
World-aware Planning Narratives Enhance Large Vision-Language Model Planner

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Large Vision-Language Models (LVLMs) show promise for embodied planning
2 tasks but struggle with complex scenarios involving unfamiliar environments and
3 multi-step goals. Current approaches rely on environment-agnostic imitation
4 learning that disconnects instructions from environmental contexts, causing models
5 to struggle with context-sensitive instructions and rely on supplementary cues
6 rather than visual reasoning during long-horizon interactions. In this work, we
7 propose World-Aware Planning Narrative Enhancement (WAP), a framework that
8 infuses LVLMs with comprehensive environmental understanding through four
9 cognitive capabilities (visual appearance modeling, spatial reasoning, functional
10 abstraction, and syntactic grounding) while developing and evaluating models
11 using only raw visual observations through curriculum learning. Evaluations on the
12 EB-ALFRED benchmark demonstrate substantial improvements, with Qwen2.5-
13 VL achieving a 60.7 absolute improvement in task success rates—particularly in
14 commonsense reasoning (+60.0) and long-horizon planning (+70.0). Notably, our
15 enhanced open-source models outperform proprietary systems like GPT-4o and
16 Claude-3.5-Sonnet by a large margin.

17 1 Introduction

18 Recent advances in Large Vision-Language Models (LVLMs) [6, 20] have expanded their applications
19 to embodied planning tasks, where agents interpret natural language instructions into a sequence of ac-
20 tions which will be further executed in interactive environments [6, 28]. These models leverage large-
21 scale pretraining on vision-language datasets to align visual inputs with textual commands, achieving
22 notable success in controlled scenarios with explicit object references and low environmental com-
23 plexity [13, 12]. However, when faced with increasingly complex real-life-like scenarios—those
24 involving unfamiliar environments, varied instruction formats, and multi-step goals—current methods
25 exhibit severe limitations in both generalization ability and reasoning consistency[26].

26 A fundamental challenge confronting current embodied planning systems lies in their environment-
27 agnostic imitation learning paradigm. In existing methodologies [13, 12], expert demonstration
28 trajectories are typically associated with simplified, environment-independent instructions (e.g., “put
29 the apple on the table”). They force models to learn direct mappings from generic instructions to action
30 sequences in an open-loop manner, disregarding the nuances of the changing surrounding environment.
31 This learning approach operates on a disjointed fashion, treating task instructions and environmental
32 contexts as distinct, unconnected elements. This hinders models from developing an integrated grasp
33 of the environment’s specific characteristics such as visual appearance, spatial relationships, object
34 functionality, and linguistic comprehension—capabilities essential for human-like task execution
35 [9]. While these models may perform adequately in standardized validation environments, their
36 performance significantly deteriorates when faced with difficult situations that demand the ability to

37 establish links between the changing surroundings and context-sensitive instructions, for example,
 38 “place the apple on the table near the television”. The limitation is becomes evident during extended
 39 sequences of interaction, where models, due to their lack of detailed environmental representations,
 40 struggle to integrate previous visual observations. Hence, they resort to supplementary environmental
 41 cues, like feedback from actions taken or indicators of task progress, instead of relying exclusively
 42 on visual input for decision-making.

43 To address these gaps, we introduce an innovative world-aware narrative enhancement approach.
 44 This method is designed to infuse LVLMs with comprehensive information about the environment.
 45 By doing so, we aim to augment the model’s understanding of the environment by incorporating
 46 relevant context from the world around it. Inspired by traditional theories of cognitive intelligence
 47 [14], our narrative enhancement focuses on collecting data that progressively cultivates four inter-
 48 connected capabilities: (1) *Visual Appearance Modeling*: Detailed capture of object textures and
 49 geometries. (2) *Spatial-Relational Reasoning*: Understanding the spatial arrangement and room
 50 layouts. (3) *Functional Abstraction Learning*: Grasping tool-object relationships and symbolic
 51 representations. (4) *Syntactic Grounding*: Interpreting complex language to resolve ambiguity. To
 52 effectively steer the data generation process while considering various dimensions, we augment
 53 existing disjointed data by integrating complementary information from the environment or semantic
 54 spaces. Secondly, our framework adheres to realistic deployment scenarios: agents receive only image
 55 observations and natural language instructions in a closed-loop manner, without auxiliary privileged
 56 environmental feedback. To incrementally equip the model to tackle cognitive challenges, we collect
 57 the data and train the model with a curriculum learning strategy, which enhances the model’s ability
 58 to formulate sophisticated strategies in a wide range of situations, markedly differentiating from
 59 traditional imitation learning techniques that do not support this level of advanced intellectual growth.

60 We evaluate our framework through extensive experiments using Qwen2.5-VL [8] and InternVL3
 61 [32] on the EB-ALFRED benchmark within EmbodiedBench [26], a challenging suite of high-
 62 level planning tasks. Our approach achieves substantial improvements over baseline methods, with
 63 Qwen2.5-VL demonstrating a 60.7 absolute improvement in average task success rates. Particularly
 64 noteworthy are the significant gains in commonsense reasoning (+60.0) and long-horizon planning
 65 (+70.0), with similar patterns observed for InternVL3. Remarkably, our enhanced open-source models
 66 outperform recent proprietary systems like GPT-4o and Claude-3.5-Sonnet by a large margin.

67 This work makes three key contributions:

- 68 • Our approach bridges the gap between high-level task instructions and the nuanced details
 69 of real-world environments by integrating contextual world knowledge into planning sys-
 70 tems. This multidimensional enhancement leverages narratives that are aware of various
 71 environmental factors, making the planning process more robust and adaptable to complex,
 72 real-life scenarios.
- 73 • We demonstrate that closed-loop embodied agents can achieve superior planning perfor-
 74 mance using only visual observations and natural language instructions without privileged
 75 environmental feedback—challenging the prevailing assumption that additional auxiliary
 76 signals are necessary for robust planning in complex environments.
- 77 • Our approach establishes new state-of-the-art benchmarks on EB-ALFRED, outperforming
 78 not only existing academic baselines by 60.7 improvement but also surpassing proprietary
 79 systems like GPT-4o and Claude-3.5-Sonnet by significant margins in challenging long-
 80 horizon planning scenarios.

