PHI: A HIERARCHICAL WORLD ACTION MODEL WITH SELECTIVE PREDICTION AND FORCE FEEDBACK FOR HUMANOID BIMANUAL MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid bimanual dexterous manipulation requires coordinated dual-arm and finger actions under evolving contact interactions. Existing world action models (WAMs) offer a promising way to model such future dynamics, but primarily optimize future representations for prediction, leaving their utility for downstream action generation implicit. We introduce **PHI** (Predictive Hierarchical Intelligence), a **hierarchical world action model** for humanoid bimanual manipulation that integrates **selective future prediction** with **force feedback**. Rather than reconstructing or preserving the complete future observation, PHI selectively learns compact future visual representations that retain information useful for action generation. A hierarchical predictive pathway first extracts task-relevant future representations from privileged future observations during training and then learns to infer them from the current interaction context at deployment. To complement visual prediction in contact-rich interactions, PHI further incorporates a temporal force encoder that captures recent force evolution and provides physical feedback about contact establishment, loading, and interaction changes. The predicted future representation and force feedback are jointly integrated to guide action generation. Across five real-world humanoid bimanual manipulation tasks, PHI achieves an aggregate success rate of 82.0%, outperforming the strongest baseline by 23.3 percentage points and up to a 30.0 percentage points gain on unseen book instances. These results demonstrate PHI's effectiveness through selective future prediction and temporal force feedback. PHI-WAM.

1 INTRODUCTION

Humanoid bimanual dexterous manipulation requires coordinating dual-arm motions and fine-grained finger actions under continuously evolving physical interactions Chen et al. (2024b); Liu et al. (2025b). This is particularly challenging in contact-rich tasks, where subtle changes in deformation, slip, separation, or contact mode can demand substantially different actions despite similar visual observations Liu et al. (2025d). Since such interaction states are only partially observable from vision Liu et al. (2025a), successful Humanoid manipulation requires both estimating the current physical interaction and anticipating how task-relevant dynamics may evolve.

Many generalist robot policies directly map the current interaction context to actions. Vision-Language-Action (VLA) models, such as OpenVLA Kim et al. (2024) and $\pi_{0.5}$ Black et al. (2024); Physical Intelligence et al. (2025), predict actions from visual observations, language instructions, and robot states Brohan et al. (2023); Octo Model Team et al. (2024). Recent World Action Models (WAMs), such as WorldVLA Cen et al. (2025) and GR-2 Cheang et al. (2024), further incorporate future prediction into action-oriented modeling. Existing approaches predominantly learn to predict or reconstruct future visual observations, either explicitly at the image or video level Ye et al. (2026); Team et al. (2026) or implicitly in a latent space Cen et al. (2025); Sun et al. (2026). While latent prediction avoids the cost of pixel-level reconstruction, the underlying predictive objective remains centered on representing the future observation Bi et al. (2025). This can lead the model to preserve visual information that is predictable but unnecessary for the current task Yuan et al. (2026), leaving

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

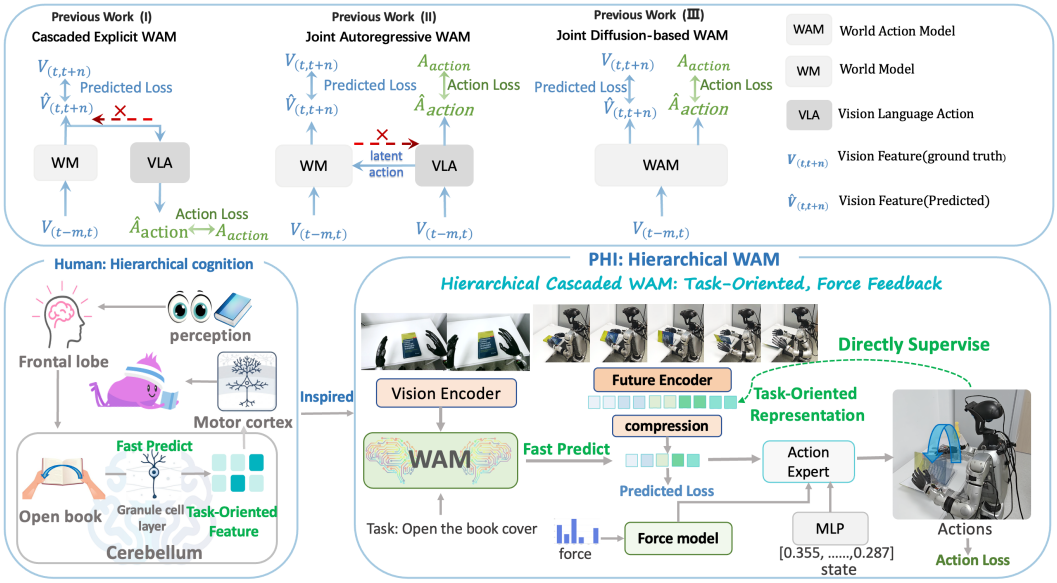


Figure 1: Comparison of future representation learning paradigms. Existing approaches learn future visual representations through video prediction, with or without action conditioning, while PHI learns future representations grounded in their utility for downstream action generation.

an important question under-specified: **what information should a future representation preserve to support downstream action generation?**

We explore this question by shifting the focus from reconstructing future observations to learning task oriented future representations. Inspired by how the brain may selectively process and represent information according to ongoing goals Nguyen & Person (2025); Gao et al. (2018), our method allows task oriented supervision to directly shape the predicted future representation, encouraging the model to selectively retain information that supports the current task. This enables a more selective and efficient representation of the future, focusing predictive capacity on task oriented information rather than reconstructing future observations in their entirety.

Based on this perspective, we introduce **PHI (Predictive Hierarchical Intelligence)**, a hierarchical World Action Model for humanoid bimanual dexterous manipulation. As illustrated in Fig.1, existing WAMs can be broadly categorized into following paradigms. Cascaded approaches first reconstruct future visual observations and then derive representations for action generation, emphasizing future visual reconstruction while also capturing task-irrelevant visual information Cheang et al. (2024). In contrast, joint autoregressive and diffusion approaches learn future representations together with action representations, enabling more efficient latent prediction while establishing an indirect connection between the learned representation and task execution Sun et al. (2026); Team et al. (2026). These paradigms provide complementary ways to connect future modeling with action generation. In contrast, the human brain provides a different perspective by forming compact predictive representations and selectively emphasizing task oriented information for motor control Nguyen & Person (2025); Gao et al. (2018).

Inspired by this principle, PHI combines efficient latent prediction with a hierarchical cascaded WAM that selectively learns future features useful for downstream action generation. PHI avoids pixel-level reconstruction and, through an **Exploit–Predict–Align** procedure, extracts future latents from privileged observations, predicts them from the current interaction context, and aligns them with downstream action supervision. For humanoid manipulation, PHI further incorporates temporal force feedback to encode force features and complement visual prediction with physical information about contact and interaction changes. The predicted future representation and force feedback jointly guide the action expert, which adopts structured pathways for different body effectors inspired by their organization in the human motor cortex Gordon et al. (2023) as detailed in Sec.3. We evaluate PHI on a real Unitree G1 humanoid across five tasks, achieving 82.0% aggregate success, compared with 58.7% for strong general-purpose VLA baselines, and up to 30.0 percentage points of improvement on unseen books. Our main contributions are summarized as follows:

- We introduce **PHI**, a hierarchical WAM that integrates selective future prediction and temporal force feedback for humanoid bimanual manipulation.
- We develop an **Exploit–Predict–Align** procedure to learn task oriented future representations from privileged future observations and align them with downstream policy.
- We introduce **temporal force feedback** to capture evolving physical interactions beyond visual observations and guide robust humanoid manipulation.

