H-WM: ROBOTIC TASK AND MOTION PLANNING GUIDED BY HIERARCHICAL WORLD MODEL

Wenyuan Chen[1,2*], **Jinbang Huang**[1*], **Oscar Pang**[1,2], **Zhiyuan Li**[1,2], **Xiao Hu**[1], **Lingfeng Zhang**[1], **Zhanguang Zhang**[1], **Mark Coates**[3], **Tongtong Cao**[1], **Xingyue Quan**[1], **Yingxue Zhang**[1]
[1] Huawei Noah's Ark Lab, [2] University of Toronto, [3] McGill University
[*] Equal contribution,

ABSTRACT

World models are becoming central to robotic planning and control as they enable prediction of future state transitions. Existing approaches often emphasize video generation or natural-language prediction, which are difficult to directly ground in robot actions and suffer from compounding errors over long horizons. Traditional task and motion planning relies on symbolic-logic world models, such as planning domains, that are robot-executable and robust for long-horizon reasoning, but typically operate independently of visual perception, preventing synchronized symbolic and perceptual state prediction. We propose a Hierarchical World Model (H-WM) that jointly predicts logical and visual state transitions within a unified bilevel framework. H-WM combines a high-level logical world model with a low-level visual world model, integrating the robot-executable, long-horizon robustness of symbolic reasoning with perceptual grounding from visual observations. The hierarchical outputs provide stable and consistent intermediate guidance for long-horizon tasks, mitigating error accumulation and enabling robust execution across extended task sequences. To train H-WM, we further introduce a robotic dataset that aligns robot motion with symbolic states, actions, and visual observations. Experiments across vision–language–action (VLA) control policies demonstrate the effectiveness and generality of the approach.

1 INTRODUCTION

Recent advances in Vision–Language–Action (VLA) models have enabled robotic systems that tightly couple multimodal perception and control via large pre-trained foundation models, achieving stronger generalization than traditional modular pipelines. However, most existing VLA methods adopt an end-to-end paradigm that maps visual observations and language instructions directly to low-level actions, and their performance degrades substantially on long-horizon tasks (Neau et al., 2025; Yang et al., 2025a). This failure is driven by compounding execution errors, ambiguous goal specifications, limited intermediate supervision, and overfitting to agent-centric representations.

A natural response to these limitations is to introduce richer intermediate guidance; however, existing approaches fall into three dominant paradigms, each with fundamental shortcomings. First, LLM-based hierarchical planners decompose tasks into subgoals or action sequences (Shi et al., 2025b; Li et al., 2025a; Wang et al., 2024b; Zhi et al., 2025; Black et al., 2024; Intelligence et al., 2025), but are fundamentally constrained by language as the intermediate interface: LLMs struggle to reason about physical constraints, and their vague, unstructured representations lead to semantic–execution misalignment (Shi et al., 2025b; Chen et al., 2025). Second, world-model-based approaches aim to provide predictive visual guidance (Shao et al., 2025; Chen et al., 2025; Xiang et al., 2025), yet existing formulations suffer from complementary limitations, particularly in long-horizon settings where compounding prediction errors degrade planning reliability. Third, classical Task and Motion Planning (TAMP) achieves long-horizon consistency through explicit logical world models for symbolic reasoning (Kaelbling & Lozano-Pérez, 2011), but relies on manually designed abstractions and engineered perception-to-symbol pipelines that are weakly aligned with raw visual observations, resulting in brittleness to perception noise and poor scalability to unstructured

Correspondence to: jinbang.huang@h-partners.com, {zhanguang.zhang, yingxue.zhang}@huawei.com

environments (Silver et al., 2021). Consequently, none of these paradigms delivers the informative, grounded, and long-horizon–robust guidance required for reliable VLA execution.

In this paper, we propose a novel hierarchical world model (H-WM) that jointly predicts logical and visual state transitions within a unified framework, enabling more effective intermediate guidance for VLA models on complex long-horizon tasks. First, we introduce a logical world model that performs long-horizon symbolic reasoning by predicting structured logical state transitions and action sequences, providing globally consistent task-level guidance while explicitly enforcing logical consistency and physical constraints. Second, we introduce a feature-embedding–based visual world model conditioned on logical states and actions, which generates a sequence of latent visual subgoals to ground symbolic plans in perceptual space. Together, the proposed hierarchical world model bridges high-level symbolic reasoning and low-level perceptual grounding, delivering informative, grounded, and long-horizon–robust guidance by combining the complementary strengths of prior approaches. To summarize, our key contributions are:

- We propose a training pipeline for **hierarchical world model** that jointly predicts world transition dynamics at multiple abstraction levels, aligning long-horizon logical state transitions with visual state evolution to support coherent task execution over extended horizons.
- We introduce a **logical world model**, realized as an LLM fine-tuned for symbolic planning, that internalizes long-horizon planning behavior and mitigates the brittleness of classical symbolic planners while delivering globally consistent and interpretable task-level guidance.
- We develop a **latent-feature visual world model** that conditions on previous observations and the logical actions and states predicted by the logical world model to generate compact latent visual subgoals, providing stable and rich visual guidance.
- We present a **systematic pipeline** that integrates hierarchical logical and visual guidance into VLA models, allowing low-level policies to flexibly condition on logical actions, logical states, visual subgoals, or their combinations for physically feasible execution.