81 2 Related Works

82 Embodied planning [4, 25] represents a critical cognitive capability for embodied agents, serving
 83 as the brain that guides physical interactions within complex environments. Early approaches to
 84 embodied planning [18, 27, 31] relied exclusively on textual environment metadata rather than
 85 visual perception, limiting their adaptability to real-world scenarios. Later visual-based methods
 86 [7, 11, 15, 16, 30] employed cascaded pipelines with external models for visual processing like
 87 semantic maps. Recent LVLM-based systems [21, 25] have begun processing visual input directly
 88 but continue to depend on unrealistic forms of environmental feedback like action success signals
 89 and task progress information that would be unavailable in genuine deployment scenarios. Our work
 90 advances the field by operating solely on raw visual observations without privileged feedback and by

91 processing complete observational history, more closely mirroring human capabilities in real-world
92 settings.

93 Methodologically, embodied planning approaches can be categorized as either training-free or training-
94 based. Training-free methods leverage prompt engineering [16, 10, 3] or multi-agent frameworks
95 with specialized roles [29, 5, 22]. Training-based approaches include supervised fine-tuning on
96 human expert demonstrations [23, 2], or methods that recognize the importance of trial-and-error
97 learning through either direct preference optimization [17, 30] or reinforcement learning [1, 24, 19].
98 Unlike these approaches that often treat task instructions and environmental contexts as disconnected
99 elements, our world-aware planning narrative enhancement framework systematically develops
100 integrated cognitive capabilities, enabling LVLMs to form sophisticated planning strategies without
101 relying on privileged cues, more closely resembling human cognitive processes in complex real-world
102 scenarios.

103 3 Methodology

104 We propose a multi-dimensional cognitive enhancement framework for embodied planning that
105 systematically develops reasoning capabilities across complementary dimensions while operating
106 solely on raw visual observations. Figure 1 illustrates our approach’s core components.

107 3.1 Problem Formulation

108 In embodied planning tasks, an agent must execute a sequence of actions $\{a_1, a_2, \dots, a_T\}$ based
109 on a natural language instruction I and egocentric visual observations $\{o_1, o_2, \dots, o_T\}$. We operate
110 within a closed-loop control framework, where each action decision depends critically on both current
111 observation and historical context. This mirrors real-world robotics scenarios where agents must
112 continuously adapt to environmental changes resulting from previous interactions. Unlike many
113 previous approaches that focus on isolated decision points, our framework explicitly models temporal
114 dependencies by conditioning each action on the complete observation history:

$$a_t = f_{\Theta}(I, \{o_1, o_2, \dots, o_t\}) \quad (1)$$

115 Where f_{Θ} represents our vision-language model that predicts the next action based on the instruction
116 and observation history. The key challenge is to develop robust planning capabilities across diverse
117 scenarios with only visual-context history. Our multi-dimensional cognitive enhancement approach
118 specifically addresses these challenges by systematically developing the agent’s ability to extract and
119 integrate relevant information across time.
120

121 3.2 Multi-Dimensional Cognitive Enhancement

122 Our complete narratives enhance framework follows the pipeline illustrated in Figure 1. Our frame-
123 work enhances cognitive capabilities across four complementary dimensions that systematically
124 develop different aspects of the embodied intelligence required for robust planning.

125 3.2.1 Instruction Augmentation

126 Given an original instruction I and expert trajectory $\tau = \{a_t\}_{t=1}^T$ with corresponding observations
127 $\{o_t\}_{t=1}^T$, we generate enhanced instructions $\{\tilde{I}_k\}$ across four cognitive dimensions:

$$\tilde{I}_k = \mathcal{M}(I, o_T; \theta_k), \quad k \in \{\text{Visual, Spatial, Functional, Syntactic}\} \quad (2)$$

128 Where \mathcal{M} is a superior vision-language model and θ_k represents dimension-specific prompting
129 strategies:
130

- 131 • **Visual Dimension:** Enhances object appearance modeling by explicitly describing visual
132 attributes critical for identification (e.g., *cone-shaped lamp*, *small square clock*).
- 133 • **Spatial Dimension:** Augments positional understanding by specifying object locations rela-
134 tive to environmental landmarks (*beside the wooden chair*) and precise spatial relationships
135 (*below the lamp base*).

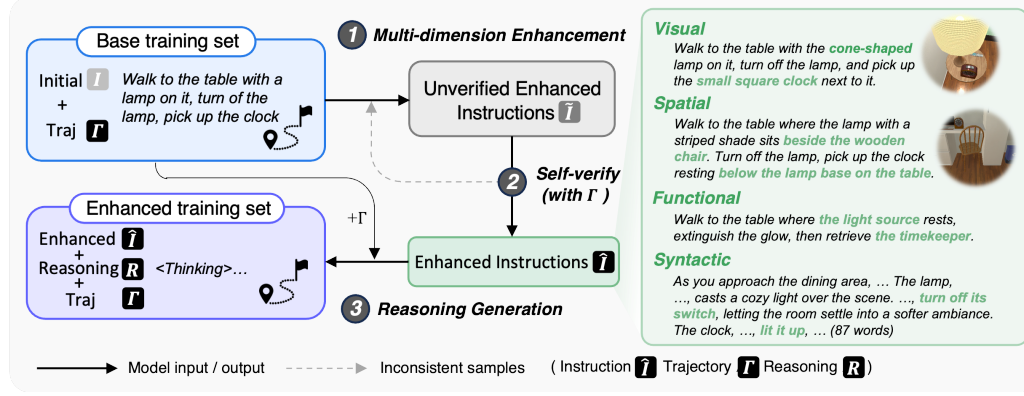


Figure 1: **Narratives enhance pipeline.** Our framework transforms base instruction-trajectory pairs through four main stages: (1) Multi-dimensional instruction enhancement, generating cognitively enriched variants; (2) Self-verification to ensure semantic consistency with original tasks; (3) Reasoning generation to provide explicit cognitive traces for each action; and (4) Construction of the enhanced training set that maintains trajectory information while adding cognitive annotations.

- **Functional Dimension:** Develops deeper object-interaction understanding by articulating affordances and functional properties (*light source*, *timekeeper*) that capture causal relationships between objects.
- **Syntactic Dimension:** Introduces linguistic complexity through narrative structures, indirect references, and contextual dependencies that require resolving ambiguity beyond literal instructions.

This augmentation process creates a structured hierarchy of cognitive demands, progressively challenging the model’s environmental reasoning capabilities.