2 Related Work

2.1 Future Modeling for Robot Action Generation

Generalist vision-language-action policies generate robot actions from visual observations, language instructions, and robot states. RT-2 transfers vision-language knowledge to robotic control Brohan et al. (2023), while Octo and OpenVLA develop general-purpose policies through large-scale multimodal pretraining and robot demonstrations Octo Model Team et al. (2024); Kim et al. (2024). π_0 combines a pretrained vision-language backbone with a flow-matching action expert, while $\pi_{0.5}$ extends this paradigm through heterogeneous co-training across robots, tasks, and data sources Black et al. (2024); Physical Intelligence et al. (2025). While these models provide strong policy backbones, they do not explicitly model future interaction representations. World models provide an alternative way to incorporate foresight into robot control. Classical latent world models learn compact dynamics for planning and control Hafner et al. (2019; 2023); Hansen et al. (2024), while robotic manipulation methods such as GR-1, UniPi, and RoboDreamer model future interactions through visual prediction or video generation Wu et al. (2023); Du et al. (2023); Zhou et al. (2024). More recent World Action Models (WAMs) connect future modeling with action generation: World-VLA models future visual observations with robot actions Cen et al. (2025), while GR-2 combines video generation with action prediction Cheang et al. (2024). Motus and MotuBrain further explore unified latent action and world action formulations Bi et al. (2025); Team et al. (2026), while Fast-WAM uses future-video modeling as training supervision without explicit future prediction at inference Yuan et al. (2026). In latent space, LAPA learns latent actions, Moto predicts latent motion tokens, and VLA-JEPA predicts future representations without pixel reconstruction Ye et al. (2024); Chen et al. (2024a); Sun et al. (2026). PHI instead learns *task-oriented future representations*, using downstream action supervision to shape the privileged future latent, which is then predicted from the current context for action generation.

2.2 Force-Aware Manipulation

Vision alone provides incomplete information for contact-rich manipulation, motivating the integration of force and tactile sensing into learning-based policies. Existing methods incorporate physical sensing in different ways: ForceMimic jointly models time-varying force and motion for contact-rich imitation learning Liu et al. (2025c), while FoAR and Reactive Diffusion Policy (RDP) combine visual observations with force or tactile feedback for reactive control He et al. (2024); Xue et al. (2025). Recent VLA-based approaches further integrate physical sensing into generalist policies, including force-aware action decoding Yu et al. (2025), force-aware task representations and hybrid force-position control Li et al. (2026b), tactile-force alignment Huang et al. (2026), and slow fast force adaptation Li et al. (2026c). These studies highlight the benefits of physical sensing for multimodal action generation and force regulation. A related line of work models the temporal structure of force signals. FM-VLA compresses force histories into memory tokens to provide non-Markovian physical context Li et al. (2026a). PHI instead uses a short history of low-dimensional force measurements to encode the *ongoing physical interaction*, complementing the predicted future visual latent. The force representation captures recent contact evolution, while the future latent provides anticipatory information about task evolution, together supplying complementary physical and predictive context for contact-rich action generation.

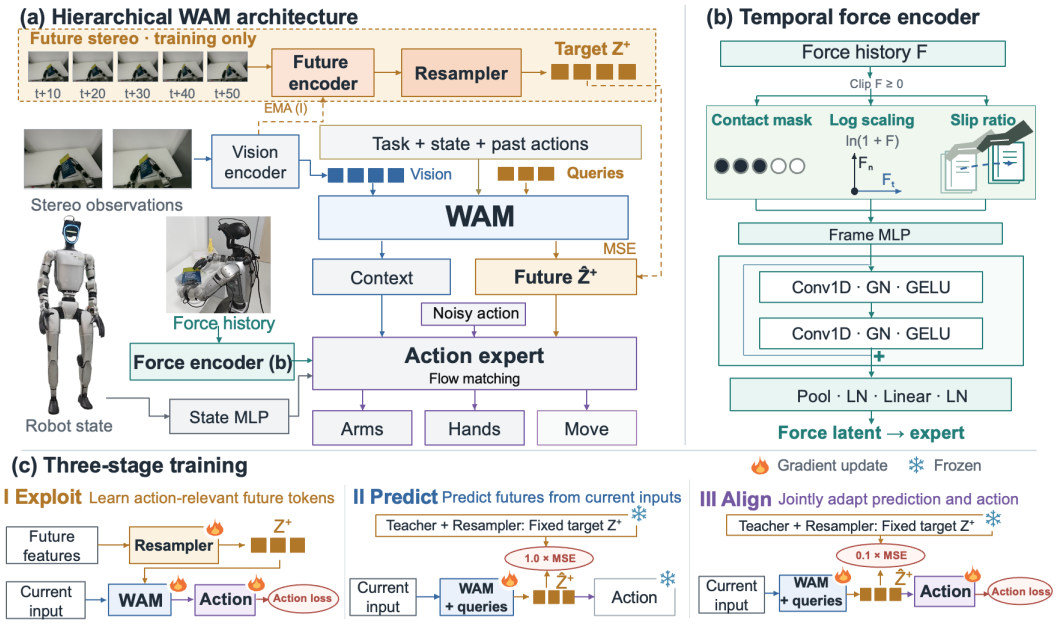


Figure 2: Overview of PHI. (a) A training-only future pathway constructs Z_t^+ from future stereo observations, while the current interaction context and future tokens condition the action policy. (b) A temporal force encoder maps force history to a physical-interaction latent. (c) The Exploit–Predict–Align training procedure.

3 Method

3.1 Problem Formulation and PHI Overview

PHI adopts a two-stage architecture consisting of a World Action Model (WAM) and a flow-matching action expert. Given the current interaction context, the WAM produces a task-relevant visual-language representation and predicts a task oriented future representation. During training, a separate *Progressive Task Oriented Future Learning* module $\mathcal{T}_\phi$ constructs the future representation target from privileged future observations, which is then used to train the causal future prediction pathway. Meanwhile, the action expert receives the representations produced by the WAM together with the robot state and recent force history, and generates a chunk of future actions. The overall architecture and training schedule are illustrated in Fig. 2.