2 Related Work

2.1 World Model

World models have emerged as a powerful paradigm for embodied intelligence, addressing challenges such as data efficiency, execution safety, and standardized evaluation (Li et al., 2025c; Long et al., 2025). Recent work has shown their effectiveness across a range of robotic settings, including policy learning (Hafner et al., 2025; Li et al., 2025b), policy evaluation (Quevedo et al., 2025; Li et al., 2025d), test-time rollout simulation (Bar et al., 2025; Yang et al., 2025b), and unified model–policy training (Zhao et al., 2025; Bi et al., 2025; Zhang et al., 2025). By predicting action outcomes, world models enable reliable long-horizon reasoning for navigation and manipulation, and can be tightly integrated with policy generation to leverage large-scale, unlabeled multimodal data (Zhao et al., 2025; Bi et al., 2025; Zhang et al., 2025; Liao et al., 2025). Recent systems further demonstrate that world models can serve as the core of robotic control architectures and be directly adapted for action generation (Kim et al., 2026). Most existing approaches focus on pixel-level world modeling, yet real-world dynamics can be captured at multiple abstraction levels, including latent features, keypoints, and object-centric representations (Xie et al., 2019; Hoque et al., 2022; Zhou et al., 2024; Assran et al., 2025; Manuelli et al., 2020; Liu et al., 2023b). Lower-level representations offer high expressivity but suffer from poor sample efficiency and limited out-of-distribution generalization (Xie et al., 2019; Hoque et al., 2022). In contrast, higher-level abstractions improve efficiency and generalization at the cost of reduced representational capacity (Manuelli et al., 2020; Liu et al., 2023b; Driess et al., 2023). To balance this trade-off, several works have explored multi-level world models. HiP (Ajay et al., 2023) grounds symbolic plans into visuomotor control via video generation and inverse dynamics, but relies heavily on pixel-space prediction and predates recent advances in VLA models. Xing et al. (2025) introduces hierarchical continuous–discrete representations, yet its plans are not grounded in executable robotic control.

2.2 VISION–LANGUAGE–ACTION MODELS

Vision–Language–Action (VLA) models enable direct mapping from visual observations and language instructions to robotic actions and have become a dominant paradigm for general-purpose robot control (Ma et al., 1778; Zhong et al., 2025). End-to-end VLA models directly generate low-level actions from multimodal inputs via large-scale pretraining (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024; Intelligence et al., 2025), but suffer from performance degradation on long-horizon tasks due to goal ambiguity and error accumulation (Yang et al., 2025a). Hierarchical VLA models mitigate this issue by introducing intermediate representations that decouple high-level planning from low-level execution (Shao et al., 2025). Prior work explores sparse geometric abstractions such as keypoints (Wu et al., 2025; Yuan et al., 2024), discrete skill libraries (Ahn et al., 2022; Driess et al., 2023), language-based subtask decomposition (Shi et al., 2025b), and visually grounded guidance including goal images, point clouds, or videos (Chen et al., 2025; Zhen et al., 2024; Patel et al., 2025). While these approaches improve long-horizon reasoning, they either lack sufficient geometric fidelity or remain sensitive to distribution shifts in generated visual guidance, limiting robustness in complex manipulation scenarios.

2.3 TASK AND MOTION PLANNING

Task and motion planning (TAMP) combines high-level symbolic reasoning with low-level motion generation, enabling robots to perform complex, multi-step tasks under physical constraints. Classical TAMP methods integrate symbolic planners with geometric motion planners to ensure feasibility (Kaelbling & Lozano-Pérez, 2011; Toussaint, 2015), but often suffer from high computational cost and poor scalability in long-horizon or cluttered environments. To address these challenges, learning-based approaches have been proposed to accelerate planning. Prior work applies supervised learning to learn heuristics or policies from demonstrations (Silver et al., 2021; Dalal et al., 2023; McDonald et al., 2022), while reinforcement learning improves adaptability in uncertain and dynamic settings (Chitnis et al., 2016; Paxton et al., 2017; Xu et al., 2021). More recently, LLM-based TAMP leverages language priors to guide symbolic reasoning and planning (Huang et al., 2022; Wang et al., 2024a; Chen et al., 2024; Li et al., 2023). Despite these advances, a key limitation remains the reliance on manually designed symbolic abstractions. Recent work learns planning domains directly from data or descriptions (Diehl et al., 2021; Kumar et al., 2023; Silver et al., 2023; Liang et al., 2024; Wong et al., 2023; Mao et al., 2023; Liu et al., 2025; Zhu et al., 2024; Huang et al., 2025b; 2026; Guan et al., 2023; Han et al., 2024; Oswald et al., 2025), but focuses merely on abstract world modeling without visual grounding. Building on these efforts, we propose a H-WM framework that jointly captures symbolic and visual dynamics, enabling simultaneous prediction of logical transitions and visual observations, which improves low-level policy learning and execution.

3 PRELIMINARY

3.1 SYMBOLIC FORMALIZATION

A PDDL planning domain is defined as $\mathcal{D} = (\mathcal{P}, \mathcal{A})$, where $\mathcal{P}$ is a finite set of predicate symbols and $\mathcal{A}$ is a set of parameterized action schemas. For a given planning problem, let $\mathcal{O} = \{o_1, \ldots, o_n\}$ denote the set of objects. Each **predicate** $p \in \mathcal{P}$ is associated with an arity k and defines a Boolean relation $p : \mathcal{O}^k \to \{0, 1\}$. Instantiating a predicate with concrete objects yields a *ground atom*. The set of all possible ground atoms is defined as

$$\mathcal{G} = \{\, p(o_1, \ldots, o_k) \mid p \in \mathcal{P},\ o_i \in \mathcal{O} \,\}.$$

A symbolic state is represented as a set of true ground atoms, $\mathcal{X} \subseteq \mathcal{G}$. An **action schema** $a \in \mathcal{A}$ is defined as

$$a = \langle \mathrm{Pre}(a),\ \mathrm{Add}(a),\ \mathrm{Del}(a) \rangle,$$

where $\mathrm{Pre}(a)$, $\mathrm{Add}(a)$, and $\mathrm{Del}(a)$ are sets of atoms denoting the preconditions, add effects, and delete effects, respectively.

