MOTION CHAIN OF THOUGHT: DISENTANGLING MOTION SEMANTICS FROM VISUAL APPEARANCE FOR ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

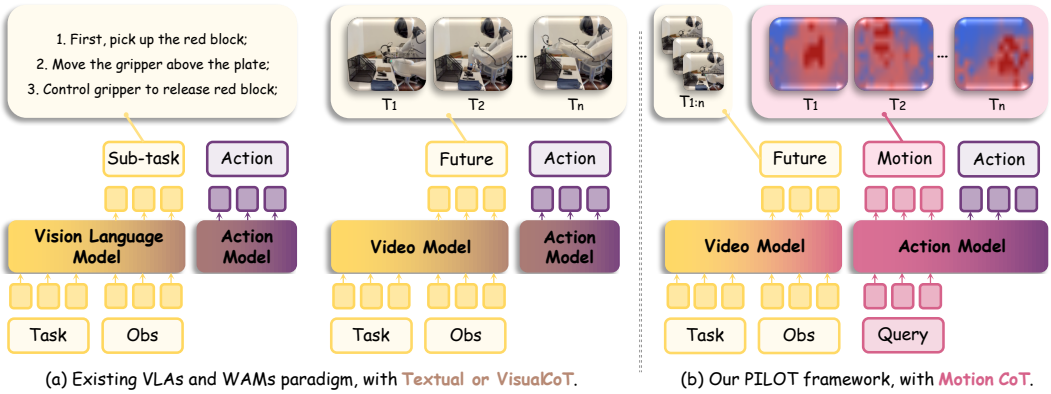


Figure 1: We propose **PILOT**, a framework that separates motion-semantic representation learning from trajectory refinement. Representational Deduction (RD) supervises learnable Motion-CoT tokens derived from observation and instruction using future representations as training targets. Motion-CoT enables the action model to focus on fine-grained trajectory refinement while alleviating representational entanglement.

ABSTRACT

World Action Models (WAMs) aim to construct a unified architecture capable of understanding world state evolution and guiding action generation. However, existing visual branches focus on predicting static visual observation, rather than reflecting potential transition information that captures the evolution of world states under motion interactions. This leads to representational entanglement between high-level physical condition evolution and low-level action trajectory generation within the Action Model, creating a structural bottleneck while weakening the predictive capability of world evolution modeling for action generation. We propose PILOT (Physical Inference for Latent Optimized Trajectories), whose core Representational Deduction (RD) bridges this gap by integrating motion chain-of-thought (CoT) guidance as a native model capability. Specifically, RD aims to encourage the action branch to explicitly model potential state transition tokens, which are retained as CoT in the reasoning space to guide fine-grained motion trajectory. Experiments demonstrate that RD not only improves the success rate and generalization ability of WAMs in complex robotic manipulation tasks but also enhances the model's physical interpretability by decoupling high-level motion semantics from low-level trajectory details. Furthermore, the abundant state transition supervision signals introduced by RD effectively alleviate the sparse supervision in action generation, enabling it to serve as an efficient few-shot real-robot fine-tuning strategy. Anonymous code and supplementary material are available on project website `https://pilot-wam-2026.github.io/`.

1 INTRODUCTION

World Models Hafner et al. (2023); Bruce et al. (2024) and Embodied Intelligence Brohan et al. (2022; 2023); Black et al. (2024); Intelligence et al. (2025); Kim et al. (2024); Xu et al. (2026)

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.

are dedicated to "understanding how the physical world evolves during human interaction" and "planning the action trajectory based on current observations," respectively.

The unification of these paradigms represents a significant step toward General Spatial Intelligence Yang et al. (2025). Recent World Action Models (WAMs) Bi et al. (2026); Lyu et al. (2026); Team et al. (2026); Li et al. (2026b;a) approach this goal by leveraging world models to predict future states and introducing the Inverse Dynamics Module (IDM) Chen et al. (2025); Liu et al. (2026); Zhang et al. (2026c) to convert these predictions into trajectory. However, future states are confined to the visual latent space constructed by frozen VAEs Pai et al. (2025); Cen et al. (2025a); Zhang et al. (2026b). This structural bottleneck causes the joint optimization objective of WAMs to primarily target visual reconstruction over trajectory generation. Guiding WAMs to focus on motion semantic representations to eliminate the visual bottleneck via an end-to-end framework remains a long-term goal.

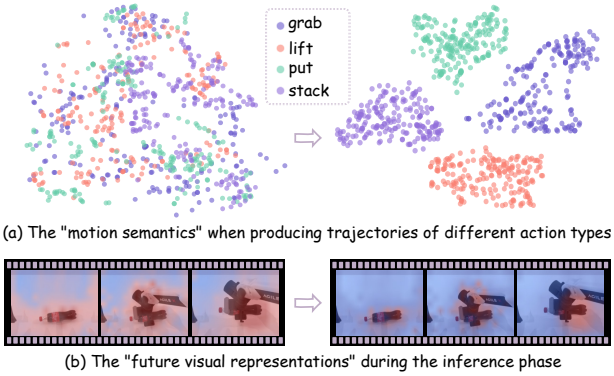


Figure 2: Visualization in (a) shows representations of existing WAMs (left) are entangled across actions, and PILOT (right) cleanly clusters by motion semantics. (b) PCA overlay shows PILOT focuses prediction on end-effector-object interaction, while baseline is distracted by background.

Recent experiments Yuan et al. (2026); Yang et al. (2026c); Sun et al. (2026b) indicate that while explicit future prediction aids training, it provides limited foresight for trajectory generation during inference. The effectiveness of WAMs stems from the future supervision to World Model, promoting the learning of physically meaningful visual priors. We attribute this to the fact that static visual representations at a single moment fail to reflect the evolution process of physical states under action interactions. Without explicit modeling of state transitions, directly generating actions through the diffusion process leads to representational entanglement, which is visualized in Fig. 2. To address this, we propose that action model should simultaneously learn high-level motion intention structures and low-level manipulation details from the same inputs.