Formally, we define the current interaction context as $c_t = (o_t, \ell, s_t)$, where o_t denotes the current visual observation, ℓ the language instruction, and s_t the robot proprioceptive state. The WAM produces the current visual-language representation and the predicted future representation:

$$\left(Z_t^{vl}, \hat{Z}_t^+\right) = \mathrm{WAM}_\theta(c_t). \tag{1}$$

Here, Z_t^{vl} denotes the current visual-language representation, while $\hat{Z}_t^+$ denotes the predicted future representation. During training, the Future Learning module $\mathcal{T}_\phi$ uses privileged future observations to construct the target future representation:

$$Z_t^+ = \mathcal{T}_\phi\left(o_{t+\Delta_1}, \ldots, o_{t+\Delta_M}\right). \tag{2}$$

In parallel, a temporal force encoder maps the recent force history $\mathcal{F}_t = \{F_{t-K+1}, \ldots, F_t\}$ to a force representation $z_t^f = E_f(\mathcal{F}_t)$. The action expert then takes the WAM representations, robot state, and force representation as inputs and generates an action chunk $A_t = \{a_t, \ldots, a_{t+H-1}\}$ through conditional flow matching, formulated as $A_t \sim \mathcal{G}_\theta(Z_t^{vl}, \hat{Z}_t^+, s_t, z_t^f)$. At inference time, the privileged future observations and $\mathcal{T}_\phi$ are removed, and $\hat{Z}_t^+$ is predicted solely from the current interaction context, enabling causal action generation.

Inspired by the effector-specific organization of the human motor cortex Gordon et al. (2023), we further structure the action expert with separate projectors for different body effectors. Specifically,

the action space is decomposed into three groups, namely arms, hands, and move, each mapped to a dedicated action projector for generating its corresponding actions. This effector-specific decomposition reduces the learning burden of the high-dimensional action space and facilitates more effective optimization of coordinated humanoid actions.

3.2 Progressive Task Oriented Future Learning

During training, PHI uses future observations as privileged information to construct a compact target representation through module $\mathcal{T}_\phi$ which consists of an visual encoder E_v^{EMA} and a Perceiver-based resampler R_ψ. For each training sample at time t, we select M future observation times $t+\Delta_1, \ldots, t+\Delta_M$. The visual encoder E_v^{EMA} first extracts visual features from the corresponding future observations. Its initial parameters are copied from the online visual encoder and subsequently updated using exponential moving average (EMA), without receiving gradients directly. The extracted features are then aggregated by the resampler R_ψ into a fixed number of latent tokens:

$$Z_t^+ = \mathcal{T}_\phi\left(o_{t+\Delta_1}, \ldots, o_{t+\Delta_M}\right) = R_\psi\left(E_v^{\mathrm{EMA}}(o_{t+\Delta_1}), \ldots, E_v^{\mathrm{EMA}}(o_{t+\Delta_M})\right), \tag{3}$$

where Z_t^+ denotes the target future representation. The resampler compresses the future visual features into a compact, task oriented representation, retaining information most useful for downstream action generation. This provides the WAM with a compact and task oriented prediction target, substantially reducing the difficulty of future representation prediction.

3.3 Temporal Force Feedback

Visual observations provide rich information about object geometry and motion, but are insufficient to fully characterize physical interactions during contact. PHI therefore encodes the recent force history $\mathcal{F}_t = \{F_{t-K+1}, \ldots, F_t\}$ to provide explicit information about contact evolution. Each force measurement contains 20 channels. We first apply non-negative clipping and logarithmic scaling to compress the dynamic range, and additionally extract a binary contact mask and tangential-to-normal force ratios to explicitly describe contact state and potential sliding behavior. The resulting features are

$$\widetilde{F}_t = \left[\log(1+\max(F_t,0)); \mathbb{I}(F_t > 1.0); F_t^{\mathrm{ratio}}\right] \in \mathbb{R}^{50}, \tag{4}$$

where F_t^{ratio} contains 10 tangential-to-normal force ratios, computed as

$$F_{t,i}^{\mathrm{ratio}} = \log\left(1 + \frac{F_{t,i}^{\mathrm{tangential}}}{F_{t,i}^{\mathrm{normal}} + \epsilon}\right). \tag{5}$$

Thus, each timestep is represented by 50 force features, combining continuous force magnitude with discrete contact information and local force balance.

A temporal encoder E_f aggregates the recent force feature sequence $\widetilde{F}_{t-K+1}, \ldots, \widetilde{F}_t$ into a force representation $z_t^f = E_f(\widetilde{F}_{t-K+1}, \ldots, \widetilde{F}_t)$, which is provided directly to the action expert to complement the predictive visual representation $\hat{Z}_t^+$ with temporally grounded physical information.

3.4 Three-Stage Training: Exploit, Predict, and Align

PHI is trained with a three-stage curriculum, namely *Exploit, Predict, and Align*, which progressively transfers task oriented information from privileged future observations to a causal future prediction pathway. We first learn an task oriented future target using privileged observations, then train the model to predict this target from the current context, and finally replace the privileged target with the predicted representation, jointly optimizing future prediction and downstream action generation.

Stage I: Exploit. We first learn the task oriented future target from privileged future observations. Since Z_t^+ is constructed from future observations during training, it provides privileged access to future information and allows the model to identify features useful for downstream action generation. Given a demonstrated action chunk A_t and a noise sample $\epsilon \sim \mathcal{N}(0, I)$, we sample a flow time τ and construct $X_t^\tau = (1-\tau)A_t + \tau\epsilon$ with target velocity $u_t = \epsilon - A_t$. The action expert is optimized with the conditional flow-matching objective

$$\mathcal{L}_{\mathrm{I}} = \mathbb{E}_{\tau,\epsilon,A_t}\left[\left\|\mathcal{G}_\theta\left(X_t^\tau, \tau, Z_t^{vl}, Z_t^+, s_t, z_t^f\right) - u_t\right\|_2^2\right]. \tag{6}$$

During this stage, the resampler R_ψ is optimized through the action objective, while the teacher visual encoder E_v^{EMA} is updated only by exponential moving average. Consequently, the target future representation $Z_t^+ = \mathcal{T}_\phi(\cdot)$ is shaped to retain task oriented information that is useful for downstream action generation.

Stage II: Predict. After learning the task oriented future target, we train the WAM to predict the same representation from the current interaction context. Specifically, given $c_t = (o_t, \ell, s_t)$, the WAM produces $\left(Z_t^{vl}, \hat{Z}_t^+\right) = \mathrm{WAM}_\theta(c_t)$, where $\hat{Z}_t^+$ is supervised by the fixed privileged target through latent distillation:

$$\mathcal{L}_{\mathrm{II}} = \left\| \hat{Z}_t^+ - \mathrm{sg}\big(Z_t^+\big) \right\|_2^2. \tag{7}$$

Here, $\mathrm{sg}(\cdot)$ denotes stop-gradient. The WAM receives only the current interaction context and never accesses future observations, while Z_t^+ provides training-time supervision. Since Z_t^+ has been shaped by downstream action generation in Stage I, Stage II learns to predict task oriented future information from the current context.

Stage III: Align. Finally, we replace the privileged future representation Z_t^+ with the predicted future representation $\hat{Z}_t^+$ and jointly optimize future prediction and action generation, while keeping the target module $\mathcal{T}_\phi$ fixed. The overall objective combines the losses from the previous two stages:

$$\mathcal{L}\mathrm{III} = \mathcal{L}\mathrm{I} + \tau \mathcal{L}\mathrm{II}, \qquad \tau = 0.1. \tag{8}$$

Here, $\mathcal{L}$I uses $\hat{Z}_t^+$ as the future representation for action generation, while $\mathcal{L}$II maintains prediction consistency with the privileged target Z_t^+.

