MetaWorld: Skill Transfer and Composition in a Hierarchical World Model for Grounding High-Level Instructions

Anonymous CVPR submission

Paper ID *****

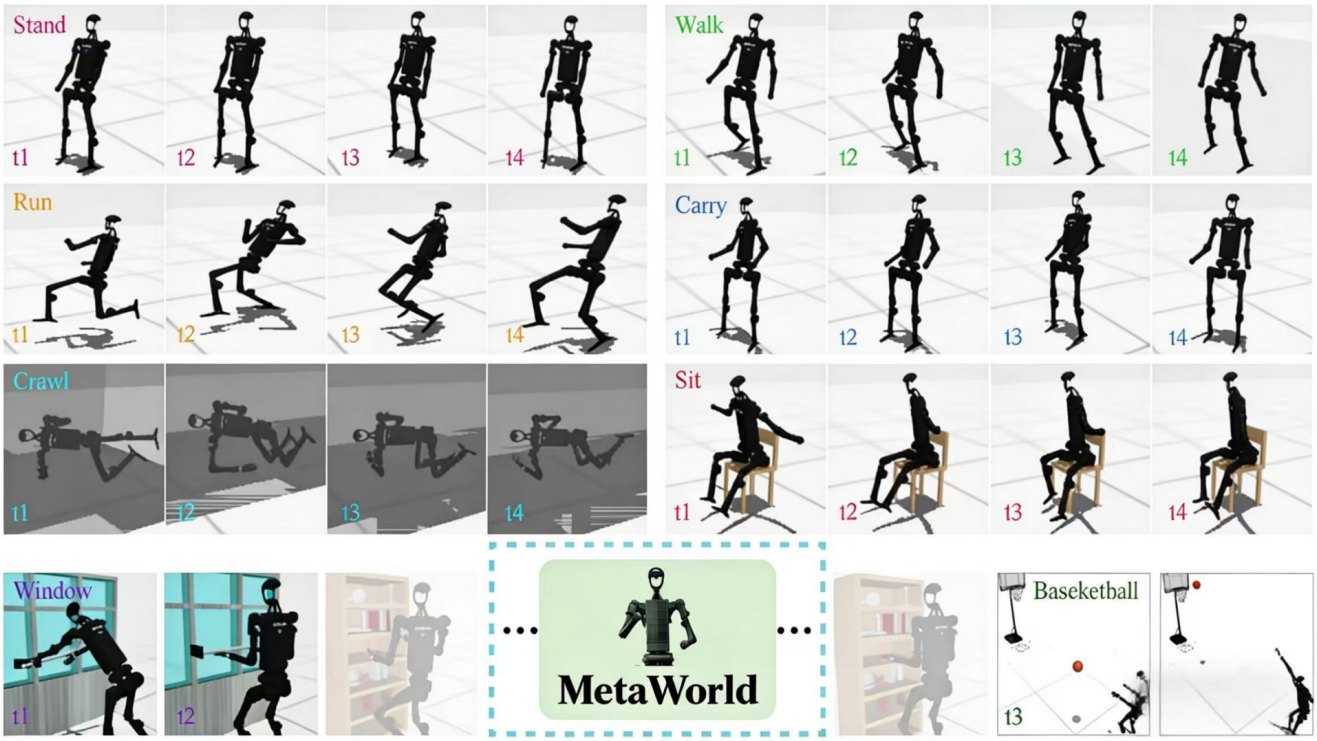

Figure 1. **MetaWorld** is a hierarchical world model framework that bridges the semantic-physical gap in robotic tasks. It leverages Vision-Language Models (VLMs) for high-level semantic task planning, integrates motion priors through a multi-expert policy library for action fusion, and achieves efficient low-level physical control via a latent dynamics model.

Abstract

Humanoid loco-manipulation requires coordinating whole-body joints for simultaneous locomotion and object interaction, yet it remains challenging due to the high-dimensional action space. Current methods face three primary bottlenecks: (1) low sample efficiency in reinforcement learning, (2) poor generalization of imitation learning in long-horizon tasks, and (3) physical inconsistency in VLM-based planning. To address these, we propose MetaWorld, a hierarchical world model that integrates semantic planning and physical control via expert policy transfer. Specifically, (i) we introduce a motion prior fusion mechanism that leverages a pre-trained expert library within a compact latent dynamics model to accelerate online adaptation; (ii) a hierarchical task decoupling strategy is employed to decompose long-term logic from local execution, ensuring robust cross-scene transfer; and (iii) the VLM is utilized as a semantic interface to map high-level instructions directly to pre-validated physical skills, bypass-

ing the symbol grounding problem and ensuring dynamic feasibility. Evaluated on challenging tasks in Humanoid-Bench, MetaWorld significantly outperforms TD-MPC2 and Dreamer V3 in both success rate and motion smoothness, achieving a 139.1% increase in average reward. Our code will be found at https://anonymous.4open.science/r/metaworld-2BF4/

1. Introduction

Humanoid robots, as quintessential embodiments of embodied intelligence, have long faced a fundamental challenge: executing semantically-driven loco-manipulation tasks within unstructured and dynamic environments. At the core of this challenge lies a significant "abstraction gap" in current robot control systems, the disconnect between high-level semantic understanding and low-level physical execution [6, 13, 16, 25]. On one hand, large-scale models represented by Vision-Language Models (VLMs) demonstrate exceptional capabilities in high-level task planning and semantic reasoning, understanding "what to do"; on the other hand, low-level control methods based on imitation learning or reinforcement learning can generate precise joint-level actions, addressing "how to do it." However, the absence of a unified and scalable framework to bridge these disparate capabilities often results in semantic plans that are decoupled from physical constraints, conversely, low-level control policies often lack the versatility to generalize across complex, composed high-level tasks.

