FROM PREDICTIVE GEOMETRY TO ACTION: LEARNING COMPLEMENTARY PREDICTIVE REPRESENTATIONS FOR EMBODIED MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective manipulation requires a policy to capture not only about the current scene geometry, but also about how that geometry will evolve under interaction. Existing approaches either augment policies with geometric information from current observations or introduce future prediction as auxiliary supervision, leaving open how predictive geometry should be represented for action generation. Therefore, we propose a two-stage framework, which learns complementary predictive representations from multi-view video clips. During pretraining, a frozen 3D foundation model provides dense geometric features that are compressed into predictive geometric representations and optimized to preserve scene structure and its temporal evolution. Action supervision is then applied to geometric transitions to extract a complementary motor representation that emphasizes action-relevant transitions. During downstream post-training, predictive geometry is transferred through geometry queries, while the motor representation guides intermediate features along the action-generation pathway. Future observations are used only as privileged supervision during training and are not required at inference time. Across LIBERO, LIBERO-Plus, SimplerEnv, and real-world bimanual and single-arm tasks, our approach consistently improves downstream policies, including an absolute gain of 10.4 percentage points on the WidowX evaluation. We further demonstrate transfer across multiple VLA architectures and WAMs, with improvements in both future-world and action prediction.

1 INTRODUCTION

Anticipating how the physical world will evolve(Assran et al., 2025; 2023) under interaction is becoming increasingly important for vision-language-action (VLA) models(Ma et al., 2026; Luo et al., 2026) and world action models (WAM)(Li et al., 2026b; Ye et al., 2026). Such predictive capability allows an agent to reason about the consequences of interaction, rather than mapping only its current observation to an immediate action. Yet effective manipulation requires more than temporal foresight alone: it also depends on a precise understanding of how the scene is spatially organized. Geometry specifies object layouts, spatial relationships, and robot-scene configurations, while dynamics describe how these structures evolve as an interaction unfolds. For action generation, these two aspects play distinct but complementary roles: geometry provides the spatial structure within which actions are executed, while its evolution provides cues about how the scene may change over time. This naturally raises a central question: how should geometric structure and its predicted evolution be represented so that they can more effectively support action generation?

Recent efforts to incorporate geometry into predictive policy learning follow two directions. As shown in Fig.1, input-centric approaches(Sun et al., 2025; Qu et al., 2025; Li et al., 2026c; Zhang et al., 2025b) augment observations with depth, point clouds, 3D coordinates, or pretrained geometric features. These methods improve spatial grounding, but use geometry as additional context for describing the currently observed scene and may rely on sensing or geometry-estimation pipelines. Prediction-centric approaches(Yang et al., 2026; He et al., 2026; Li et al., 2025a;b; Su et al., 2026) instead supervise models with future RGB-D observations, future geometric states, or other world-prediction targets, explicitly exposing the policy to future scene structure. However, such targets are represented as dense observations or holistic latent states, requiring the learned representation to

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.

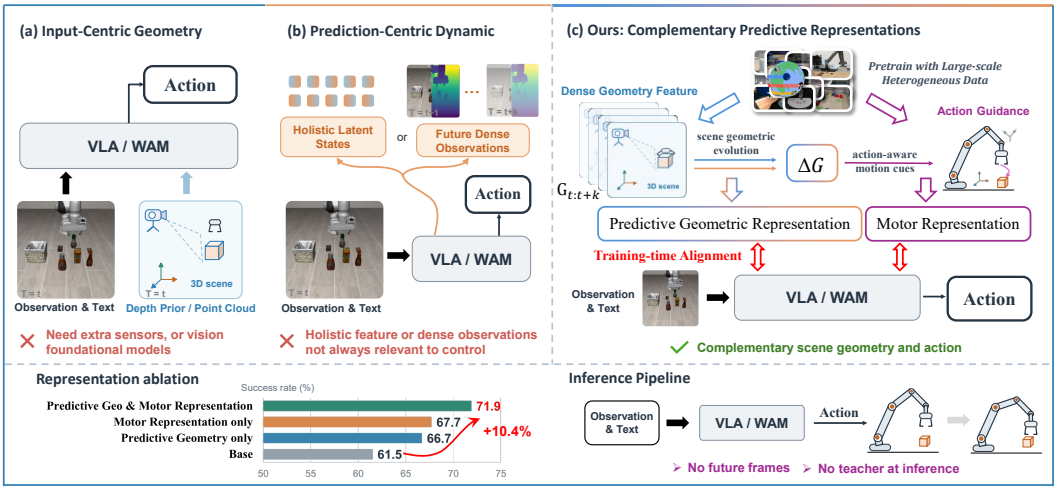


Figure 1: Comparison of geometry-guidance paradigms for policy learning. (a) Input-centric methods augment policies with spatial information derived from current observations. (b) Prediction-centric methods use future-world or predictive targets as additional supervision. (c) Our approach learns a predictive geometric representation from dense geometric features and derives a complementary motor representation from the evolution of scene geometry. The two representations are transferred through role-specific guidance pathways during training and are not required at inference time.

retain geometric information whose relevance to control may vary substantially. More broadly, these approaches primarily organize geometry to represent the scene and its evolution. They leave less explicit how action-relevant information should be distinguished from geometry needed for a coherent world representation. This creates a gap between predictive geometric modeling and action-oriented motor representation learning.

A natural way to bridge this gap is to recognize that predictive geometry and action generation place different preferences on learned representations. Predictive geometric modeling benefits from preserving a sufficiently complete and coherent account of scene structure. Action generation, in contrast, benefits from emphasizing geometric variations that are informative for motor commands while reducing the influence of static or action-irrelevant scene content. Rather than forcing these roles to be served in exactly the same representation, we introduce two complementary representations with distinct but related purposes: a *predictive geometric representation* that provides a compact and structured description of scene geometry and its evolution, and a *motor representation* that emphasizes action-aware information from geometric transitions. The former provides structured predictive context, while the latter focuses this predictive information toward control.

Building on this observation, we propose the two-stage framework, which learns complementary predictive representations (CPR) from large-scale heterogeneous interaction videos(Collaboration, 2023; Jin et al., 2026; Liu et al., 2024; Goyal et al., 2017; Chen et al., 2025) and transfers them to downstream policies. As illustrated in Fig. 1, our framework first constructs a predictive geometric representation from dense geometric features and then extracts action-aware information from the temporal evolution of scene geometry across video clips, providing complementary guidance for downstream action generation. Specifically, as shown in Fig.2, during representation pretraining, a frozen 3D foundation model (Wang et al., 2025) extracts dense geometric features from multi-view video observations, which are progressively compressed by a hierarchical query-based encoder into compact temporal states. Geometry reconstruction, masked representation prediction, and geometric transition prediction encourage the predictive geometric representation to preserve scene structure while capturing its evolution over time. We then apply action supervision to transitions between consecutive predictive geometric states, yielding a motor representation that emphasizes action-relevant geometric transitions. The predictive geometric representation therefore captures structured information about the evolving scene, while the motor representation focuses on transitions that are more directly informative for action. During post-training, as shown in Fig.3, these two representations are transferred through role-specific supervision pathways. Dedicated geometry queries are trained to infer the predictive geometric representation from current-time inputs, while intermediate features along the action-generation pathway are aligned with the motor representation. Importantly, future

observations and the pretrained representation model are used only to construct privileged supervision during training and are not required at inference time. We evaluate CPR on both multiple VLA architectures and WAM across diverse simulated and real-world manipulation benchmarks, demonstrating consistent improvements across different robot embodiments and action-generation paradigms. Our contributions are summarized as follows:

- We identify complementary roles for predictive geometric and motor representations in predictive policy learning: predictive geometric modeling benefits from preserving structured scene geometry, while action generation benefits from emphasizing control-relevant geometric transitions.
- We develop a two-stage framework that learns and transfers these complementary predictive representations. Hierarchical geometric tokenization and predictive objectives learn geometric states, while action supervision extracts action-aware information from their transitions; the two representations are transferred through geometry-query and motor-alignment pathways.
- Extensive experiments across LIBERO, LIBERO-Plus, SimplerEnv, and real-world bimanual and single-arm manipulation tasks demonstrate consistent improvements across policy architectures and robot embodiments. CPR yields a 10.4% absolute gain on the WidowX evaluation and improves future-world and action prediction in WAM.