Binding an action schema a with a tuple of objects $(o_1, \ldots, o_j) \in \mathcal{O}^j$ produces a *ground action* $a(o_1, \ldots, o_j)$. A ground action is applicable in state $\mathcal{X}^t$ if

$$\mathrm{Pre}(a) \subseteq \mathcal{X}^m.$$

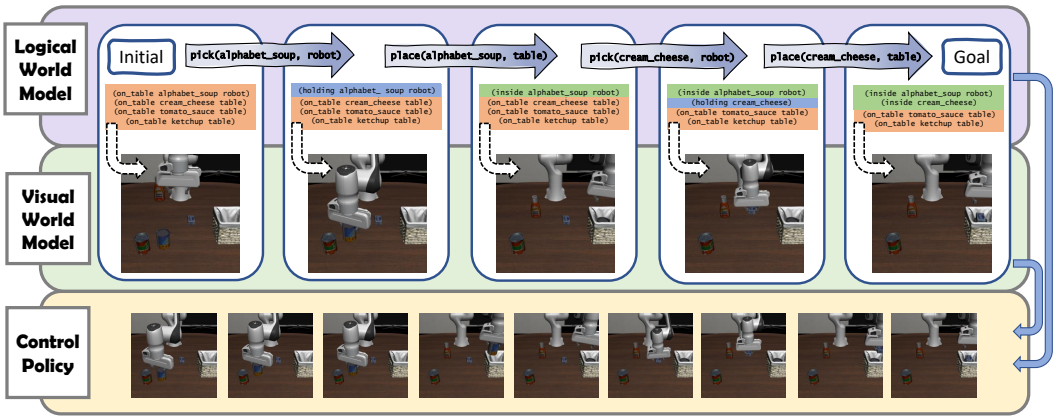

Figure 1: The proposed hierarchical world model jointly captures state transitions in both symbolic and perceptual spaces to guide robot motion policies. The logical world model performs long-horizon reasoning to predict logical states and action sequences. Conditioned on the previous observation, executed action, and predicted next logical state, the visual world model generates perceptually grounded sub-goal images that translate symbolic decisions into visual sub-goals. The low-level policy consumes the logical action, logical state, visual sub-goal, or their combination to produce continuous robot motions, enabling consistent task execution while maintaining physical feasibility.

Executing an applicable ground action induces a deterministic state transition

$$\mathcal{X}^{m+1} = \big(\mathcal{X}^m \setminus \mathrm{Del}(a)\big) \cup \mathrm{Add}(a).$$

3.2 Visual Representation

We consider a long-horizon task decomposed into a sequence of subtasks in logical space, indexed by m. At subtask step m, the robot receives a visual observation obs_m, represented as an RGB image, and a logical action prompt a_m describing the intended subtask. The symbolic logical state at step m is denoted by $\mathcal{X}_m$. The robot joint configuration is denoted by q_m. Let $\phi(\cdot)$ denote a fixed vision encoder that maps an image to a d-dimensional latent feature space $\mathcal{F} \subset \mathbb{R}^d$. The encoded visual representation of an image g is given by $f = \phi(g)$. For each subtask, we define a goal image g_{goal} and its corresponding latent feature representation $f_{\text{goal}} = \phi(g_{\text{goal}})$. The visual world model predicts a latent goal feature $f_{\text{pred}} \in \mathcal{F}$ conditioned on $(obs_m, a_m, \mathcal{X}_m, q_m)$. The predicted feature is used as fixed visual guidance for low-level control throughout the execution of the current subtask.

4 Methodology

4.1 Logic World Model

The logical world model enables long-horizon reasoning in symbolic space. Classical TAMP frameworks typically rely on hand-crafted Planning Domain Definition Language (PDDL) domains to explicitly represent symbolic states, actions, and transitions (Kaelbling & Lozano-Pérez, 2011). However, such symbolic planners are brittle under imperfect perception, where inaccuracies in logical state estimation often lead to cascading planning failures. To improve robustness, we learn symbolic planning dynamics directly from data using LLMs. We construct a curated version of the LIBERO dataset by annotating, at each step m, the logical state $\mathcal{X}_m$ and action a_m, synchronized with visual observations obs_m and robot configurations q_m, forming a unified representation $\mathcal{S}_m = \langle \mathcal{X}_m, a_m, obs_m, q_m \rangle$. From each episode, we extract the symbolic action sequence $\tau = \{a_0, a_1, \ldots, a_{\text{goal}}\}$ together with the complete sequence of intermediate logical states $\{\mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_{\text{goal}}\}$. These symbolic state–action trajectories are transformed into chain-of-thought (CoT) explanations that explicitly articulate state transitions, precondition satisfaction, and goal progression, following (Huang et al., 2025a). A base LLM is fine-tuned on these CoT traces to internalize symbolic transition dynamics and planning behavior, yielding a logical world model M_L.