3.5 Multimodal Action Generation and Causal Inference

At deployment, the WAM takes the current context c_t and produces the current visual-language representation Z_t^{vl} and the predicted future representation $\hat{Z}_t^+$. The recent force history $\mathcal{F}_t$ is encoded into z_t^f, and the robot state s_t is directly provided to the action expert. The privileged future-observation pathway and target module $\mathcal{T}_\phi$ are removed, ensuring causal inference. The flow-matching action expert then generates an action chunk A_t, which is executed using Real-Time Chunking (RTC) Black et al. (2025). RTC asynchronously predicts the next action chunk while executing the current one, using action inpainting to smoothly connect consecutive chunks.

4 EXPERIMENTS

We conduct experiments to answer the following three questions: **(1)** Can PHI improve policy performance as the complexity of physical interaction increases? **(2)** Can PHI generalize to unseen objects and visual conditions? **(3)** What makes the learned future representation useful for humanoid bimanual dexterous manipulation?

4.1 Experimental Setup

Robot Platform. We evaluate all methods on a Unitree G1 humanoid with 29 degrees of freedom, equipped with two 6-DoF BrainCo dexterous hands with tactile sensing and a head-mounted binocular RGB camera. The policy predicts an action chunk of $H = 50$ steps in a receding-horizon manner. Actions are tracked by ULC, a unified loco-manipulation controller that converts root velocity, body height, torso, and dual-arm commands into whole-body joint targets Sun et al. (2025).

Task Suite and Dataset. We evaluate five real-robot tasks spanning different levels of physical interaction: *Walk to Table*, *Place Bottle*, *Open Book*, *Place Cube*, and *Place Book Straight*. As shown in Fig. 3, the suite progresses from no-contact locomotion and object grasping to sustained bimanual contact and friction-regulated manipulation. Table 1 summarizes the key characteristics of each task, including the required contact mode and bimanual coordination. We use three contact levels to describe the increasing interaction requirements: L1 for no intentional object contact, L2 for grasping or sustained contact, and L3 for friction-regulated contact. Representative tactile traces are shown in Fig. 11 (Appendix A.3). These traces characterize the interaction patterns in our suite.

For each task, we collect 100 demonstration trajectories through VR teleoperation. The collected data follow the LeRobot format Cadene et al. (2024), including synchronized observations, robot states, actions, and tactile measurements. From tactile measurements, we extract 20-dimensional normal and tangential force features from both hands, which are included in the 49-dimensional robot state and used to construct a 10-step force history for the temporal force encoder.

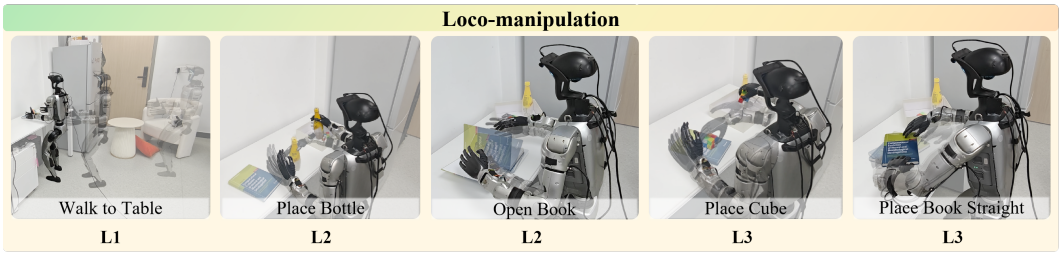


Figure 3: Real-robot task suite used for evaluation. The five tasks span different levels of physical interaction, from no-contact locomotion and object grasping to sustained bimanual contact and friction-regulated manipulation.

Table 1: Task characteristics and contact complexity. L1–L3 indicate increasing levels of physical interaction required for successful task completion.

Task	Group	Arms	Necessary contact mode
Walk to Table	L1	–	No intentional
Place Bottle	L2	Unimanual	Grasp
Open Book	L2	Bimanual	Sustained support
Place Cube	L3	Unimanual	Friction-regulated grasp
Place Book Straight	L3	Bimanual	Sliding and rotation

Baselines and Evaluation Protocol. We compare PHI with two representative vision-language-action policies: π_0 Black et al. (2024) and $\pi_{0.5}$ Physical Intelligence et al. (2025). All methods use the same training data, robot platform, task definitions, action space, controller, and evaluation protocol. For in-domain evaluation, each task is evaluated over 30 real-robot trials with randomized initial conditions, using the same initial-condition sequence across methods. For *Place Book Straight*, the table is divided into six regions with five trials per region, while for *Place Bottle*, trials are evenly distributed across three scene configurations.

4.2 Main Results

Table 2: Success rates (%) on the real-robot task suite.

Method	Walk to Table	Place Bottle	Open Book	Place Cube	Place Book Straight	Avg.
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	**100.0** (30/30)	10.0 (3/30)	46.7 (14/30)	36.7 (11/30)	43.3 (13/30)	47.3
π_0 (Black et al., 2024)	**100.0** (30/30)	66.7 (20/30)	50.0 (15/30)	40.0 (12/30)	36.7 (11/30)	58.7
PHI (Ours)	**100.0** (30/30)	**76.7** (23/30)	**86.7** (26/30)	**70.0** (21/30)	**76.7** (23/30)	**82.0** (+23.3)

Table 2 summarizes the real-robot results over 150 trials per method. PHI achieves an overall success rate of 82.0%, substantially higher than π_0 (58.7%), corresponding to an absolute improvement of 23.3 persentage points. The gains are most pronounced on contact-rich manipulation tasks. On the two bimanual book tasks, PHI reaches 86.7% on *Open Book* and 76.7% on *Place Book Straight*, outperforming both baselines. PHI also improves performance on *Place Bottle* and *Place Cube*, indicating that the benefit extends beyond bimanual manipulation to force-sensitive object interaction.

Task-Specific Performance Gains. Figure 4 summarizes the per-task gains. Across increasingly challenging tasks, PHI consistently exhibits larger gains over both baselines, particularly on contact-rich manipulation tasks. In contrast, all methods achieve 100% on *Walk to Table*, which requires no intentional hand–object contact. Figure 8 (Appendix) shows selected successful PHI executions across the five tasks.

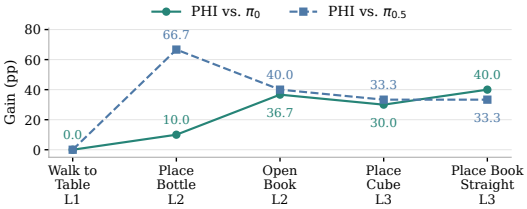


Figure 4: PHI's gains over π_0 and $\pi_{0.5}$.

4.3 Generalization to Unseen Objects and Visual Conditions

We next evaluate PHI under visual and physical distribution shifts in scene composition and object identity. Specifically, we consider visual distractors that alter scene composition and unseen objects that differ in appearance and physical interaction properties.