Current mainstream approaches exhibit distinct limitations in addressing this problem. Although end-to-end reinforcement learning is theoretically capable of discovering optimal policies, it suffers from extreme **sample inefficiency** due to the expansive search space of humanoid whole-body coordination [24]. Imitation learning, while effective at acquiring natural motion patterns, typically yields **poor generalization** when extrapolating limited demonstrations to unseen, long-horizon scenarios [21]. Furthermore, direct application of VLMs to robot control faces severe **physical inconsistency**; without kinematic or dynamic constraints, the generated plans often fail to ground symbols into executable physical actions. While existing literature explores the integration of these methods, most approaches remain restricted to simplistic task concatenation or a loose coupling of independent modules, failing to establish a truly unified, hierarchical interaction architecture [18, 20, 23].

To systematically address these challenges, this paper presents **MetaWorld**, a hierarchical world model-based robot control framework that synergistically combines the semantic reasoning of VLMs, motion priors from imitation learning, and online adaptation mechanisms from model-based reinforcement learning. **The main contributions of this paper include:**

- **Addressing Sample Efficiency via Motion Prior Fusion:** We propose a hierarchical world model that leverages a pre-trained multi-expert policy library. By synthesizing actions within a compact **latent dynamics model**, our framework achieves efficient reuse of skills and significantly accelerates online adaptation.
- **Ensuring Physical Consistency via Semantic Interfaces:** We establish a reliable interaction paradigm using VLMs as **semantic interfaces**. By mapping high-level instructions directly to pre-validated physical skills, we bypass the **symbol grounding** problem and ensure the dynamic feasibility of generated motions.
- **Enhancing Generalization via Hierarchical Decoupling:** We introduce a modular architecture that decouples long-horizon semantic logic from local physical execution. This two-stage framework enables robust cross-scene migration and achieves superior generalization across diverse, unstructured environments.

2. Related Works

2.1. World Models in Robotics

World models facilitate policy learning by predicting environment state transitions, emerging as a critical tool for enhancing sample efficiency and generalization in robotics [17]. Mainstream approaches, such as the Dreamer series [7–10], perform policy optimization through latent space prediction, while model-based reinforcement learning methods like TD-MPC2 [11] further achieve high-performance control in complex dynamic environments. However, most existing world models lack hierarchical depth, hindering the decomposition of high-level semantic tasks into executable physical action sequences, thereby limiting their direct application to long-horizon, multi-step embodied tasks [5]. The hierarchical world model architecture proposed in this paper introduces explicit separation between the semantic planning layer and physical execution layer, enabling the world model to simultaneously handle semantic parsing of language instructions and dynamic prediction of physical environments, effectively bridging the gap between task planning and action generation.

2.2. Skill Transfer Learning

Transfer learning facilitates robotic generalization through the systematic reuse of pre-acquired skills. Current robot control approaches include: domain adaptation (e.g., domain randomization) to bridge the sim-to-real gap [14]; policy fine-tuning (e.g., progressive networks) for local adjustments [3]; and meta-learning (e.g., MAML) for fast few-shot adaptation [4, 15]. However, these methods face two major bottlenecks: a heavy reliance on target-domain data or multi-stage training, both of which limit their practical deployment; and a failure to achieve real-time, millisecond-