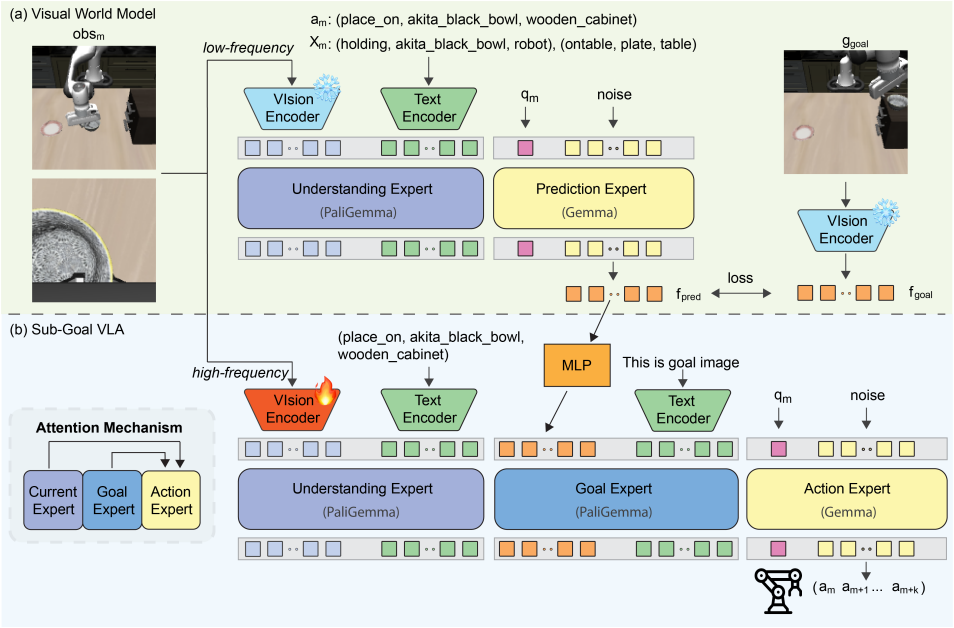

Figure 2: (a) Overview of the proposed visual world model. The model comprises an understanding expert that processes the current scene and language prompt, and a prediction expert that generates goal-state encoded feature representations to guide low-level motion control. (b) Overview of the proposed sub-goal VLA. The understanding expert encodes multimodal perceptual features, while the goal expert processes the predicted goal-state representations. Integrating explicit goal-state information enables more accurate planning and execution in complex environments.

During inference, M_L is reused in a dual role. As M_L^{search}, it proposes candidate symbolic action sequences along with their induced logical state transitions; as M_L^{reward}, it evaluates partial trajectories by scoring their logical consistency and goal alignment. Candidate plans generated by M_L^{search} are ranked by M_L^{reward}, effectively treating the learned logical world model as both a planner and a structured reward function. This actor–critic-style reuse enables efficient and robust long-horizon planning without reliance on hand-engineered symbolic domains.

4.2 Visual World Model

The visual world model provides precise and stable guidance for long-horizon low-level motion control by explicitly aligning symbolic task structure with latent visual representations. Unlike prior world models that generate goal images or videos through unconstrained iterative prediction, where errors can accumulate across generations, we predict latent visual features that are strictly conditioned on the current logical instruction. This symbolic conditioning constrains the space of valid visual goals, reducing error propagation common in unconstrained iterative generative models. In addition, the visual world model avoids long-horizon visual rollouts by predicting only task-critical subgoal representations rather than full video sequences, further improving stability and reliability. By anchoring these sparse latent predictions to logical state transitions, the model ensures consistent and robust grounding of symbolic plans in perceptual space.

As shown in Fig. 2(a), the visual world model comprises an understanding expert and a prediction expert that together map symbolic information to a visually grounded subgoal. At subtask step m, the understanding expert encodes the current observation obs_m together with the logical action a_m and logical state $\mathcal{X}_m$, producing a joint representation that explicitly associates the symbolic task transition with its visual context. This representation defines the intended subgoal conditioned on both the current scene and the symbolic plan. Conditioned on this joint representation and the current robot configuration q_m, the prediction expert outputs a latent visual subgoal feature f_{pred} in a shared feature space. Rather than generating full images or videos, the model predicts a compact

representation corresponding to the visual appearance of the scene upon successful completion of the current subtask. The prediction is implemented via an iterative denoising process to capture uncertainty while remaining constrained by the symbolic input. During training, supervision is provided by synchronizing each logical state, logical action and the terminal key frame. The predicted feature f_{pred} is aligned with the ground-truth feature f_{goal}, obtained by encoding the corresponding goal image using the same frozen vision encoder. We optimize this alignment using the sliced Wasserstein loss (Tanguy et al., 2025), which encourages distribution-level consistency and yields more stable training than an ℓ_2 objective. At inference time, the visual world model is queried once per subtask, and the resulting latent goal feature is held fixed throughout execution, serving as a stable perceptual target for low-level motion control.

4.3 Pipeline for VLA under Hierarchical World Model Guidance

The sub-goal VLA operates as a hierarchical-world-model-guided low-level policy within our framework. Rather than acting solely on instantaneous perception, it executes actions under structured guidance from both the logical and visual world models, which provide symbolic task constraints and perceptually grounded subgoals, respectively. This hierarchical conditioning enables the VLA to maintain consistency with long-horizon task structure while remaining responsive to local visual feedback.

As illustrated in Fig. 2(b), the sub-goal VLA consists of three experts: an understanding expert, a goal expert, and an action expert. Given the current observation obs_m, logical action prompt a_m, and logical state $\mathcal{X}_m$, the understanding expert encodes visual inputs using a SigLIP vision encoder (Zhai et al., 2023) and fuses them with symbolic task information to form a multimodal representation of the current scene. In parallel, the goal expert receives the latent visual subgoal f_{pred} generated by the visual world model, explicitly representing the desired perceptual outcome associated with the current symbolic transition. The action expert then conditions jointly on the current scene representation, the logical task specification, and the latent visual subgoal to generate a sequence of low-level action chunks $\hat{a}_{m:m+k}$, further conditioned on the robot state q_m. All experts are implemented as decoder-only Transformer architectures (Beyer et al., 2024; Team et al., 2024) initialized from large-scale pretrained models. To integrate world-model guidance into action generation, we introduce a structured attention mechanism in which the action expert attends to both experts to synthesize task intent, visual state, and goal constraints. Reverse information flow is explicitly disallowed to preserve hierarchical structure and stabilize training. The policy is trained end-to-end using a flow-matching objective (Lipman et al., 2023), enabling reliable execution under hierarchical world-model guidance.

