Environmental Understanding Vision-Language Model for Embodied Agent

Jinsik Bang Jaeyeon Bae Donggyu Lee Siyeol Jung Taehwan Kim
UNIST
{bang, qowodussla, leedongkyu2019, siyeol, taehwankim}@unist.ac.kr
https://eu-ea.github.io

Abstract

Vision-language models (VLMs) have shown strong perception and reasoning abilities for instruction-following embodied agents. However, despite these abilities and their generalization performance, they still face limitations in environmental understanding, often failing on interactions or relying on environment metadata during execution. To address this challenge, we propose a novel framework named Environmental Understanding Embodied Agent (EUEA), which fine-tunes four core skills: 1) object perception *for identifying relevant objects, 2)* task planning *for generating interaction subgoals, 3)* action understanding *for judging success likelihood, and 4)* goal recognition *for determining goal completion. By fine-tuning VLMs with EUEA skills, our framework enables more reliable task execution for instruction-following. We further introduce a recovery step that leverages these core skills and a group relative policy optimization (GRPO) stage that refines inconsistent skill predictions. The recovery step samples alternative actions to correct failure cases, and the GRPO stage refines inconsistent skill predictions. Across ALFRED tasks, our VLM significantly outperforms a behavior-cloning baseline, achieving an 8.86% improvement in average success rate. The recovery and GRPO stages provide an additional 3.03% gain, further enhancing overall performance. Finally, our skill-level analyses reveal key limitations in the environmental understanding of closed- and open-source VLMs and identify the capabilities necessary for effective agent–environment interaction.*

1. Introduction

Environmental understanding is a core capability for embodied agents, enabling them to perceive, interpret, and interact with the environment to achieve given tasks. Recently, large language models (LLMs) and vision-language models (VLMs) have demonstrated impressive performance in the field of embodied agents [2, 7, 24, 38, 39], showing strong understanding and reasoning capabilities. However, despite their demonstrated abilities and generalization performance, they still face limitations in achieving environmental understanding. Some prior embodied agent approaches [11, 12] rely on complex modular pipelines that require separate components, and other work [43] depends on environment metadata (e.g., object IDs, masks) to execute actions. On the other hand, end-to-end models often lack explicit environmental interpretation capabilities [36–38]. As a result, execution errors lead to task failures. To overcome these failures, existing recovery methods rely on the environment to automatically validate outcomes [34] and predefined failure types [10], or they depend on textual feedback [30], which lacks the visual grounding essential for embodied agents.

To address these limitations, we propose the Environmental Understanding Embodied Agent (EUEA), a novel framework that bridges explicit skill modeling with end-to-end learning. Unlike approaches requiring complex module combinations, EUEA internalizes four core skills within a single VLM, without separate module modeling: 1) *object perception* for identifying relevant objects, 2) *task planning* for generating interaction subgoals, 3) *action understanding* for judging success likelihood, and 4) *goal recognition* for determining goal completion. This explicit skill-level supervision not only enables the model to perform visual perception and decision-making within a single architecture but also provides interpretability of its capabilities. Leveraging this unified formulation, we introduce a sampling-based recovery step without additional training that corrects failed interactions. Furthermore, we propose a group relative policy optimization (GRPO) [29] stage that refines inconsistent skill predictions via internal skill reward functions, thereby improving interaction performance.

Embodied agents require navigation and interaction; the latter is particularly challenging due to reliance on hand-engineered components [21], making it difficult to generalize. By integrating strong VLM perception and reasoning with our core skills, EUEA enables end-to-end interaction for instruction-following. Our skill-based experiments also reveal limitations in existing closed- and open-source VLMs [4, 8, 22, 53] and highlight the capabilities necessary for effective agent-environment interaction in instruction-following. In summary, our contributions are as follows:

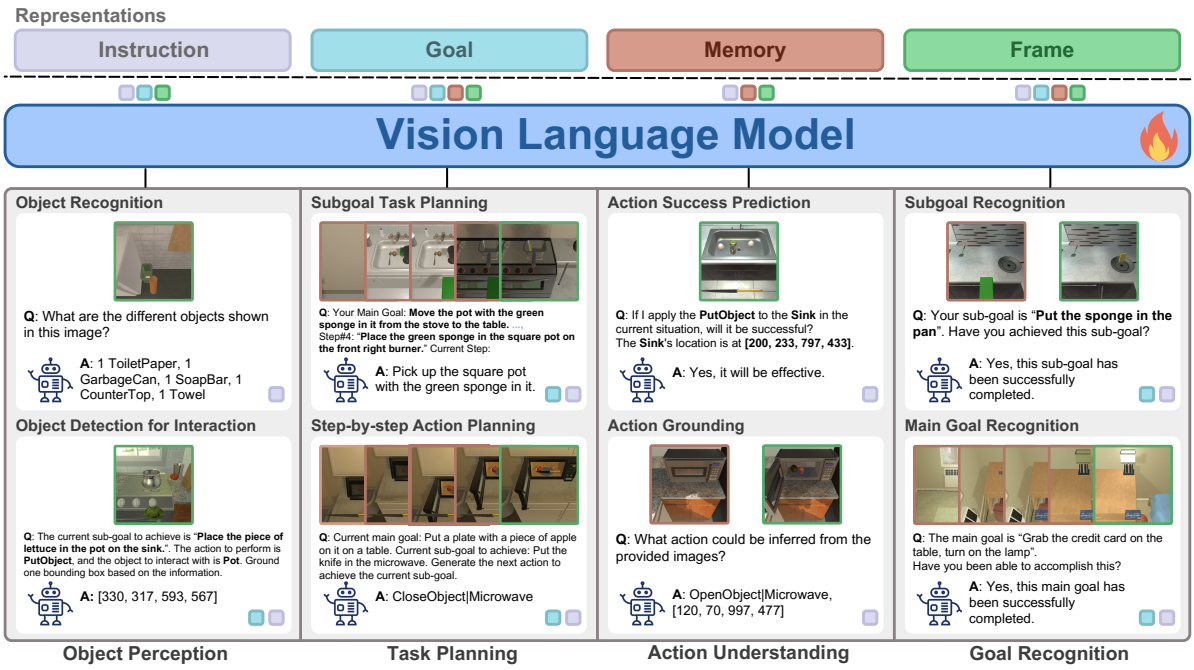

Figure 1. **Overview of the four core skills of EUEA to enhance VLM's environmental understanding and interaction.** Each core skill consists of two sub-skills, and we fine-tune the VLM in a single stage using the data from all skills. The colored boxes in each skill example indicate the representations. Additional each skill's examples template can be found in the supplementary material, Sec. B.