Robustness to Visual Distractors. We evaluate PHI under changes in scene composition by placing irrelevant objects with varying appearances and positions around the target object, while keeping the target and manipulation objective unchanged. As shown in Fig. 5(a), PHI achieves 70.0% success under visual clutter, compared with 30.0% for π_0 and 10.0% for $\pi_{0.5}$, versus 76.7% in-domain. The scenes below panel (a) show one in-domain layout and three clutter examples.

Generalization to Unseen Objects. We further evaluate object-level generalization using three books that never appear in the training data: *Linear Algebra Done Right*, *Federated Learning*, and *Deep Learning*. The held-out books differ from the training objects in cover appearance, thickness, and material properties, including compliance. These variations introduce both visual and physical distribution shifts, as changes in thickness and compliance can alter the forces and contact conditions required for grasping, bracing, sliding, and opening. We evaluate both *Place Book Straight* and *Open Book* using the same three held-out books, with 10 trials per book. As shown in Fig. 5(b,c), PHI achieves 60.0% success on *Place Book Straight*, compared with 30.0% for $\pi_{0.5}$ and 23.3% for π_0, and 56.7% on *Open Book*, compared with 33.3% for both baselines. These results correspond to absolute improvements of 30.0 and 23.4 percentage points over the strongest baseline, respectively. Together with the visual-distractor experiments, these results show that PHI generalizes to changes in scene composition, object identity, and physical properties. This suggests that its future representations capture task-relevant information for manipulation under varied visual and physical conditions.

4.4 Ablation Studies

We conduct ablation studies to examine the contributions of temporal force conditioning and predictive future latents. We evaluate both components on *Place Book Straight*, which requires continuous contact, bimanual coordination, and fine-grained interaction control. Each ablation removes one component while keeping the remaining architecture and training procedure unchanged.

Effect of Temporal Force Encoding. We remove the temporal force encoder and its history representation z_t^f, while retaining the current force measurements in s_t and keeping the Future Learning module $\mathcal{T}_\phi$ unchanged. As shown in Table 3, this leads to a 13.4 percentage points drop on *Place Book Straight*, which requires continuous force regulation during sliding. Without force modeling, the robot more frequently fails to move the book because the fingers either fail to establish contact or press the book too tightly. However, the overall manipulation behavior remains correct, resulting in a relatively limited drop in success rate. This suggests that force-history modeling mainly improves contact regulation during execution rather than action planning. The force traces for all tasks are provided in Fig. 11 (Appendix).

Table 3: Ablation success rates (%).

Variant	Place Book Straight
PHI	**76.7**(23/30)
w/o Temporal Force Encoder	63.3(19/30)
w/o Future Learning Module	50.0(15/30)

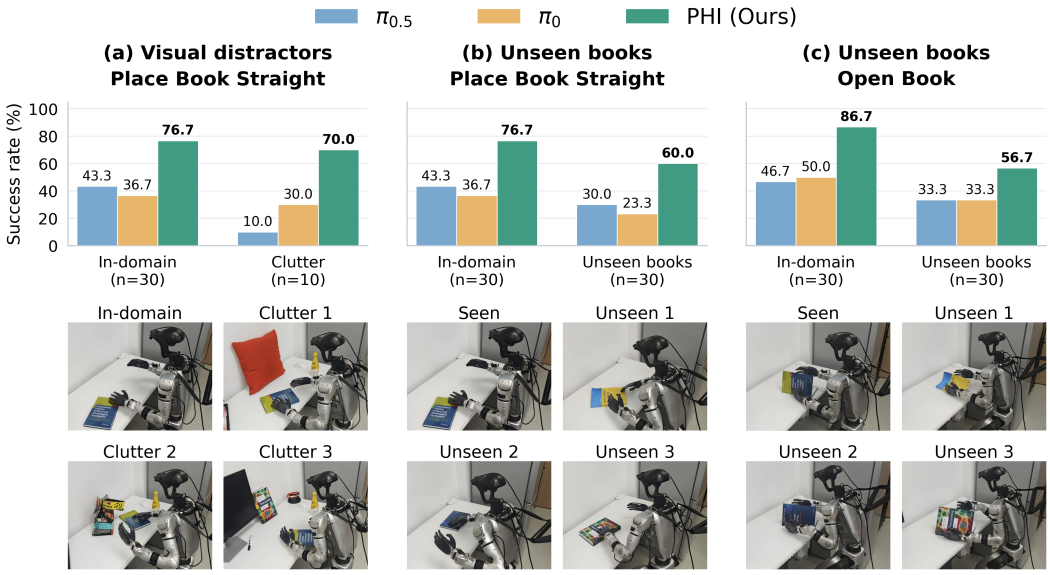


Figure 5: Generalization to visual distractors and unseen books.

Effect of Predictive Future Latents. We remove the Future Learning module $\mathcal{T}_\phi$ while retaining force conditioning. As shown in Table 3, this causes a 26.7 percentage points drop on *Place Book Straight*. In our experiments, the robot shows clear degradation in manipulation planning: after moving the book into position, it often stalls when a subsequent rotation is required or rotates in the wrong direction. These results show that the future features extracted by our module provide useful information for planning subsequent actions and selecting appropriate manipulation behaviors.

Future-Latent Alignment. Figure 6 visualizes predicted and target future-latent tokens. The target is extracted from five future stereo observations, and both features are projected onto a shared principal component analysis (PCA) basis. The first two components explain 36.7% and 21.2% of the total variance, respectively. Diamonds and triangles denote target and predicted tokens, with lines connecting matched queries; colors indicate per-query MSE over 2048 feature dimensions. The average cosine similarity is 0.9690 across 512 queries, indicating strong alignment between the predicted and target features. Figure 10(Appendix) further reports per-query original-space errors for the same sample.

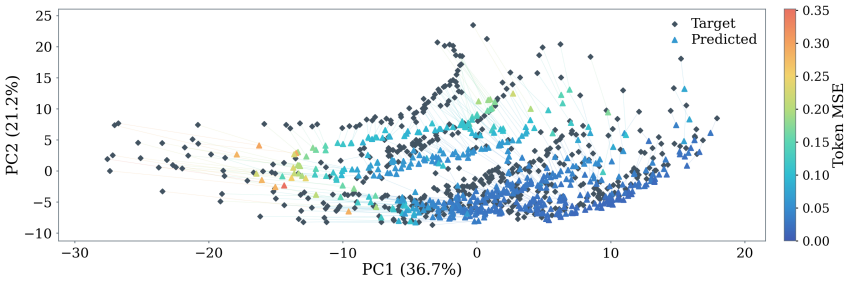


Figure 6: Joint PCA of predicted and target future-latent tokens for a single exported frame. Lines connect matched queries; colors indicate per-query MSE in the original feature space.

5 Conclusion

We presented PHI, a hierarchical World Action Model for humanoid bimanual dexterous manipulation. PHI learns future representations shaped by downstream action supervision and combines them with temporal force features for action generation. Real-robot experiments on the Unitree G1 demonstrate improved performance across bimanual manipulation tasks and distribution shifts, highlighting the value of task oriented future representations for humanoid manipulation.

REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model, 2025. URL https://arxiv.org/abs/2512.13030.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. URL https://arxiv.org/abs/2410.24164.