How to supervise the model to learn the motion semantic representations that contain state transition information? The existing WAM Ye et al. (2026b); Gao et al. (2026); Cen et al. (2025b); Feng et al. (2025) can precisely predict what will happen in the future, but still doesn't know how to do it. Most relevant works are Latent Action Models (LAMs), which typically capture pixel-level differences and reconstruct future frames from the current frame via IDM Ye et al. (2025); Bu et al. (2025); Routray et al. (2025). However, this kind of visual-aware-only supervision, fails to prompt the model to understand what action conditions led to the variation in the visual state. Instead of deriving latent motions directly from consecutive frames, we employ them exclusively as supervision, guiding the model to summarize physical state transitions and motion semantics within learnable tokens. We propose not to pursue to capture motion-related state transition semantics from pixel variations between two frames, but rather to use variations in physical spatial states as supervisions.

In this paper, we propose Physical Inference for Latent Optimized Trajectories **(PILOT)**, an end-to-end framework grounded in the Representational Deduction **(RD)** mechanism that bridges the gap between static visual prediction and dynamic action manipulation by establishing motion semantic CoT reasoning as an innate capability of WAMs. Architecturally, PILOT decouples the World Model into understanding and generation pipelines to substitute explicit future image prediction with the Motion-CoT during inference, thereby accelerating deployment speed while improving both simulation and real-world. Specifically, the Action Model derives high-level Motion-CoT tokens from the current observation, instruction, and proprioceptive state; future observations are used only as training supervision for learning predictive state-transition representations. Subsequently, we freeze the pre-trained VJEPA-as-Encoder Assran et al. (2025); Mur-Labadia et al. (2026); Maes et al. (2026) to capture deterministic features tightly coupled with action conditions, and train a Causal Dynamics Engine **(CDE)** to predict future spatial positions, trajectories, shapes, and interaction relationships conditioned on motion semantic CoT, rather than probabilistic underlying visual rendering details.

Meanwhile, these motion semantics are retained as context tokens in the reasoning space, allowing the subsequent diffusion process to concentrate on refining fine-grained end-effector trajectories and achieve representational decoupling. This brings an additional benefit: Motion-CoT serves as an extra supervision signal, alleviating the sparse supervision problem faced by the action model due to the low information density of action signals. Our contributions can be summarized as follows.

- We propose PILOT, a novel WAM framework that decouples motion intention from trajectory generation via a Representational Deduction mechanism, and internalizes latent state transition information as Motion CoT to alleviate representational entanglement.
- We introduce a Causal Dynamics Engine for future-representation supervision and asymmetric attention in the Action-Perceiver that prevents noised actions from influencing motion-semantic queries, enabling action inference without pixel-level future reconstruction.
- PILOT achieves 97.9%, 80.3%, and 58.3% success on LIBERO, LIBERO-Plus, and RoboCasa-GR1, respectively, and 82.5% average success on the eight standard real-world tasks, with 91.6% lower inference latency compared to predict-then-act paradigms.

2 Related Work

World Models for Robotic Manipulations Recently, WAMs Cen et al. (2025b); Li et al. (2026b;a); Cen et al. (2025a); Zhang et al. (2026b;d) have emerged as a promising paradigm that leverages world model to predict future, thereby guiding motion manipulation. Existing video-model-based WAMs Kim et al. (2026) generally follow two paradigms. The serial-paradigm first predicts future observation and then reconstructs actions via an inverse dynamic Pai et al. (2025); Chen et al. (2025); Zhou et al. (2024), while the parallel-paradigm jointly models the visual-action space through shared attention mechanism (MoT) Bi et al. (2026); Lyu et al. (2026); Yuan et al. (2026) or joint denoising Ye et al. (2026b); Gao et al. (2026). Unlike generative pixel reconstruction strategies, joint embedding predictive frameworks (JEPA) Maes et al. (2026); Assran et al. (2025); Mur-Labadia et al. (2026) focus on predicting dynamic features such as object spatial positions and trajectories. Some studies Nilaksh et al. (2026); Su et al. (2026) incorporated JEPA-as-encoder to obtain abundant representations. Most relevant to our work is VLA-JEPA Sun et al. (2026b), which introduces JEPA to extract latent states, addressing the challenges of pixel reconstruction bias and future information leakage.

However, purely visual supervision fails to provide state transition information. This inherent discrepancy causes the deviation of co-training from the motion semantics, resulting in fragile transferability and inefficient fine-tuning; conversely, completely removing visual supervision and relying solely on sparse action supervision leads to representation collapse of video model. To address this, we propose Representational Deduction, which bridges the aforementioned gap by incorporating additional motion semantics and physical state transition information.

Latent Action Learning for Robotics LAMs Routray et al. (2025); Liu et al. (2026); Bruce et al. (2024) aim to pre-train backbones on large-scale human videos lacking action labels, thereby establishing prior knowledge of latent motion. Existing approaches Ye et al. (2025); Bu et al. (2025); Chen et al. (2025); Zhang et al. (2026c); Yang et al. (2026b) typically align vision and manipulation representations using Q-former Li et al. (2023) and unified codebook Zhang et al. (2026a); Ma et al. (2026b); Chen et al. (2025; 2026) to integrate robotic manipulation into the Next-Token-Prediction framework.

Despite the promise of LAMs, their performance is hindered by noise from camera-induced background variations and future frame leakage, which encourages trivial encoding of future frames rather than learning state transitions Sun et al. (2026b). These issues cause structural bottlenecks in learning control-oriented latent actions. Recent works mitigate this via latent constraints Luo et al. (2026b) or multi-encoder-query Su et al. (2026). However, the LAMs inherently aligns latent actions with visual space rather than actionable control signals. PILOT fundamentally avoids these defects by ensuring that future visual information is used exclusively as a supervision target, not as a source for capturing latent motion representations, thereby guaranteeing that the model focuses on dynamic representations independent of camera motion and background changes.

