set included six newly annotated scenarios built exclusively for this competition, comprising 4,745 held-out task instructions, over 3,200 LTL goals, and nearly 12,000 action steps.

	Validation	**Holdout**
#task instruction	438	4745
#goal	1474	3204
- #state	493	1360
- #relation	819	1196
- #action	162	648
#trajectory	438	4745
- #step	4420	11840
- avg. step	10.09	8.76

Table 1: Dataset statistics.

2.3. Fine-Grained Evaluation Metrics

A major contribution of this challenge was the introduction of LTL to evaluate alternative trajectories and fine-grained metrics to capture specific reasoning bottlenecks:

- **Goal Evaluation:** Evaluated via an F1 logic matching score, assessing whether generated LTL predicates accurately capture *State Goals* (e.g., `is_open`), *Relation Goals* (e.g., `next_to`), and *Action Goals* (e.g., `touch`).

- **Trajectory Feasibility:** Used a simulated environment checker to detect both *Grammar Errors* (parsing failures, hallucinated objects) and *Runtime Errors*. Runtime errors were further categorized into *Missing Steps*, *Additional Steps*, *Wrong Order*, and *Affordance Violations* (e.g., trying to grasp an un-grabbable object).
- **Transition Logic Accuracy & Planner Success:** For transition modeling, models were scored using bipartite graph matching on predicted PDDL logic forms, followed by a Planner Success Rate metric which tested if a symbolic planner could successfully utilize the LLM-generated rules.

2.4. Baselines and Starter Kit

To lower the barrier to entry and establish competitive performance ceilings, the organizing team provided a comprehensive starter kit[1] with detailed instructions[2]. The kit included data loaders, the local evaluation API, prompt templates, and sample submissions generated by the `Qwen3-14B` model (Yang et al., 2025). During our preliminary baseline testing, we observed a wide spread in capabilities. For instance, among proprietary models, `o1-preview` (OpenAI, 2024) established the highest average performance baseline (74.9% on BEHAVIOR), demonstrating strong capabilities in test-time reasoning for subgoal decomposition and action sequencing. `Claude-3.5-Sonnet` (Anthropic, 2024) showed particular strength in goal interpretation. Among open-weight baselines, `Llama-3-70B` (Llama Team et al., 2024) and `Mixtral-8x22B` (Mistral AI, 2024) proved highly competitive but struggled severely with spatial relation transitions and long-horizon affordance tracking. These baselines provided reference points that participants were challenged to surpass through innovative methods.

3. Participation Statistics and Organization

To ensure a seamless, accessible, and rigorous competition, the challenge was hosted on the EvalAI platform[3] (Yadav et al., 2019) to encourage broad participation while mitigating risks of data contamination and leaderboard overfitting.

Demographics. The challenge attracted substantial interest from the natural language processing, computer vision, and robotics communities. In total, 342 participants from 24 countries or regions registered for the competition, forming 58 active teams. The participant pool exhibited a balanced composition, with 76% affiliated with academic institutions and 24% from industry research laboratories.

Engagement. To facilitate communication and collaboration, the organizing committee maintained an active Slack workspace[4] for announcements, discussions, and issue tracking. Over the course of the competition, more than 4,500 submissions were evaluated across the four ability modules, reflecting the intensive iterative development efforts of participating teams, particularly those achieving top performance.

4. Winning Solutions and Innovative Approaches

1. https://drive.google.com/file/d/1d-SfGfp109NjRVhY8tFWrNu_KyWpLJbB/view?usp=sharing
2. https://neurips25-eai.github.io/participate
3. https://eval.ai/web/challenges/challenge-page/2621/overview
4. https://join.slack.com/t/eaichallengen-xzk8506/shared_invite/zt-3cm0dms1t-tvylNsamaZQgC03k0kBB9g