- We propose a novel EUEA framework that enhances the VLM's environmental understanding by integrating four core skills without separate module modeling. Through this unified, end-to-end learning approach, we enable the VLM to perceive, make decisions, and perform tasks, thereby effectively improving interaction performance.
- We further introduce a sampling-based recovery step that leverages EUEA skills and a GRPO refinement stage. The recovery step uses sampling to recover failure actions, and the GRPO stage refines inconsistent skill predictions. These provide an additional 3.03% improvement in average task success on ALFRED [31], beyond the 8.86% gain from supervised fine-tuning.
- We provide the skill datasets of 1.24M and 3.7M total samples and the skill evaluation benchmark using ALFRED [31] and LangR [36], along with a method for constructing skill datasets. Additionally, we evaluate the skill performance of existing VLMs and demonstrate that our approach achieves superior environmental understanding.

2. Related Work

Environmental Understanding. Environmental understanding is fundamental for successful embodied tasks [17, 23, 45, 51]. Prior works has enhanced this capability through object-based representations [15] and auxiliary modules [52]. In vision-language navigation, detection-based scene representations [6], episodic memory for exploring area [5], semantic voxel maps, and multi-view 3D reconstruction [25] improve spatial reasoning and navigation accuracy. More recent works [27] also identifies objects using visual tokens and defines action generation as skills, though they depend on separate object predictors and generate only the next action, without the ability to anticipate future state changes. Despite these advances, prior approaches [9, 16, 33, 42] generally improve environmental understanding through few-shot examples. $\pi_{0.5}$ [17] performs instruction-following in real environments through open-world generalization, but it lacks explicit environmental interpretation capabilities. In contrast, our EUEA framework defines the entire instruction-following process as a set of skills based on a partially observable Markov decision process (POMDP) [19], and integrates them into a single VLM via fine-tuning. This enables the VLM to perform end-to-end learning and to handle various functions such as perception, goal recognition, action planning, and future state anticipation in a unified structure, without requiring complex module combinations. Through this unified design, our method provides an integrated pathway toward robust environmental understanding.

Interaction-Level Correction and Refinement. Recent works has explored improving embodied agents by correcting errors during interaction through action-level repair [10] or planning-time refinement [42]. Existing methods recovers from failure by modifying the generated code using an inputted reason inferred from a fine-tuned model [10], detect errors using self-feedback without visual grounding [30], or recover the interaction at the subgoal level using fixed plans [48], symbolic state verification [49]. Other LLM-based planning methods refine the error by environment feedback

[34]. While these methods refine the errors, they generally depend on predefined failure types [10], external environment feedback [34], detects failure using self-feedback based on action executability but it does not consider visual information, thus cannot determine whether the action was actually successful [40]. In contrast, our approach recovers failed actions through sampling and further refines the model's responses in the GRPO [29] stage, generating alternative or more confident action selections. This self-correction mechanism enables recovery without predefined failure types or external feedback, relying instead on the model's learned environmental understanding. While prior studies [29, 46] have primarily applied group-based reinforcement learning (RL) to enhance reasoning ability, we employ it to further correct actions in our framework. Recent work [13] has also utilized group-based RL to stabilize multi-turn LLM agents. Inspired by this work, we adopt a GRPO refinement stage to stabilize EUEA skills and further improve task performance. Our method not only achieves performance gains without additional training via recovery step, but also yields further improvements through additional reinforcement.

3. Method

Inspired by embodied datasets and benchmarks [18, 23, 26, 41, 47], we focus on equipping VLMs with grounded environmental understanding for instruction-following tasks. To achieve this, we define four core skills that are necessary for step-by-step interaction based on a reward-free POMDP [19]. At each time step t, the VLM receives only a single image frame f_t and its accumulated past memory $\mathcal{M}$. Inferring all task-relevant information, such as visible objects v_t from this partial view, the VLM selects the current action $a_t \in \mathcal{A}$, an interaction target o_t, and a bounding box b_t to achieve the main goal $g \in \mathcal{G}$. After the action is executed, the entire interaction is recorded as a new memory state $m_t = (f_t, v_t, p_t, a_t, o_t, b_t, r_t, sg_t)$. The memory $\mathcal{M}$ consists of these observations over time and includes a total of T steps, represented as $\mathcal{M} = \{m_1, m_2, \ldots, m_T\}, m_t \in \mathcal{M}, \forall t \in \{1, 2, \ldots, T\}$. Here, p_t is the agent's position, r_t indicates the result of action at step t including succeeded or failed, and sg_t is the subgoal corresponding to step t. Examples from our EUEA skills are shown in Figure 1. We additionally introduce a recovery step via sampling, leveraging EUEA skills to enable the VLM to recover from failures. In Sec. 3.3, we further introduce a GRPO stage that refines inconsistent skill predictions.

3.1. Environmental Understanding Skills

The instruction, denoted as I_{SKILL}, is tailored for each skill to produce the intended outcome and includes eight distinct skills: object recognition (I_{OR}), object detection (I_{OD}), subgoal task planning (I_{STP}), step-by-step action planning (I_{SAP}), action success prediction (I_{ASP}), action grounding (I_{AG}), main goal recognition ($I_{GR_{main}}$), and subgoal recognition ($I_{GR_{sub}}$). These instructions are provided in the supplementary material, Sec. B.