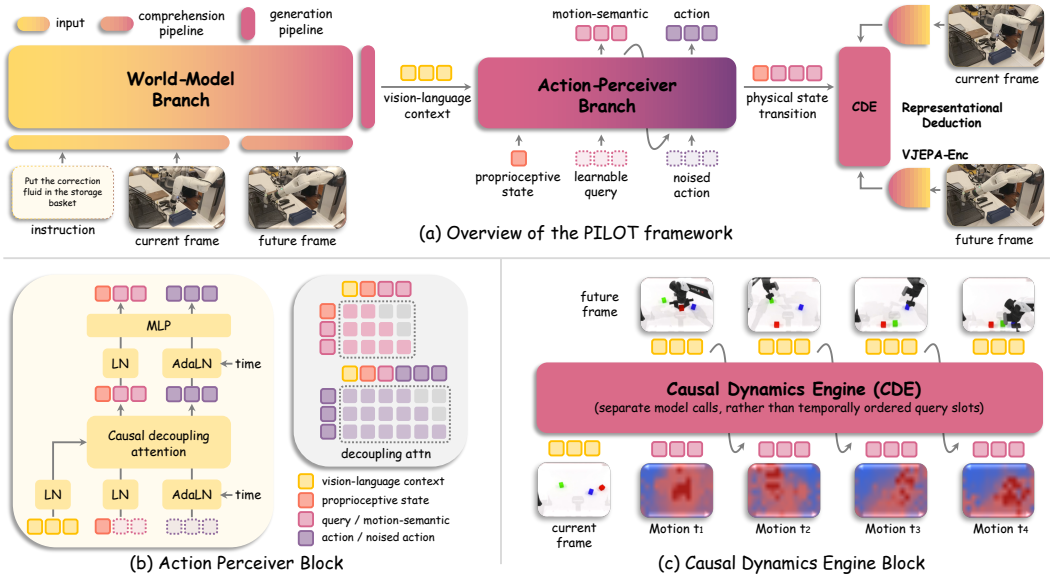


Figure 3: Overview of the **PILOT** framework. In the World-Model (a), current observations and instructions are encoded into a unified context. Subsequently, the Action-Perceiver employs learnable query tokens to distill motion-semantic context via a causally-decoupled attention mechanism (b), which guides the flow-matching action decoder. Simultaneously, we leverage a Causal Dynamics Engine (c) to supervise the motion semantics by learning physical state transition information. The labels Motion t_1-t_4 in (c) denote representations obtained from separate model calls, rather than temporally ordered query slots within a single forward pass.

3 METHOD

As illustrated in Fig. 3, PILOT combines a Cosmos-Predict2.5 world model (NVIDIA et al., 2025) for vision–language context and auxiliary future prediction, a Perceiver-style Action Model (Jaegle et al., 2021) for motion-semantic queries and flow-matching action decoding, and an RD branch for VJEPA2-AC future-representation supervision (Assran et al., 2025).

3.1 PROBLEM FORMULATION AND NOTATION.

At time t, the policy receives an RGB observation o_t, instruction ℓ, and proprioceptive state $s_t \in \mathbb{R}^{d_s}$, and predicts an H-step action chunk $\boldsymbol{a}_t = (a_t, \ldots, a_{t+H-1}) \in \mathbb{R}^{H \times d_a}$. Training additionally uses a future observation $o_{t+\Delta}$ from the same trajectory. With context $\boldsymbol{c} = \mathrm{WM}_\theta(o_t, \ell)$, policy is

$$p_\Theta(\boldsymbol{a}_t \mid o_t, \ell, s_t) = \int p_\Theta(\boldsymbol{a}_t \mid \boldsymbol{m}, \boldsymbol{c}, s_t)\, p_\Theta(\boldsymbol{m} \mid \boldsymbol{c}, s_t)\, \mathrm{d}\boldsymbol{m}, \tag{1}$$

where $\boldsymbol{m} = f_\Theta(\boldsymbol{c}, s_t) \in \mathbb{R}^{K \times d}$ is deterministic, so its conditional distribution is a point mass. The action decoder uses motion semantics, visual context, and proprioceptive state together.

3.2 MOTION-COT FROM LEARNABLE QUERIES

Vision-language context. In the comprehension pipeline, we repurpose Cosmos-Predict2.5 as a vision-language encoder. The video tokenizer encodes observation o_t into clean latents and the text encoder embeds instruction ℓ; a single conditional forward pass yields the context sequence

$$\boldsymbol{c} \;=\; \mathrm{WM}_\theta(o_t, \ell) \;\in\; \mathbb{R}^{N_c \times d_c}, \tag{2}$$

where WM_θ denotes the trainable world-model parameters, and N_c and d_c are the number and dimension of context tokens. The pre-trained video prior in WM_θ provides visual and physical cues for downstream action generation.

Learnable motion-semantic queries. We introduce K learnable queries $\boldsymbol{Q} \in \mathbb{R}^{K \times d_h}$ in a Perceiver-style flow-matching head, with initialization and internal width $d_h = 768$. The final slot projection produces Motion-CoT outputs of width $d = 1024$. Motion-CoT denotes these learned

semantic slots: their indices have no prescribed temporal ordering and are distinct from the H action steps. At a sampled flow time τ, the noised action chunk $\boldsymbol{a}_t^\tau = (1-\tau)\,\boldsymbol{\epsilon} + \tau\,\boldsymbol{a}_t$ is embedded into action tokens $\boldsymbol{A}^\tau \in \mathbb{R}^{H\times d_h}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The state s_t is mapped into a state token $\boldsymbol{S} \in \mathbb{R}^{1\times d_h}$, and the latent sequence $[\boldsymbol{S}; \boldsymbol{Q}; \boldsymbol{A}^\tau]$ attends to $\boldsymbol{c}$ through a stack of L blocks.

Causally-decoupled attention mechanism. To prevent motion semantics from absorbing diffusion noise, we decouple the attention inside each Perceiver block. Denoting the non-action latents as $\boldsymbol{Z} = [\boldsymbol{S}; \boldsymbol{Q}]$, each block applies

$$\boldsymbol{Z} \leftarrow \boldsymbol{Z} + \mathrm{Attn}\big(\boldsymbol{Z},\, [\boldsymbol{c}; \boldsymbol{Z}]\big), \tag{3}$$

$$\boldsymbol{A}^\tau \leftarrow \boldsymbol{A}^\tau + \mathrm{Attn}\big(\boldsymbol{A}^\tau,\, [\boldsymbol{c}; \boldsymbol{Z}; \boldsymbol{A}^\tau]\big), \tag{4}$$

where $\boldsymbol{Z}$ attends to itself and visual context, while action tokens also attend to $\boldsymbol{Z}$. Here, "causally-decoupled" describes the directed information flow: queries condition actions, while reverse attention from noised actions to queries is blocked. Time-conditioned AdaLN is applied only to action tokens, and $\boldsymbol{Z}$ uses plain layer normalization to remain invariant to the flow time τ. After the final block and the $d_h \to d$ output projection, we read out the query slots as the motion-semantic context:

$$\boldsymbol{O} = \mathrm{Act.Per.}(\boldsymbol{Z}, \boldsymbol{A}^\tau, \boldsymbol{c}; \tau) \;\in\; \mathbb{R}^{(1+K+H)\times d}, \qquad \boldsymbol{m} = \boldsymbol{O}[:,\, 1{:}K{+}1,\, :] \;\in\; \mathbb{R}^{K\times d}. \tag{5}$$