3.2.2 Semantic Consistency Verification

To ensure augmented instructions maintain task equivalence, we implement a verification mechanism:

$$\mathcal{C}(\tilde{I}) = 4 \leq \sum_{i=1}^5 \mathbb{I} \left(\mathcal{V}(I, \tilde{I}, \{o_i : i \in \{1 \cdots T\}\}; \phi_i) = 1 \right) \quad (3)$$

Where $\mathcal{V}(\cdot; \phi_i)$ represents the verification function that checks if the generated instruction \tilde{I} represents the same intention as I , $\mathbb{I}(\cdot)$ is an identification function and evaluates to 1 if the input expression is true. Instructions fail to meet the threshold will trigger a regeneration of the instruction to maintain dataset quality.

3.2.3 Stepwise Reasoning Generation

For each action a_t in trajectory τ , we generate explicit reasoning r_t that captures the cognitive process linking observation to action:

$$r_t = \mathcal{M}(\tilde{I}, o_t, \{(\cdot, a_t)\} \cup \{(r_i, a_i) : i \in \{1, \cdots, t-1\}\}) \quad (4)$$

The t -th step of reasoning is left out for the vision-language model \mathcal{M} to predict. These reasoning annotations serve as intermediate supervision signals that help the model develop explicit cognitive processes that might otherwise remain implicit, including environmental state tracking, object relationship inferences, and action preconditions.

3.3 Curriculum Learning Framework

Our training procedure follows a three-stage curriculum that gradually increases cognitive complexity:

$$\Theta^* = \arg \min_{\Theta} \sum_{s=1}^3 \mathbb{E}_{(\tau, I) \sim \mathcal{D}_s} \mathcal{L}_{\text{CE}}(f_{\Theta}(\{o_{1:t}\}, I), a_t) \quad (5)$$

where f_{Θ} represents the vision-language model being optimized and \mathcal{L}_{CE} is the cross-entropy loss for action prediction. The curriculum stages are:

1. **Base Stage** (\mathcal{D}_1): Training on original instruction-trajectory pairs to establish foundational action mapping capabilities.
2. **Environmental Understanding Stage** (\mathcal{D}_2): Incorporating visual and spatial augmentations to develop perceptual grounding and scene comprehension.
3. **Conceptual Reasoning Stage** (\mathcal{D}_3): Introducing functional and syntactic augmentations to develop higher-order reasoning about object relations and ambiguous references.

This progressive training scheme aligns with cognitive development theories, allowing the model to first master perception-action correspondences before tackling more abstract semantic relationships. Importantly, our framework operates under strict partial observability constraints—providing only egocentric RGB observations without privileged information—to ensure real-world applicability.

4 Experiments

We conduct comprehensive experiments to evaluate our framework’s effectiveness in enhancing embodied planning capabilities. Our analysis focuses on both overall performance metrics and fine-grained cognitive capabilities across diverse task contexts.

4.1 Experimental Settings

Dataset We construct an enhanced corpus comprising 80,875 instruction-trajectory pairs derived from the original 16,145 ALFRED trajectories through our multi-dimensional augmentation approach. The dataset is structured across four cognitive dimensions: Visual (appearance attributes), Spatial (positional relationships), Functional (interaction affordances), and Syntactic (referential complexity). This structured enhancement enables systematic evaluation of specific cognitive capabilities crucial for embodied agents.

Models For instruction augmentation and reasoning generation described in Section 3, we employ Qwen2.5-VL-72B-Instruct as the teacher model. We evaluate our framework on two foundation model series: Qwen2.5-VL (Qwen2.5-VL-8B-Instruct) [8] and InternVL3 (InternVL3-8B) [32], representing state-of-the-art vision-language architectures with distinct pretraining approaches.

Evaluation We evaluate on the EB-ALFRED benchmark from EmbodiedBench [25], which provides refined evaluation protocols over the original ALFRED benchmark, including streamlined action spaces and higher-quality language instructions. Beyond the standard Success Rate (SR) metric, we also use the standard deviation to quantify a model’s ability to maintain consistent performance across varying task complexities:

$$\text{STD} = \sqrt{\frac{1}{6} \sum_{c \in \mathcal{C}} (SR_c - \overline{SR})^2} \quad (6)$$

Where \mathcal{C} represents the six task categories (Base, Common, Complex, Visual, Spatial, and Long), and \overline{SR} is the mean success rate across all categories. Lower STD values indicate more balanced capabilities across different task types, while higher values suggest uneven performance that excels in some categories but struggles in others, representing lower robustness. This metric provides crucial insight into model robustness beyond aggregate performance measures.

4.2 Main results

As demonstrated in Table 1, our curriculum learning framework establishes new state-of-the-art performance across all task categories while operating under strictly realistic observation constraints.

Table 1: Performance comparison on EmbodiedBench (EB-ALFRED). Results show success rates () across task categories. Models marked with † indicate our proposed approach and its variants. Best results in **bold**.

Model	Avg.	STD↓	Base	Common	Complex	Visual	Spatial	Long
<i>Proprietary Models (Original open-loop setting with action feedback)</i>								
GPT-4o	56.3	7.8	64	54	68	46	52	54
Claude-3.5-Sonnet	64.0	8.6	72	66	76	60	58	52
Gemini-1.5-Pro	62.3	7.8	70	64	72	58	52	58
Gemini-2.0-flash	52.3	6.2	62	48	54	46	46	58
Gemini-1.5-flash	39.3	10.6	44	40	56	42	26	28
GPT-4o mini	24.0	13.0	34	28	36	24	22	0
<i>Open-Source Models</i>								
InternVL2.5-78B-MPO	40.0	4.5	48	36	42	40	40	34
Qwen2.5-VL-72B-Ins	39.7	6.3	50	42	42	36	34	34
Qwen2-VL-72B-Ins	33.7	4.8	40	30	40	30	32	30
Llama-3.2-90B-Vision-Ins	32.0	10.1	38	34	44	28	32	16
InternVL2.5-38B-MPO	25.7	4.7	30	20	20	28	32	24
InternVL2.5-38B	23.3	9.0	36	30	36	22	14	26
Llama-3.2-11B-Vision-Ins	13.7	7.4	24	8	16	22	6	6
InternVL2.5-8B-MPO	7.7	4.3	12	6	14	6	6	2
Qwen2.5-VL-7B-Ins	4.7	3.9	10	8	6	2	0	2
Qwen2-VL-7B-Ins	1.7	2.3	6	0	2	0	0	2
InternVL3-8B	10.7	7.6	20	12	20	8	2	2
<i>Proprietary Models (Under harder closed-loop settings)†</i>								
GPT-4o	26	2.8	30	26	28	22	26	24
Claude-3.5-Sonnet	57.3	13.4	62	62	64	40	42	74
<i>Our Approach with InternVL3-8B†</i>								
InternVL3-8B	6	4.7	12	8	10	2	0	4
+ Basic Reasoning	46.0	18.4	58	16	56	46	34	66
+ Full Augmentation	57.0	5.5	62	52	60	58	48	62
+ Curriculum Learning	61.0	7.2	66	56	66	58	50	70
<i>Our Approach with Qwen2.5-VL-7B-Ins†</i>								
Qwen2.5-VL-7B-Ins	2	2.5	6	2	4	0	0	0
+ Basic Reasoning	47.0	14.0	64	22	48	50	44	54
+ Full Augmentation	58.0	6.8	60	62	62	46	54	64
+ Curriculum Learning	62.7	6.3	66	62	70	56	52	70