2 Related Work

Predictive Representations for VLA and WAM. Predictive modeling provides VLA models and WAMs with foresight beyond the current observation. Existing approaches introduce future information either through observation prediction or compact predictive representations. Observation-level methods predict future RGB observations, videos, or world states to support action generation (Cen et al., 2025; Zhang et al., 2025a; Lv et al., 2025; Yuan et al., 2026; Agarwal et al., 2026), while latent-space approaches compress future information into compact conditions or visuomotor representations (Xu et al., 2026; Su et al., 2026; Sun et al., 2026). Recent methods (Lee et al., 2026b; Guo et al., 2026; Ma et al., 2026; Yan et al., 2026; Lee et al., 2026a; Wang et al., 2026b), also explore structured predictive signals such as dynamic, spatial, or semantic information to connect world modeling with control. However, existing predictive representations vary in how geometric structure and action-relevant temporal transitions are organized. Our work focuses specifically on predictive geometry: we learn a structured representation that preserves scene geometry and its evolution, and derive a complementary motor representation from geometric transitions.

Geometric Learning for Robotic Manipulation. Geometric information has been increasingly incorporated into robotic policies to improve spatial grounding. Input-centric approaches (Qu et al., 2025; Sun et al., 2025; Bhat et al., 2025; Zhang et al., 2025b) augment current observations with depth, point clouds, 3D coordinates, or spatial features, providing direct access to the geometry of the observed scene. More recent prediction-centric approaches (Li et al., 2025b; He et al., 2026; Yang et al., 2026; Su et al., 2026; Li et al., 2026a) instead use geometric or spatiotemporal prediction as auxiliary supervision, including depth prediction, future robot trajectories, and 3D motion representations. These approaches demonstrate the value of geometry for predictive policy learning, but geometric structure and action relevance may favor different types of information: preserving scene geometry requires broad structural fidelity, whereas action generation benefits from transitions that are informative for motor behavior. Our work addresses this distinction by learning complementary predictive geometric and motor representations, and transferring them separately to downstream VLA and WAM policies through dedicated interfaces.

3 Method

We present a two-stage framework for learning complementary predictive representations and transferring them to downstream policies. As illustrated in Fig. 2, during *(i) Complementary Predictive Representation Pretraining*, continuous multi-view video clips are first converted into compact predictive geometric representations that preserve the spatial structure of the scene while capturing its temporal evolution. We then use the transitions between these predictive geometric representations, together with temporally aligned action supervision, to extract action-aware information into a complementary motor representation. During *(ii) Representation-Guided Policy Post-training*, as shown

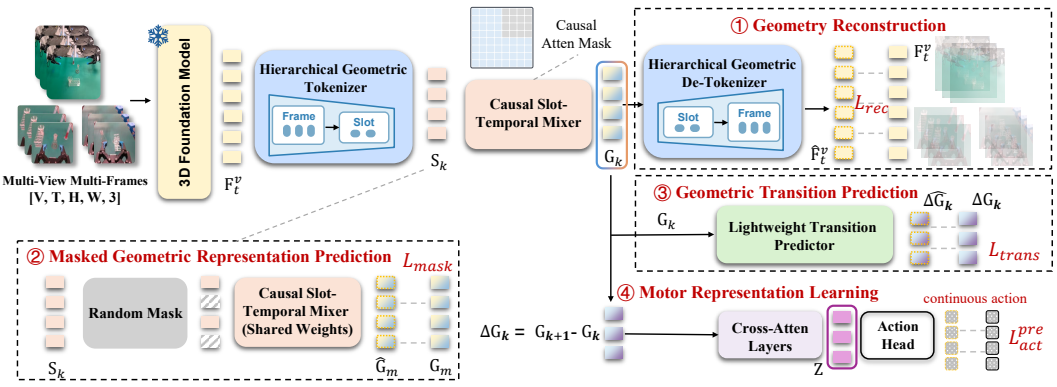


Figure 2: Overview of complementary predictive representation pretraining. A frozen 3D foundation model extracts dense geometric features from multi-view video observations, which are compressed by the Hierarchical Geometric Tokenizer and temporally modeled by a causal Slot-Temporal Mixer to produce the predictive geometric representation. Geometry reconstruction, masked geometric representation prediction, and geometric transition prediction jointly shape this representation to preserve scene structure and capture its temporal evolution. The resulting geometric transitions are used with action supervision to extract an action-aware motor representation, providing complementary predictive and motor information for policy learning.

in Fig.3, the predictive geometric representations and motor representations are transferred to the policy through two dedicated supervision pathways. Geometry queries are trained to infer structured predictive geometric information from current-time inputs, while intermediate features along the action-generation pathway are aligned with the motor representation.

3.1 Complementary Predictive Representation Pretraining

Hierarchical Geometric Tokenization. As shown in Fig.2, we first transform dense multi-view observations into a compact and temporally organized sequence of geometric states. A frozen pretrained 3D foundation model (Wang et al., 2025) is applied independently to each frame-view pair O_t^v, where $t \in T$ and $v \in V$ denote the number of frames and camera views, respectively, producing dense geometric features $\mathbf{F}_t^v$. These features provide rich spatial information, but directly modeling all geometric tokens across views and time incurs substantial computational overhead. We therefore introduce a hierarchical query-based encoder that progressively compresses the representation along the view and temporal dimensions.

At each time step t, geometric features from all camera views are concatenated along the token dimension and augmented with learnable view embeddings to preserve view-specific information. A set of learnable geometric queries then cross-attends to the resulting multi-view tokens, producing a compact frame-level geometric state $\mathbf{H}_t$. We subsequently aggregate the frame-level states along the temporal dimension. Inspired by (Wan et al., 2025), the first state $\mathbf{H}_0$ is retained without temporal aggregation and serves as a temporal anchor, while each complete group of four subsequent frame-level states is compressed into one temporal slot. This produces K temporal slots, where $K = 1 + \lfloor \frac{T-1}{4} \rfloor$. Temporal positional embeddings are introduced before temporal aggregation to preserve the ordering of the underlying frames. We denote the resulting tokenized geometric sequence as $\mathbf{S} = \{\mathbf{S}_k\}_{k=1}^K$. This hierarchical design substantially reduces the spatiotemporal token budget while maintaining an explicitly organized representation of scene geometry over time.

Predictive Geometric Representation Learning. Although the tokenized geometric sequence $\mathbf{S}$ provides compact descriptions of individual temporal segments, each state alone does not explicitly encode how the interaction evolves over time. As shown in Fig.2, we therefore process the sequence using a causal Slot-Temporal Mixer, implemented as a stack of causal attention blocks. At temporal slot k, the representation attends only to states at or before k, such that each output summarizes the geometric history available up to that point. The resulting sequence forms the predictive geometric representations $\mathbf{G} = \{\mathbf{G}_k\}_{k=1}^K$.

We shape $\mathbf{G}$ with three complementary objectives, each emphasizing a different property of predictive geometric information. Geometry reconstruction preserves the spatial information inherited

from the frozen 3D foundation model; masked geometric representation prediction encourages temporal consistency and contextual inference across geometric states; and geometric transition prediction explicitly models local transitions between consecutive states. Together, these objectives encourage $\mathbf{G}$ to remain structurally informative while capturing predictable patterns in the evolution of scene geometry.

1) Geometry Reconstruction. Hierarchical tokenization substantially compresses the dense geometric features and may consequently discard information about the underlying scene structure. To preserve such information, we attach a reconstruction decoder to the compact geometric pathway. The decoder mirrors the encoding hierarchy by first expanding the temporal slots into frame-level representations and subsequently reconstructing the geometric features associated with each camera view, yielding $\widehat{\mathbf{F}}_t^v$. Let Ω denote the set of valid frame-view pairs. We optimize

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{(t,v)\in\Omega} \text{SmoothL1}\left(\widehat{\mathbf{F}}_t^v, \text{sg}(\mathbf{F}_t^v)\right), \tag{1}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator. By reconstructing features from the frozen 3D encoder rather than raw observations, this objective encourages the compact representation to retain geometrically informative scene structure despite the substantial token compression.