Model	Goal Interpretation		Action Sequencing				Subgoal Decomposition				Transition Modeling				Average Perf.	
	F_1		*Task SR*		*Execution SR*		*Task SR*		*Execution SR*		F_1		*Planner SR*		*Module SR*	
	V	*B*	*V*	*B*	*V*	*B*	*V*	*B*	*V*	*B*	*V*	*B*	*V*	*B*	*V*	*B*
AxisTilted2	**65.4**	**99.6**	**82.6**	**98.0**	**94.3**	**99.0**	78.7	**97.0**	91.2	**98.0**	**100.0**	**100.0**	99.8	**99.0**	**81.7**	**98.5**
SingaX	64.5	86.2	81.9	85.0	92.0	91.0	79.3	79.0	89.1	86.0	99.5	98.9	**99.9**	**99.0**	81.4	87.3
CtrlAct	48.2	83.2	71.6	85.0	83.6	96.0	73.3	96.0	89.9	96.0	95.1	97.7	97.8	97.0	72.4	90.4
nju-lamda12	57.9	85.0	73.5	81.0	84.9	91.0	74.7	80.0	86.6	88.0	98.8	99.8	**99.9**	**99.0**	76.4	86.3
*Gemini-2.5-Pro**	*37.8*	*85.4*	*67.6*	*72.0*	*81.9*	*80.0*	*71.3*	*59.0*	*88.3*	*65.0*	*45.1*	*58.1*	*91.0*	*96.0*	*61.2*	*73.4*

Table 2: Final evaluation results of winning teams. * denotes the zero-shot baseline.

The challenge received a diverse array of submissions, ranging from fully autonomous closed-loop agents to highly optimized, distilled lightweight models. A recurring theme across the top submissions was the recognition that frontier LLMs do not lack general knowledge, but rather fail due to a human-simulator mismatch, a failure to adhere to strict physical constraints, formatting rules, and implicit state dependencies. To bridge this gap, the winning teams employed techniques spanning evaluator-guided distillation, iterative prompt induction, linguistic rule enforcement, and test-time reflection.

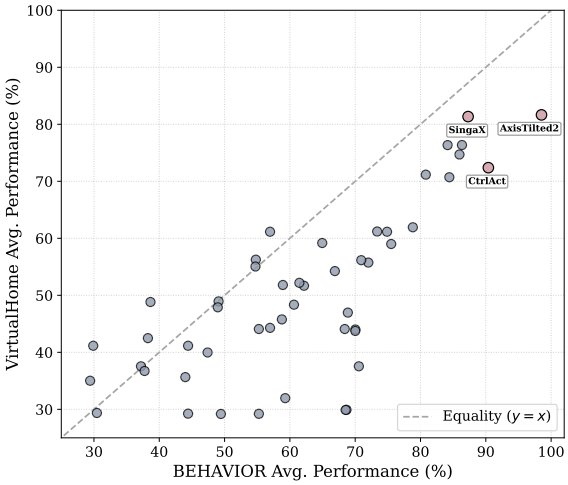


Figure 3: The overwhelming majority of models cluster beneath this line, highlighting the systematic difficulty gap of VirtualHome.

4.1. First Place: Evaluator-Guided LLM Distillation (Team AxisTilted2)

Team AxisTilted2 (Pradeep and Sreekala, 2025) secured first place by demonstrating that small, task-specialized open-source models can match or outperform frontier LLMs when trained on high-quality, evaluator-verified data. Their approach treated each EAI module as a supervised sequence-to-sequence task, utilizing a two-stage "evaluator-in-the-loop" distillation pipeline.

Evaluator-in-the-Loop Data Construction. Instead of relying solely on ground-truth annotations, the team generated synthetic training data using a cyclic refinement process. A frontier model (e.g., GPT-5-mini) generated an initial candidate output (such as an action sequence or subgoal list). This output was then scored by the official EAI local evaluator. If the evaluator returned errors (e.g., missing preconditions, affordance violations), the error log was appended to the prompt, and the model was asked to refine its answer. This process was repeated up to eight times, yielding a highly robust dataset of prompts and verified correct responses.