203 The Qwen2.5-VL implementation achieves a 13.5× improvement in average success rate (from 4.67
204 to 62.67) with 14% improvement compared with baseline method, surpassing even GPT-4o (56.3)
205 and approaching Gemini-1.5-Pro (62.3) under harder settings — without having access to privileged
206 environmental information. Similar performance gains are observed with InternVL3 models, which
207 improve from 10.67 to 61.0.

208 Notably, our approach maintains a competitive Standard Deviation (STD) of 6.3, lower than many
209 proprietary models such as Claude-3.5-Sonnet (8.6) and Gemini-1.5-flash (10.6), indicating more
210 balanced capabilities across diverse task categories. This represents a substantial improvement over
211 our basic reasoning approach, which exhibits a high STD of 14.0, suggesting uneven performance
212 that handles some categories well but struggles significantly with others. The progression from basic
213 reasoning (STD=14.0) to curriculum learning (STD=6.3) demonstrates how our multi-dimensional
214 enhancement approach systematically builds more balanced cognitive capabilities.

215 **Enhanced Environmental Cognition** Our framework demonstrates substantial improvements
216 in environmental understanding capabilities, evidenced by performance across specialized task
217 categories:

1. **Visual Perception:** For InternVL3, the success rate improves from 46 (baseline) to 58 on tasks requiring the identification of objects based on appearance (e.g., distinguishing objects by color, shape, or pattern).
2. **Spatial Reasoning:** Performance rises from 34 to 50 (InternVL3) on tasks requiring precise positioning and relational understanding (e.g., “place the object to the left of the sink”).
3. **Semantic Grounding:** For Qwen2.5-VL, commonsense reasoning accuracy increases from 22 to 62, enabling effective handling of instructions requiring real-world knowledge and functional understanding of objects.
4. **Referential Resolution:** The model successfully resolves ambiguous references (e.g., “that container mentioned before”) through contextual reasoning with accuracy raising from 48 to 70 (Qwen2.5-VL), overcoming a critical limitation in conventional vision-language systems.

These improvements underscore the importance of structured environmental modeling beyond simple action prediction for robust embodied planning.

Further generalization on long-horizon planning Our approach achieves remarkable 70 success rate on long-horizon tasks—those requiring 15+ sequential actions—representing a 35-fold improvement over baseline models and matching Claude-3.5-Sonnet’s performance in this challenging category. Notably, while proprietary models show mixed results under the stricter closed-loop setting (without action feedback), our framework maintains consistent performance across all settings.

This exceptional capability stems from two key innovations, (1) **Full Temporal Context:** Our training paradigm incorporates complete observation histories rather than isolated frames, enabling the model to capture causal relationships between actions and environmental changes across extended sequences. (2) **Multi-dimensional Knowledge Integration:** The four cognitive dimensions of our data enhancement approach collectively enable the model to maintain coherent world representations throughout extended task execution.

The dramatic performance disparity between GPT-4o (24 in closed-loop vs. 54 in open-loop settings) and Claude-3.5-Sonnet (74 in closed-loop vs. 52 in open-loop) on long-horizon tasks is particularly revealing. While GPT-4o struggles without action feedback, likely falling into repetitive error patterns, Claude-3.5-Sonnet actually improves under closed-loop constraints, suggesting superior error recognition and recovery capabilities. Our framework (70 success rate) approaches Claude’s closed-loop performance without requiring proprietary model access, validating our hypothesis that proper environmental modeling serves as a critical foundation for compositional task planning. This confirms that closed-loop settings, despite being more challenging, better reflect the requirements for successful long-horizon planning in real-world scenarios.

5 Analysis

5.1 Self-Directed Enhancement Potential

To explore autonomous data augmentation capabilities, we investigate a self-directed enhancement approach where models independently select augmentation strategies based on task descriptions. This implicit enhancement contrasts with our explicit curriculum learning framework.

Table 2: Comparison between explicit and self-directed enhancement approaches

Method	Avg	STD↓	Base	Common	Complex	Visual	Spatial	Long
Basic Reasoning	47.0	14.0	64	22	48	50	44	54
Explicit (Standard)	58.0	6.8	60	62	62	46	54	64
Explicit (Curriculum)	62.7	6.3	66	62	70	56	52	70
Self-Directed (Standard)	54.7	7.7	60	50	62	44	50	62
Self-Directed (Curriculum)	56.7	7.2	66	48	62	52	52	60

As shown in Table 2, self-directed enhancement achieves moderate success (56.7 average), but exhibits notable limitations compared to our explicit framework (62.7). The self-directed approach demonstrates reasonable environmental perception capabilities—matching explicit methods in Visual

(52 vs. 56) and Spatial (52 vs. 52) categories. However, it shows significant deficiencies in semantic reasoning: Commonsense understanding (48 vs. 62) and Long-horizon planning (60 vs. 70).

While the self-directed curriculum approach improves performance consistency (STD=7.2) compared to the basic reasoning baseline (STD=14.0), it remains less balanced than our explicit curriculum method (STD=6.3). This suggests that while the model can autonomously develop more uniform capabilities across tasks, it still benefits from structured guidance in developing complementary cognitive skills. The model shows particular weakness in commonsense reasoning (48), indicating a tendency toward surface-level linguistic modifications rather than deeper semantic understanding.