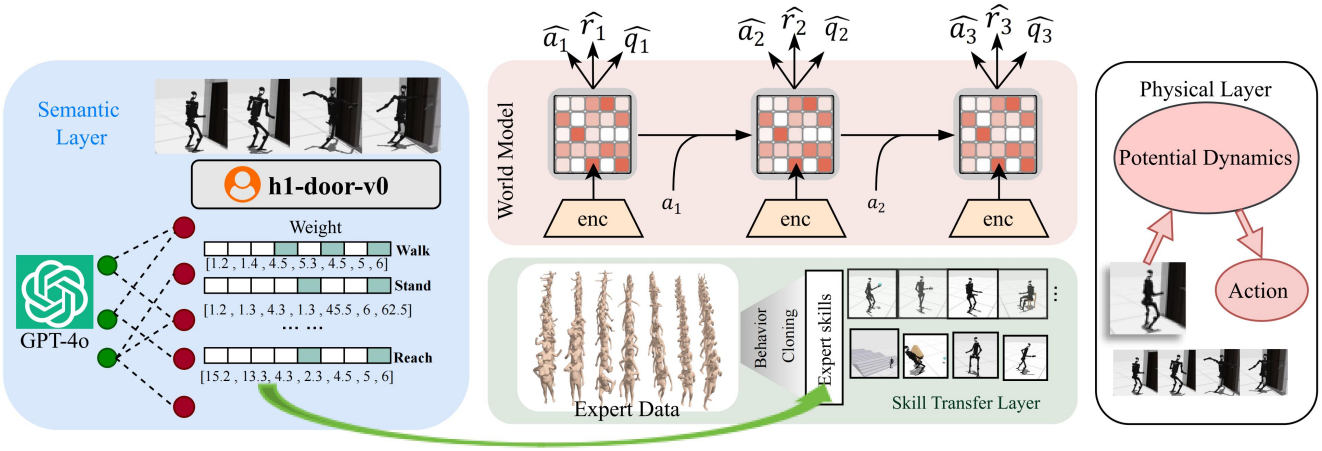

Figure 2. This illustrates the three-tier architecture of the **MetaWorld** framework: the semantic layer parses observation into executable skill sequences via a vision-language model; the skill transfer layer integrates expert policy priors through a hierarchical world model and enables dynamic adaptation; the physical layer performs precise control in a compact state space using a latent dynamics model.

level adaptation in the face of dynamic disturbances. To address these issues, we propose a dynamic expert selection and motion prior fusion mechanism. By constructing a multi-expert policy library, our approach dynamically selects the most relevant strategies within a Model Predictive Control (MPC) framework to enable real-time policy adjustment. This approach maintains the efficiency of expert policies while significantly enhancing adaptability and generalization in unstructured environments.

2.3. VLM Applications in Real-time Task Parsing

Vision-Language Models (VLMs) have demonstrated remarkable proficiency in open-vocabulary perception, scene understanding, and high-level reasoning, and have been widely applied to high-level task planning and semantic guidance in robotics. Representative works such as VIMA and RT-2 attempt to directly use VLMs to generate robot action sequences or serve as front-ends for symbolic planners [2, 13]. However, these methods commonly face the "symbol grounding" problem: VLM-generated plans often ignore the robot's kinematic and dynamic constraints as well as environmental physics, resulting in infeasible plans. In this work, we mitigate the physical limitations of VLMs by restricting their role to high-level semantic parsing and mapping their outputs onto a set of pre-validated, physically feasible expert policies. Concurrently, our framework leverages the semantic understanding and task decomposition strengths of VLMs to process open-ended environmental semantics and achieve dynamic replanning through closed-loop feedback during training, guiding complex tasks to reference fundamental expert policies.

2.4. Transfer Learning and Policy Composition

Transfer learning aims to leverage knowledge from source tasks to accelerate learning in target domains, a paradigm particularly vital for high-DOF systems like humanoids [1, 12]. Previous works have explored policy composition by linearly blending pre-trained primitives or using gating networks to select specialized experts. However, these methods often suffer from performance degradation when facing long-horizon loco-manipulation tasks that require seamless transitions between locomotion and manipulation [26]. MetaWorld advances this field by integrating policy transfer directly into a hierarchical world model. Unlike conventional distillation-based methods that compress multiple experts into a single student policy, our approach employs a **dynamic expert selection** mechanism within a latent dynamics space. By fusing motion priors from a multi-expert library, MetaWorld can compose novel motor skills on-the-fly, ensuring that the transferred knowledge is not only semantically relevant but also physically consistent with the robot's hardware constraints.

3. Method

3.1. MetaWorld Hierarchical Architecture

The central idea of **MetaWorld** is to decompose the robotic control problem into two distinct layers: a semantic planning layer responsible for interpreting task intent, and a physical execution layer responsible for generating physically feasible actions. This hierarchical design can be formalized as:

$$\pi(a_t \mid s_t, \mathcal{T}) = \pi_{\text{phys}}(a_t \mid s_t, \pi_{\text{sem}}(\mathcal{T})) \quad (1)$$

where π_{sem} maps the task description $\mathcal{T}$ to a semantic plan, and π_{phys} generates specific actions a_t based on the current

state s_t and the semantic plan. This architecture enables independent optimization of semantic understanding and physical control components while maintaining overall optimality. The optimization objective is to maximize the expected cumulative reward $J(\pi) = \mathbb{E}[\sum \gamma^t r(s_t, a_t)]$ where hierarchical optimization effectively addresses the distinct challenges at the semantic and physical levels.

3.2. Semantic Planning and Symbol Grounding

Definition 1: *Symbol Grounding Error* refers to the state-action divergence between high-level semantic intentions and low-level physical feasibility. In our framework, we bound this error by projecting VLM-generated semantic sub-goals into a pre-trained *expert-action manifold*. Unlike end-to-end methods that generate actions in unconstrained spaces, MetaWorld ensures that even under semantic ambiguity, the resulting movement remains within a kinematically stable prior, thereby maintaining physical consistency.

The semantic planning layer employs a Vision-Language Model (VLM) to map natural language task descriptions to expert policy weights. Unlike traditional methods, we constrain the VLM output to an expert weight vector $\mathbf{w}$ rather than direct actions:

$$\mathbf{w} = f_{\text{VLM}}(\mathcal{T}, \mathcal{E}) \tag{2}$$

The key innovation lies in treating the expert library as a **functional basis**. By transforming the symbol grounding problem into a **continuous manifold blending** of expert policies, we enable the generation of hybrid behaviors. The VLM generates expert weights through carefully engineered prompts, with the response R processed by a parsing function to obtain normalized weights $w_i = \exp(\text{extract}_i(R))/\sum_j \exp(\text{extract}_j(R))$. Since each expert policy π^i_{exp} is physically feasible, the generated semantic plan $\pi_{\text{sem}}(\mathcal{T}) = \sum_i w_i \pi^i_{\text{exp}}$ naturally satisfies physical constraints. The symbol grounding error is bounded within the range of expert policy differences, significantly outperforming direct action generation methods.