Flow-matching action decoding. Action tokens after the last block are projected back to the action space and supervised with the conditional flow-matching velocity objective:

$$\mathcal{L}_{\mathrm{act}} \;=\; \mathbb{E}_{\tau,\boldsymbol{\epsilon}} \big\| \mathrm{Proj}_{\mathrm{act}}\big(\boldsymbol{O}[:,\, K{+}1\,{:},\, :]\big) - (\boldsymbol{a}_t - \boldsymbol{\epsilon}) \big\|^2, \tag{6}$$

where $\mathrm{Proj}_{\mathrm{act}}$ is an MLP mapping action tokens to the action dimension d_a. At inference, Euler integration starts from Gaussian noise, conditioned on $\boldsymbol{m}$, $\boldsymbol{c}$, and s_t.

3.3 Representational Deduction with a CDE

RD trains $\boldsymbol{m}$ to support prediction of a future representation. The CDE models a directed current-to-future mapping conditioned on motion semantics and robot state.

Frozen state representations. Given the current and future observations $(o_t, o_{t+\Delta})$, frozen VJEPA2-AC encoder E_ψ produces per-frame patch representations $\boldsymbol{r}_t{=}E_\psi(o_t), \boldsymbol{r}_{t+\Delta}{=}E_\psi(o_{t+\Delta})$, where each representation lies in $\mathbb{R}^{P\times d_r}$, with P patch tokens and token dimension d_r. The image encoder provides the representation targets; action conditioning belongs to the pretrained predictor. RD uses the encoded future observation only as a training target, while queries are computed from current context and state.

Causal dynamics engine. The CDE G_ξ is initialized from the pre-trained VJEPA2-AC predictor and remains trainable. Instead of using the original action and state inputs of the predictor, we condition the CDE on learned motion semantics and robotic state. Two lightweight linear projectors first map the motion semantics and proprioceptive state to encoder feature dimensionality:

$$\tilde{\boldsymbol{m}} = \boldsymbol{W}_m\,\boldsymbol{m} \in \mathbb{R}^{K\times d_r}, \qquad \tilde{\boldsymbol{s}} = \boldsymbol{W}_s\,s_t \in \mathbb{R}^{1\times d_r}. \tag{7}$$

Inside the CDE, these conditioning tokens are further mapped from $d_r{=}1408$ to the predictor's internal width, $d_p{=}1024$, and concatenated with the projected current-frame patch features. The predictor outputs are then mapped back to $d_r{=}1408$ to obtain the future patch-feature grid:

$$\hat{\boldsymbol{r}}_{t+\Delta} \;=\; G_\xi\big(\boldsymbol{r}_t\,;\, \tilde{\boldsymbol{m}},\, \tilde{\boldsymbol{s}}\big) \;\in\; \mathbb{R}^{P\times d_r}. \tag{8}$$

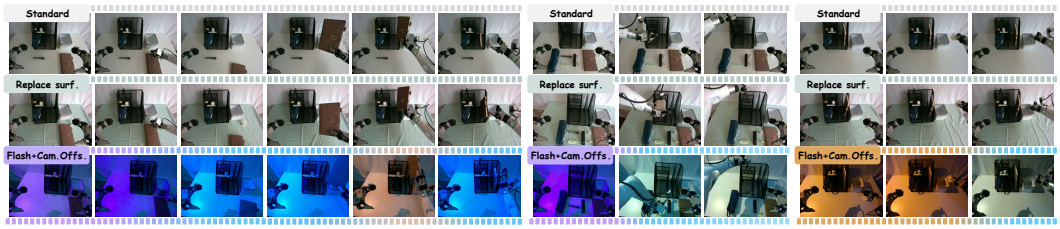
Conditioning tokens are prepended to current-frame patches, and one $P \times d_r$ future representation is returned. The K query slots jointly condition this prediction rather than represent successive future states.

representational deduction loss. We supervise the CDE by matching its prediction to the frozen encoder representation of the ground-truth future frame:

$$\mathcal{L}_{\mathrm{RD}} \;=\; \mathrm{SmoothL1}\big(\hat{\boldsymbol{r}}_{t+\Delta},\, \boldsymbol{r}_{t+\Delta}\big). \tag{9}$$

Gradients from $\mathcal{L}_{\mathrm{RD}}$ propagate through the CDE and the projector $\boldsymbol{W}_m$ to the learnable queries, encouraging $\boldsymbol{m}$ to encode transition information that is predictive of future VJEPA2-AC representations. This objective supplies transition-oriented representation supervision alongside the action loss and the auxiliary future-frame objective.

Figure 4: Real-world evaluation on Agibot-G1. Panels show temporal snapshots of (a) placing a book upright in a storage slot, (b) inserting a pen into a slot, and (c) placing trash in a box. Rows compare Standard, Replace surface (Desk), and Flash+Camera Offset conditions, illustrating execution under changed surfaces, lighting, and viewpoints. Quantitative success rates are reported in Table 4.



(a) Place the book vertically in the rightmost slot of storage. (b) Insert the pen into the top slot. (c) Put the trash in the box.

3.4 World-Model Branch and Training Objective

Future-frame latent prediction. In the generation pipeline, the world model retains its generative capability through an auxiliary future-frame objective. We encode the short clip $[\, o_t,\, o_{t+\Delta}\,]$ into clean video latents, using the current-frame latents as conditions and the future-frame latents as prediction targets. Flow-matching noise is applied only to the target latents, and the world model is trained to predict the corresponding target velocity:

$$\mathcal{L}_{\text{wm}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}} \big\| \, \text{WM}_\theta(\boldsymbol{z}^\tau;\, o_t, \ell, \tau) - (\boldsymbol{\epsilon} - \boldsymbol{z}^*) \, \big\|^2_{\odot \boldsymbol{M}_{\boldsymbol{z}^\tau}}, \tag{10}$$

where WM_θ denotes the trainable world-model parameters, $\boldsymbol{z}^*$ are the clean target latents, $\boldsymbol{z}^\tau$ is their noised version, and $\boldsymbol{M}_{\boldsymbol{z}^\tau}$ restricts the loss to future-frame positions. This objective preserves the instruction-grounded future prior of the pre-trained generator, while the generation pipeline is disabled during inference to avoid the visual bottleneck.