Object Perception. The object perception skill consists of object recognition (OR) and object detection (OD), both of which support interaction. OR identifies objects present in the observed image, while OD detects a bounding box for an object relevant to a given current goal. These skills allow a VLM to recognize and localize them with a bounding box, enabling low-level control interaction. OR and OD are respectively formulated by the following equations:

$$v_t = \pi_\theta(I_{OR}, f_t) \tag{1a}$$

$$b_t = \pi_\theta(I_{OD}, sg_t, a_t, o_t, f_t) \tag{1b}$$

Where π_θ denotes the VLM-based agent, and b_t refers to the bounding box for the object o_t.

Task Planning. Task planning is divided into subgoal task planning (STP) and step-by-step action planning (SAP) skills. STP generates subgoals to achieve a given main goal, while SAP generates an action to achieve a given current goal. These skills enable the VLM not only to identify the given main goal and plan interaction subgoals, but also to generate actions by taking k past memory information, allowing it to interact with the environment as a performer to achieve the task. These skills formulated by the following equations:

$$sg_n = \pi_\theta(I_{STP}, \mathcal{M}) \tag{2a}$$

$$a_t, o_t = \pi_\theta(I_{SAP}, f_t, v_t, p_t, sg_t, m_{t-k:t-1}) \tag{2b}$$

Where sg_n denotes the nth subgoal generated by STP.

Action Understanding. We assume that meaningful environmental understanding requires a VLM to anticipate the outcomes of its actions and to describe the resulting state changes in the environment. To support this capability, we design three skills: action success prediction (ASP), future situation captioning (FSC), and action grounding (AG). ASP predicts whether an action will succeed or fail in the current situation. FSC is an extended ASP skill that allows VLM to describe the expected changes resulting from a given action. AG allows VLM to predict which action was executed between given two images. To build these skills, especially ASP and FSC, the dataset must contain sufficient failure examples, but expert demonstrations primarily include successful interactions. To address this imbalance, we generate failure-rich data through random exploration, where the VLM performs zero-shot random actions within the discrete action space. This process yields diverse successful and failed interactions across scenes. FSC data is additionally constructed by generating captions that explain the differences between pre- and post-action images using both expert trajectories and random-exploration memory as shown in the Figure 2. These skills are designed to predict action success, allowing the VLM to implicitly learn the constraints of interacting with the environment, thereby enhancing its

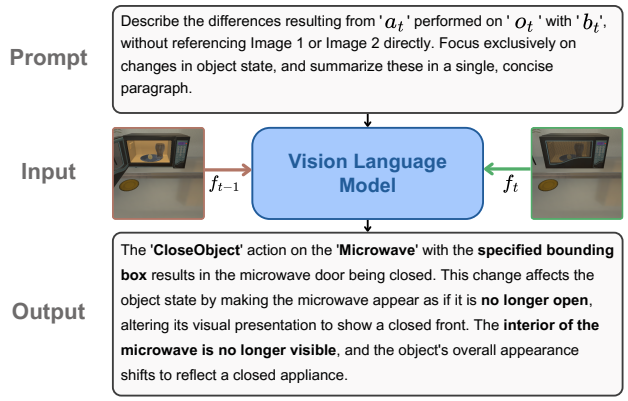

Figure 2. **Approach for generating data for environmental understanding.** We construct future situation captioning dataset that enables the prediction of captions describing changes between two images f_{t-1} and f_t, from a single image f_{t-1}.

understanding of the environment. ASP, FSC, and AG are respectively formulated by the following equations:

$$ASP_t = \pi_\theta(I_{ASP}, a_t, o_t, b_t, f_t) \quad (3a)$$

$$FSC_t = \pi_\theta(I_{FSC}, a_t, o_t, b_t, f_t) \quad (3b)$$

$$a_{t-1}, o_{t-1}, b_{t-1} = \pi_\theta(I_{AG}, f_{t-1:t}) \quad (3c)$$

Where ASP_t represents the predicted result of an action at step t, including "Yes" or "No", while FSC_t denotes a caption describing the predicted outcome of the action at step t. I_{FSC} is the instruction input to FSC.

Goal Recognition. Goal recognition is the skill that enables an agent to determine whether a subgoal or main goal has been achieved. The VLM learns to determine the achievement of the subgoal based on the main goal, or the achievement of the main goal, enhancing its understanding of the given task. GR_{main} and GR_{sub} are respectively formulated by the following equations:

$$GR_{main} = \pi_\theta(I_{GR_{main}}, \mathcal{M}) \quad (4a)$$

$$GR_{sub} = \pi_\theta(I_{GR_{sub}}, sg_t, a_{t-n:t}, f_{t-n:t}) \quad (4b)$$

Where GR_{main} represents the predicted result for the main goal, including "Yes" or "No", while GR_{sub} similarly represents the predicted result for the subgoal, also including "Yes" or "No". $a_{t-n:t}$ and $f_{t-n:t}$ denote the actions and frames corresponding to the past n steps of information and current step t from the given subgoal sg_t, including step t. This formulation allows the VLM to recognize both its ability to accomplish the overarching main goal and its progress in achieving intermediate subgoals, thereby improving overall task understanding. In particular, this GR_{sub} enables the model to decide when to move to the next subgoal.