Specialized Distillation and Scaffolding. Using the curated data, the team fine-tuned various Qwen3 models using QLoRA (Dettmers et al., 2023). Their results revealed a striking difference between the two simulator environments:

- *BEHAVIOR:* Because BEHAVIOR tasks are strictly bounded, a remarkably small Qwen3-0.6B model, fine-tuned on compact prompt templates, nearly saturated the benchmark (achieving 98.0% task success on Action Sequencing and 99.5% on Transition Modeling).

- *VirtualHome:* Due to higher semantic diversity and longer horizons, VirtualHome required larger models and advanced scaffolding. The team fine-tuned a Qwen3-32B model using Domain Adaptation (Gururangan et al., 2020) and Cross-Task Learning (Pfeiffer et al., 2020). Furthermore, they introduced a *learned LLM-as-an-evaluator*, a Qwen3-32B model explicitly trained on the trajectory of corrections from the data generation phase to judge and refine candidate outputs at inference time.

AxisTilted2's success proved that carefully internalizing the evaluator's implicit criteria via Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) is vastly more compute-efficient at inference time than complex agentic scaffolding. Their code and models are available on GitHub[5,6,7] and HuggingFace[8].

4.2. Second Place: Iterative Prompt Induction (Team SingaX)

Team SingaX (Niu et al., 2025) took a training-free approach, achieving second place by focusing on automated prompt optimization and multi-model consensus. Their core insight was that LLM reasoning failures are highly systematic and can be preemptively avoided by injecting explicit "safeguards" derived from historical error logs into the system prompt.

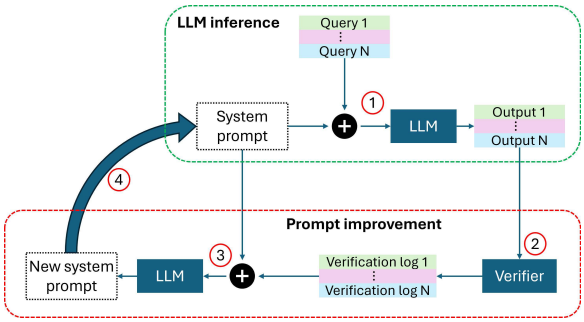


Figure 4: Iterative Prompt Induction Framework proposed by SingaX.

Safeguard Generation via Prompt Induction. The team developed an iterative prompt induction framework illustrated in Figure 4. During the development phase, baseline solutions were evaluated using the EAI local verifier to generate detailed JSON error logs. These logs were fed into a secondary "Safeguard Generation LLM" utilizing GPT-5. This meta-agent was tasked with synthesizing high-level, generalizable rules from the specific failures. For example, if the logs showed frequent affordance errors regarding closed containers, the Safeguard LLM generated the rule: *"If CLOSED: OPEN before interacting"*. These guardrails were then injected into the system prompt for the primary answering LLM. This method successfully prevented the model from repeating known logical inconsistencies without requiring weight updates.

Multi-Model Best-of-N (BoN). To further enhance robustness, SingaX deployed a heterogeneous Best-of-N strategy. Rather than oversampling a single model, they queried three distinct model families (Qwen, Gemini, and GPT) using the safeguard-optimized prompts. This created a diverse candidate pool. A strong LLM verifier was then prompted to score each candidate based on syntactic validity, environment constraints, and goal plausibility. By selecting the highest-scoring candidate across different model families, the team effectively mitigated the correlated failure modes typically seen in single-model sampling.

4.3. Third Place: Linguistic Guidance and Domain Alignment (Team CtrlAct)

5. https://github.com/CinnamonRolls1/inference-central
6. https://github.com/CinnamonRolls1/eai-extractor
7. https://github.com/spsanps/eai-experiments
8. https://huggingface.co/AxisTilted2

Team CtrlAct (Xiao et al., 2025) secured third place by systematically decomposing the embodied planning pipeline into linguistic grounding and procedural reasoning, applying targeted interventions to each component.