2) Masked Geometric Representation Prediction. To further capture temporal dependencies among geometric states, we introduce a slot-level masked prediction objective that requires missing geometric states to be inferred from the available temporal context. Specifically, we randomly sample a subset of temporal slots $\mathcal{M} \subseteq \{2, \ldots, K\}$ for masking. A full-sequence target branch processes all temporal slots through the Slot-Temporal Mixer to produce the target representation $\mathbf{G}$. In parallel, a context branch processes the masked temporal slots using the same weight-shared Slot-Temporal Mixer. A lightweight predictor then uses the resulting contextual representations to estimate the target states at the masked positions, producing $\widehat{\mathbf{G}}_m$ for each $m \in \mathcal{M}$. The objective is

$$\mathcal{L}_{\text{mask}} = \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \text{SmoothL1}\left(\widehat{\mathbf{G}}_m, \text{sg}(\mathbf{G}_m)\right) + \lambda_{\text{vc}}\mathcal{R}_{\text{vc}}(\mathbf{G}), \tag{2}$$

where $\mathcal{R}_{\text{vc}}$ denotes a variance-covariance regularizer. Its variance term maintains sufficient dispersion across feature dimensions, while its covariance term discourages redundant correlations between dimensions. Together, they prevent degenerate latent solutions and promote an informative predictive state space.

3) Geometric Transition Prediction. Masked representation prediction encourages contextual inference among temporal states, but does not directly constrain the direction in which the latent geometry evolves. We therefore explicitly model local geometric transitions. For two consecutive predictive states, we define the latent transition as

$$\Delta\mathbf{G}_k = \mathbf{G}_{k+1} - \mathbf{G}_k, \qquad k = 1, \ldots, K-1, \tag{3}$$

and denote the transition sequence by $\Delta\mathbf{G} = \{\Delta\mathbf{G}_k\}_{k=1}^{K-1}$. A lightweight transition predictor takes the current predictive state $\mathbf{G}_k$ and estimates its subsequent latent change $\widehat{\Delta\mathbf{G}}_k$. We optimize

$$\mathcal{L}_{\text{trans}} = \frac{1}{K-1} \sum_{k=1}^{K-1} \text{SmoothL1}\left(\widehat{\Delta\mathbf{G}}_k, \text{sg}(\Delta\mathbf{G}_k)\right). \tag{4}$$

Predicting the residual transition rather than reconstructing another complete geometric state focuses this objective on how the latent scene configuration changes over time. Together with geometry reconstruction and masked representation prediction, transition prediction encourages $\mathbf{G}$ to capture both the spatial structure of the scene and its predictable temporal evolution.

Motor Representation Learning. The predictive geometric representation is designed to retain a broad and structured account of scene geometry and its evolution. While such information is useful for predictive modeling, not every geometric variation is equally informative for control. In particular, $\Delta\mathbf{G}$ may contain transitions induced by robot motion and object interaction together with variations that are only weakly related to the corresponding motor commands. We therefore use the geometric transitions as a motion-centric source and introduce action supervision to emphasize

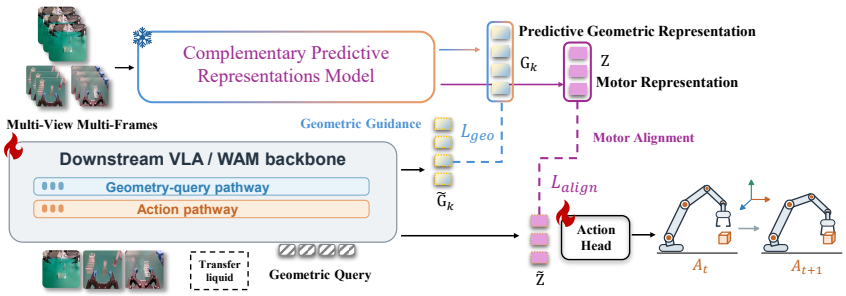


Figure 3: Overview of representation-guided policy post-training. The frozen pretrained model provides the predictive geometric representation $\mathbf{G}$ and motor representation $\mathbf{Z}$ as privileged supervision. The two representations are transferred through role-specific pathways: $\mathbf{G}$ supervises dedicated geometry queries, while $\mathbf{Z}$ aligns with intermediate features along the action-generation pathway.

the component of these transitions that is informative for action generation. Specifically, a set of learnable motor queries cross-attends to the transition tokens in $\Delta\mathbf{G}$, which serve as keys and values. The resulting features form the motor representation $\mathbf{Z}$. An action prediction head maps $\mathbf{Z}$ to a predicted action chunk $\widehat{\mathbf{A}}$, which is supervised by the ground-truth action chunk $\mathbf{A}$:

$$\mathcal{L}_{\text{act}}^{\text{pre}} = \left\| \widehat{\mathbf{A}} - \mathbf{A} \right\|_1 . \tag{5}$$

The complete representation-pretraining objective is

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{act}}^{\text{pre}}, \tag{6}$$

where λ_{mask} and λ_{trans} balance the corresponding objectives.

3.2 Representation-Guided Policy Post-Training

We next transfer the complementary information encoded by the predictive geometric and motor representations to a downstream policy. As shown in Fig.3, for each post-training example, the frozen pretrained model processes a multi-view training clip containing the current and future observations and extracts the target predictive geometric representation $\mathbf{G}$ and motor representation $\mathbf{Z}$. These future-informed representations serve exclusively as privileged supervision, while the downstream policy continues to receive only its standard current-time inputs. Let O_t denote the current observation, l the language instruction, and s_t the robot state. The policy backbone processes (O_t, l, s_t) together with a set of learnable geometry queries. In addition to the intermediate features used by the policy for action generation, the geometry queries produce a sequence of predicted geometric representation $\widetilde{\mathbf{G}} = \{\widetilde{\mathbf{G}}_k\}_{k=1}^{K}$. We align these representations with the frozen targets $\mathbf{G}$ using

$$\mathcal{L}_{\text{geo}} = \frac{1}{K}\sum_{k=1}^{K} \text{SmoothL1}\left(\widetilde{\mathbf{G}}_k, \text{sg}(\mathbf{G}_k)\right), \tag{7}$$

which encourages the model to capture the current scene configuration together with its anticipated geometric evolution. The motor representation is transferred through a different interface, rather than requiring the policy to explicitly recover $\mathbf{Z}$ through an additional prediction output, we use $\mathbf{Z}$ as a guidance target for intermediate features along the action-generation pathway. Let $\widetilde{\mathbf{Z}}$ denote the intermediate policy features, we align these features with the frozen motor representation target using a cosine-distance objective:

$$\mathcal{L}_{\text{align}} = 1 - \cos\left(\widetilde{\mathbf{Z}}, \text{sg}(\mathbf{Z})\right). \tag{8}$$

Finally, we combine the two representation-guidance objectives with the native action-generation objective of the downstream policy:

$$\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{act}}^{\text{post}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}, \tag{9}$$

where $\mathcal{L}_{\text{act}}^{\text{post}}$ denotes the action-generation objective of the policy backbone. The representation guidance therefore does not replace the backbone's native action modeling objective, but complements it with predictive geometric and motor supervision. At inference time, neither future observations nor the pretrained representation model is required.

Table 1: Comparison of different policies on LIBERO and LIBERO-Plus.

Methods	LIBERO					LIBERO-Plus
	Object	Spatial	Goal	Long	Avg.	Total
OpenVLA-OFT (Kim et al., 2025)	98.4	97.6	97.9	94.5	97.1	69.6
π_0 (Black et al., 2024)	98.8	96.8	95.8	85.2	94.2	53.6
GeoVLA(Sun et al., 2025)	99.0	98.4	96.6	96.6	97.7	-
3D-CAVLA (Bhat et al., 2025)	96.8	98.2	98.2	96.1	98.1	-
4D-WAM(Yang et al., 2026)	99.5	98.8	**98.6**	**97.4**	**98.6**	71.0
ELAN4D(He et al., 2026)	98.8	98.2	96.8	94.2	97.0	78.2
Ours w/o CPR	98.2	98.2	96.2	93.6	96.6	75.0
Ours	**100.0**	**99.0**	98.2	97.0	**98.6**	**81.7**

Table 2: Comparison of different policies on WidowX Robot.