Skill Dataset Construction. We use two instruction-following benchmarks to construct our skill datasets: ALFRED [31], built on AI2-THOR [20], and LangR rearrangement benchmark [36], built on Habitat 2.0 [35]. ALFRED consists of 8,055 expert demonstrations, 25,743 human annotations, and 428k image-action pairs, while LangR includes 150k episodes and 55k instructions. From these benchmarks, we collect environmental observations from image-action pairs and annotations, and design prompt templates to construct the four core skill datasets that support environmental understanding. To assess generalization to unseen environments, we additionally sample 31 of the 108 ALFRED training scenes and 42 of the 84 LangR training scenes. Using this data, we build four skill datasets containing 382k training, 858k validation, and 5.3k evaluation samples for ALFRED, and 906k training, 2.8M validation, and 4.4k evaluation samples for LangR. Across ALFRED and LangR, the fine-tuning sets include 108k and 237k samples for object perception, 25k and 144k for task planning, 217k and 471k for action prediction, and 29k and 52k for goal recognition. The evaluation sets contain 1k and 1k samples for object perception, 1.2k and 2k for task planning, 2.1k and 1k for action prediction, and 1k and 0.5k for goal recognition. The full data templates and collection pipeline are detailed in the supplementary material, Sec. B.

3.2. Recovery Step via Sampling

Despite utilizing fine-tuned skills such as step-by-step action planning (SAP) and object detection (OD) to interact with the environment, interaction failures can still occur. To address this, we introduce a **Recovery Step**, which is triggered when action fails. In this step, the VLM agent π_θ samples alternative actions using SAP skill. The sampled output differs from the failed one but remains aligned with the given goal, guided by the corresponding instruction I_{SAP}. This alternative actions are selected based on the score equation:

$$s_{t,i} = -\log \pi_\theta(a_{t,i}, o_{t,i} \mid I_{SAP}, m_{t-k:t}) \quad (5)$$

Where i is sampling index, and the score $s_{t,i}$ is the negative log-likelihood of the SAP skill from Equation 2b. Since the π_θ generates actions conditioned on the given goal, the predicted probability can be interpreted as goal-achievement confidence, allowing π_θ to select the most promising action for the given goal.

To recover from failure, we conduct n SAP to generate new candidate actions $a_{t,i}$ with objects $o_{t,i}$. When π_θ assigns a high probability to a sampled action-object pair, the score $s_{t,i}$ is close to zero. If all sampled action–object pairs correspond to the same previously failed action, we instead perform OD n times and compute the score s_t for the previously failed action–object pair using Equation 5 similarly. After sampling n times, π_θ selects either an action-object pair a_{new}, o_{new} or a bounding box b_{new} with the lowest score to interact with the environment. If an action-object pair is selected, we perform OD to obtain the corresponding bounding box b_{new}. This step ensures recovery from two types of failure scenarios through sampling without additional training.

3.3. GRPO Refinement Stage

Although supervised fine-tuning (SFT) on the our four skills improves environmental understanding, the VLM can still produce incorrect decisions. To address this issue, we introduce a GRPO [29] refinement stage that reduces inconsistent responses through rule-based rewards. This framework enables the VLM to further enhance its understanding of the environment.

Reward Function. We define reward functions using task-specific correctness metrics, including metrics based on intersection of union (IoU), the Jaccard index, and action-sequence order. The total reward function is given by: $R_{total} = R_{OP} + R_{TP} + R_{AU} + R_{GR}$, where R_{OP}, R_{TP}, R_{AU}, and R_{GR} are the reward functions for the four core skills: object perception (R_{OP}), task planning (R_{TP}), action understanding (R_{AU}), and goal recognition (R_{GR}). Since each input instance corresponds to a single skill, the rewards for all other skills are set to zero. R_{OP} uses the Jaccard index to reward correct predictions for the OR subskill and applies an IoU based reward for bounding boxes in OD. R_{TP} assigns rewards based on the correctness of the predicted action-object pair in SAP and the correctness of the predicted action sequence order in STP. R_{AU} rewards correct success predictions in ASP and FSC. In AG, it combines the correctness of the action-object pair used in SAP and OD with an IoU weighted reward for the bounding box. Finally, R_{GR} provides rewards based on the correctness of both main goal and subgoal predictions. The detailed reward values and the complete formulation of all reward functions are provided in supplementary material, Sec. C.

Dataset construction and Strategy. We use the previously created validation set to collect eight response samples for all data and then filter out the ambiguous cases. Our sampling strategy proceeds as follows: 1) Sampling eight responses for each data instance. 2) Counting the number of correct responses c for each instance. 3) Selecting instance whose normalized standard deviation of rewards, exceeds a threshold τ. This allows us to construct a compact dataset of around 10k instances on ALFRED [31], consisting only of cases where the model shows uncertainty, without using the entire validation set. Even with this smaller but compact dataset, this stage improve the VLM's decision-making ability by refining ambiguous responses.

4. Experiments

4.1. Experimental Setup

Training Details. We use InternVL3-8B [53] as our main VLM for ALFRED [31] and LangR [36], due to their strong multi-modal understanding, scalability, and support for multiple image inputs. To assess the contribution of our environmental understanding skills, we also train a behavior cloning (BC) baseline in which all skills are removed except OD, SAP, GR_{sub}. In SFT stage, we full fine-tune all VLMs for 1 epoch each using 8 A100 80GB GPUs, following the their training scripts [8, 50, 53]. The vision encoder is frozen, while the MLP and LLM components are fine-tuned. We set the batch size to 128, the max sequence length to 8192. In the GRPO stage, we fine-tune InternVL3-8B with LoRA [14] for 5 out of 10 epochs due to early stopping, using 2 A100 80GB GPUs, following DAPO [46] hyperparameter setting. We set the batch size to 64 and the max sequence length to 8192.