Nevertheless, the 56.7 success rate achieved without human-designed enhancement rules suggests promising avenues for autonomous improvement in future work.

5.2 Ablation Study

We conduct a systematic component analysis to isolate the contributions of individual elements within our framework. Table 3 presents the performance across the controlled ablation variants using Qwen2.5-VL-7B as the base model.

Table 3: Ablation analysis of framework components

Configuration	Avg	STD↓	Base	Common	Complex	Visual	Spatial	Long
Base Model	4.7	3.9	10	8	6	2	0	2
+ Basic Reasoning	47.0	14.0	64	22	48	50	44	54
+ Visual/Spatial Only	46.7	17.1	60	16	56	46	42	60
+ Partial Reasoning	54.0	9.3	62	46	64	52	40	60
+ Complete Reasoning	58.0	6.8	60	62	62	46	54	64
+ Curriculum (Full)	62.7	6.3	66	62	70	56	52	70

The basic reasoning baseline achieves 47.0 average success but exhibits highly inconsistent performance across task categories (STD=14.0), revealing fundamental limitations of naive instruction enhancement. Models trained with this approach overfit to explicit action descriptions without developing balanced environmental understanding, leading to particularly poor performance on commonsense tasks (22 vs. 62 with full framework).

Restricting enhancement to visual and spatial dimensions (+Visual/Spatial Only) maintains similar average performance (46.7) but actually increases performance variance (STD=17.1), with a dramatic gap between the highest (60 Long) and lowest (16 Common) categories. This confirms that perceptual enhancements alone are insufficient for developing comprehensive cognitive capabilities.

The progression from partial to complete reasoning replacement shows the importance of coherent action sequencing. Partial replacement improves average performance (54.0) and reduces inconsistency (STD=9.3), but still shows substantial variance across categories. In contrast, complete reasoning replacement achieves both higher average performance (58.0) and significantly more balanced capabilities (STD=6.8), demonstrating coherent modeling of environment-action relationships.

The full curriculum framework further elevates both average performance (62.7) and consistency (STD=6.3), particularly in long-horizon scenarios (70 success, +12 over Complete Reasoning), validating our progressive skill acquisition hypothesis. Notably, the framework maintains this balanced performance profile despite a 13.5× increase in overall success rate compared to the base model, indicating systematic improvement across all cognitive dimensions rather than specialized gains in particular categories.

5.3 Case Study

To provide qualitative insights into our model’s capabilities, we analyze representative examples of embodied planning and reasoning. Our case studies focus on (1) cognitive process transparency in complex multi-step planning and (2) the effectiveness of our multi-dimensional augmentation approach.

Figure 2 illustrates the execution of a seemingly simple instruction: “Place a chilled apple section into the bin.” This instruction contains implicit procedural requirements—the apple must be sliced and chilled—necessitating a sequence of steps beyond merely disposing of an apple.



Figure 2: **Reasoning process visualization for complex instruction execution.** The figure shows our model executing the instruction “Place a chilled apple section into the bin.” The model successfully decomposes this seemingly simple instruction into 18 distinct actions across three phases, demonstrating robust planning capabilities. Reasoning annotations highlight five critical cognitive abilities: task decomposition, functional understanding, situational awareness, object property reasoning, and commonsense knowledge application. This example illustrates how our model maintains coherent planning over a long horizon (18 steps) while handling implicit requirements not explicitly stated in the instruction (e.g., the apple must be chilled before disposal).

Our Qwen2.5-VL model with curriculum learning successfully decomposes this task into 18 distinct actions across three phases. First, the model identifies necessary tools, demonstrating task decomposition by recognizing that sectioning requires a knife. Second, it exhibits situational awareness, notably in Reasoning 6 where it places the knife down before handling the apple, showing safety awareness. Most importantly, the model correctly infers the need to refrigerate the apple to fulfill the “chilled” requirement—demonstrating both functional understanding of appliances and commonsense knowledge application. This example highlights improvements achieved through our curriculum learning approach, as baseline models consistently failed to maintain coherence across extended sequences, typically omitting the critical chilling step.

6 Conclusion

In this paper, we address fundamental limitations in embodied agents where current LVLMs struggle with complex real-life scenarios due to their environment-agnostic imitation learning paradigm. Previous approaches treat task instructions and environmental contexts as disconnected elements, leading to poor generalization in unfamiliar environments and inconsistent reasoning during multi-step tasks. We introduce a world-aware narrative enhancement approach that systematically develops four cognitive capabilities: visual appearance modeling, spatial-relational reasoning, functional abstraction learning, and syntactic grounding. Our experiments on EB-ALFRED demonstrate substantial improvements with Qwen2.5-VL and InternVL3, achieving up to 60.7 higher success rates and outperforming even proprietary models like GPT-4o in challenging scenarios. Our approach

proves that high-performance embodied planning is possible using only raw visual observations without privileged feedback, establishing a new state-of-the-art for embodied AI systems in complex, real-life-like environments.

References

- [1] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. *ArXiv*, abs/2302.02662, 2023. URL <https://api.semanticscholar.org/CorpusID:256615643>.
- [2] Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Dongbin Zhao, and He Wang. Robogpt: an intelligent agent of making embodied long-term decisions for daily instruction tasks, 2024. URL <https://arxiv.org/abs/2311.15649>.
- [3] Jacky Liang, Wenlong Huang, F. Xia, Peng Xu, Karol Hausman, Brian Ichter, Peter R. Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500, 2022. URL <https://api.semanticscholar.org/CorpusID:252355542>.
- [4] Yang Liu, Weixing Chen, Yongjie Bai, Guanbin Li, Wen Gao, and Liang Lin. Aligning cyber space with physical world: A comprehensive survey on embodied AI. *CoRR*, abs/2407.06886, 2024. doi: 10.48550/ARXIV.2407.06886. URL <https://doi.org/10.48550/arXiv.2407.06886>.
- [5] Jinjie Mai, Jun Chen, Bing chuan Li, Guocheng Qian, Mohamed Elhoseiny, and Bernard Ghanem. Llm as a robotic brain: Unifying egocentric memory and control. *ArXiv*, abs/2304.09349, 2023. URL <https://api.semanticscholar.org/CorpusID:258212642>.
- [6] OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.
- [7] Alexander Pashevich, Cordelia Schmid, and Chen Sun. Episodic transformer for vision-and-language navigation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15922–15932, 2021. URL <https://api.semanticscholar.org/CorpusID:234482879>.
- [8] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- [9] Lawrence Shapiro and Shannon Spaulding. Embodied Cognition. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2025 edition, 2025.
- [10] Suyeon Shin, Sujin Jeon, Junghyun Kim, Gi-Cheon Kang, and Byoung-Tak Zhang. Socratic planner: Inquiry-based zero-shot planning for embodied instruction following. *ArXiv*, abs/2404.15190, 2024. URL <https://api.semanticscholar.org/CorpusID:269302975>.
- [11] Keisuke Shirai, Cristian Camilo Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. Vision-language interpreter for robot task planning. *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2051–2058, 2023. URL <https://api.semanticscholar.org/CorpusID:264935138>.
- [12] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Motlaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks, 2020. URL <https://arxiv.org/abs/1912.01734>.