Overall objective. The three branches are trained jointly with a single weighted objective:

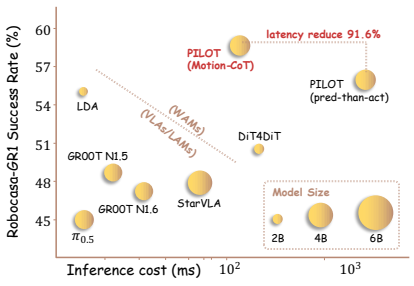
$$\mathcal{L} \;=\; \mathcal{L}_{\text{act}} \;+\; \lambda_{\text{wm}}\, \mathcal{L}_{\text{wm}} \;+\; \lambda_{\text{RD}}\, \mathcal{L}_{\text{RD}}, \tag{11}$$

where λ_{wm} and λ_{RD} balance the auxiliary future-frame and Representational Deduction losses against the action objective. In practice, the VJEPA2-AC encoder is frozen; the world model and CDE are fine-tuned from their pre-trained weights; and the Action Model together with the two projectors is trained from scratch.

Inference. At test time, the world model encodes (o_t, ℓ) into $\boldsymbol{c}$, the state-conditioned queries produce $\boldsymbol{m}$, and the flow-matching head generates the action chunk from $(\boldsymbol{m}, \boldsymbol{c}, s_t)$. The RD branch and future-frame prediction serve only as training-time supervision and incur no additional inference cost; the generative decoder can be optionally invoked for future-frame visualization.

4 Experiments

Figure 5: RoboCasa-GR1 success rate and inference latency comparison. Marker size denotes model size.



Implementation Details Our codebase is built upon the StarVLA project (Ye et al., 2026a). PILOT is implemented with three cooperative branches. The **World-Model branch** builds on a pre-trained Cosmos-Predict2.5 video diffusion transformer, which simultaneously serves as a visual-language encoder (providing a 196×2048 context) and a trainable latent future-frame generator. The **Action-Model branch** uses $K = 64$ learnable query tokens with initialization and internal width $d_h = 768$, projected to Motion-CoT outputs of width $d = 1024$, and a DiT-based flow-matching action decoder with 20 denoising steps. The **Representational Deduction branch** employs a frozen VJEPA2-AC encoder that produces current and future patch-level representations (256×1408), and a lightweight Causal Dynamics Engine that predicts future representations conditioned on motion semantics and proprioceptive state.

4.1 SIMULATION BENCHMARKS

LIBERO. LIBERO Liu et al. (2023) is a tabletop manipulation benchmark with a Franka Panda robot, featuring four task suites of increasing complexity: Spatial, Object, Goal, and Long-horizon.

As shown in Table 1, PILOT achieves **97.9%** average success rate with approximately 2.5B trainable parameters. This success rate surpasses all baselines, including Motus (97.7%) and $\pi_{0.5}$ (96.9%). Notably, PILOT excels on the challenging Goal (98.1%) and Long-horizon (97.2%) suites, where understanding the *intent* of manipulation is critical for long-horizon planning. This validates our central thesis: decoupling motion semantics from trajectory generation via Representational Deduction effectively alleviates representational entanglement.

Table 1: Evaluation on LIBERO (Liu et al., 2023). Success rates (%) on Spatial, Object, Goal, and Long-horizon suites. Size denotes parameter count (B: billions); † denotes trainable parameters for PILOT, with its parameter breakdown in Appendix B.2. Avg. averages the four suite scores. Baselines are published results. VLA/LAM/WAM group vision-language-, latent-, and world-action models.

Group	Method	Size	Spatial	Object	Goal	Long	**Avg.**
VLA	$\pi_{0.5}$	3.5B	98.8	98.2	98.0	92.4	96.9
	GR00T-N1.6	3.3B	97.5	98.5	97.5	94.4	97.0
	OpenVLA-OFT	7B	97.6	98.4	97.9	94.5	97.1
LAM	LAPA*	7B	73.8	74.6	58.8	55.4	65.7
	UniVLA	7B	96.5	96.8	95.6	92.0	95.2
	Mantis	5.8B	98.8	99.2	94.4	94.2	96.7
	VLA-JEPA	3B	96.2	99.6	97.2	95.8	97.2
WAM	F1	4B	98.2	97.8	95.4	91.3	95.7
	Fast-WAM	6B	98.2	100.0	97.0	95.2	97.6
	Motus	8B	96.8	99.8	96.6	97.6	97.7
	PILOT (Ours)	**2.5B**†	97.1	99.2	98.1	97.2	97.9

LIBERO-Plus. As reported in Table 2, PILOT achieves **80.3%** condition-count-weighted success rate, exceeding PokeVLA (79.3%) by 1.0 percentage point in the displayed comparison. The improvement is particularly pronounced on Robot (70.0% vs. 46.1%) and Layout (82.0% vs. 77.2%), suggesting that the motion-semantic tokens learned by Representational Deduction encode *transition dynamics* rather than superficial appearance, making the policy inherently more robust to structural perturbations.

Table 2: Evaluation on LIBERO-Plus (Fei et al., 2025). Success rates (%) under Camera, Robot, Language, Light, Background, Noise, and Layout perturbations. Policies trained on standard LIBERO are evaluated by zero-shot transfer, testing robustness beyond the training distribution. VLA/LAM/WAM denote vision-language-, latent-, and world-action models; Size is parameter count, with † denoting trainable parameters for PILOT (Appendix B.2), and Avg. the condition-count-weighted mean recomputed from the displayed category scores. The appendix specifies the aggregation weights.