Task Evaluation. To evaluate whether the learned skills improve VLM's task performance, we evaluate our models on ALFRED [31], which provides clearly defined state changes for robust task evaluation, including long-horizon interaction tasks. We follow ALFWorld [32] by using 134 instruction-following tasks grounded in ALFRED states, and, motivated by EMMA [43] which emphasizes the value of human-annotated free-form instructions, we augment them with additional free-form variants, yielding 429 evaluation tasks in total. These tasks are categorized into six type: Examine in Light (Look), Pick&Place (Pick), Pick Two&Place (Pick Two), Clean&Place (Clean), Cool&Place (Cool), and Heat&Place (Heat). We exclude LangR [36] from task evaluation because its trajectories often include interactions with non-visible objects, creating scenarios that are incompatible with our setting where the agent may act only on visible observations. We construct an instruction-following pipeline where VLM takes only instructions and observed images from the environment as inputs, while storing and utilizing intermediate outputs in memory to perform sequentially. We categorize the given subgoals into two types: **Navigation** and **Interaction**. In the **Navigation** phase, the agent follows PDDL expert actions while gathering environmental information through OR. During the **Interaction** phase, the agent generates actions to achieve subgoals through SAP, subsequently predicts the target object and its corresponding bounding box through OD, and executes the interaction with the environment based on these predictions. We treat an action as failed when the image does not change, which fits discrete-action settings with distinct visual transitions. This cycle continues until all given subgoals are completed, using GR_{sub}. We set k to 4 for the memory input and n to 10 for the recovery step.

Skill Evaluation. To evaluate how effective VLMs are in generating interactions for embodied AI, we propose an evaluation benchmark based on four key skills, each skill has two metrics: **1) Object Grounding** evaluates object identification accuracy based on count and inclusion correctness. **2) Object Detection** measures bounding box accuracy using Intersection over Union (IoU). **3) Planning** evaluates generated and ground-truth subgoals via cosine similarity with a BERT-based transformer [28]. **4) Step-by-Step** checks if both the action and object match the ground truth. **5) Action Prediction** evaluates "Yes" or "No" response accuracy for

Table 1. **Comparison of task success rates with instruction-following VLM Agents.** * indicates that they are from the reported results [31, 43].

VLM Agent	Task Success Rate						
	Avg.	Look	Pick	Pick Two	Clean	Cool	Heat
EMMA* [43]	67.83	66.67	71.95	75.93	65.31	55.56	71.80
Human Performance* [31]	91.00	-	-	-	-	-	-
BC (InternVL3-8B)	74.59	88.89	73.17	57.41	62.24	96.83	75.64
Ours (SFT)	83.45	**90.74**	**86.59**	75.93	65.31	**98.41**	**91.03**
Ours (GRPO)	**85.78**	**90.74**	85.37	**85.19**	**74.49**	**98.41**	87.18

Table 2. **Comparison of recovery methods for task evaluation.** We evaluate task performance by comparing different recovery methods. *Env Feedback* adds external environment feedback when it failed.

Recovery Method	Success Rate		Goal Condition	
	SFT	GRPO	SFT	GRPO
Ours	83.45	85.78	88.42	90.17
w/ Env feedback	**85.78**	85.78	**90.09**	90.17
w/ Recovery Step	**85.78**	**86.48**	89.74	**90.48**

Table 3. **Skill evaluation results of closed- and open-source VLMs.** We evaluate both closed- and open-source models using the evaluation dataset generated from four skills utilizing ALFRED [31] and LangR [36]. **Detection** is measured using IoU, **Planning** is evaluated using a BERT-based transformer [28] with cosine similarity between subgoals, and all other skills are evaluated based on accuracy. Navigation* represents three additional sub-skills for navigation. Ours* indicates the InternVL3-8B VLM fine-tuned with ALFRED skills.

ALFRED	Model	Object Perception		Goal Recognition		Action Understanding		Task Planning	
		Grounding	Detection	Main	Sub	Prediction	Grounding	Planning	Step-by-step
Closed	GPT-5	51.45	24.34	92.20	73.80	81.48	28.86	0.801	71.52
	GPT-o3	57.28	29.55	94.80	80.60	86.75	31.76	0.796	77.41
	Claude-4.5-Sonnet	38.05	1.54	90.80	65.80	51.60	62.55	0.812	46.15
	Gemini-2.5-Flash	45.00	43.15	88.60	66.80	69.45	32.53	0.799	74.80
	Gemini-2.5-Pro	63.53	60.75	86.20	80.20	85.43	31.08	0.819	85.76
Open	LLaVA-OneVision-7B [22]	22.19	18.19	69.60	68.00	71.05	44.88	0.695	6.87
	InternVL2.5-8B [8]	30.90	8.79	67.00	76.20	68.70	47.39	0.608	0.82
	InternVL3-8B [53]	33.70	26.68	71.80	77.40	68.80	48.26	0.742	4.42
	Qwen2.5-VL-7B [4]	28.15	1.82	84.00	73.00	48.78	49.61	0.764	39.44
	Qwen3-VL-8B [3]	48.99	51.79	89.40	80.20	76.50	67.18	0.815	45.34
	BC (InternVL3-8B)	**78.01**	73.94	71.80	83.00	63.25	9.27	0.630	**98.53**
	Ours (InternVL3-8B)	75.84	**81.73**	**99.40**	**98.60**	**96.80**	**89.09**	**0.894**	98.20
LangR	**Model**	**Object Perception**		**Goal Recognition**		**Action Understanding**		**Task Planning**	
		Grounding	Detection	Main		Prediction	Grounding	Navigation*	Step-by-step
Closed	GPT-5	35.07	12.35	87.40		62.50	28.29	60.53	88.02
	GPT-o3	35.17	20.77	95.80		72.96	27.30	59.67	85.57
	Claude-4.5-Sonnet	24.96	0.00	64.40		38.27	21.50	43.80	89.00
	Gemini-2.5-Flash	30.43	24.08	94.40		71.94	27.55	59.20	77.26
	Gemini-2.5-Pro	45.68	59.77	95.40		82.14	30.35	38.80	97.07
Open	LLaVA-OneVision-7B [22]	11.15	8.52	53.00		42.60	0.19	39.53	40.34
	InternVL2.5-8B [8]	25.99	1.12	75.00		44.00	12.59	47.53	30.32
	InternVL3-8B [53]	29.03	8.62	74.60		48.00	13.10	52.90	69.68
	Qwen2.5-VL-7B [4]	21.64	0.00	64.00		36.22	22.04	57.07	79.46
	Qwen3-VL-8B [3]	38.41	29.94	91.00		63.27	27.22	60.60	72.13
Cross	Ours* (ALFRED-tuned)	26.81	58.35	86.60		77.81	47.52	-	77.06
	BC (InternVL3-8B)	**86.59**	88.93	**100.00**		38.78	14.14	73.67	98.04
	Ours (InternVL3-8B)	86.50	**93.20**	**100.00**		**100.00**	**49.84**	**74.07**	**99.27**