3.3. Dynamic Adaptation Mechanism

To address dynamic environmental changes, we introduce a **Zero-shot Behavioral Compositor** that performs state-aware expert selection. Based on the current state s_t, we construct a selection probability distribution:

$$p(i \mid s_t) = \frac{\exp(\phi(s_t)^\top \psi(\pi^i_{\text{exp}}))}{\sum_{j=1}^{K} \exp(\phi(s_t)^\top \psi(\pi^j_{\text{exp}}))} \tag{3}$$

where ϕ is a state encoding function and ψ is an expert feature extraction function. The VLM-generated semantic weights w_i are fused with the dynamic selection probabilities $p(i|s_t)$ to obtain the final weights $\tilde{w}_i(s_t, \mathcal{T}) = \alpha w_i + (1-\alpha)p(i|s_t)$. This fusion mechanism diverges from classical discrete option-selection by enabling **continuous interpolation** between experts. The parameter $\alpha \in [0, 1]$ controls the relative importance of semantic planning versus state awareness, enabling a balance between task consistency and environmental adaptability. The reference expert action $a_{\text{ref}} = \sum_i \tilde{w}_i(s_t, \mathcal{T})\pi^i_{\text{exp}}(s_t)$ provides a high-quality initial solution for the physical execution layer.

3.4. Physical Execution and Online Optimization

The physical execution layer employs the TD-MPC2 algorithm, constructing a latent dynamics model for Model Predictive Control (MPC). Observations o_t are encoded into latent states $z_t = f_{\text{enc}}(o_t)$ and state evolution is predicted through the dynamics model $z_{t+1} = f_{\text{dyn}}(z_t, a_t)$. The MPC optimization problem solves for the optimal action sequence over a future horizon H:

$$\mathbf{a}^*_{t:t+H-1} = \arg \max_{\mathbf{a}_{t:t+H-1}} \mathbb{E} \left[\sum_{k=0}^{H-1} \gamma^k r(z_{t+k}, a_{t+k}) + \gamma^H V(z_{t+H}) \right] \tag{4}$$

the expert-guided action a_{ref} is incorporated into the optimization objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TD}} + \lambda \|a_t - a_{\text{ref}}\|^2$ where the Temporal-Difference (TD) learning loss $\mathcal{L}_{\text{TD}} = \mathbb{E}[\|Q(z_t, a_t) - (r_t + \gamma Q(z_{t+1}, \pi(z_{t+1})))\|^2]$ ensures accurate value function estimation. This design maintains online adaptation capabilities while leveraging expert knowledge to accelerate the learning process.

3.5. Theoretical Analysis and Implementation

Based on contraction mapping theory, we prove algorithm convergence under appropriate parameter selection. The value iteration operator $\mathcal{T}$ satisfies the contraction property $\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$ guaranteeing convergence of the value function to the optimal solution. Compared to traditional methods, the sample complexity is reduced from $\mathcal{O}(|\mathcal{S}||\mathcal{A}|/[(1-\gamma)^2\epsilon^2])$ to $\mathcal{O}(1/[(1-\gamma)^2\epsilon^2] + K)$ demonstrating the efficiency advantage of knowledge reuse.

4. Experiment

We evaluate our method on HumanoidBench [22], a comprehensive benchmark for humanoid locomotion and manipulation. For locomotion, we select three tasks: walk, stand, and run; for manipulation, we include door opening. The locomotion tasks are learned via imitation learning from the AMASS [19] expert dataset, shaped with trajectory-tracking reward signals. Other skills in the base expert policy repository (e.g., reach) are inherited from TD-MPC2 and are therefore excluded from evaluation. While our architecture is designed with a comprehensive library of eight specialized experts—namely *walk*, *carry*, *sit*, *crawl*,