Group	Method	Size	Cam.	Rob.	Lang.	Light	Back.	Noise	Layout	**Avg.**
VLA	Spatial Forcing	7B	20.1	13.4	40.9	29.1	33.4	25.7	39.3	28.5
	π_0-FAST	3.3B	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
	OpenVLA-OFT	7B	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
	PokeVLA	1.22B	84.7	46.1	84.8	94.6	82.6	89.8	77.2	79.3
LAM	UniVLA	7B	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
WAM	WorldVLA	7B	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
	Fast-WAM	6B	16.4	44.5	68.9	78.2	53.7	37.7	60.7	50.0
	PILOT (Ours)	**2.5B**†	76.0	70.0	84.0	92.0	85.0	78.0	82.0	80.3

Table 3: RoboCasa-GR1: Category success rates (%). Comparison of 10 methods across 24 tasks; parentheses give task counts. PILOT uses 50 rollouts/task; all averages are recomputed from the full task table, with categories weighted by task count. PnP denotes pick-and-place; other groups denote novel source locations. Best and second-best values are shaded per row; ties share a shade. D4D: DiT4DiT; UWM: UWM-XL; QG: QwenGR00T; VP: VP-VLA; Phys.: PhysBrain-4B; Lang.: LangForce. Full task results are in Appendix B.14.

Category	GR00T N1.5	GR00T N1.6	D4D	UWM	QG	LDA	VP	Phys.	Lang.	**PILOT**
PnP Close (6)	45.3	24.2	50.3	30.7	50.3	56.3	54.3	52.7	55.7	57.0
Cuttingboard (5)	46.4	56.9	57.6	17.4	52.8	64.2	60.8	52.0	53.2	62.4
Placemat (4)	45.5	51.9	39.0	10.0	38.0	47.8	54.5	48.0	48.0	48.5
Tray (5)	48.8	55.1	46.4	17.2	39.2	53.4	46.0	46.0	47.2	64.8
Plate (4)	56.5	57.6	60.0	16.3	58.5	53.0	53.5	62.5	58.5	57.0
Average	48.2	47.6	50.8	19.3	47.8	55.4	53.8	52.0	52.6	58.3

RoboCasa-GR1. As shown in Table 3, PILOT achieves **58.3%** average success over 24 tasks with 50 rollouts each, exceeding the strongest baseline in the displayed comparison, LDA (55.4%), by 2.9 points. Dark gray marks its best overall score and its leading PnP Close (57.0%) and Tray (64.8%) results; light gray marks its second-best Cuttingboard score (62.4%). The largest category gain is on Tray (64.8% vs. 55.1%), an improvement of 9.7 points. Within Tray, success reaches 92% for Pot, 76% for Cardboardbox, and 72% for Plate, covering multiple target receptacles. Placemat and Plate average 48.5% and 57.0%, respectively. Cuttingboard-to-Pan and Cuttingboard-to-Pot reach 86% and 78%, while Placemat-to-Basket reaches 78%. The full table in Appendix B.14 marks PILOT best on 10 individual tasks and second-best on one. These task-level comparisons complement the category averages; the overall score weights each of the 24 tasks equally.

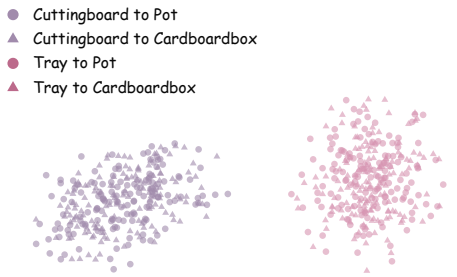


Figure 6: RoboCasa-GR1 representations. T-SNE visualization of Motion-CoT token embeddings across four RoboCasa-GR1 pick-and-place tasks with different source locations and target containers. Marker colors indicate source (Cuttingboard/Tray); circles/triangles indicate target (Pot/Cardboardbox).

4.2 Real-World Evaluation

Standard Tasks. Each of the eight standard desk tasks is evaluated over 50 rollouts. Table 4 compares PILOT with π_0, GR00T-N1.7, and $\pi_{0.5}$. PILOT leads on six tasks: Pen, Eraser, Charger, Pencil Case, Stapler, and Trash. Gains over $\pi_{0.5}$ reach 32 points on Eraser, 18 on Stapler and Trash, 16 on Pencil Case, and 12 on Charger. Compared with π_0, PILOT improves all eight standard tasks by 16–50 points, with the largest gain on Eraser. Pen reaches 90%, providing the full-data reference below.

Generalization Transfer. All three experiments use the **Pen task**, fixing the manipulation objective while changing visual conditions. PILOT reaches 72% under Flash+Camera Offset and 68% on the replaced Desk, each 10 points above $\pi_{0.5}$. On Color, it reaches 64%, versus 62% for π_0. These scores compare directly with the clean Pen reference. Relative to that reference, PILOT retains 80.0% under F+C and 75.6% under Desk, compared with 70.5% and 65.9% for $\pi_{0.5}$, respectively. This task-matched comparison measures transfer across changes in illumination, viewpoint, and tabletop appearance.

Few-Shot Fine-Tuning. This experiment also uses the **Pen task** on Agibot-G1, with 10% of its demonstrations. PILOT achieves **78%**, versus 68% for π_0, 66% for $\pi_{0.5}$, and 30% for GR00T-N1.7, exceeding the strongest baseline by 10 points. It retains 86.7% of its full-data Pen performance (78% vs. 90%); $\pi_{0.5}$ retains 75.0% (66% vs. 88%). Their gap increases from 2 to 12 points, supporting the value of motion-semantic representations with fewer demonstrations.

Table 4: Real-world evaluation on Agibot-G1. Success rates (%) for eight standard desk tasks. All three generalization experiments—F+C (strobe lighting with camera offset), Desk (replaced surface), and Color (changed object colors)—and the FS (10%) few-shot experiment are conducted on the **Pen task**. FS uses 10% of the Pen demonstrations; Avg. is the arithmetic mean within each setting. Era.: eraser; CF: correction fluid; Chg.: charger; PC: pencil case; Sta.: stapler; Tr.: trash. Best results in each column are **bold**.