action success. **6) Action Grounding** verifies whether the model correctly identifies action, object, and bounding box across images. **7) Goal Recognition (Main & Sub)** evaluates yes/no response accuracy for goal completion. We evaluate our fine-tuning dataset generated with ALFRED [31] and LangR [36] using both open source VLMs [4, 8, 22, 53] and closed state-of-the-art VLMs [1]. Additionally, in LangR, we define three sub-skills for navigation: 1) identifying where to go in order to perform actions such as "Pick" or "Place" in rearrangement tasks, 2) selecting the target object, and 3) selecting the destination.

Due to space limits, we provide detailed descriptions of the sub-skills, evaluation prompts for all skills, and the pipeline algorithm in the supplementary material, Secs. B–E.

4.2. Result

Task Evaluation. Table 1 compares our VLMs with the BC baseline on task evaluation using ALFRED [31]. Our VLMs show significant improvements on most tasks over the BC baseline, achieving average task-success gains of 8.86% and 10.96% in the SFT and GRPO stages, respectively. This indicates that the four core skills enhance environmental

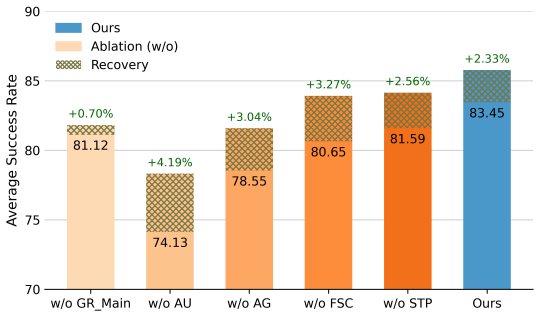

Figure 3. **Ablation study on our proposed skills for task evaluation.** We evaluate task performance by conducting an ablation study on the proposed skills that impact environmental understanding. AG refers to action grounding, FSC to future situation captioning, AU to action understanding, STP to subgoal task planning, and GR_{Main} to main goal recognition, respectively.

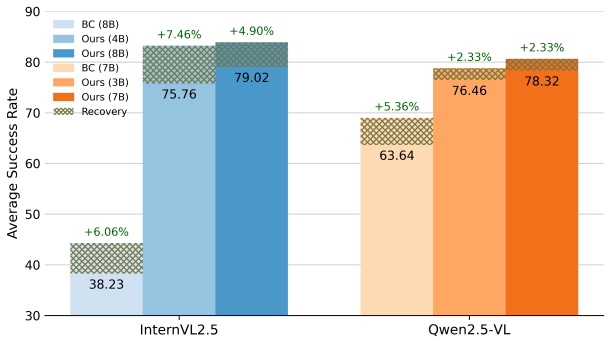

Figure 4. **Comparison of VLM backbones on task evaluation.** We evaluate task performance using the InternVL2.5-series and Qwen2.5-VL-series as backbones in SFT stage.

understanding, with the GRPO stage further refines inconsistent responses, where correct and incorrect predictions alternate, leading to even stronger environmental understanding. Although EMMA [43] extends ALFWorld [32] to a vision-based setting, it relies on high-level interactions and environment metadata, and operates without subgoals. In contrast, our model depends solely on visual observations and low-level actions given subgoals, making the interaction setting more demanding while still achieving strong fine-grained performance.

Comparison of Recovery Methods. We conduct a comparative analysis to evaluate the effect of our recovery method on task performance compared to the baseline. The baseline recovery relies on additional external environment feedback after failures to guide recovery. We include this baseline for comparison, as prior works either utilize environmental feedback directly [34], train models through such feedback [40], or define failure types [10] that can be attributed to environmental feedback. Table 2 shows that our recovery step achieves the highest task success rate for both SFT and GRPO stage, especially outperforming oracle feedback in GRPO stage. This indicates that, rather than relying on external environmental feedback, sampling based on the agent's environmental understanding can achieve comparable or even superior performance.

Skill Evaluation. Table 3 presents the evaluation results of the skill datasets using ALFRED and the photorealistic LangR [36] benchmarks. Across both benchmarks, Gemini-2.5-Pro achieves the strongest zero shot performance in most skills despite not being trained on either dataset. Its strong reasoning ability improves skill prediction, particularly for goal recognition and planning. In goal recognition, the performance gap between Gemini-2.5-Pro and GPT-o3, and our fine-tuned model is only 4.6% in both skill benchmarks, which is consistent with prior works [30, 34, 42, 44] showing the strong planning capabilities of LLMs and VLMs.

However, most zero-shot VLMs still lack environmental understanding, resulting in limited accuracy in action success prediction, action grounding, and object perception in both skill benchmarks. In ALFRED's step-by-step action planning, Gemini-2.5-Pro reaches 85.76 percent accuracy, meaning that nearly 14% of individual actions fail, which often leads to task failure. This challenge becomes even greater as the action space increases. In contrast, our model, fine tuned on the four proposed skills, significantly improves performance across all evaluations. By enhancing EUEA skills, it outperforms the baselines and demonstrates stronger performance on the skills that support instruction-following, even in real world settings. Although BC shows similar or slightly better results on some skills, these differences do not lead to stronger task performance. These results highlight the need for additional training to improve environmental understanding, ensuring VLMs function as agents.