Kevin Black, Manuel Y. Galliker, and Sergey Levine. Real-time execution of action chunking flow policies, 2025. URL https://arxiv.org/abs/2506.07339.

Anthony Brohan, Noah Brown, Justice Carbajal, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *Proceedings of the Conference on Robot Learning*, 2023. URL https://arxiv.org/abs/2307.15818.

Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Khalil Meftah, Maxime Ellerbach, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. Worldvla: Towards autoregressive action world model, 2025. URL https://arxiv.org/abs/2506.21539.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Yi Chen, Yuying Ge, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for robot manipulation. *arXiv preprint arXiv:2412.04445*, 2024a. URL https://arxiv.org/abs/2412.04445.

Yuanpei Chen, Yiran Geng, Fangwei Zhong, Jiaming Ji, Jiechuang Jiang, Zongqing Lu, Hao Dong, and Yaodong Yang. Bi-dexhands: Towards human-level bimanual dexterous manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2804–2818, 2024b. doi: 10.1109/TPAMI.2023.3339515.

Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023. URL https://arxiv.org/abs/2302.00111.

Zhenyu Gao, Courtney Davis, Alyse M Thomas, Michael N Economo, Amada M Abrego, Karel Svoboda, Chris I De Zeeuw, and Nuo Li. A cortico-cerebellar loop for motor planning. *Nature*, 563(7729):113–116, 2018.

Evan M. Gordon, Roselyne J. Chauvin, Andrew N. Van, Aishwarya Rajesh, Ashley Nielsen, Dillan J. Newbold, Charles J. Lynch, Nicole A. Seider, Samuel R. Krimmel, Nicole R. Scheidter, Julia Monk, Ryland L. Miller, Athanasia Metoki, David F. Montez, Annie Zheng, Immanuel Elbau, Thomas Madison, Tomoyuki Nishino, Michael J. Myers, Sydney Kaplan, Carolina Badke D'Andrea, Damion V. Demeter, Matthew Feigelis, Julian S. B. Ramirez, Ting Xu, Deanna M. Barch, Christopher D. Smyser, Cynthia E. Rogers, Jan Zimmermann, Kristen N. Botteron, Joshua R. Pruett, Jon T. Willie, Peter Brunner, Benjamin P. Kay, Conor Liston, Deanna J. Greene, Jarod L. Roland, Jonathan D. Power, Conor Liston, Steven E. Petersen, Marcus E. Raichle, Timothy O. Laumann, Damien N. Fair, and Nico U. F. Dosenbach. A somato-cognitive action network alternates with effector regions in motor cortex. *Nature*, 617(7960):351–359, 2023. doi: 10.1038/s41586-023-05964-2.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*, pp. 2555–2565, 2019. URL https://arxiv.org/abs/1811.04551.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. URL https://arxiv.org/abs/2301.04104.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2024. URL https://arxiv.org/abs/2310.16828.

Zihao He, Hongjie Fang, Jingjing Chen, Hao-Shu Fang, and Cewu Lu. FoAR: Force-aware reactive policy for contact-rich robotic manipulation. *arXiv preprint arXiv:2411.15753*, 2024. URL https://arxiv.org/abs/2411.15753.

Yuzhe Huang, Pei Lin, Wanlin Li, Daohan Li, Jiajun Li, Jiaming Jiang, Chenxi Xiao, and Ziyuan Jiao. TaF-VLA: Tactile-force alignment in vision-language-action models for force-aware manipulation. *arXiv preprint arXiv:2601.20321*, 2026. URL https://arxiv.org/abs/2601.20321.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, et al. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. URL https://arxiv.org/abs/2406.09246.

Ruicheng Li, Qixiu Li, Ruichun Ma, Yu Deng, Lin Luo, Zhiying Du, Jianfeng Xiang, Huizhi Liang, Ruicheng Wang, Jiaolong Yang, and Baining Guo. FM-VLA: Force-based memory for vision-language-action models in contact-rich manipulation. *arXiv preprint arXiv:2607.18231*, 2026a. URL https://arxiv.org/abs/2607.18231.

Yang Li, Zhaxizhuoma, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, Qiaojun Yu, Cewu Lu, Yu Qiao, and Jiangmiao Pang. ForceVLA2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation. *arXiv preprint arXiv:2603.15169*, 2026b. URL https://arxiv.org/abs/2603.15169.

Yao Li, Peiyuan Tang, Wuyang Zhang, Chengyang Zhu, Yifan Duan, Weikai Shi, Xiaodong Zhang, Zijiang Yang, Jianmin Ji, and Yanyong Zhang. FAVLA: A force-adaptive fast-slow vla model for contact-rich robotic manipulation. *arXiv preprint arXiv:2602.23648*, 2026c. URL https://arxiv.org/abs/2602.23648.

Qingtao Liu, Yu Cui, Zhengnan Sun, Gaofeng Li, Jiming Chen, and Qi Ye. Vtdexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: A diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Wenhai Liu, Junbo Wang, Yiming Wang, Weiming Wang, and Cewu Lu. ForceMimic: Force-centric imitation learning with force-motion capture system for contact-rich manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1105–1112, 2025c. doi: 10.1109/ICRA55743.2025.11128061. URL https://arxiv.org/abs/2410.07554.

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references. In *The Thirteenth International Conference on Learning Representations*, 2025d.

Katrina P Nguyen and Abigail L Person. Cerebellar circuit computations for predictive motor control. *Nature Reviews Neuroscience*, 26(9):538–553, 2025.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. URL https://arxiv.org/abs/2405.12213.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL https://arxiv.org/abs/2504.16054.

Jingwen Sun, Wenyao Zhang, Zekun Qi, Shaojie Ren, Zezhi Liu, Hanxin Zhu, Guangzhong Sun, Xin Jin, and Zhibo Chen. VLA-JEPA: Enhancing vision-language-action model with latent world model. *arXiv preprint arXiv:2602.10098*, 2026. URL https://arxiv.org/abs/2602.10098.

Wandong Sun, Luying Feng, Baoshi Cao, Yang Liu, Yaochu Jin, and Zongwu Xie. ULC: A unified and fine-grained controller for humanoid loco-manipulation. *arXiv preprint arXiv:2507.06905*, 2025. URL https://arxiv.org/abs/2507.06905.

Motubrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, Louis Liu, Mengchen Cai, Rongxu Cui, Ruowen Zhao, Runqing Wang, Shuhe Huang, Yao Feng, Yinze Rong, Zeyuan Wang, and Jun Zhu. Motubrain: An advanced world action model for robot control, 2026. URL https://arxiv.org/abs/2604.27792.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. URL https://arxiv.org/abs/2312.13139.

Han Xue, Jieji Ren, Wendi Chen, Gu Zhang, Yuan Fang, Guoying Gu, Huazhe Xu, and Cewu Lu. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2503.02881.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. URL https://arxiv.org/abs/2410.11758.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, Wenqiang Zhang, and Cewu Lu. ForceVLA: Enhancing vla models with a force-aware moe for contact-rich manipulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2505.22159.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination?, 2026. URL https://arxiv.org/abs/2603.16666.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. URL https://arxiv.org/abs/2404.12377.