4.4 Subtask Completion and Transition Prediction

In our hierarchical world model framework, long-horizon tasks are decomposed into a sequence of subtasks, requiring reliable prediction of subtask completion to coordinate transitions and regenerate world-model guidance for VLAs. We introduce a subtask completion prediction module that monitors execution progress and signals when the current subtask t has been achieved, triggering the system to advance to subtask $t{+}1$ and re-query the visual world model to generate a new latent subgoal f_{pred}. The completion predictor is implemented on top of the understanding expert. Given the current observation obs_m and action prompt a_m, a dedicated `[CLS]` token is jointly processed with multimodal features to aggregate global execution progress via cross-attention. The resulting token representation is passed to a lightweight classification head to predict subtask completion, enabling stable and synchronized transitions across the hierarchical pipeline.

5 Experiments

5.1 Training Dataset Generation

5.1.1 LIBERO-Logic Dataset

To train our hierarchical world model, we construct a subgoal-augmented version of the LIBERO dataset, building upon the standard LIBERO benchmark (Liu et al., 2023a). The resulting LIBERO-Logic dataset provides frame-level synchronization between robot state, visual observations, logical

states, and logical actions. Annotations are obtained through a two-stage labeling process. First, we automatically replay each manipulation episode and apply a set of pre-designed predicate classifiers to infer logical states and actions at every timestep. Second, we manually screen the automatically labeled data to correct annotation errors and ensure consistency. This process yields over 980k high-quality image–logic pairs.

5.2 Benchmark and evaluation

We evaluate H-WM and a range of VLA baselines on LIBERO-LoHo, a long-horizon benchmark derived from the standard LIBERO benchmark (Liu et al., 2023a). LIBERO-LoHo consists of five task types, each constructed by extending the task horizon to approximately twice that of the original LIBERO tasks by requiring the robot to manipulate a larger number of objects. This benchmark is designed to systematically assess long-horizon planning performance. All models are trained on the LIBERO-Logic dataset and evaluated on LIBERO-LoHo to ensure a controlled and fair comparison. Performance is measured using Q-Score (QS) and Success Rate (SR), where QS denotes the fraction of completed sub-goals relative to all sub-goals, and SR measures the percentage of tasks fully completed.

5.3 Baselines

We evaluate the effectiveness of H-WM guidance for VLAs against a diverse set of baseline models, including end-to-end VLA approaches such as π_0 (Black et al., 2024) and $\pi_{0.5}$ (Intelligence et al., 2025). We also include a hierarchical planning baseline that follows a common practice of natural-language–based task decomposition with VLA-based motion execution (Shi et al., 2025a). These baselines represent prevailing paradigms for long-horizon VLA control. Through these comparisons, we aim to demonstrate the advantages of H-WM in providing effective intermediate guidance for VLAs on complex long-horizon tasks.

5.4 Ablation Studies

To isolate the contribution of bilevel guidance beyond logic-only reasoning, we conduct a dedicated ablation study in which the visual guidance is unused and only the logical prediction is retained. Under this setting, the VLA receives high-level symbolic plans solely through predicted logical state transitions, without access to intermediate visual subgoals or perceptual guidance. This controlled ablation allows us to rigorously analyze the role of visual guidance in grounding symbolic plans into perceptual observations, supporting reliable subtask execution, reducing error accumulation across subtask boundaries, and improving robustness on long-horizon tasks under partial observability.

6 Result

Table 1: Performance on the LIBERO-LoHo benchmark. For each task, we report Q-Score (QS) and Success Rate (SR) (in %). The best results are shown in **bold**, and the second-best results are underlined.

Methods	**Tasks (QS / SR)**					**Average**
	Task 1	Task 2	Task 3	Task 4	Task 5	
H-WM-guided $\pi_{0.5}$ (ours)	**98.0/94.0**	**86.7/60.0**	**74.0/46.0**	**70.7/42.0**	**95.0/82.0**	**84.9/64.8**
Logic-guided $\pi_{0.5}$	95.3/86.0	84.7/58.0	54.7/16.0	39.3/4.0	92.0/78.0	73.2/48.4
Language-guided $\pi_{0.5}$	84.7/54.0	80.7/42.0	68.0/24.0	41.3/4.0	59.5/10.0	66.8/26.8
$\pi_{0.5}$	66.0/4.0	73.3/24.0	54.7/4.0	44.7/0.0	38.0/0.0	55.3/6.4
π_0	54.0/0.0	62.0/28.0	44.0/0.0	31.3/0.0	34.0/0.0	45.1/5.6

Table 1 shows that H-WM-guided $\pi_{0.5}$ consistently outperforms all baseline methods, achieving substantially higher Q-Score and Success Rate across all tasks. This improvement demonstrates the effectiveness of rich bilevel guidance, where logical planning and visual grounding jointly provide informative intermediate supervision for long-horizon execution. Compared to language-guided