4.3. Analyses

Impact of Task Performance under Skill Ablation. We analyze the impact of the proposed skills on task evaluation. Figure 3 presents an ablation study on main goal recognition (GR_{Main}), action understanding (AU), action grounding (AG), future situation captioning (FSC), and subgoal task planning (STP), which are not explicitly used in the evaluation pipeline. Removing each skill individually results in task success rate drops of 2.33% for GR_{Main}, 9.32% for AU, 4.9% for AG, 1.86% for FSC, and 2.8% for STP, corresponding Ours SFT stage. These results show that removing any skill drops performance, with AU having the largest impact. This demonstrates that each skill contributes meaningfully to task evaluation and that improved prediction of environment changes leads to higher task success rate.

Analysis of Task Performance across VLM Backbones. Figure 4 presents the task performance results of our method after SFT using different VLM backbones, InternVL2.5-series [8] and Qwen2.5-VL-series [4]. Our approach improves the average success rate over the respective BC baselines for both series. Furthermore, even with models that

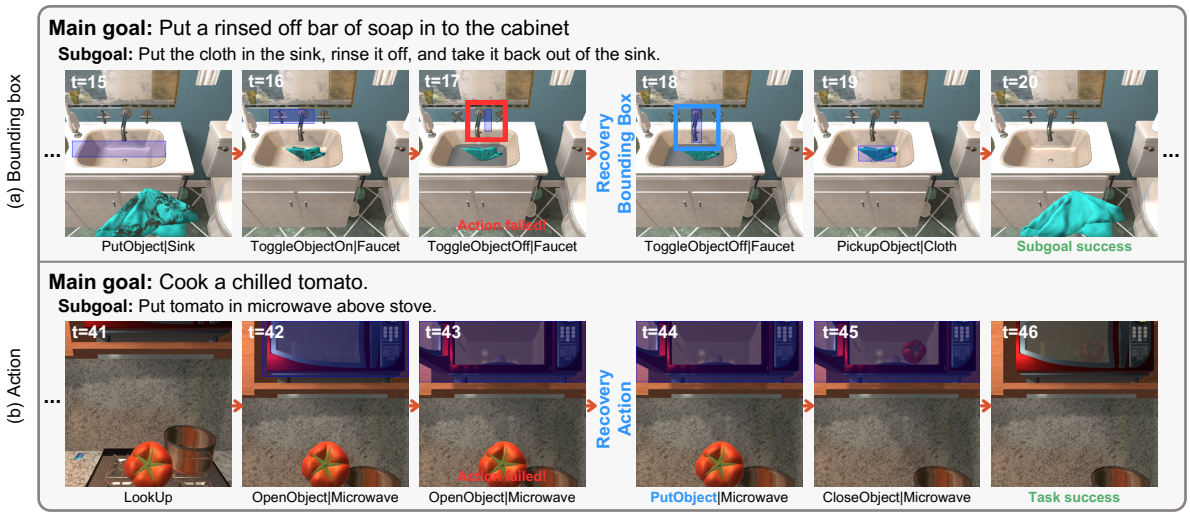

Figure 5. **Analysis of cases where the recovery step resolves a failed interaction.** (a) shows when an incorrect detection is corrected, allowing the action to succeed. (b) shows when an interaction fails, and an alternative action completes the given task successfully.

have about 50% fewer parameters, InternVL2.5-4B and Qwen2.5-VL-3B, our method achieves notable performance gains of 12.84% and 37.53%, respectively, compared to their BC baselines. These results demonstrate that our approach effectively enhances average task performance and remains robust across different backbone scales.

Out-of-Distribution (OOD) Skill Evaluation. We evaluate the VLM fine-tuned on the ALFRED skill dataset using the LangR skill benchmark to evaluate cross-environment generalization. Although both datasets contain Pick&Place interactions, ALFRED has more diverse action space, whereas LangR focuses only on simple, photorealistic Pick&Place rearrangement scenes. To address vocabulary mismatches, we map actions and objects using a BERT-based transformer [28] (e.g., "Pick" to "PickupObject", "CoffeeTable" to "Table"). As shown in Table 3, the ALFRED-tuned model performs substantially lower than the LangR-tuned model, which is expected given the distribution gap, but it still outperforms the zero-shot InternVL3-8B on all skills except object grounding. This result shows that the EUEA skills transfer even under significant distribution shift, indicating promising OOD generalization.

Qualitative Results on Recovery Step. Figure 5 demonstrates how our recovery step helps the VLM complete tasks despite failures. In Figure 5 (a), the model initially generates an incorrect bounding box for an action-object interaction ($t = 17$) but corrects it at $t = 18$, achieving the subgoal and completing the task. In Figure 5 (b), when a failure occurs during execution, the VLM generates a new action to complete the task. We observe that during multi-step interactions, the VLM can repeat previous actions, possibly due to token distribution flattening during SFT. The recovery step mitigates this by generating both bounding boxes and appropriate subsequent actions, improving task performance.

We provide additional failure cases, including analyses of how removing individual skills affects performance on each task, and present evaluation results on the validation set in LangR, a performance analysis on data scaling, and an ablation study on the memory input and recovery step count in the supplementary material, Sec. F.

5. Conclusion

In this study, we propose a novel framework EUEA, which fine-tunes four core skills to VLM that enable agents enhance their environmental understanding. Our skill benchmark shows that existing VLMs still exhibit limited environmental understanding without specific environment skill fine-tuning. Additionally, we introduce a recovery step that selects an alternative action through sampling when an interaction fails, and a GRPO stage that reduces inconsistent responses. These approaches allow the model to successfully complete tasks, achieving a high success rate of 86.48%. Our findings highlight the importance of environmental understanding, skill learning, the recovery step, and the additional refinement stage in developing more reliable end-to-end agents.