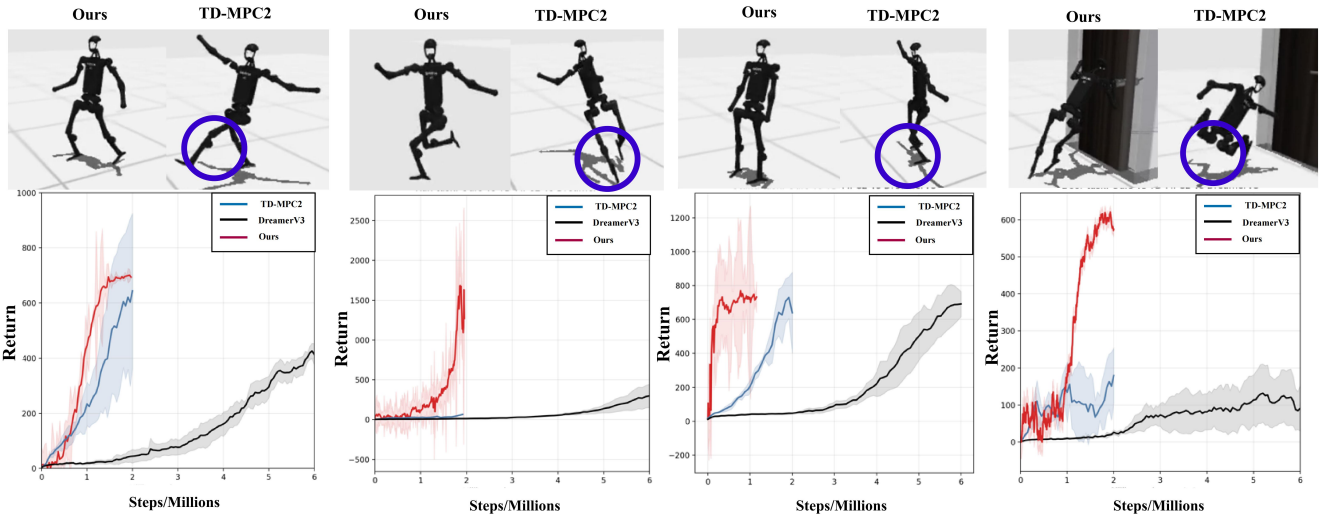

Figure 3. We evaluate our method on locomotion (run, walk, stand) and manipulation (door) tasks against baselines TD-MPC2 and DreamerV3.

Task	Metric	DreamerV3	TD-MPC2	Ours	Imp. (%)
Stand	Ret. ↑ (± Std)	699.3 ± 62.7	749.8 ± 54.3	$\mathbf{793.4 \pm 13.5}$	9.5
	Conv. ↓ (M)	5.5	1.9	**0.8**	78.3
Walk	Ret. ↑ (± Std)	428.2 ± 14.5	644.2 ± 162.3	$\mathbf{701.2 \pm 7.6}$	30.77
	Conv. ↓ (M)	6.0	1.8	**1.4**	64.1
Run	Ret. ↑ (± Std)	298.5 ± 84.5	66.1 ± 4.7	$\mathbf{1689.9 \pm 13.6}$	826.0
	Conv. ↓ (M)	6.0	2.0	**1.9**	52.5
Door	Ret. ↑ (± Std)	165.8 ± 50.2	179.8 ± 52.9	$\mathbf{680.0 \pm 50.0}$	293.5
	Conv. ↓ (M)	9.0	2.0	**1.7**	69.1
Avg.	Ret. ↑	398.0	410.0	**966.1**	139.1
	Conv. ↓ (M)	6.6	1.9	**1.5**	64.7

Table 1. Performance comparison on locomotion and manipulation tasks. All values are reported to one decimal place. The **Ours** column is highlighted in light blue, and the rightmost column shows the percentage improvement.

stairs, *stand*, *run*, and *reach*—this evaluation focuses on a representative subset: *walk*, *stand*, and *run*. These tasks are selected to rigorously benchmark the model's fundamental locomotion stability and its ability to manage high-dynamic transitions.Our policy library consists of $N = 8$ specialized experts (e.g., walk, run, stand, etc.). Each expert is a neural policy trained on thousands of motion clips from the AMASS dataset, rather than a single episode mapping. This ensures that each primitive expert possesses sufficient generalization within its specific modal domain (e.g., various walking speeds and gaits).

4.1. Baseline

We evaluate our method against two state-of-the-art model-based reinforcement learning algorithms to demonstrate its superiority: TD-MPC2 [11], a scalable model-predictive control framework that utilizes local trajectory optimization with learned implicit world models, and DreamerV3 [9], a robust actor-critic algorithm that learns behaviors within a latent imagination space across diverse domains. These

baselines represent the current frontier in sample efficiency and asymptotic performance, providing a rigorous benchmark for our proposed components.

Distinction from Classical Hierarchical Control. While hierarchical planning conditioned on task priors has been explored, our approach diverges from traditional *discrete option-selection* frameworks. Instead of switching between isolated controllers, MetaWorld treats the expert library as a **continuous functional basis**. By leveraging VLM-driven dynamic weights, we enable the **emergence of hybrid behaviors** that are not present in any single expert's training distribution. This formulation transforms the hierarchical interface from a simple router into a generative action compositor, significantly enhancing the zero-shot adaptability in non-stationary environments.

4.2. Experimental setup

Our architecture employs the following hyperparameters: the VLM temperature parameter $\tau = 0.3$ controls the stochasticity in expert weight generation; the expert guidance weight $\lambda = 0.05$ balances TD-learning and expert priors; the MPC planning horizon is set to $H = 3$ to ensure real-time performance; the discount factor $\gamma = 0.99$ trades off long-term rewards; the learning rate is fixed at 0.001 to ensure stable convergence; the batch size of 256 balances training efficiency and memory usage; and the fusion coefficient $\alpha = 0.7$ regulates the relative importance of semantic planning and state adaptation. For task success criteria, we follow the definitions provided in HumanoidBench.

4.3. Validation of Hierarchical Architecture Effectiveness

Purpose of the experiment. In this section, we evaluate the effectiveness of our hierarchical architecture against traditional model-based reinforcement learning methods. We first assess the performance of the base expert policy across three locomotion tasks: walk, run, and stand. Subsequently, on the walk task, we further test the model's capability to perform complex manipulation tasks after dynamic adaptation. To quantitatively evaluate the performance, we record the cumulative *Return* to measure the task completion quality and the number of *Convergence Steps* to reflect the sample efficiency. Specifically, a higher Return indicates superior task execution, while fewer Steps demonstrate the model's ability to achieve optimal performance with minimal environment interactions, highlighting the efficiency of our hierarchical architecture.