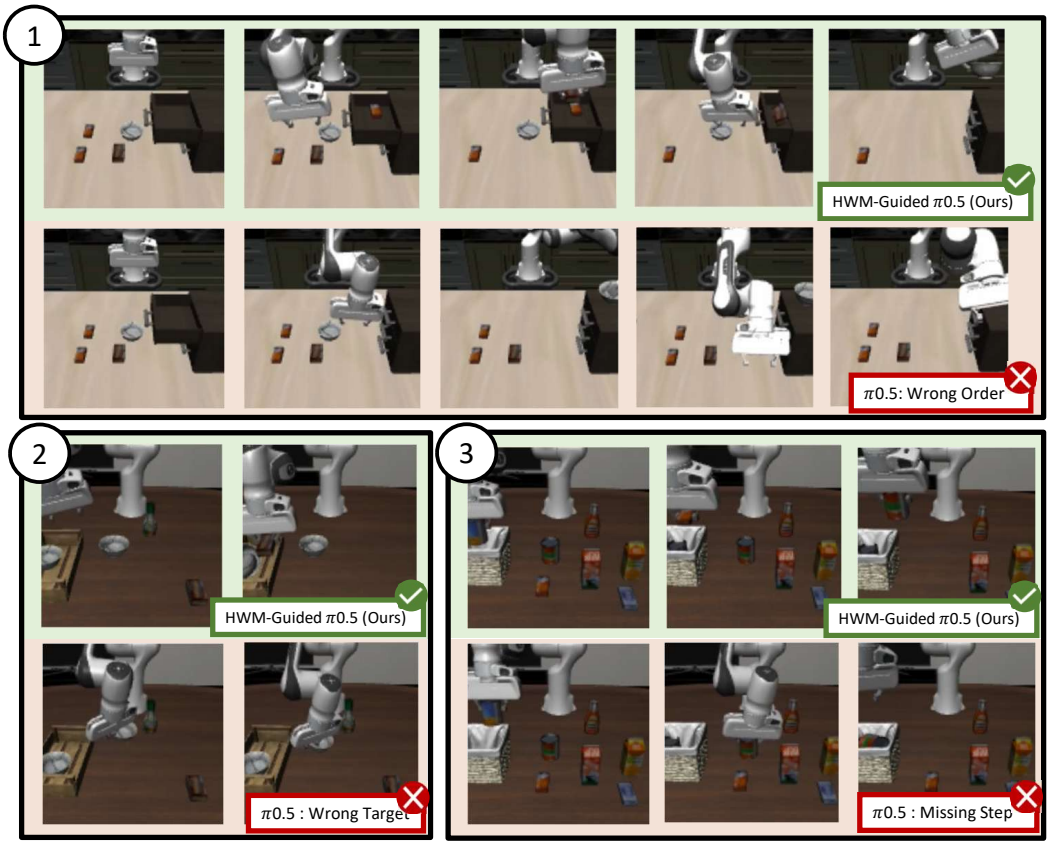

Figure 3: Case Study of vanilla $\pi_0.5$ and our H-WM-guided policy on long-horizon tasks.Case 1: Vanilla $\pi_0.5$ fails to reason over the full task horizon and prematurely closes the cabinet before placing the required target object, whereas our method successfully completes the task using bilevel guidance.Case 2: Vanilla $\pi_{0.5}$ selects an incorrect object, while our approach correctly identifies and manipulates the target, benefiting from future visual subgoal guidance that provides strong cues about target appearance and state transitions.Case 3: Vanilla $\pi_{0.5}$ omits critical intermediate steps due to incomplete task understanding; in contrast, logical state transitions combined with explicit visual grounding provide step-by-step guidance, enabling successful task execution.

$\pi_{0.5}$, the proposed H-WM guidance yields markedly stronger performance, indicating that multi-modal guidance grounded in symbolic structure and visual subgoals is significantly more effective than guidance derived from natural language alone.

The ablation results further highlight the necessity of bilevel guidance: while logic-only guidance already improves over unguided and language-guided baselines, incorporating visual guidance consistently leads to higher success rates, particularly on perceptually challenging tasks. This confirms that visual world modeling provides accurate and actionable grounding for symbolic plans, enabling better alignment between symbolic constraints and perceptual execution. Moreover, logic-guided models outperform language-based task decomposition, as the logical world model produces more reliable task decompositions through structured action segmentation, explicit symbolic constraints, and stricter alignment with robot physics. In contrast, natural language decomposition suffers from ambiguity and imprecise semantics, whereas structured symbolic representations eliminate vagueness and enforce execution-consistent planning.

6.1 Case Study of VLA Failure Modes and H-WM Improvements

We identify three representative failure modes exhibited by unguided VLA models on long-horizon tasks and illustrate how H-WM guidance effectively mitigates each of them.

Failure Mode 1: Temporal misordering of subtasks. In Case 1, the task requires a strict execution order: placing the butter at the back and chocolate pudding in cabinet drawer and closing it, and finally placing the bowl on the cabinet. Vanilla $\pi_{0.5}$ fails to reason over the full task horizon and prematurely closes the cabinet before all required objects are placed, leading to task failure. In contrast, H-WM guidance enforces structured symbolic state transitions and visually grounded subgoals, ensuring that prerequisite actions are completed before irreversible state changes, thereby enabling successful task completion.

Failure Mode 2: Incorrect target selection under visual ambiguity. Case 2 involves placing the left bowl, the salad dressing and chocolate pudding into a wooden tray. Without guidance, vanilla $\pi_{0.5}$ selects incorrect objects due to ambiguity in visual appearance and insufficient anticipation of future task requirements. H-WM guidance resolves this issue by providing future visual subgoals that encode target object shape and placement transitions, allowing the policy to correctly identify and manipulate the intended objects.