Limitation and Future Work. This study evaluates task performance and skill evaluation in a discrete environment, which may limit a comprehensive understanding of state transitions. Future work should extend this approach to continuous environments by incorporating video-based input, enabling better tracking of environmental changes. Additionally, while this study focuses on interaction, low-level navigation remains a crucial challenge for end-to-end agents. Moreover, current models, including ours, often rely on intuitive predictions rather than explicit reasoning, which may lead to misunderstandings of the environment. We believe that integrating reasoning mechanisms that utilize past information and predict future outcomes would enhance decision-making capabilities.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-25442824, AI Star Fellowship Program (Ulsan National Institute of Science and Technology & No.RS-2020-II201336, Artificial Intelligence graduate school support (UNIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-24683548).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 1

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 6

[4] 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. 1, 6, 7

[5] Valts Blukis, Chris Paxton, Dieter Fox, Animesh Garg, and Yoav Artzi. A persistent spatial semantic representation for high-level natural language instruction execution. In *Conference on Robot Learning*, pages 706–717. PMLR, 2022. 2

[6] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258, 2020. 2

[7] William Chen, Suneel Belkhale, Suvir Mirchandani, Oier Mees, Danny Driess, Karl Pertsch, and Sergey Levine. Training strategies for efficient embodied reasoning. *arXiv preprint arXiv:2505.08243*, 2025. 1

[8] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 5, 6, 7

[9] 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. 2

[10] Jiafei Duan, Wilbert Pumacay, Nishanth Kumar, Yi Ru Wang, Shulin Tian, Wentao Yuan, Ranjay Krishna, Dieter Fox, Ajay Mandlekar, and Yijie Guo. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. *arXiv preprint arXiv:2410.00371*, 2024. 1, 2, 3, 7

[11] Inoue et al. Prompter: Utilizing large language model prompting for a data efficient embodied instruction following. *arXiv*, 2022. 1

[12] Zhao et al. Epo: Hierarchical llm agents with environment preference optimization. *EMNLP*, 2024. 1

[13] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025. 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[15] Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. Are you looking? grounding to multiple modalities in vision-and-language navigation. *arXiv preprint arXiv:1906.00347*, 2019. 2

[16] 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*, pages 9118–9147. PMLR, 2022. 2

[17] Physical Intelligence, 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 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, 2025. 2

[18] Md Mofijul Islam, Alexi Gladstone, Riashat Islam, and Tariq Iqbal. EQA-MX: Embodied question answering using multimodal expression. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[19] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998. 2, 3

[20] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 4

[21] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. 1

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 6

[23] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li

Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2025. 2, 3

[24] 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. *arXiv preprint arXiv:2209.07753*, 2022. 1

[25] Rui Liu, Wenguan Wang, and Yi Yang. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16328, 2024. 2

[26] 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, Alexander Sax, and Aravind Rajeswaran. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16488–16498, 2024. 3

[27] Georgios Pantazopoulos, Malvina Nikandrou, Amit Parekh, Bhathiya Hemanthage, Arash Eshghi, Ioannis Konstas, Verena Rieser, Oliver Lemon, and Alessandro Suglia. Multitask multimodal prompted training for interactive embodied task completion. *arXiv preprint arXiv:2311.04067*, 2023. 2

[28] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. 5, 6, 8

[29] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1, 3, 5

[30] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023. 1, 2, 7

[31] 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*, pages 10740–10749, 2020. 2, 4, 5, 6

[32] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020. 5, 7

[33] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2998–3009, 2023. 2

[34] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models. *Advances in neural information processing systems*, 36:58202–58245, 2023. 1, 3, 7

[35] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 4

[36] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 2, 4, 5, 6, 7

[37] Andrew Szot, Bogdan Mazoure, Harsh Agrawal, R Devon Hjelm, Zsolt Kira, and Alexander Toshev. Grounding multimodal large language models in actions. *Advances in Neural Information Processing Systems*, 37:20198–20224, 2024.

[38] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. From multimodal llms to generalist embodied agents: Methods and lessons. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10644–10655, 2025. 1

[39] 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*, 2023. 1

[40] Hanlin Wang, Chak Tou Leong, Jian Wang, and Wenjie Li. E2cl: exploration-based error correction learning for embodied agents. *arXiv preprint arXiv:2409.03256*, 2024. 3, 7

[41] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai, 2023. 3

[42] 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 multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 2, 7

[43] Yijun Yang, Tianyi Zhou, Kanxue Li, Dapeng Tao, Lusong Li, Li Shen, Xiaodong He, Jing Jiang, and Yuhui Shi. Embodied multi-modal agent trained by an llm from a parallel textworld. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26275–26285, 2024. 1, 5, 6, 7

[44] 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. 7

[45] Minjong Yoo, Jinwoo Jang, Wei-Jin Park, and Honguk Woo. Exploratory retrieval-augmented planning for continual embodied instruction following. *Advances in Neural Information Processing Systems*, 37:67034–67060, 2025. 2

[46] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforce-

ment learning system at scale, 2025. *URL https://arxiv.org/abs/2503.14476*, 2025. 3, 5

[47] Min Zhang, Xian Fu, Jianye Hao, Peilong Han, Hao Zhang, Lei Shi, Hongyao Tang, and Yan Zheng. Mfe-etp: A comprehensive evaluation benchmark for multi-modal foundation models on embodied task planning, 2024. 3

[48] Yichi Zhang and Joyce Chai. Hierarchical task learning from language instructions with unified transformers and self-monitoring. *arXiv preprint arXiv:2106.03427*, 2021. 2

[49] Yichi Zhang, Jianing Yang, Jiayi Pan, Shane Storks, Nikhil Devraj, Ziqiao Ma, Keunwoo Yu, Yuwei Bao, and Joyce Chai. Danli: Deliberative agent for following natural language instructions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1280–1298, 2022. 2

[50] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. 5

[51] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. In *European Conference on Computer Vision*, pages 129–148. Springer, 2024. 2

[52] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10012–10022, 2020. 2

[53] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, 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. 1, 5, 6