Methods	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
OpenVLA-OFT(Kim et al., 2025)	12.5	4.2	4.2	**100.0**	30.2
π_0(Black et al., 2024)	25.0	16.7	12.5	29.2	20.9
SpatialVLA(Qu et al., 2025)	16.7	25.0	29.2	**100.0**	42.7
FALCON(Zhang et al., 2025b)	62.5	41.7	20.8	**100.0**	56.3
QDepth-VLA(Li et al., 2025b)	82.0	57.5	39.6	95.0	68.5
WoG(Su et al., 2026)	79.2	50.0	33.0	91.7	63.5
Ours w/o CPR	83.3	58.3	54.2	50.0	61.5
Ours	**95.8**	**70.8**	**70.8**	50.0	**71.9**

4 EXPERIMENT

We design our experiments to answer three questions:

(Q1) Does CPR guidance improve downstream manipulation performance across diverse settings?

(Q2) Does CPR generalize across policy paradigms while improving predictive modeling?

(Q3) How should CPR be organized and transferred for effective policy learning?

Experimental setup. We evaluate CPR guidance across simulated and real-world manipulation settings, covering task suites, robot embodiments, and downstream policies. Unless otherwise specified, we use QwenOFT (Yang et al., 2025; Kim et al., 2025) as the downstream policy baseline and denote this baseline as *Ours w/o CPR*. Our evaluation includes LIBERO (Liu et al., 2024), LIBERO-Plus (Fei et al., 2025), SimplerEnv (Li et al., 2024b), and four real-world manipulation tasks. Our real-world evaluation consists of three bimanual tasks, *remove a stopper*, *pick up tubes into box*, and *transfer liquid into an erlenmeyer*, together with one single-arm task, *make a burger*. These tasks cover bimanual coordination, object transfer, liquid manipulation, and multi-stage manipulation, providing challenges for evaluating perception and action generation in physical environments. We compare against policy baselines without predictive-representation guidance, including OpenVLA-OFT (Kim et al., 2025) and π_0 (Black et al., 2024). We further compare with input-centric approaches that incorporate geometric information into the policy input, including GeoVLA (Sun et al., 2025), 3D-CAVLA (Bhat et al., 2025), SpatialVLA (Qu et al., 2025), and FALCON (Zhang et al., 2025b). In addition, we include prediction-centric approaches that provide future or predictive information to policy learning, including 4D-WAM (Yang et al., 2026), ELAN4D (He et al., 2026), QDepth-VLA (Li et al., 2025b), and WoG (Su et al., 2026). Implementation details and training configurations are provided in the Appendix.F and B.

4.1 **(Q1):** SIMULATION AND REAL-WORLD RESULTS

Simulation Results on LIBERO and LIBERO-Plus. We first evaluate CPR on manipulation and under distribution shifts using LIBERO and LIBERO-Plus. As shown in Tab. 1, CPR improves the average LIBERO success rate from 96.6% to 98.6%, an absolute gain of 2.0 percentage points, with notable improvements on the *Spatial* and *Long* suites. We further evaluate the policy zero-shot on LIBERO-Plus without adaptation, where CPR increases the overall task-count-weighted success rate from 75.0% to 81.7%. The gains on both benchmarks indicate that complementary predictive guidance benefits manipulation under both nominal and shifted environments, providing structured geometric information together with action-aware cues for downstream control.

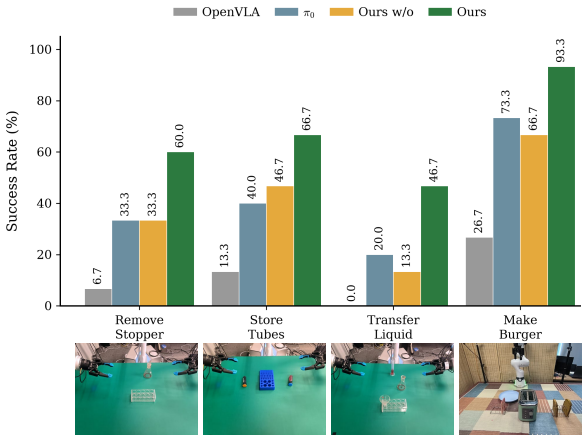


Figure 4: Performance comparison on real-robot manipulation tasks.

Simulation Results on WidowX. We next evaluate the method in SimplerEnv using the WidowX robot. As reported in Tab. 2, CPR

guidance raises the average success rate from 61.5% to 71.9%, corresponding to an absolute gain of 10.4 percentage points. The improvement on *Stack Cube* is especially informative because successful stacking requires accurate reasoning about the relative configuration of the manipulated cube and its target, as well as coordinated actions as the scene evolves. The gain on this task is consistent with the intended roles of the two representations: predictive geometric guidance provides structured information about scene configuration and its evolution, while motor guidance emphasizes action-aware transitions associated with manipulation.

Real-world evaluation. We evaluate CPR on four real-world manipulation tasks shown in Fig. 4. The CPR-guided model consistently outperforms the baselines across tasks. The improvements are particularly pronounced on fine-grained manipulation tasks, such as *Remove Stopper* and *Transfer Liquid*, where precise spatial reasoning and action execution are required. By providing fine-grained geometric perception together with predictive information about future states, CPR enables the policy to better anticipate object configurations and execute more precise interactions. These results demonstrate that the complementary predictive guidance learned by CPR translates from simulation to real-world manipulation, particularly in tasks requiring accurate and fine-grained control.

4.2 **(Q2):** Transfer Across Diverse Policy Paradigms

Transfer to Diverse VLA Models. To examine whether CPR depends on a particular downstream policy architecture, we integrate the same representation guidance into two additional VLA models, GR00T-style (Bjorck et al., 2025) and π_0-style models (Black et al., 2024). The representation pretraining procedure and the roles of the two representations remain unchanged, while their transfer interfaces are adapted to the corresponding downstream architectures. As shown in Fig. 5, adding the same alignment objective improves the average success rate of GR00T from 62.5% to 74.0%. Similarly, π_0 improves from 59.4% to 68.8%, corresponding to a gain of 9.4 percentage points. Together with the gains observed for OFT, these results indicate that the proposed guidance is compatible with substantially different VLA architectures and is not tied to a single action-generation mechanism.

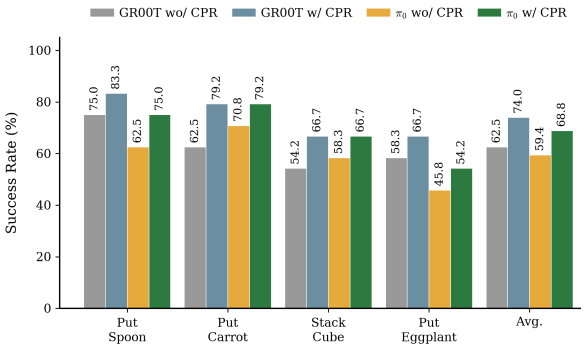


Figure 5: Transfer of CPR to different downstream VLA models on the WidowX robot.

Transfer to World Action Models. We further investigate whether complementary predictive representations remain useful for the WAM paradigm. We integrate CPR into FastWAM (Yuan et al., 2026), which jointly models future world evolution and robot actions. To preserve the respective roles of the two representations, we introduce an additional geometry DiT branch in parallel with the original world and action modeling branches for predictive geometric modeling, while using the motor representation to provide action-aware guidance to the action branch. As shown in Tab. 3, incorporating CPR improves the average LIBERO success rate of FastWAM from 97.6% to 98.5%. Although FastWAM already explicitly models future world evolution, these results suggest that structured predictive geometry and action-aware guidance can provide complementary information beyond the original world-action modeling objective. Combined with the results across three VLA architectures, this experiment further demonstrates the broad compatibility of CPR across substantially different policies. More details are provided in the Appendix.C.2.