Method	Standard Tasks								Avg.	Generalization			Avg.	FS (10%)
	Pen	Era.	CF	Chg.	PC	Sta.	Book	Tr.		F+C	Desk	Color		
π_0	74	46	42	66	28	64	52	74	55.8	46	54	62	54.0	68
GR00T-N1.7	38	52	30	64	24	74	80	82	55.5	18	24	28	23.3	30
$\pi_{0.5}$	88	64	**86**	82	40	70	**84**	76	73.8	62	58	**70**	63.3	66
PILOT (Ours)	**90**	**96**	70	**94**	**56**	**88**	72	**94**	**82.5**	**72**	**68**	64	**68.0**	**78**

4.3 Ablation Studies

Table 5 shows the results of progressively adding components. Starting from a bare backbone without future prediction or Motion-CoT adding future-frame prediction improves LIBERO by 2.3 points, indicating that the auxiliary future-frame objective provides useful visual grounding. Replacing future-frame prediction with Motion-CoT alone yields a gain (0.9 point), indicating that distilled semantic tokens are a more effective information bottleneck than pixel-level generation. Adding the Representational Deduction branch with Causal Dynamics Engine supervision further boosts benchmarks (1.6 points), validating that explicit dynamics prediction regularizes motion-semantic tokens toward physically meaningful transitions. Enabling causally-decoupled attention further improves LIBERO success by 0.8 percentage points, supporting the benefit of isolating motion-semantic queries from noised action tokens while allowing action tokens to attend to both the queries and visual context. Finally, combining future-frame prediction with the full Motion-CoT pipeline yields the best results, demonstrating that the two complementary objectives-pixel-level visual grounding and semantic-level transition prediction-jointly achieve optimal performance.

Figure 7: PCA RGB pseudo-color visualization of Causal Dynamics Engine predicted future representations.

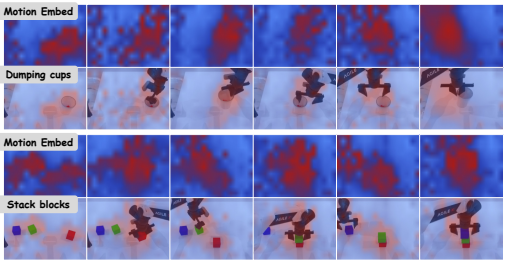


Table 5: LIBERO component ablation (success %). Fut. Pred: future-frame prediction; MotionCoT: queries; RD: Representational Deduction; CDE: Causal Dynamics Engine; Decoup.: decoupled attention. Checks/crosses: enabled/disabled; dashes: not applicable.

Fut. Pred	MotionCoT	RD	CDE	Decoup.	LIBERO
×	×	–	–	–	91.3
✓	×	–	–	–	93.6
×	✓	×	×	×	94.5
×	✓	✓	✓	×	96.1
×	✓	✓	✓	✓	96.9
✓	✓	✓	✓	✓	**97.9**

4.4 Representational Analysis

CDE Predictive Representation Quality The Causal Dynamics Engine predicts future VJEPA2-AC representations conditioned on current state and Motion-CoT tokens. In Figure 6, we compare four RoboCasa-GR1 pick-and-place tasks with different source locations and target containers. The visualization suggests that the CDE can capture patterns in the evolution of future latent representations.

PCA RGB pseudo-color visualization. We apply PCA to reduce the 1408-dimensional predicted tokens to RGB channels for a 16×16 pseudo-color image. Figure 7 shows that the CDE predictions produce coherent spatial structures that closely match the ground-truth future PCA colors, while differing from the current state. The spatial coherence of the predicted pseudo-color images confirms that the CDE does not merely predict token-level averages but captures structured, region-aware state transitions.

5 Conclusion and Limitation

Conclusion. We present PILOT, which combines future-representation supervision of Motion-CoT with noise-isolated action decoding. Current-context queries guide trajectory refinement alongside visual features, while auxiliary prediction branches are omitted at inference. The evaluations examine manipulation success, robustness under selected shifts, and data-efficient adaptation on Agibot-G1. These results support transition-oriented representation learning as a practical complement to action supervision.

Limitation. The current implementation uses a fixed number of motion queries and a fixed future-prediction interval. Adapting these choices to task duration may improve flexibility. Moreover, our visualizations are qualitative and do not assign a human-readable meaning to individual query slots. Developing lightweight diagnostics for these representations is a useful direction for future work.

AI USE STATEMENT

Generative AI assisted manuscript preparation, including LaTeX formatting, layout adaptation, citation and consistency checks, prose revision, and removal of repeated exposition.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are specified in Section 3. The appendix provides training and inference pseudocode, architecture and optimization hyperparameters, evaluation protocols, dataset descriptions, and complete task-level RoboCasa-GR1 results (Section B.14).

REFERENCES

Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezhen Song. Geoaware-vla: Implicit geometry aware vision-language-action model. *arXiv:2509.14117*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *CVPR*, pp. 35101–35113, 2026.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv:2212.06817*, 2022.

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. *arXiv preprint arXiv:2307.15818*, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv:2505.06111*, 2025.

Jun Cen, Siteng Huang, Yuqian Yuan, Kehan Li, Hangjie Yuan, Chaohui Yu, Yuming Jiang, Jiayan Guo, Xin Li, Hao Luo, et al. Rynnvla-002: A unified vision-language-action and world model. *arXiv:2511.17502*, 2025a.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv:2506.21539*, 2025b.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, et al. Lawam: Latent world action models for efficient dynamics-aware robot policies. *arXiv:2606.15768*, 2026.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19752–19763, 2025.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv:2510.13626*, 2025.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv:2507.12898*, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv:2602.06949*, 2026.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv:2301.04104*, 2023.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi_0.5: a vision-language-action model with open-world generalization. *arXiv:2504.16054*, 2025.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, pp. 4651–4664, 2021.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv:2406.09246*, 2024.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv:2502.19645*, 2025.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, et al. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv:2601.16163*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv:2601.21998*, 2026a.

Shalfun Li, Victor Yao, Charles Yang, Truth Qu, Regis Cheng, Ryan Yu, Howard Lu, Newton Von, Vincent Chen, Yohann Tang, et al. Wall-wm: Carving world action modeling at the event joints. *arXiv:2606.01955*, 2026b.

Wei Li, Renshan Zhang, Rui Shao, Jie He, and Liqiang Nie. Cogvla: Cognition-aligned vision-language-action models via instruction-driven routing & sparsification. *Advances in neural information processing systems*, 38:137646–137675, 2026c.