A Appendix

A.1 Visualization of Main Results

Figure 7 visualizes the success rates reported in Table 2 to facilitate comparison across methods and tasks. Figure 8 presents representative PHI execution sequences captured from a third-person view across the five real-robot tasks: *Walk to Table*, *Place Bottle*, *Open Book*, *Place Cube*, and *Place Book Straight*. The sequences show successful task completion across both locomotion and contact-rich bimanual manipulation, demonstrating consistent execution across the evaluated tasks.

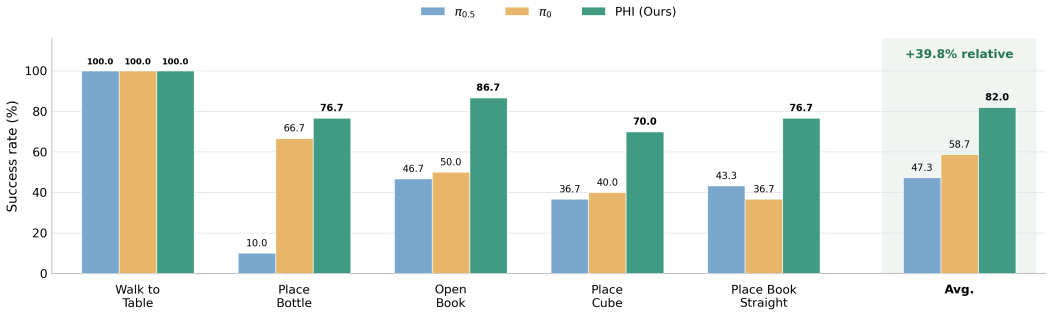


Figure 7: Success rates across five real-robot tasks, with 30 trials per task and method. Avg. denotes the mean success rate across tasks. Bold labels indicate the best result in each column.

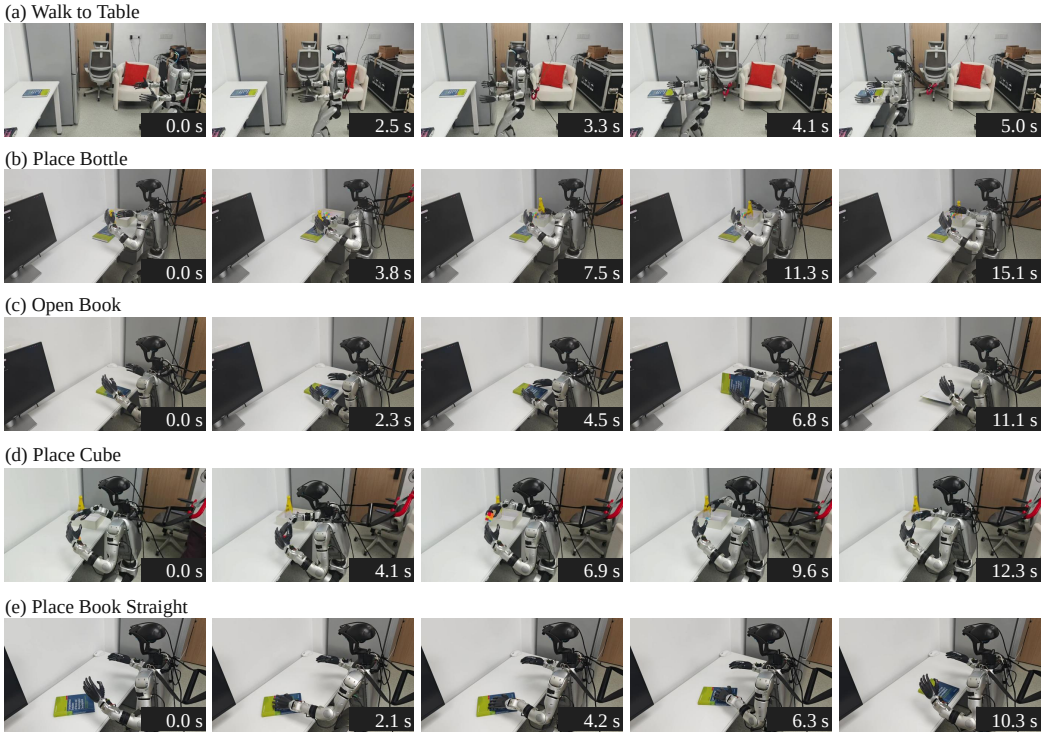


Figure 8: Selected PHI execution sequences across the five tasks.

A.2 Future-Target Attention and Alinement.

Figure 9 visualizes the final-layer attention of the future Perceiver when aggregating privileged future stereo observations to construct the target representation. In the example, attention is concentrated on the manipulated object and interacting hands, indicating that the target representation emphasizes task-relevant visual regions for future manipulation.

Figure 10 compares the predicted and target future representations for a single frame. In (a), the predicted latent is visualized across query tokens and pooled channels, while (b) shows the corresponding privileged target latent using the same color scale, allowing their feature patterns to be directly compared. In (c), the per-query MSE is computed in the original feature space to quantify the discrepancy between the two representations. The predicted and target latents exhibit similar feature patterns, with a mean MSE of 0.0632 and a mean cosine similarity of 0.9690, indicating that the predicted representation closely aligns with the privileged future target.

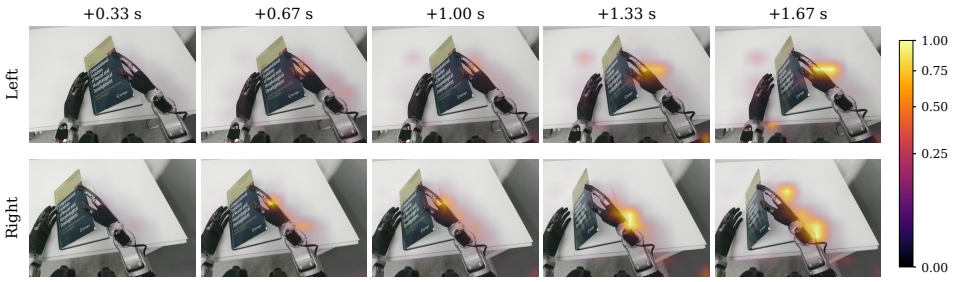


Figure 9: Final-layer attention of the future Perceiver over privileged future stereo observations, with rows denoting cameras and columns denoting future time offsets.