Table 3: Transfer of CPR to FastWAM on the LIBERO.

Methods	Object	Spatial	Goal	Long	Avg.
FastWAM wo/ CPR	**100.0**	98.2	97.2	95.2	97.6
FastWAM w/ CPR	**100.0**	**99.0**	**97.4**	**96.8**	**98.5**

4.3 **(Q3):** Analysis of Representation Design and Transfer

Effect of Complementary Predictive Representation Guidance. To examine complementary contributions of the two predictive representations, we conduct an ablation study on the WidowX

Table 4: Ablation of CPR guidance on the WidowX evaluation. PGR and MR denote the predictive geometric and motor representations, respectively.

	PGR	MR	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
(a)			83.3	58.3	54.2	50.0	61.5
(b)	✓		79.2	66.7	66.7	**54.2**	66.7
(c)		✓	91.7	62.5	62.5	**54.2**	67.7
(d)	✓	✓	**95.8**	**70.8**	**70.8**	50.0	**71.9**

Table 5: Ablation of representation organization on the WidowX evaluation.

Methods	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
Baseline	83.3	58.3	54.2	**50.0**	61.5
Unified	91.7	**70.8**	62.5	45.8	67.7
Complementary	**95.8**	**70.8**	**70.8**	**50.0**	**71.9**

evaluation, with results reported in Tab. 4. Variant (a) corresponds to the VLA baseline without representation guidance. Introducing the predictive geometric representation in variant (b) improves the average success rate from 61.5% to 66.7%, yielding a gain of 5.2%. Variant (c) instead transfers the motor representation by aligning it with the intermediate features used for action generation. This variant improves the average success rate to 67.7%, outperforming the baseline by 6.2%. This result supports motor representation in providing action-aware predictive guidance to the action-generation pathway. Finally, combining both representations in variant (d) achieves the best average success rate of 71.9%, corresponding to a 10.4% absolute improvement over the baseline. Combining predictive geometric and motor representation allows the downstream policy to benefit from predictive geometric information and action-aware dynamics, leading to the strongest performance.

Effect of Representation Organization. We examine whether assigning predictive geometry and action-aware information to complementary representations is beneficial. In the unified variant, the motor branch is removed, and the predictive geometric representation is jointly supervised by the action objective and aligned with intermediate features of the downstream VLA. As shown in Tab. 5, the unified representation improves the average success rate from 61.5% to 67.7%, but remains below our complementary design at 71.9%, which indicates that providing complementary representations for structured predictive geometry and action-aware information is more effective than asking a single representation to serve both roles under training setup. To further probe encoded information, we evaluate the pre-trained representations zero-shot on LIBERO-90(Liu et al., 2024). As shown in Fig. 6, the motor representation concentrates more samples in the low action-error region, while the predictive geometric representation better preserves scene structure in depth probing, where the pseudo labels are generated by DepthAnything(Lin et al., 2025). The latter recovers more faithful object geometry and relative depth relationships compared with the unified representation. Together, these results suggest that assigning distinct roles to predictive geometric and motor representations helps preserve both geometric fidelity and action-relevant information, supporting the complementary design of CPR. More details are provided in Appendix A.3.

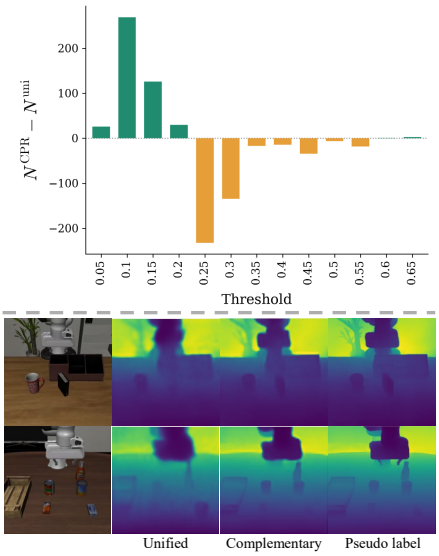


Figure 6: Probing analysis of representation organization on LIBERO-90. The top compares action-error distributions between the CPR motor and unified representations, while the bottom shows depth-probing results.

5 Conclusion

We study complementary predictive representation learning to connect evolving scene geometry with action generation. A predictive geometric representation captures structured scene evolution, while a motor representation emphasizes action-relevant geometric transitions; the two are transferred to downstream policies through dedicated guidance pathways without requiring future observations at inference time. Experiments across simulated and real-world manipulation settings demonstrate consistent gains across VLA architectures and WAMs, including improvements in both manipulation performance and future world/action prediction. These results highlight the importance of organizing predictive geometric information according to its role in downstream control.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

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 preprint arXiv:2506.09985*, 2025.

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Hydra: Hybrid robot actions for imitation learning. In *Conference on Robot Learning*, pp. 2113–2133. PMLR, 2023.

Vineet Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 3d cavla: Leveraging depth and 3d context to generalize vision language action models for unseen tasks. *arXiv preprint arXiv:2505.05800*, 2025.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint 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 preprint 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 preprint arXiv:2212.06817*, 2022.

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 preprint arXiv:2505.06111*, 2025.

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 preprint arXiv:2506.21539*, 2025.

Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley ur5 demonstration dataset. *URL https://sites. google. com/view/berkeley-ur5/home*, 2023.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint arXiv:2210.10047*, 2022.

Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J Lim. Clvr jaco play dataset, 2023. *URL https://github. com/clvrai/clvr_jaco_play_dataset*.

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021a.

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint arXiv:2109.13396*, 2021b.

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 preprint arXiv:2510.13626*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Zeyuan He, Bowen Yang, Zhirui Fang, Keru Zhou, Lei Jiang, Jingjing Qian, Fan Mo, Junchi Yan, Philip Torr, Xiu Li, et al. Elan4d: Embodiment-centric 4d supervision for vision-language-action models via plug-and-play adaptation. *arXiv preprint arXiv:2605.30484*, 2026.

Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, 44(10-11):1863–1891, 2025.

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *conference on Robot Learning*, pp. 991–1002. PMLR, 2022.

Shi Jin, Yuntian Wang, Yuhui Duan, Di Wu, Gaoqi Dong, Xiaohang Liu, Xiaotong Li, Hongfei Jia, Zehao Zhang, Tianyu Wang, et al. Umi-bench 1.0: An open and reproducible real-world benchmark for tabletop robotic manipulation with umi data. *arXiv preprint arXiv:2606.10382*, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

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 preprint arXiv:2406.09246*, 2024.

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

Sangoh Lee, Sangwoo Mo, and Wook-Shin Han. Act with intent: Distilling behavior intent for vision-language-action models. *arXiv preprint arXiv:2608.23478*, 2026a.

Seungjae Lee, Yoonkyo Jung, Jusuk Lee, Jonghun Shin, Amir Hossein Shahidzadeh, Yao-Chih Lee, H Jin Kim, Jia-Bin Huang, and Furong Huang. $\backslash\mu_0$: A scalable 3d interaction-trace world model. *arXiv preprint arXiv:2606.13769*, 2026b.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025a.

Hao Li, Ziqin Wang, Zi-han Ding, Shuai Yang, Yilun Chen, Yang Tian, Xiaolin Hu, Tai Wang, Dahua Lin, Feng Zhao, et al. Robointer: A holistic intermediate representation suite towards robotic manipulation. *arXiv preprint arXiv:2602.09973*, 2026a.

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 preprint arXiv:2601.21998*, 2026b.

Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024a.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024b.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, et al. Wam4d: Fast 4d world action model via spatial register tokens. *arXiv preprint arXiv:2606.14048*, 2026c.

Yixuan Li, Yuhui Chen, Mingcai Zhou, Haoran Li, Zhengtao Zhang, and Dongbin Zhao. Qdepth-vla: Quantized depth prediction as auxiliary supervision for vision-language-action models. *arXiv preprint arXiv:2510.14836*, 2025b.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, 44(10-11):1727–1742, 2025.

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 preprint arXiv:2605.00078*, 2026.

Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multistage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 40:1476–1491, 2024.

Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, 44(4):592–606, 2025.

Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.

Haoxiang Ma, Junhao Cai, Xiaoxu Xu, Hao Li, Yuyin Yang, Yang Tian, Jiafei Cao, Hongrui Zhu, Zherui Qiu, Yuqiang Yang, et al. Internvla-a1. 5: Unifying understanding, latent foresight, and action for compositional generalization. *arXiv preprint arXiv:2607.04988*, 2026.

Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1048–1055. IEEE, 2019.

Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. *arXiv preprint arXiv:2210.01911*, 2022.

Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. *arXiv preprint arXiv:2210.11435*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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 preprint arXiv:2501.15830*, 2025.

Gabriel Quere, Annette Hagengruber, Maged Iskandar, Samuel Bustamante, Daniel Leidner, Freek Stulp, and Jörn Vogel. Shared control templates for assistive robotics. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 1956–1962. IEEE, 2020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Erick Rosete-Beas, Oier Mees, Gabriel Kalweit, Joschka Boedecker, and Wolfram Burgard. Latent plans for task-agnostic offline reinforcement learning. In *Conference on Robot Learning*, pp. 1838–1849. PMLR, 2023.

Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multi-modal task specifications. *arXiv preprint arXiv:2309.14320*, 2023.

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 preprint arXiv:2602.22010*, 2026.

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.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on robot learning*, pp. 1723–1736. PMLR, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Junke Wang, Qihang Zhang, Shuai Yang, Yiming Luo, Yujun Shen, Zuxuan Wu, Yu-Gang Jiang, and Yinghao Xu. Repwam: World action modeling with representation visual-action tokenizers. *arXiv preprint arXiv:2606.13674*, 2026a.

Ziqin Wang, Hao Li, Weijun Wang, Junhao Cai, Jia Zeng, Yilun Chen, Jiangmiao Pang, and Si Liu. Robointer1. 5: A holistic intermediate representation suite for embodied world modeling and robotic manipulation. *arXiv preprint arXiv:2607.18709*, 2026b.

Xiaoxu Xu, Hao Li, Jinhui Ye, Yilun Chen, Jia Zeng, Xinyi Chen, Linning Xu, Dahua Lin, Weixin Li, and Jiangmiao Pang. Futurevla: Joint visuomotor prediction for vision-language-action model. *arXiv preprint arXiv:2603.10712*, 2026.

Ge Yan, Kris Wu, and Xiaolong Wang. ucsd kitchens dataset, 2023.

Haodong Yan, Junfeng Li, Junjie He, Zhide Zhong, MingMing Yu, Wenxuan Song, Jiaguan Zhu, Yangyang Zheng, Yuqiao Du, Jiadi You, et al. Robust-wam: Bridging generative pretraining and semantic foresight in world-action models. *arXiv preprint arXiv:2608.05903*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026.

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 preprint arXiv:2602.15922*, 2026.

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

Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025a.

Zhengshen Zhang, Hao Li, Yalun Dai, Zhengbang Zhu, Lei Zhou, Chenchen Liu, Dong Wang, Francis EH Tay, Sijin Chen, Ziwei Liu, et al. From spatial to actions: Grounding vision-language-action model in spatial foundation priors. *arXiv preprint arXiv:2510.17439*, 2025b.

Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot, 2023a.

Yifeng Zhu, Peter Stone, and Yuke Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters*, 7(2):4126–4133, 2022.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pp. 1199–1210. PMLR, 2023b.

APPENDIX CONTENTS

A Additional Experiments **16**
A.1 **(Q1):** Simulation and Real-World Results 16
A.2 **(Q2):** Transfer Across Diverse Policy Paradigms 17
A.3 **(Q3):** Analysis of Representation Design and Transfer 20

B Training Details **26**
B.1 Datasets 26
B.2 Training Protocol 27

C Model Details **28**
C.1 Hierarchical Geometric Tokenization 28
C.2 Details of FastWAM 29

D Limitations and Future Work **30**

E Simulation and Real-World Evaluation Details **30**
E.1 Simulation Evaluation Details 30
E.2 Real-World Robot Evaluation Details 31

F Details of Compared Methods **33**

G AI Use Statement **35**

Table 6: Comparison of different VLA models on LIBERO-Plus. Success rates (%) are reported under the seven perturbation types. Total denotes the official task-count-weighted mean over all 10,030 perturbed tasks, rather than an unweighted mean of the seven categories.

Methods	Cam.	Rob.	Lang.	Light	Back.	Noise	Layout	Total
OpenVLA-OFT(Kim et al., 2025)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0 (Black et al., 2024)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
WorldVLA(Cen et al., 2025)	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
UniVLA(Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	43.9
4D-WAM(Yang et al., 2026)	45.2	64.3	**90.6**	**94.3**	57.7	69.1	79.2	71.0
ELAN4D(He et al., 2026)	63.7	**70.7**	77.8	89.8	91.4	79.9	**81.4**	78.2
Ours w/o CPR	47.6	59.2	86.5	94.2	**94.9**	76.1	78.8	75.0
Ours	**77.5**	60.8	88.6	86.1	89.1	**95.6**	77.2	**81.7**

A Additional Experiments

A.1 (Q1): Simulation and Real-World Results

Full LIBERO-Plus Results. We provide the detailed results on the seven perturbation types of LIBERO-Plus in Tab. 6, which provide a more detailed view of the robustness gains reported in the main experiments.

Real-World Out-of-Distribution Generalization. We further evaluate the ability of CPR to generalize to unseen visual and spatial conditions without additional training. Starting from the same real-world tasks described above, we introduce task-specific perturbations that change the observation distribution while preserving the underlying manipulation objective. Specifically, as shown in Fig.8, for *Remove Stopper*, we shift the position of the erlenmeyer flask to alter the relative geometry between the two arms and the manipulated object. For *Store Tubes*, we randomly vary the height of the target box, requiring the policy to adapt to changes in the target geometry. For *Transfer Liquid*, we introduce a randomly moving toy car as a dynamic visual distractor during execution, increasing the difficulty of separating task-relevant objects from irrelevant scene motion. For *Make a Burger*, we add multiple irrelevant objects to the workspace, introducing additional visual clutter.

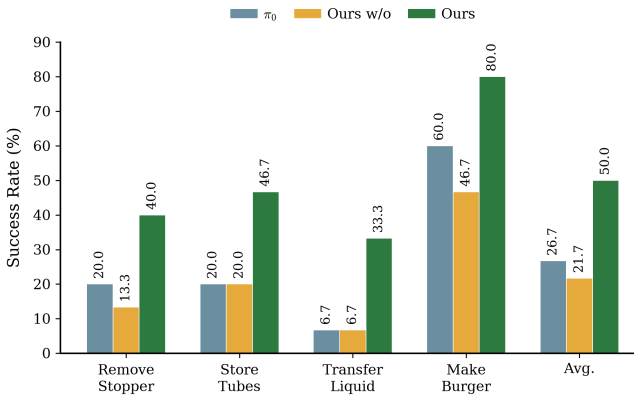


Figure 7: Out-of-Distribution success rates on real-world tasks under visual and spatial perturbations.

These perturbations respectively introduce changes in object placement, target geometry, dynamic visual context, and scene clutter, while keeping the underlying task semantics unchanged. As shown in Fig. 7, CPR achieves the highest success rate on all four tasks and improves the overall average success rate from 21.7% for the baseline without CPR to 50.0%. The consistent improvement under these unseen conditions suggests that complementary predictive representation guidance can provide useful geometric and action-aware information beyond the nominal training distribution, helping the downstream policy maintain effective manipulation behavior under changes in visual and spatial context.