Result. As illustrated in Table 1 and Fig. 3, the **MetaWorld** framework exhibits overwhelming superiority across four representative tasks, achieving an average return improvement of 139.1%. In basic locomotion tasks, particularly the remarkable **826.0% boost in the Run task**, the performance gains stem from the *manifold constraints* imposed by the imitation-learned expert library on the high-dimensional action space. By transforming continuous action search into a weighted composition of expert trajectories, our architecture provides the VLM planner with a physically feasible and low-dimensional action subspace, effectively mitigating the *curse of dimensionality* inherent in humanoid control.

For the complex Door task, the **293.5% improvement** validates our hierarchical innovation: the VLM acts as a semantic decoder to decompose ambiguous instructions into structured sub-goal sequences, while the dynamic selection mechanism orchestrates corresponding motion primitives

Table 2. Success rate evaluation of base expert policies and complex manipulation tasks. We evaluate each policy using checkpoints at 2.0M steps across 10 independent trials. Our method significantly outperforms state-of-the-art MBRL baselines.

Method	Locomotion (Success / 10)			Manipulation
	Stand	Walk	Run	Door (Success / 10)
DreamerV3 [**?**]	2/10	2/10	1/10	0/10
TD-MPC2 [11]	3/10	3/10	2/10	1/10
Ours	**9/10**	**9/10**	**9/10**	**8/10**

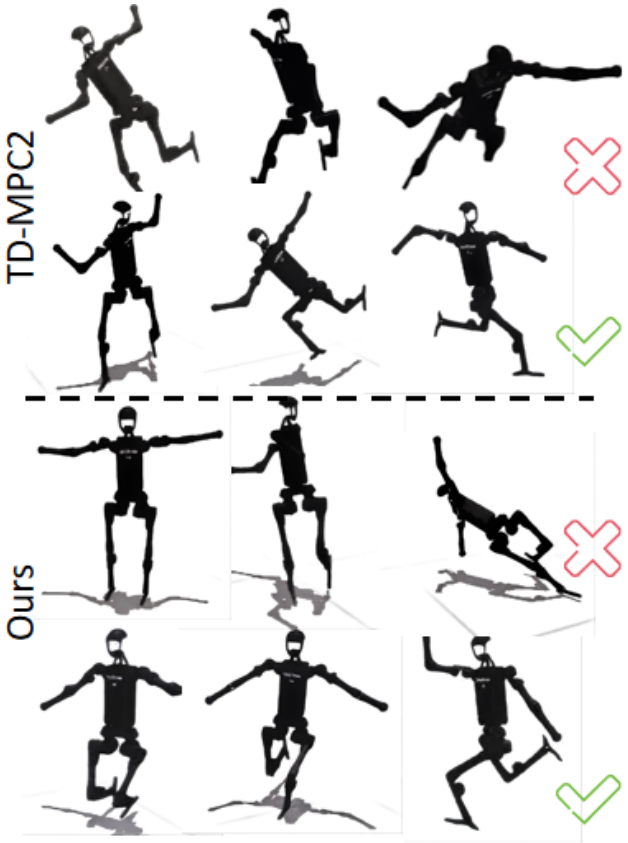

Figure 4. We visualized our method against baselines TD-MPC2.