A.3 Tactile Traces During Task Execution

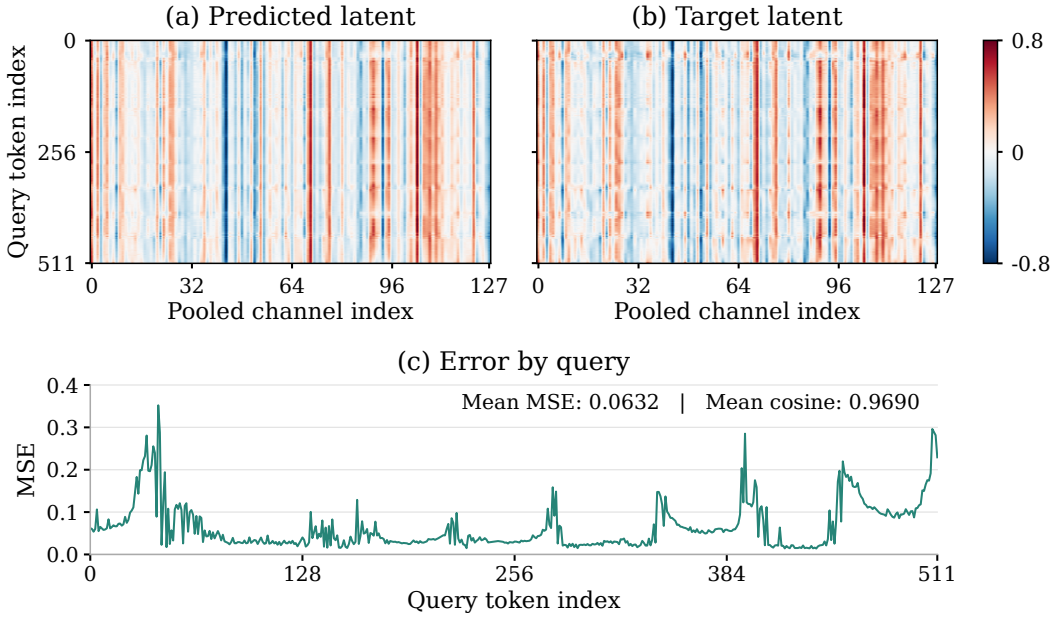


Figure 10: Single-frame future-latent comparison. (a,b) Predicted and target representations with shared color limits, clipped at the pooled absolute-value 99.5th percentile. (c) Per-query MSE in the original feature space.

Figure 11 shows selected successful execution segments for the five tasks, ordered from locomotion without object contact to contact-rich manipulation. Each row pairs the normal-force channels of both hands with five approximately equally spaced camera frames. Among these tasks, *Place Book Straight* exhibits the most complex force variations. The robot first uses frictional contact with one hand to pull the book toward itself, and then uses frictional interactions with both hands to rotate and reorient the book. The corresponding force traces show distinct and coupled variations between the two hands, reflecting the changing contact conditions throughout the manipulation.

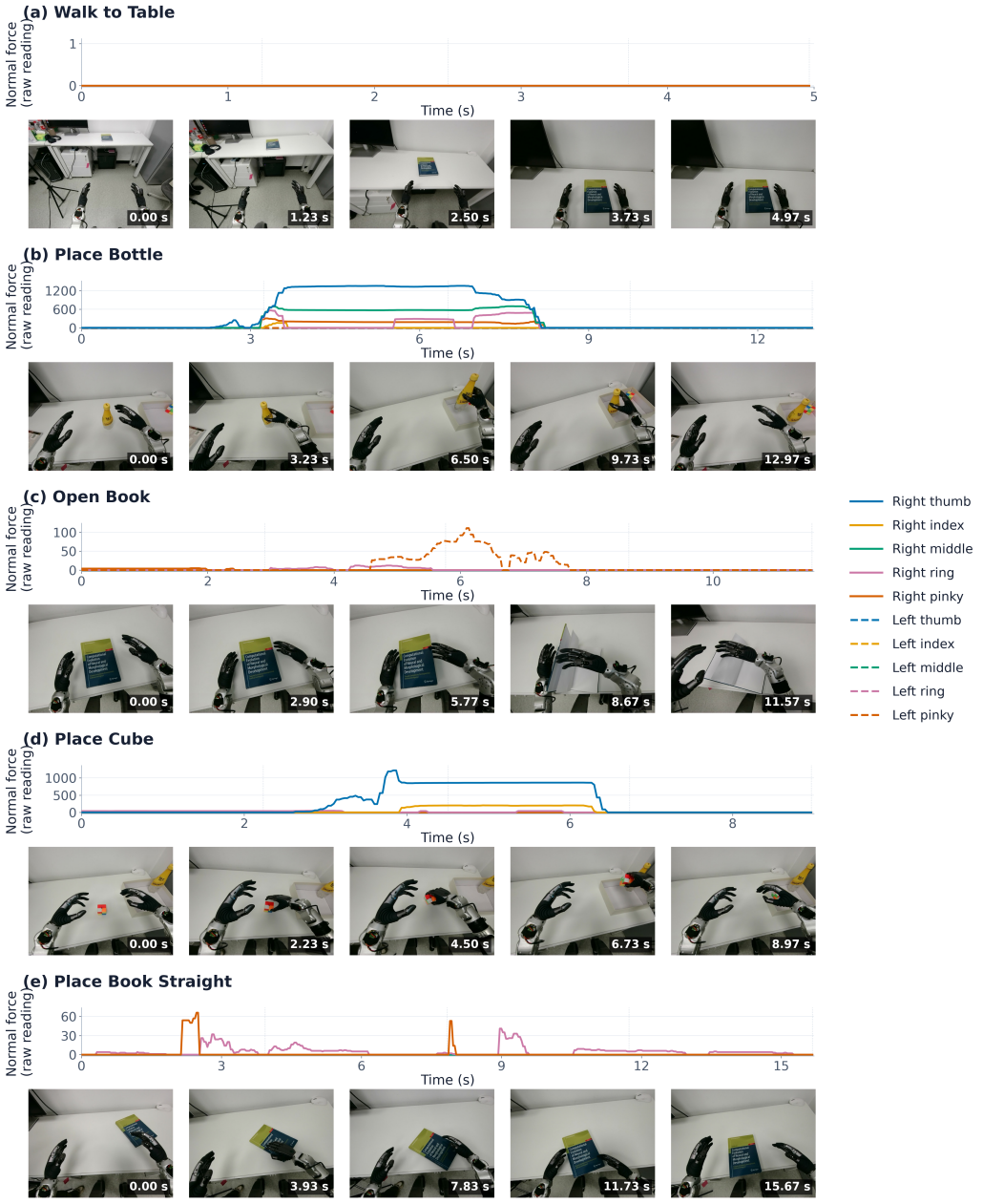


Figure 11: Normal-force traces and corresponding camera frames for five selected execution segments.

A.4 Key Training Hyperparameters

Table 4 summarizes the key training hyperparameters used for PHI, including the optimization settings, learning rates for different training stages, action and future-prediction loss weights, and model input and output configurations.

Table 4: Key training hyperparameters.

Hyperparameter	**Setting**
Batch size	6
Training steps	500K
Maximum token length	768
Action chunk length	50
Action history length	10
Number of future frames / stride	5 / 10
Training precision	BFloat16
Stage I learning rates	WAM: 5×10^{-5}; vision encoder: 2.5×10^{-5}; future resampler: 5×10^{-5}; action expert: 5×10^{-5}
Stage II learning rates	WAM: 5×10^{-5};
Stage III learning rates	WAM: 2.5×10^{-5}; action expert: 1×10^{-5}
Action loss weights	$\lambda_{\text{arm}} = 1.5$, $\lambda_{\text{hand}} = 2.0$, $\lambda_{\text{move}} = 0.5$
Future loss weight	$\lambda_{\text{future}} = 1.0$ (Stage II); $\lambda_{\text{future}} = 0.1$ (Stage III)
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay $= 0.01$
Learning-rate schedule	1,000-step warmup, cosine decay, gradient clipping $= 1.0$