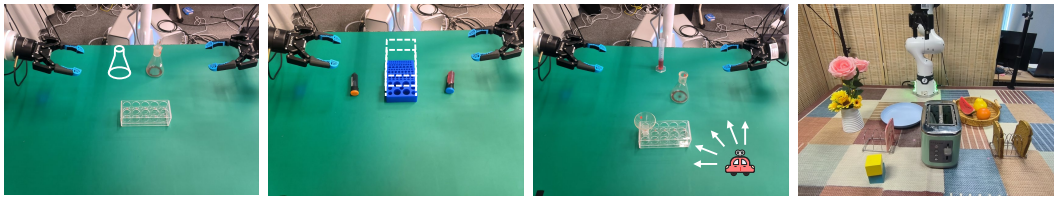


Figure 8: Visualization of the real-world out-of-distribution perturbations. From left to right: *Remove Stopper*, with randomly shifted flask positions; *Store Tubes*, with randomly varied target-box heights; *Transfer Liquid*, with a randomly moving toy car as a dynamic visual distractor; and *Make Sandwich*, with additional irrelevant objects added to the workspace. These perturbations introduce changes in object placement, target geometry, dynamic visual context, and scene clutter while preserving the underlying manipulation tasks.

Table 7: Zero-shot future-world and action prediction performance of FastWAM on LIBERO-90. PSNR and SSIM are higher-is-better, while LPIPS and action ℓ_1 error are lower-is-better.

Model	PSNR↑	SSIM↑	LPIPS↓	Action ℓ_1 ↓
FastWAM w/o CPR	19.67	0.764	0.229	0.124
FastWAM w/ CPR	**22.91**	**0.803**	**0.186**	**0.075**

A.2 **(Q2):** Transfer Across Diverse Policy Paradigms

Quantitative Analysis of FastWAM. Beyond downstream task success in Sec.4.2, we examine whether CPR improves the predictive capabilities of WAMs themselves. We freeze the FastWAM models trained on LIBERO and evaluate them on LIBERO-90 without additional adaptation. Future visual prediction is evaluated using PSNR, SSIM, and LPIPS, while action prediction is measured by the ℓ_1 error between predicted and ground-truth actions. As shown in Tab. 7, CPR consistently improves both future visual prediction and action prediction. PSNR increases from 19.67 to 22.91, SSIM improves from 0.764 to 0.803, and LPIPS decreases from 0.229 to 0.186. The action ℓ_1 error decreases from 0.124 to 0.075, corresponding to a relative reduction of approximately 39.5%. These results provide evidence that the proposed representation guidance improves not only downstream manipulation performance but also the predictive quality of the underlying world-action model. The improvement in visual prediction is consistent with the structured geometric supervision provided by the predictive geometric representation, while the reduction in action error is consistent with the action-aware information emphasized by the motor representation. Together, they provide complementary predictive guidance for jointly modeling future world evolution and robot actions.

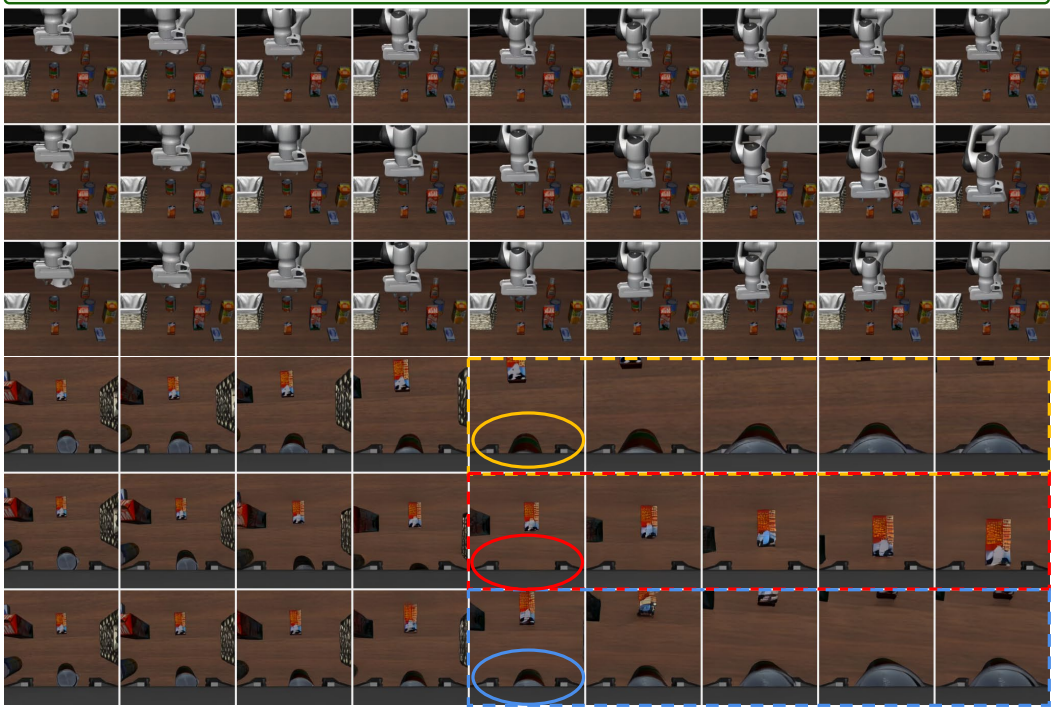
Qualitative Analysis of FastWAM. We provide additional qualitative comparisons to further examine the improvements observed in the zero-shot world- and action-prediction evaluation on LIBERO-90. As illustrated in Fig. 9, the differences between FastWAM with and without CPR are evident in both geometric fidelity and action execution.

In Fig. 9(a), we consider a scene in which the robot is preparing to grasp a can. Without complementary predictive representation guidance, the predicted future frames exhibit a noticeable discrepancy in the estimated geometric configuration, particularly in the relative position between the gripper and the target object. As a consequence, the generated gripper trajectory does not reach the appropriate grasp location, resulting in an action prediction that is inconsistent with the target interaction. With CPR guidance, the predicted frames more accurately preserve the geometric configuration of the object and gripper, allowing the generated trajectory to approach the can at a more appropriate location. This qualitative example is consistent with the role of the predictive geometric representation in providing structured geometric information for future scene prediction. We observe similar differences in more temporally extended interactions, as shown in (b) and (c). Without CPR, the generated videos may exhibit missing or ambiguous object structures and progressively deviate from the expected scene evolution. At the same time, the predicted actions can become inconsistent with the intended manipulation sequence. For example, as shown in (c), in a drawer-opening interaction, the baseline predictions may fail to clearly represent the drawer's changing configuration and do not adequately capture the corresponding opening motion. In contrast, FastWAM with CPR maintains

more coherent object configurations over time and produces predictions that more faithfully reflect the evolving manipulation process.

These observations provide a qualitative counterpart to the quantitative improvements reported in Tab. 7. The predictive geometric representation encourages the model to preserve structured geometric information during future-world modeling, which can improve the fidelity of object shape, relative spatial configuration, and scene evolution in the generated frames. Meanwhile, the motor representation emphasizes action-aware information from geometric transitions, providing guidance that is more directly associated with the execution of the corresponding manipulation. The combination therefore allows the model to better maintain both the evolving geometric state of the scene and the action sequence associated with that evolution. Overall, the qualitative results suggest that CPR improves the consistency between predicted future geometry and predicted robot actions. Rather than improving only the visual appearance of individual future frames, the complementary guidance helps the model maintain a more coherent correspondence between *what the scene is expected to become* and *what action should be taken as it evolves*. This observation is consistent with the complementary roles of the predictive geometric and motor representations in our framework.

(a) Pick up the tomato sauce and put it in the basket.