Shijie Lian, Bin Yu, Xiaopeng Lin, Laurence Tianruo Yang, Zhaolong Shen, Changti Wu, YuZhuo Miao, Cong Huang, Kai Chen, et al. Langforce: Bayesian decomposition of vision language action models via latent action queries. In *ICML*, 2026.

Xiaopeng Lin, Shijie Lian, Bin Yu, Ruoqi Yang, Zhaolong Shen, Changti Wu, Yuzhuo Miao, Yurun Jin, Yukun Shi, Jiyan He, et al. Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence. *arXiv:2512.16793*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*, 36:44776–44791, 2023.

Mengya Liu, Baoxiong Jia, Jiangyong Huang, Jingze Zhang, and Siyuan Huang. Lara: Latent action representation alignment for vision-language-action models. *arXiv:2606.07100*, 2026.

Songming Liu, Lingxuan Wu, Bangguo Li, et al. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *ICLR*, volume 2025, pp. 29982–30009, 2025.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable vla pretraining in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35047–35058, 2026a.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-h0. 7: A latent world-action model from egocentric videos. *arXiv:2605.00078*, 2026b.

Qi Lv, Weijie Kong, Hao Li, et al. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv:2509.06951*, 2025.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv:2602.12215*, 2026.

Teli Ma, Jia Zheng, Zifan Wang, et al. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv:2603.10448*, 2026a.

Xiangkai Ma, Lekai Xing, Han Zhang, Wenzhong Li, and Sanglu Lu. Unifying perception and action: A hybrid-modality pipeline with implicit visual chain-of-thought for robotic action generation. In *CVPR*, pp. 22380–22390, 2026b.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv:2603.19312*, 2026.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv:2603.14482*, 2026.

Nilaksh, Saurav Jha, Artem Zholus, and Sarath Chandar. Reconstruction or semantics? what makes a latent space useful for robotic world models. *arXiv:2605.06388*, 2026.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, et al. World simulation with video foundation models for physical ai. *arXiv:2511.00062*, 2025.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv:2512.15692*, 2025.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, et al. Fast: Efficient action tokenization for vision-language-action models. *arXiv:2501.09747*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv:2501.15830*, 2025.

Sandeep Routray, Hengkai Pan, Unnat Jain, Shikhar Bahl, and Deepak Pathak. Vipra: Video prediction for robot actions. *arXiv:2511.07732*, 2025.

Yue Su, Sijin Chen, Haixin Shi, Mingyu Liu, Zhengshen Zhang, Ningyuan Huang, Weiheng Zhong, Zhengbang Zhu, Yuxiao Liu, and Xihui Liu. World guidance: World modeling in condition space for action generation. *arXiv:2602.22010*, 2026.

Guoheng Sun, Tingting Du, Kaixi Feng, Chenxiang Luo, Xingguo Ding, Zheyu Shen, Ziyao Wang, Yexiao He, and Ang Li. Rocket: Residual-oriented multi-layer alignment for spatially-aware vision-language-action models. *arXiv:2602.17951*, 2026a.

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:2602.10098*, 2026b.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, James Li, Jiabao Liu, Jingrui Pang, Kiro Jing, et al. Motubrain: An advanced world action model for robot control. *arXiv:2604.27792*, 2026.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv:2302.00482*, 2023.

Zixuan Wang, Yuxin Chen, Yuqi Liu, et al. Vp-vla: Visual prompting as an interface for vision-language-action models. *arXiv:2603.22003*, 2026.

Sheng Xu, Ruixing Jin, Huayi Zhou, Bo Yue, Guanren Qiao, Yunxin Tai, Yueci Deng, Kui Jia, and Guiliang Liu. From reaction to anticipation: Proactive failure recovery through agentic task graph for robotic manipulation. *arXiv:2605.11951*, 2026.

Fuxiang Yang, Donglin Di, Lulu Tang, Xuancheng Zhang, Lei Fan, Hao Li, Wei Chen, Tonghua Su, and Baorui Ma. Chain of world: World model thinking in latent motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6675–6684, 2026a.

Jiange Yang, Yansong Shi, Haoyi Zhu, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. Como: Learning continuous latent motion from internet videos for scalable robot learning. In *CVPR*, pp. 42352–42363, 2026b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, pp. 10632–10643, 2025.

Yi Yang, Xueqi Li, Yiyang Chen, Jin Song, Yihan Wang, Zipeng Xiao, Jiadi Su, You Qiaoben, Pengfei Liu, and Zhijie Deng. Mantis: A versatile vision-language-action model with disentangled visual foresight. In *CVPR*, pp. 42505–42515, 2026c.

Jinhui Ye, Ning Gao, Senqiao Yang, Jinliang Zheng, Zixuan Wang, Yuxin Chen, Pengguang Chen, Yilun Chen, Shu Liu, and Jiaya Jia. Starvla-α: Reducing complexity in vision-language-action systems. *arXiv preprint arXiv:2604.11757*, 2026a.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *ICLR*, volume 2025, pp. 28213–28239, 2025.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv:2602.15922*, 2026b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv:2603.16666*, 2026.

Chubin Zhang, Jianan Wang, Zifeng Gao, Yue Su, Tianru Dai, Cai Zhou, Jiwen Lu, and Yansong Tang. Clap: Contrastive latent action pretraining for learning vision-language-action models from human videos. *arXiv:2601.04061*, 2026a.

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *NeurIPS*, 38:24195–24228, 2026b.

Wenyao Zhang, Bozhou Zhang, Zekun Qi, Wenjun Zeng, Xin Jin, and Li Zhang. Disentangled robot learning via separate forward and inverse dynamics pretraining. *arXiv preprint arXiv:2604.16391*, 2026c.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, et al. Imagewam: Do world action models really need video generation, or just image editing? *arXiv:2606.19531*, 2026d.

Yupeng Zheng, Xiang Li, Songen Gu, Yuhang Zheng, Shuai Tian, Weize Li, Linbo Wang, Senyu Fei, Pengfei Li, Yinfeng Gao, et al. Pokevla: Empowering pocket-sized vision-language-action model with comprehensive world knowledge guidance. *arXiv:2604.20834*, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv:2404.12377*, 2024.

Chuning Zhu, Raymond Yu, Siyuan Feng, et al. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv:2504.02792*, 2025.