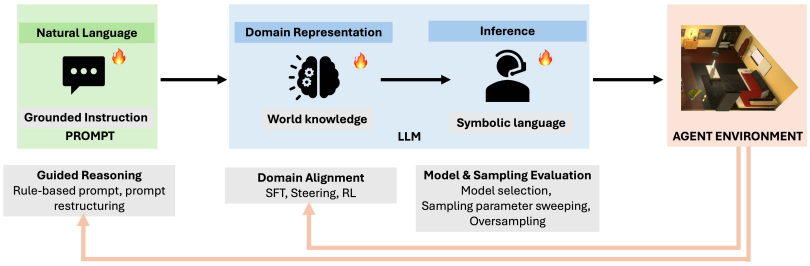


Figure 5: The action generation pipeline proposed by CtrlAct.

Rule-Based Linguistic Guidance. CtrlAct observed that errors in Goal Interpretation frequently stemmed from linguistic ambiguity rather than poor spatial reasoning. Through error analysis, they designed rule-based linguistic prompts to handle edge cases, such as distinguishing between telic and atelic verbs (e.g., recognizing that the instruction "wash clothes" does not necessarily entail the end state "clothes are clean" depending on task boundaries). By explicitly mapping surface linguistic forms to the structured task ontology in the prompt, they significantly reduced hallucination and misinterpretation errors.

Domain Alignment vs. Planning. For procedural tasks, the team contrasted Supervised Fine-Tuning (Ouyang et al., 2022) with PPO (Schulman et al., 2017) and Activation Steering (Zou et al., 2023). They found that SFT was highly effective for Transition Modeling where the model learns static domain-specific knowledge, achieving near-perfect accuracy with minimal data. However, for long-horizon planning, SFT improvements were marginal. Crucially, CtrlAct reported that applying RL and Activation Steering degraded sequential reasoning performance. RL destabilized the model's temporal ordering, indicating that simple reward-shaping is insufficient for the complex causal dependencies of realistic embodied planning. Instead, they relied heavily on a standard Best-of-N oversampling strategy, which proved highly effective for the Action Sequencing module, raising success rates from 79% to 85% simply by scaling inference compute. Their code is available on GitHub[9].

4.4. Most Innovative Approach: Re^2 Agent (Team nju-lamda12)

The Most Innovative Approach award was granted to Team nju-lamda12 (Chen et al., 2025) for their **Reflective and Re-execution (Re^2) Agent**. While other teams primarily used the local evaluator offline to generate synthetic training data or tune prompts, nju-lamda12 built a closed-loop, interactive agent architecture that mimics trial-and-error learning during task execution, as shown in Figure 6.

Knowledge-Driven Prompting and Checklists. The Re^2 Agent initializes with a strong knowledge-driven prompt. For Subgoal Decomposition, they introduced a "Critical Preconditions Checklist" that forces the LLM to explicitly check container accessibility, tool activation, and instrument requirements before proposing a step.

Reflection and Re-execution. The core innovation lies in the agent's interactive loop. Upon generating a candidate action sequence, the agent simulates execution. If an error occurs (e.g., a missing precondition), the agent receives the environmental feedback, enters a *Reflection* phase to abstract the failure into a temporal logic rule, and *Re-executes* the task with the newly injected constraint. For instance, in Transition Modeling, their verifier-