- [13] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=0IOX0YcCdTn>.
- [14] R. S. Siegler. Children’s learning. *American Psychologist*, 60(8):769–778, 2005. doi: 10.1037/0003-066X.60.8.769.
- [15] 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. *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2022. URL <https://api.semanticscholar.org/CorpusID:252519594>.
- [16] Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2986–2997, 2022. URL <https://api.semanticscholar.org/CorpusID:254408960>.
- [17] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization for LLM agents. *CoRR*, abs/2403.02502, 2024. doi: 10.48550/ARXIV.2403.02502. URL <https://doi.org/10.48550/arXiv.2403.02502>.
- [18] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplaner: Adaptive planning from feedback with language models. *ArXiv*, abs/2305.16653, 2023. URL <https://api.semanticscholar.org/CorpusID:258947337>.
- [19] Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazouze, Walter Talbott, Katherine Metcalf, Natalie Mackraz, Devon Hjelm, and Alexander Toshev. Large language models as generalizable policies for embodied tasks. *ArXiv*, abs/2310.17722, 2023. URL <https://api.semanticscholar.org/CorpusID:264555578>.
- [20] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024. URL <https://api.semanticscholar.org/CorpusID:268297180>.
- [21] Siyin Wang, Zhaoye Fei, Qinyuan Cheng, Shiduo Zhang, Panpan Cai, Jinlan Fu, and Xipeng Qiu. World modeling makes a better planner: Dual preference optimization for embodied task planning. *CoRR*, abs/2503.10480, 2025. doi: 10.48550/ARXIV.2503.10480. URL <https://doi.org/10.48550/arXiv.2503.10480>.
- [22] Zidan Wang, Rui Shen, and Bradly C. Stadie. Wonderful team: Zero-shot physical task planning with visual llms. 2024. URL <https://api.semanticscholar.org/CorpusID:271533474>.
- [23] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *ArXiv*, abs/2307.01848, 2023. URL <https://api.semanticscholar.org/CorpusID:259342896>.
- [24] Jingkan Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Octopus: Embodied vision-language programmer from environmental feedback. In *European Conference on Computer Vision*, 2023. URL <https://api.semanticscholar.org/CorpusID:263909250>.
- [25] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *ArXiv*, abs/2502.09560, 2025. URL <https://api.semanticscholar.org/CorpusID:276317279>.

- [26] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025. URL <https://arxiv.org/abs/2502.09560>.
- [27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.
- [28] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S. Yu. Large language models for robotics: A survey, 2023. URL <https://arxiv.org/abs/2311.07226>.
- [29] Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, and Chuang Gan. Building cooperative embodied agents modularly with large language models. *ArXiv*, abs/2307.02485, 2023. URL <https://api.semanticscholar.org/CorpusID:259342833>.
- [30] Qi Zhao, Haotian Fu, Chen Sun, and George Dimitri Konidaris. Epo: Hierarchical llm agents with environment preference optimization. *ArXiv*, abs/2408.16090, 2024. URL <https://api.semanticscholar.org/CorpusID:272146208>.
- [31] Zirui Zhao, Wee Sun Lee, and David Hsu. Large language models as commonsense knowledge for large-scale task planning. *ArXiv*, abs/2305.14078, 2023. URL <https://api.semanticscholar.org/CorpusID:258841057>.
- [32] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>.

447 7 Technical Appendices and Supplementary Material

448 A Limitation

449 While our framework demonstrates significant advancements in embodied planning, several limita-
450 tions warrant further discussion. First, our world-aware narrative enhancement operates primarily
451 at the symbolic action level (e.g., "pick up knife"), lacking explicit modeling of continuous control
452 parameters. This abstraction necessitates future integration with low-level controllers to enable
453 practical deployment requiring force modulation and precise trajectory optimization. Second, while
454 our experiments demonstrate effectiveness in household environments based on the ALFRED dataset,
455 the framework’s generalizability to industrial settings or outdoor scenarios with dynamic obstacles
456 remains unverified. These domains may present distinct challenges in spatial reasoning and adaptive
457 navigation that our current implementation does not address. Third, in our approach we mainly
458 focus on enhancing user instruction and step-wise reasoning, limiting the fine-tuned model’s capacity
459 for mid-execution error correction. This constraint suggests the need for more dynamic enhance
460 mechanisms in future iterations.

461 B Training Details

462 B.1 Narrative Enhancement

463 Our instruction enhancement pipeline employs the Qwen2.5-VL-72B-Instruct model under
464 manufacturer-recommended configurations, utilizing temperature sampling ($\tau=1.0$) with nucleus
465 filtering (top-p=0.7) for balanced diversity and coherence. This process generates both narrative-
466 augmented instructions and associated step-wise rationales, maintaining strict alignment with the
467 original ALFRED action trajectories through constrained decoding mechanisms.

468 B.2 Model Fine-tuning

469 We implement full-parameter optimization on two vision-language architectures: Qwen2.5-VL-72B-
470 Instruct and InternVL3-8B. The training regime employs AdamW optimization with base learning
471 rate $\eta = 110^{-5}$, 10% linear warmup, and cosine decay scheduling over 3 epochs. Experiments utilize
472 contrastive context windows (16k/32k tokens) with per-device batch size 4, distributed across 8xA100-
473 80GB nodes via tensor parallelism. The complete training cycle requires 14 hours per model variant,
474 aggregating to 800 A100 GPU-hours when accounting for ablation studies and hyperparameter tuning.
475 Memory optimization is achieved through Flash Attention v2 and BF16 mixed-precision training,
476 maintaining numerical stability while maximizing hardware utilization.