Failure Mode 3: Omission of critical intermediate steps. In Case 3, the task requires placing alphabet soup, butter, and tomato sauce into a basket. Vanilla $\pi_{0.5}$ omits necessary intermediate actions, reflecting an incomplete understanding of the task goal and its decomposition. By contrast, the logical world model decomposes the task into explicit symbolic steps, while the visual world model grounds each step in perceptual observations, providing clear step-by-step guidance that prevents action omission and supports reliable long-horizon execution.

7 LIMITATIONS

Although the proposed Hierarchical World Model shows strong performance on guiding long-horizon manipulation tasks, several limitations remain. H-WM introduces additional model components and training stages compared to standalone VLA models, which increases training complexity and computational cost. In addition, the logical world model relies on structured logical state representations. While this design enables interpretable and globally consistent planning, it requires logical supervision or data augmentation during training and additional engineering effort.In addition, H-WM implicitly assumes that the task can be meaningfully modeled in a symbolic logical space. Tasks that are difficult to express with discrete predicates or require highly continuous, deformable, or contact-rich reasoning may be less suited to this formulation and would require extensions to the logical representation.

8 FUTURE WORK

The mentioned limitations highlight several promising directions for future research. Future work will focus on improving training efficiency and reducing reliance on explicit logical supervision, with the goal of lowering the overall engineering and data requirements of the framework. We also plan to incorporate additional sensory modalities, such as depth and tactile feedback, to enhance spatial reasoning and robustness in contact-rich settings. Finally, we aim to evaluate the proposed approach across a broader range of environments, object categories, and real-world task distributions to better assess its scalability, robustness, and general applicability.

9 CONCLUSION

We presented a hierarchical world model that jointly predicts logical and visual world dynamics to provide bilevel guidance for VLA models on long-horizon robotic tasks. The logical level captures global task structure and long-term dependencies, while the visual level grounds symbolic plans into visually meaningful latent subgoals, enabling stable guidance under extended planning horizon. Experiments demonstrate that the proposed H-WM-based hierarchical guidance substantially improves the long-horizon performance of VLA models compared to their base counterparts. In addition, the approach exhibits clear advantages over LLM-based hierarchical planning baselines across a range of tasks. These results indicate that hierarchical world models are an effective and scalable approach for bridging symbolic reasoning and perceptual grounding in VLA systems, and that bilevel world-model guidance serves as a powerful form of structured guidance for VLA models for long-horizon tasks.

REFERENCES

Michael Ahn et al. Do as i can, not as i say: Grounding language in robotic affordances, 2022. URL https://arxiv.org/abs/2204.01691.

Anurag Ajay et al. Compositional foundation models for hierarchical planning. *Advances in Neural Information Processing Systems*, 36:22304–22325, 2023.

Mido Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Amir Bar et al. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Lucas Beyer et al. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Hongzhe Bi et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Kevin Black et al. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023. URL https://arxiv.org/abs/2307.15818.

Haonan Chen et al. Goal-vla: Image-generative vlms as object-centric world models empowering zero-shot robot manipulation. 2025. URL https://arxiv.org/abs/2506.23919.

Yongchao Chen et al. Prompt optimization in multi-step tasks (promst): Integrating human feedback and heuristic-based sampling. 2024.

Rohan Chitnis et al. Guided search for task and motion plans using learned heuristics. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 447–454. IEEE, 2016.

Murtaza Dalal et al. Imitating task and motion planning with visuomotor transformers. *arXiv preprint arXiv:2305.16309*, 2023.

Maximilian Diehl et al. Automated generation of robotic planning domains from observations. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

Danny Driess et al. Palm-e: An embodied multimodal language model, 2023. URL https://arxiv.org/abs/2303.03378.

Lin Guan et al. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. In *Proc. Adv. Neural Inf. Proc. Systems*, 2023.

Danijar Hafner et al. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Muzhi Han et al. Interpret: Interactive predicate learning from language feedback for generalizable task planning. In *Robotics: Science and Systems (RSS)*, 2024.

Ryan Hoque et al. Visuospatial foresight for physical sequential fabric manipulation. *Autonomous Robots*, 46(1):175–199, 2022.

Jinbang Huang, Zhiyuan Li, Zhanguang Zhang, Xingyue Quan, Jianye Hao, and Yingxue Zhang. Plan2evolve: Llm self-evolution for improved planning capability via automated domain generation. *arXiv preprint arXiv:2509.21543*, 2025a.

Jinbang Huang et al. Automated planning domain inference for task and motion planning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12534–12540. IEEE, May 2025b.

Jinbang Huang et al. One demo is all it takes: Planning domain derivation with LLMs from a single demonstration. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wenlong Huang et al. Inner monologue: Embodied reasoning through planning with language models. 2022.

Physical Intelligence et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. In *2011 IEEE international conference on robotics and automation*, pp. 1470–1477. IEEE, 2011.

Moo Jin Kim et al. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Moo Jin Kim et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Nishanth Kumar et al. Learning efficient abstract planning models that choose what to predict. In *Conference on Robot Learning*, pp. 2070–2095. PMLR, 2023.

Boyi Li et al. Interactive task planning with language models. In *Proc. 2nd Workshop on Language and Robot Learning*, 2023.

Boyi Li et al. Interactive task planning with language models, 2025a. URL https://arxiv.org/abs/2310.10645.

Shuang Li et al. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025b.

Xinqing Li et al. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025c.

Yaxuan Li et al. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025d.

Yichao Liang et al. VisualPredicator: Learning abstract world models with neuro-symbolic predicates for robot planning. *arXiv preprint arXiv:2410.23156*, 2024.

Yue Liao et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Yaron Lipman et al. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Bo Liu et al. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023a.

Weiyu Liu et al. Learning compositional behaviors from demonstration and language. In *Proceedings of The 8th Conference on Robot Learning*, 2025.