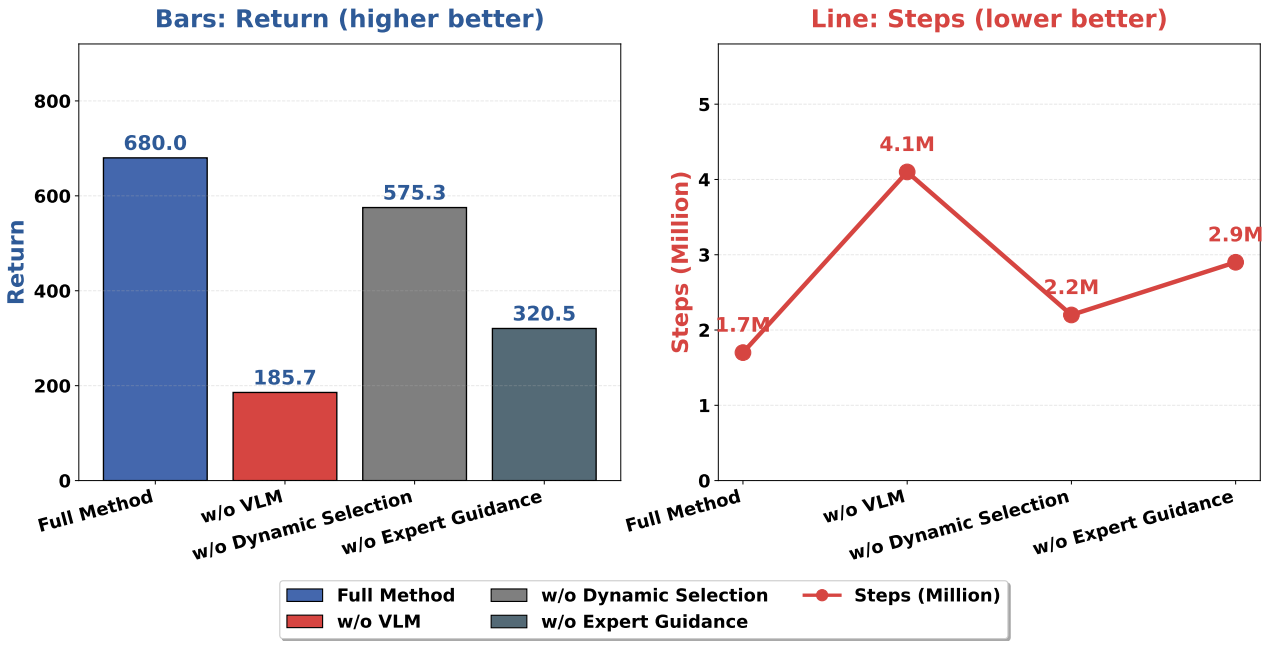

Figure 5. Ablation study results on the Door task (maximum reward and convergence steps).

for execution. The pivotal insight is that complex manipulation is achieved through the *decoupling and recomposition* of high-level reasoning and low-level motor priors. This strategy not only resolves the long-standing *symbol grounding dilemma* but also bypasses the *sample efficiency bottleneck* typical of end-to-end RL paradigms, offering a robust new paradigm for the seamless integration of logical inference and reactive control in intelligent robotics.

While the use of motion priors inherently accelerates exploration, the primary challenge addressed here is the *temporal composition* of these priors. The efficiency gain stems not just from the experts themselves, but from the latent dynamics model's ability to smoothly transition between discrete expertise domains, a task that remains prohibitively slow for flat RL architectures.

4.4. Evaluation of Task Success Rate and Stability.

To rigorously evaluate the stability and success rate of our base expert policies and complex task execution, we conduct assessments using model checkpoints saved at 2.0M environment steps. For the base locomotion experts, a trial is considered successful if the humanoid maintains its balance without falling for a duration of 30s. For complex manipulation tasks, success is defined by the successful completion of the door-opening sequence. We evaluate each policy across 10 independent trials with different random seeds and report the total number of successful completions to reflect the robustness of each method.

Result. The quantitative results in Table 2 and Fig. 4reveal the core advantages of our hierarchical architecture in complex humanoid control. First, *the inferior performance of DreamerV3 and TD-MPC2 in base locomotion tasks can be attributed to the challenges of exploration in high-dimensional state spaces.* Conventional Model-based RL algorithms attempt to learn world models and whole-body control policies from scratch, leading to a failure to maintain dynamic balance within the 2.0M sampling budget. In contrast, our method leverages **expert guidance** to constrain the policy search within a high-performance manifold, achieving superior sample efficiency.

Second, *the performance gap in the "Door" task (*80% *vs.* 10%*) highlights the necessity of coupling semantic reasoning with motor execution.* While baselines often lose balance while attempting manipulation, our VLM-driven layer dynamically adapts expert weights based on visual feedback, maintaining center-of-mass stability during complex sequences. This demonstrates that our **semantic-guided dynamic selection mechanism** is pivotal for solving long-horizon tasks that require both precision and stability.

4.5. Ablation study and stability analysis

Purpose of the experiment. To rigorously quantify the contribution of each module within the MetaWorld framework, we perform a systematic series of ablation experiments. First, we ablate semantic planning by replacing the VLM-based planner with a randomly initialized MLP-based weighting layer, where expert weights remain learnable via RL gradients but lack semantic guidance. Next, we ablate dynamic expert selection by fixing $\alpha = 1.0$ during train-

ing, thereby disabling state-aware expert selection. Finally, we ablate skill-expert guidance by setting $\lambda = 0$ in the TD-MPC2 optimization objective, removing expert action guidance. We present experimental observations and analysis for each ablation.

Result. As shown in Table 3 To systematically decouple the contributions within the **MetaWorld** framework, we conduct a series of ablation experiments on the challenging "Door" task. As illustrated in Fig. 5, the catastrophic 72.7% performance collapse when replacing the VLM with a naive learnable weighting layer reveals the framework's absolute reliance on *high-level task decomposition*; even with learnable parameters, without semantic grounding, the agent fails to discover the complex manifold of expert combinations required to translate abstract symbols into executable sub-goal sequences. This confirms that VLM-based planning provides a critical *semantic warm-start* for solving the symbol grounding problem. Furthermore, the expert guidance module serves as a critical **behavioral anchor**, where its **52.9% contribution** demonstrates that in high-DOF humanoid systems, naive exploration often fails due to complex contact dynamics. By providing a *behavioral manifold constraint*, expert policies restrict the search space to kinematically feasible regions, facilitating a vital synergy between imitation-based priors and online RL. Lastly, while dynamic expert selection contributes a relatively smaller **15.4%** to the success rate, it provides essential **closed-loop resilience**, enabling reactive recovery against external perturbations or distribution shifts. Ultimately, the framework's superiority stems from the seamless integration of semantic reasoning, behavioral priors, and dynamic adaptation.

Table 3. Ablation Study: Contribution of each component to the performance.