477 C Data Generation Process

478 C.1 Dataset Statistics

479 Our enhanced dataset comprises 80,875 instruction-trajectory pairs derived from 16,145 original
480 ALFRED trajectories, systematically expanded through two complementary approaches. The first
481 expansion preserves the original trajectory structure while adding four specialized narrative en-
482 hancements (visual, spatial, functional and syntactic), yielding four distinct subsets each containing
483 16,145 samples. Unlike conventional datasets that provide only sparse instructions and atomic
484 action sequences, our framework enriches each trajectory with: (1) step-wise observation images
485 capturing environmental states, and (2) step-wise reasoning annotations detailing action rationales
486 and preconditions.

487 Additionally, we provide a comparative dataset of 32,290 samples featuring implicit-instruction
488 augmentation. This contrastive set employs self-supervised prompting techniques where models
489 autonomously determine enhancement requirements through preliminary attention patterns, rather
490 than receiving explicit annotation guidelines. This dual-structure design enables systematic evaluation
491 of both human-guided and model-induced enhancement strategies, while maintaining parity in
492 environmental complexity and task diversity with the original ALFRED distribution.

493 C.2 Prompt Templates for Instruction Augmentation

494 Here are the prompt templates used for generating the enhanced instructions across the four cognitive
495 dimensions.

496 Visual Dimension Prompt

Visual Dimension Prompt:

Enhance the following instruction by adding spatial descriptions based on the image. Keep it natural and concise.

Examples:

1. Original: "Put two spray bottles in the cabinet"
Enhanced: "Place two cylindrical green cleaning spray bottles in the wooden cabinet."

2. Original: "Put a knife in a container"
Enhanced: "Put a 20cm silver chef knife into the blue rectangular plastic container."

Now enhance:

Original: {human_instruction}
Enhanced:

Figure 3: Prompt for instruction visual enhancement

497 Spatial Dimension Prompt

Spatial Dimension Prompt

Enhance the following instruction by adding multi-layered spatial descriptions based on the image. Refer to objects through their positional relationships with 2-3 adjacent landmarks. Keep descriptions natural and concise.

Examples:

1. Original: "Put two spray bottles in the cabinet"
Enhanced: "Put two spray bottles in the white cabinet under the stainless steel sink against the wall"

2. Original: "Put a knife in a container"
Enhanced: "Place the chef's knife holder on the granite countertop, positioned to the right of the refrigerator"

Now enhance:

Original: {human_instruction}
Enhanced:

Figure 4: Prompt for instruction spatial enhancement

498 **Functional Dimension Prompt**

Functional Dimension Prompt

Enhance the following instruction by adding functional descriptions based on world knowledge. Replace one object or its placement with an indirect reference while keeping the sentence natural and concise.

Examples:

1. Original: "Put two spray bottles in the cabinet"
Enhanced: "Place two items used for misting surfaces inside the cabinet."

2. Original: "Put a knife in a container"
Enhanced: "Insert an object commonly used for cutting into a container designed for safekeeping."

Now enhance:
Original: {human_instruction}
Enhanced:

Figure 5: Prompt for instruction functional enhancement

499 **Syntactic Dimension Prompt**

Syntactic Dimension Prompt

Enhance the following instruction into a more elaborate version by adding contextual details and symbolic substitutions. Replace objects and locations with contextual references (e.g., pronouns or implied terms) and include irrelevant but plausible background information. Keep sentences concise and avoid adding new actions.

Examples:

1. Original: "Put two spray bottles in the cabinet"
Enhanced: "The spray bottles are on the shelf, already cleaned from yesterday's use. Move them to the cabinet for storage."

2. Original: "Put a knife in a container"
Enhanced: "That knife on the counter just finished slicing vegetables. Place it in the container to keep the edge protected."

3. Original: "Put washed lettuce in the refrigerator"
Enhanced: "There's a lettuce in the sink—we've prepped enough for dinner. Wash it and store there to keep it fresh."

Now enhance:
Original: {human_instruction}
Enhanced:

Figure 6: Prompt for instruction syntactic enhancement

500 C.2.1 Verification Prompt for Semantic Consistency

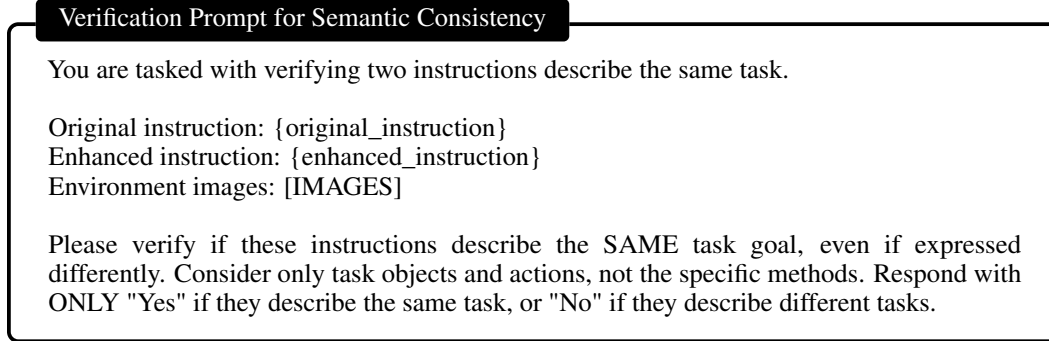


Figure 7: Verification Prompt for Semantic Consist

501 C.2.2 Reasoning Generation Prompt

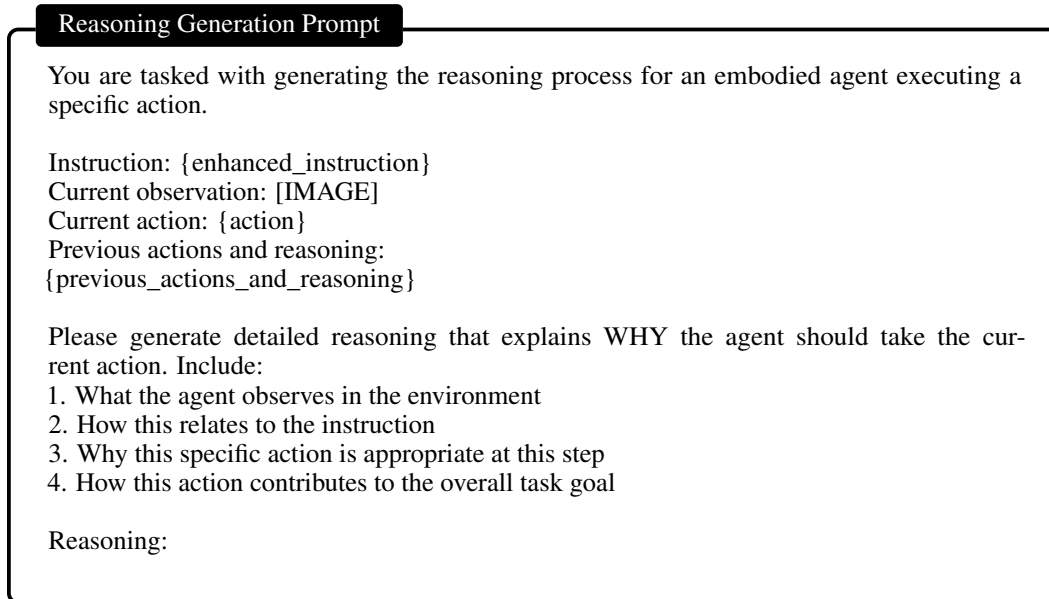


Figure 8: Reasoning Generation Prompt

502 D Broader Impact

503 Our framework enhances embodied AI’s environmental reasoning capabilities, offering potential
504 benefits for assistive robotics and cognitive science research through human-like task decomposition.
505 Improved household automation could empower vulnerable populations, while the explicit modeling
506 of environmental semantics may advance embodied intelligence studies.

507 The approach remains limited in three critical aspects: (1) Operating at symbolic action levels without
508 continuous control parameter modeling, requiring integration with low-level controllers for real-world
509 force modulation; (2) Unverified generalization beyond household environments to industrial/outdoor
510 settings with dynamic obstacles and specialized navigation demands; (3) Restricted mid-execution
511 error correction capacity due to the current enhancement paradigm’s focus on pre-planned reasoning.
512 These limitations highlight essential directions for future work to bridge simulation-to-reality gaps
513 and enable robust real-world deployment.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: The claims in the abstract and introduction accurately reflect the paper's contributions. We identify the limitation of environment-agnostic planning in current LVLMS and propose a world-aware enhancement framework that systematically develops four cognitive dimensions (visual, spatial, functional, and syntactic).

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitation is discussed in Appendix Part with a separate "Limitations" section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Code and model generation configs are provided in supplementary materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code is provided in supplementary materials and will release soon.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Setting and Details are discussed in section 4.1 and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We follow the evaluation settings in Embodiedbench, which does not contain error bar settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96 CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Experiments compute resources part is discussed in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We carefully follow NeurIPS Code of Ethics in our research.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This part is discussed in the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the dataset from ALFRED (<https://github.com/askforalfred/alfred>) and evaluation benchmark from Embodied-bench. (<https://github.com/EmbodiedBench/EmbodiedBench>)

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

774 • For existing datasets that are re-packaged, both the original license and the license of
775 the derived asset (if it has changed) should be provided.
776 • If this information is not available online, the authors are encouraged to reach out to
777 the asset’s creators.

778 **13. New assets**

779 Question: Are new assets introduced in the paper well documented and is the documentation
780 provided alongside the assets?

781 Answer: [\[Yes\]](#)

782 Justification: All assets are discussed in main paper and Appendix.

783 Guidelines:

784 • The answer NA means that the paper does not release new assets.
785 • Researchers should communicate the details of the dataset/code/model as part of their
786 submissions via structured templates. This includes details about training, license,
787 limitations, etc.
788 • The paper should discuss whether and how consent was obtained from people whose
789 asset is used.
790 • At submission time, remember to anonymize your assets (if applicable). You can either
791 create an anonymized URL or include an anonymized zip file.

792 **14. Crowdsourcing and research with human subjects**

793 Question: For crowdsourcing experiments and research with human subjects, does the paper
794 include the full text of instructions given to participants and screenshots, if applicable, as
795 well as details about compensation (if any)?

796 Answer: [\[NA\]](#)

797 Justification: The paper does not involve crowdsourcing nor research with human subjects.

798 Guidelines:

799 • The answer NA means that the paper does not involve crowdsourcing nor research with
800 human subjects.
801 • Including this information in the supplemental material is fine, but if the main contribu-
802 tion of the paper involves human subjects, then as much detail as possible should be
803 included in the main paper.
804 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
805 or other labor should be paid at least the minimum wage in the country of the data
806 collector.

807 **15. Institutional review board (IRB) approvals or equivalent for research with human**
808 **subjects**

809 Question: Does the paper describe potential risks incurred by study participants, whether
810 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
811 approvals (or an equivalent approval/review based on the requirements of your country or
812 institution) were obtained?

813 Answer: [\[NA\]](#)

814 Justification: The paper does not involve crowdsourcing nor research with human subjects.

815 Guidelines:

816 • The answer NA means that the paper does not involve crowdsourcing nor research with
817 human subjects.
818 • Depending on the country in which research is conducted, IRB approval (or equivalent)
819 may be required for any human subjects research. If you obtained IRB approval, you
820 should clearly state this in the paper.
821 • We recognize that the procedures for this may vary significantly between institutions
822 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
823 guidelines for their institution.
824 • For initial submissions, do not include any information that would break anonymity (if
825 applicable), such as the institution conducting the review.

826 **16. Declaration of LLM usage**
827 Question: Does the paper describe the usage of LLMs if it is an important, original, or
828 non-standard component of the core methods in this research? Note that if the LLM is used
829 only for writing, editing, or formatting purposes and does not impact the core methodology,
830 scientific rigorousness, or originality of the research, declaration is not required.
831 Answer: [Yes]
832 Justification: Details are discussed in Section 3.
833 Guidelines:
834 • The answer NA means that the core method development in this research does not
835 involve LLMs as any important, original, or non-standard components.
836 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
837 for what should or should not be described.