Ziang Liu et al. Model-based control with sparse neural dynamics. *Advances in Neural Information Processing Systems*, 36:6280–6296, 2023b.

Xiaoxiao Long et al. A survey: Learning embodied intelligence from physical simulators and world models. *arXiv preprint arXiv:2507.00917*, 2025.

Yueen Ma et al. A survey on vision-language-action models for embodied ai (2024). *arXiv preprint arXiv:2405.14093*, 1778.

Lucas Manuelli et al. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. *arXiv preprint arXiv:2009.05085*, 2020.

Jiayuan Mao et al. Learning reusable manipulation strategies. In *7th Annual Conference on Robot Learning*, 2023.

Michael James McDonald et al. Guided imitation of task and motion planning. pp. 630–640. PMLR, 2022.

Maëlic Neau et al. Grasp-vla: Graph-based symbolic action representation for long-horizon planning with vla policies. *arXiv preprint arXiv:2511.04357*, 2025.

James Oswald et al. Large language models as planning domain generators. In *Proc. Int. Conf. on Automated Planning and Scheduling*, 2025.

Shivansh Patel et al. Robotic manipulation by imitating generated videos without physical demonstrations, 2025. URL https://arxiv.org/abs/2507.00990.

Chris Paxton et al. Combining neural networks and tree search for task and motion planning in challenging environments. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6059–6066. IEEE, 2017.

Julian Quevedo et al. Evaluating robot policies in a world model. *arXiv preprint arXiv:2506.00613*, 2025.

Rui Shao et al. Large vlm-based vision-language-action models for robotic manipulation: A survey, 2025. URL https://arxiv.org/abs/2508.13073.

Lucy Xiaoyang Shi et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models, 2025a. URL https://arxiv.org/abs/2502.19417.

Lucy Xiaoyang Shi et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025b.

Tom Silver et al. Planning with learned object importance in large problem instances using graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11962–11971, 2021.

Tom Silver et al. Predicate invention for bilevel planning. In *Proc. AAAI Conf. on Artificial Intelligence*, 2023.

Eloi Tanguy et al. Properties of discrete sliced wasserstein losses. *Mathematics of Computation*, 94 (353):1411–1465, 2025.

Gemma Team et al. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295.

Marc Toussaint. Logic-geometric programming: An optimization-based approach to combined task and motion planning. In *IJCAI*, pp. 1930–1936, 2015.

Shu Wang et al. Llm3: Large language model-based task and motion planning with motion failure reasoning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12086–12092. IEEE, 2024a.

Shu Wang et al. Llm3:large language model-based task and motion planning with motion failure reasoning, 2024b. URL https://arxiv.org/abs/2403.11552.

Lionel Wong et al. Learning adaptive planning representations with natural language guidance. *arXiv [cs.AI]*, 2023.

Zhenyu Wu et al. Momanipvla: Transferring vision-language-action models for general mobile manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1714–1723, 2025.

Jiannan Xiang et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

Annie Xie, Frederik Ebert, Sergey Levine, and Chelsea Finn. Improvisation through physical understanding: Using novel objects as tools with visual foresight. *arXiv preprint arXiv:1904.05538*, 2019.

Eric Xing et al. Critiques of world models. *arXiv preprint arXiv:2507.05169*, 2025.

Danfei Xu et al. Deep affordance foresight: Planning through what can be done in the future. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 6206–6213. IEEE, 2021.

Yi Yang et al. Lohovla: A unified vision-language-action model for long-horizon embodied tasks. *arXiv preprint arXiv:2506.00411*, 2025a.

Yuncong Yang et al. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025b.

Wentao Yuan et al. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Xiaohua Zhai et al. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

Wenyao Zhang et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025.

Qingqing Zhao et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025.

Haoyu Zhen et al. 3d-vla: A 3d vision-language-action generative world model, 2024. URL https://arxiv.org/abs/2403.09631.

Peiyuan Zhi et al. Closed-loop open-vocabulary mobile manipulation with gpt-4v, 2025. URL https://arxiv.org/abs/2404.10220.

Yifan Zhong et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.

Gaoyue Zhou et al. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Wang Zhu et al. Language models can infer action semantics for symbolic planners from environment feedback. *arXiv [cs.AI]*, 2024.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

This paper is focusing on the study of LLM training, in which large language models are the primary object of investigation and a diverse range of models are evaluated through experiments.

Beyond their role as the target of study, LLMs were employed as tools for grammar correction and stylistic refinement, including improvements in clarity and readability. LLMs were not involved in the generation of research ideas, the design or implementation of experiments, the analysis of data, or the interpretation of findings. The authors have full responsibility for the originality, accuracy, and integrity of all scientific content reported in this work.

B LIBERO-LOHO TASK SPECIFICATIONS

We summarize the long-horizon (LOHO) tasks used in our experiments. For each task, we report the natural language instruction provided to the agent and the total number of atomic action steps required to complete the task.

Table 2: Summary of long-horizon tasks used in evaluation.

Task ID	**Task Description**	**# Length**
Task 1	Put front butter and the chocolate pudding into the wooden cabinet top drawer and then close it.	5
Task 2	Pick up the alphabet soup, then the butter, then the tomato sauce, and place each one into the basket.	6
Task 3	Pick up the alphabet soup, then the cream cheese, then the butter, and put each one into the wooden tray.	6
Task 4	Pick up the black bowl on the left, then the salad dressing, then the chocolate pudding, and put each one into the wooden tray.	6
Task 5	Put the butter at the back and the chocolate pudding into the cabinet top drawer, and then close the cabinet top drawer. Place the black bowl on top of the cabinet.	7