9. https://github.com/omics-ai/ctrlact

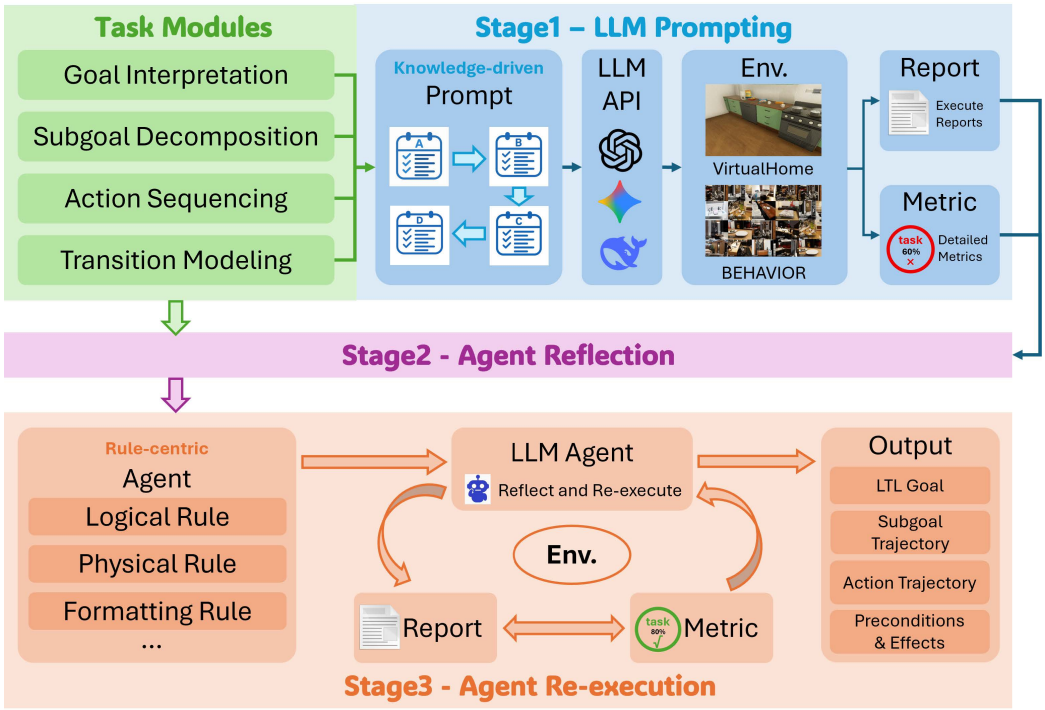


Figure 6: The Re^2 Agent framework proposed by nju-lamda12.

corrector explicitly repairs illegal quantifications (e.g., swapping `exists` for `forall`) and prunes hallucinated objects on the fly.

By dynamically balancing environmental feedback with task constraints, the Re^2 Agent achieved an impressive 86.35% overall score on the BEHAVIOR benchmark. This approach successfully highlighted the necessity of moving beyond static, single-shot prompting toward agents capable of autonomous, online error recovery. Their code is available on GitHub[10].

5. Lessons Learned

The NeurIPS 2025 Embodied Agent Interface Challenge provided a rigorous testing ground for evaluating how Large Language Models handle physical constraints, causal dependencies, and spatial reasoning. By analyzing the successes and failures across all participating teams, several critical lessons emerged for the broader embodied AI community.

Test-Time Compute is Essential for Embodied Planning. A consensus across the top submissions is that LLMs are inherently weak at single-shot, autoregressive planning over long horizons. However, they are exceptionally strong at *verifying* plans and reflecting on errors. Teams that utilized test-time compute, such as multi-model Best-of-N sampling (Team SingaX) or interactive reflection and re-execution (Team nju-lamda12), drastically outperformed those relying on zero-shot generation. The ability to simulate a trajectory, catch missing preconditions, and iteratively refine the output is now a mandatory paradigm for reliable embodied reasoning.

10. https://github.com/chenyang126/Re-2-Agent

Environment's Physics Should Be a Center Focus of Finetuning. Team AxisTilted2 demonstrated that massive parameter counts are not necessary for embodied planning if the model is fine-tuned on high-quality, evaluator-verified data. A 0.6B parameter model successfully saturated the BEHAVIOR benchmark modules by internalizing the simulator's specific constraints. However, as Team CtrlAct noted, Supervised Fine-Tuning is highly effective for *domain knowledge* (e.g., Transition Modeling) but scales poorly to complex, dynamic *planning* (e.g., Subgoal Decomposition), where test-time search remains superior.