Method	Return (↑)	Steps (M, ↓)
Full Method	**680.0**	**1.7**
w/o VLM	185.7	4.1
w/o Dynamic Selection	575.3	2.2
w/o Expert Guidance	320.5	2.9

5. Limitations

Despite its superior performance, MetaWorld exhibits several limitations that warrant further investigation. First, the current **expert policy generation** relies on a basic imitation learning module with trajectory-matching reward mechanisms, which lacks the ability to dynamically weigh fine-grained motion errors at the joint level. Second, the **expert policy selection** mechanism employs a weighted fusion approach rather than a truly intelligent routing architecture, such as a Mixture-of-Experts (MoE) system. This limits the precision of skill composition and the fluidity of semantically-conditioned switching between disparate behaviors. Third, the framework's **generalization and scalability** remain constrained; it lacks explicit few-shot adaptation capabilities for novel, complex task combinations and does not yet address potential gradient interference during multi-skill coordination. These factors may hinder its performance when scaling to even more diverse and unstructured real-world environments.

6. Conclusion

This study presents **MetaWorld**, a hierarchical world model framework designed to bridge the semantic-physical gap in humanoid robot control. By integrating VLM-based semantic planning, dynamic expert policy transfer, and latent dynamics models, MetaWorld establishes a cohesive pipeline from high-level reasoning to low-level physical execution. Experimental results on **Humanoid-Bench** demonstrate that our framework achieves a 139.1% improvement in average reward, significantly enhancing both task completion efficiency and motion coherence. Moving forward, we plan to iteratively upgrade the framework by incorporating dynamic reward shaping for imitation learning, an MoE-based semantic-aware routing mechanism, and enhanced few-shot transfer mechanisms. These advancements will move MetaWorld toward a more robust, adaptive, and scalable solution for the complex loco-manipulation demands of next-generation humanoid robots.

References

[1] Marina Y. Aoyama, Sethu Vijayakumar, and Tetsuya Narita. Few-shot transfer of tool-use skills using human demonstrations with proximity and tactile sensing. 2025. 3

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *Science Robotics*, 2023. 3

[3] Lukas Fehring, Theresa Eimer, and Marius Lindauer. Growing with experience: Growing neural networks in deep reinforcement learning. In *2025 Multi-disciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2025. 2

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017. 2

[5] Kentaro Fujii and Shingo Murata. Real-world robot

control by deep active inference with a temporally hierarchical world model. *IEEE Robotics and Automation Letters (RA-L)*, 2025. 2

[6] Haoran Geng, Songlin Wei, Congyue Deng, Bokui Shen, He Wang, and Leonidas Guibas. SAGE: Bridging Semantic and Actionable Parts for GEneralizable Articulated-Object Manipulation under Language Instructions. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 2

[7] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[8] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations (ICLR)*, 2021.

[9] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025. 5

[10] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. 2

[11] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 5, 6

[12] M. Hou, K. Hindriks, A. E. Eiben, and K. Baraka. Robot policy transfer with online demonstrations: An active reinforcement learning approach, 2025. 3

[13] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Fei-Fei Li. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. In *Conference on Robot Learning (CoRL)*, 2024. 2, 3

[14] Yuankun Jiang, Chenglin Li, Wenrui Dai, Junni Zou, and Hongkai Xiong. Variance reduced domain randomization for reinforcement learning with policy gradient. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 2

[15] M. Kayaalp, S. Vlaski, and A. Sayed. Dif-maml: Decentralized multi-agent meta-learning. *IEEE Open Journal of Signal Processing*, 3:71–93, 2022. 2

[16] Lars Kunze, Tobias Roehm, and Michael Beetz. Towards semantic robot description languages. In *2011 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5589–5595, 2011. 2

[17] Chenhao Li, Andreas Krause, and Marco Hutter. Robotic world model: A neural network simulator for robust policy optimization in robotics. In *Proceedings of The 8th Conference on Robot Learning (CoRL)*, 2025. 2

[18] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. Hamster: Hierarchical action models for open-world robot manipulation. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[19] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[20] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenlong Gao, and Hao Dong. Omnimanip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[21] Silvia Sapora, Gokul Swamy, Chris Lu, Yee Whye Teh, and Jakob Nicolaus Foerster. Evil: Evolution strategies for generalisable imitation learning. In *International Conference on Machine Learning (ICML)*, 2024. 2

[22] Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024. 4

[23] Yutong Shen, Hangxu Liu, Lei Zhang, Penghui Liu, Ruizhe Xia, Tianyi Yao, and Tongtong Feng. Detach: Cross-domain learning for long-horizon tasks via mixture of disentangled experts, 2025. 2

[24] Kai Tai Song and Hsiang Hsi Chen. Hagrasp: Hybrid action grasp control in cluttered scenes using deep reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 2

[25] Zichen Zhang, Yunshuang Li, Osbert Bastani, Abhishek Gupta, Dinesh Jayaraman, Yecheng Jason Ma, and Luca Weihs. Universal visual decomposer: Long-horizon manipulation made easy. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6973–6980. IEEE, 2024. 2

[26] Xinghao Zhu, Yuxin Chen, Lingfeng Sun, Farzad Niroui, Simon Le Cleac'h, Jiuguang Wang, and Kuan Fang. Relic: Versatile loco-manipulation through flexible interlimb coordination. In *Conference on Robot Learning (CoRL)*, 2025. 3