The Human-Simulator Mismatch Requires Explicit Safeguards. LLMs are trained on human text, which relies heavily on conversational implicature. A human-level instruction says "put the milk in the fridge", omitting the steps such as "walk to the fridge", "open the door", and "ensure your hand is free". Simulators and real-world robots, however, require these explicit intermediate steps. Teams found that injecting explicit safeguards and physical checklists into the system prompt effectively forced the LLMs to bridge this semantic gap, drastically reducing affordance and missing-step errors.

Standard Language-Centric Alignment Techniques Struggle in Embodied Contexts. Methods that traditionally improve standard NLP benchmarks such as naive RL via reward shaping or Activation Steering frequently degraded performance in embodied tasks. Team CtrlAct found that RL interventions often destabilized temporal coherence and precondition ordering. This suggests that the causal dependencies in physical environments are too complex for simple scalar reward signals, necessitating richer, hierarchical supervision or structured symbolic feedback.

6. Future Challenges and Improvements

While the competition revealed major advances in symbolic embodied reasoning, it also exposed the limits of current paradigms. Pushing toward truly generalist robotic agents will require future benchmark iterations to address the following challenges.

From Abstract Symbols to Multimodal Inputs. The current Embodied Agent Interface operates on text-based, object-centric abstractions (e.g., a JSON list of scene objects and relations). Real-world deployment requires agents to perceive the environment directly. The next evolution of this challenge must integrate Vision-Language Models (VLMs) and Vision-Language-Action (VLA) architectures, requiring agents to interpret raw egocentric visual streams, ground semantic concepts in pixels, and maintain spatial memory without relying on an oracle state text prompt.

Transitioning to Continuous Control. Currently, the action space consists of discrete, symbolic operations (e.g., `GRASP(apple)`). To bridge the sim-to-real gap, the interface must evolve to support continuous kinematic control. Future tasks should evaluate whether an LLM/VLM can output parameter-rich action primitives (e.g., 7-DoF end-effector poses) or successfully orchestrate a low-level continuous control policy.

Handling Stochasticity and Noisy Environments. The current simulators execute valid symbolic actions with 100% success. In reality, actions fail probabilistically in scenarios such as objects slipping, doors jamming, and navigation paths becoming obstructed. Future benchmarks must introduce environmental stochasticity, testing an agent's ability to monitor its own execution, recognize physical failures, and dynamically replan on the fly without user intervention.

Acknowledgments

We gratefully acknowledge AIX for sponsoring the NeurIPS 2025 Embodied Agent Interface Challenge. We extend our sincere thanks to all participants for their innovative contributions. We also thank the creators of the BEHAVIOR and VirtualHome simulators, whose foundational work made this benchmark possible.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2022.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2018.

Anthropic. Claude 3.5 sonnet model card addendum, 2024. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mark Ellison, Juliet Farahmand, Michel Friedman, Ignacio Glass, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

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*, 2023.

Yang Chen, Hong-Jie You, Jie-Jing Shao, Xiao-Wen Yang, Ming Yang, Yu-Feng Li, and Lan-Zhe Guo. Re^2 Agent: Reflection and Re-execution Agent for Embodied Decision Making. In *NeurIPS 2025 Challenge on Foundation Models for Embodied Agents*, 2025.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models. *arXiv*, 2024. doi: 10.48550/arxiv.2411.08027.

Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. LOTA-bench: Benchmarking language-oriented task planners for embodied agents. In *International Conference on Learning Representations (ICLR)*, 2024.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. RoboTHOR: An open simulation-to-real embodied AI platform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3164–3174, 2020.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sadaf Ghaffari and Nikhil Krishnaswamy. Exploring failure cases in multimodal reasoning about physical dynamics. *Proceedings of the AAAI Symposium Series*, 3:105–114, 2024. doi: 10.1609/aaaiss.v3i1.31189.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, 2020.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. *arXiv*, 2023. doi: 10.48550/arxiv.2305.14992.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*. PMLR, 2022a.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv*, 2022b. doi: 10.48550/arxiv.2207.05608.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning (CoRL)*, pages 1769–1782. PMLR, 2022c.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv*, 2022. doi: 10.48550/arxiv.2205.11916.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. In *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Yaoru Li, Shunyu Liu, Tongya Zheng, and Mingli Song. Parallelized planning-acting for efficient llm-based multi-agent systems. *arXiv*, 2025. doi: 10.48550/arxiv.2503.03505.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE, 2023.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qian Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. Llm+p: Empowering large language models with optimal planning proficiency. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023a.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. AgentBench: Evaluating LLMs as agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724*, 2023b.

Llama Team, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Drew McDermott, Malik Ghallab, Adele Howe, Craig Knoblock, Ashwin Ram, Manuela Veloso, Daniel Weld, and David Wilkins. PDDL-the planning domain definition language. *Technical Report CVM-98-003/DCS-TR-1165, Yale Center for Computational Vision and Control*, 1998.

Aoran Mei, Guo-Niu Zhu, Huaxiang Zhang, and Zhongxue Gan. Replanvlm: Replanning robotic tasks with visual language models. *IEEE Robotics and Automation Letters*, 2024.

Mistral AI. Cheaper, better, faster, stronger: Mixtral 8x22b, 2024. URL https://mistral.ai/news/mixtral-8x22b/.

Xinyuan Niu, Zhiliang Chen, Vernon Toh, Yanchao Li, Zhengyuan Liu, and Nancy F Chen. Technical report: Team singax for embodied agent interface challenge@ neurips 2025. In *NeurIPS 2025 Challenge on Foundation Models for Embodied Agents*, 2025.

OpenAI. Openai o1 system card, 2024. URL https://cdn.openai.com/o1-system-card-20241205.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. TEACh: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025, 2022.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, 2020.

Amir Pnueli. The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (SFCS 1977)*, pages 46–57. IEEE, 1977.

Chinmayan Pradeep and Sanjayan Pradeep Kumar Sreekala. Evaluator-guided llm distillation for embodied agent decision-making. In *NeurIPS 2025 Challenge on Foundation Models for Embodied Agents*, 2025.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8494–8502, 2018.

Yujia Qin, Shihao Shi, Zhiyong Qi, Binyu Zhao, Lei Huang, Xin Ma, Qiao Qi, Star Hwang, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Michael Milford, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9339–9347, 2019.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv*, 2023. doi: 10.48550/arxiv.2302.04761.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *arXiv*, 2023. doi: 10.48550/arxiv.2303.17580.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749, 2020.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Marta Skreta, Zihan Zhou, Jia Lin Yuan, Kourosh Darvish, Alán Aspuru-Guzik, and Animesh Garg. Replan: Robotic replanning with perception and language models. *arXiv preprint arXiv:2401.04157*, 2024.

Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning (CoRL)*, pages 477–490. PMLR, 2022.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zhengliang Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu, and Shu Zhang. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv*, 2024. doi: 10.48550/arxiv.2401.04334.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv*, 2023b. doi: 10.48550/arxiv.2302.01560.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, Xiaojian Ma, and Yitao Liang. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv*, 2023c. doi: 10.48550/arxiv.2311.05997.

Qingyang Xiao, Bo Su, Ling Sun, Zhu Zhu, and Thai Le. Ctrlact: Grounding llms to bridge the gap between embodied instruction and action. In *NeurIPS 2025 Challenge on Foundation Models for Embodied Agents*, 2025.

Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. TravelPlanner: A benchmark for real-world planning with language agents. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Deshraj Yadav, Rishabh Jain, Harsh Agrawal, Prithvijit Chattopadhyay, Taranjeet Singh, Akash Jain, Shiv Singh, Stefan Lee, and Dhruv Batra. EvalAI: Towards better evaluation

systems for ai agents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 133–138, 2019.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Shuyan Zhou, Frank F Xu, Hao Yao, Zhiwei Liu, Fangxiaoyu Tai, Evia Weinberger, Mengdi Wang, Ruohao Zhang, Gabriel Shum, Huan Zhang, et al. WebArena: A realistic web environment for building autonomous agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.