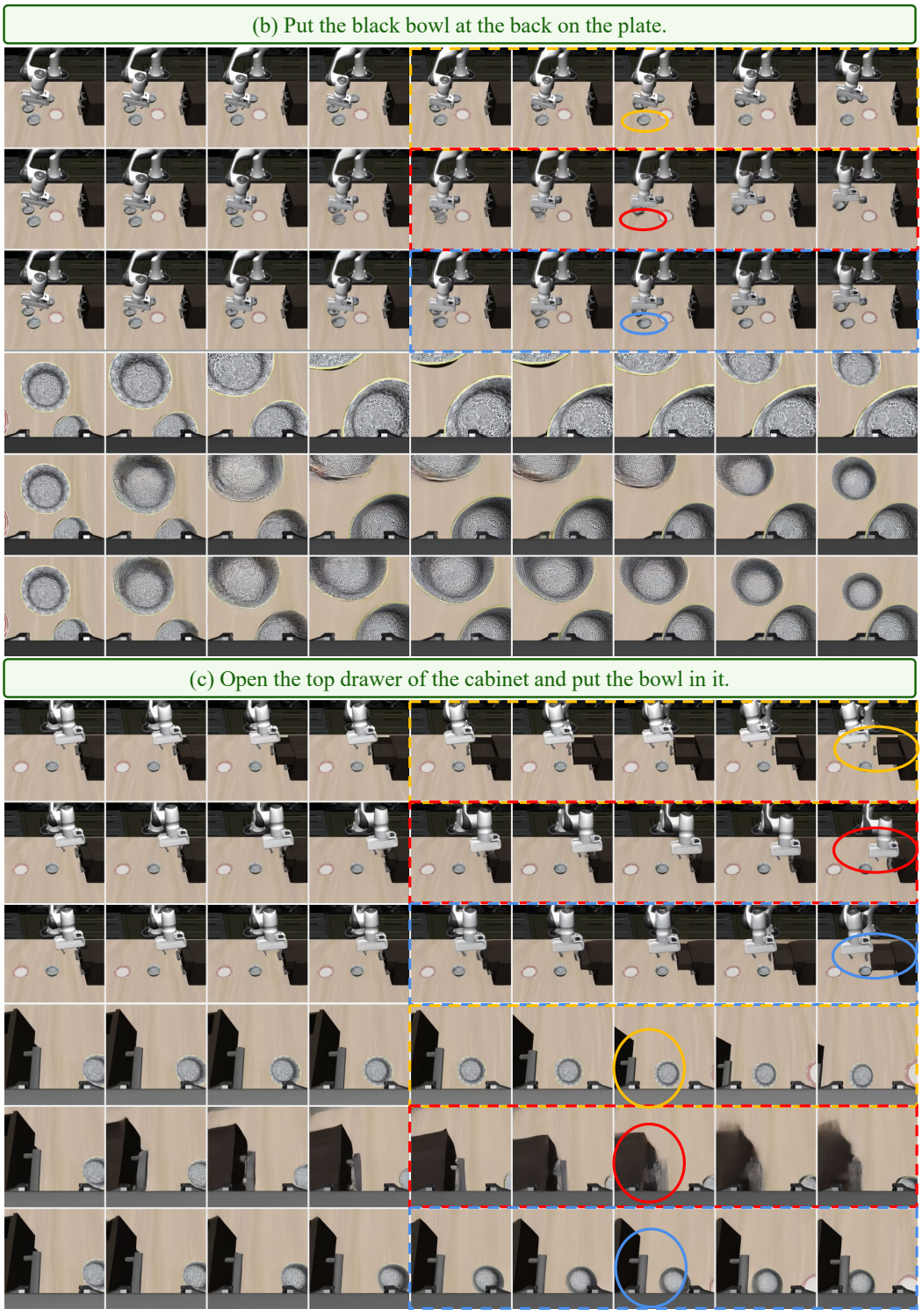


Figure 9: Qualitative comparison of zero-shot future world prediction on LIBERO-90. Each example contains a main camera view (upper half) and a wrist camera view (lower half), with columns progressing over nine predicted future frames. Within each view, the three rows from top to bottom correspond to ground-truth observations (GT), FastWAM without CPR guidance (w/o CPR), and FastWAM with CPR guidance (w/ CPR).

A.3 **(Q3):** Analysis of Representation Design and Transfer

More Details of Representation Organization. We provide additional details of the representation-organization analysis reported in Sec. 4.3. For action probing, we visualize the number of samples falling into each ℓ_1-loss bin for the different representation variants. As shown in Fig. 10, the overlapping bar charts provide the full error distributions underlying the action-probing analysis in the main paper, offering a more detailed view of how each representation distributes across low- and high-error regions.

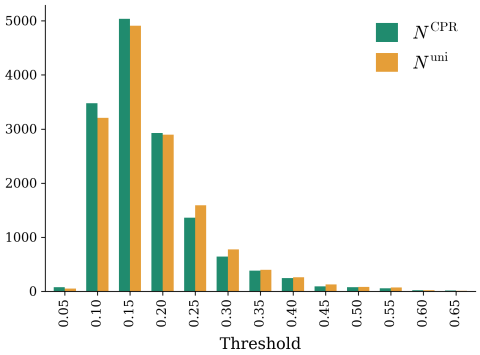
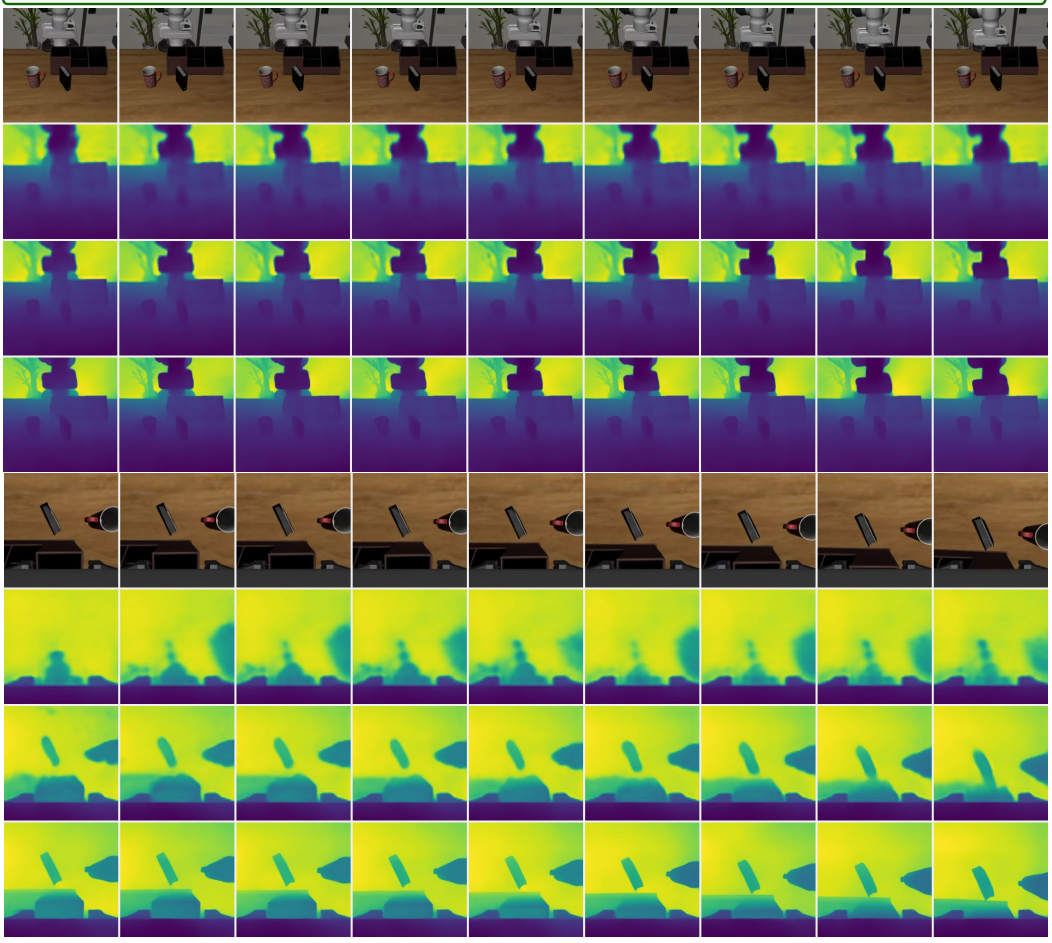


Figure 10: Distribution of action-probing experiment.

We further provide qualitative results for the depth-probing experiment. As shown in Fig. 11, the depth predictions obtained from the predictive geometric representation more faithfully recover the spatial structure of the corresponding scenes than those obtained from the unified representation. In particular, the predicted maps better preserve object boundaries, relative depth relationships, and major scene structures across the future frames. These qualitative observations complement the quantitative probing results and provide additional evidence that the predictive geometric representation retains richer geometric information than the unified representation.

(a) Pick up the book and place it in the right compartment of the caddy.



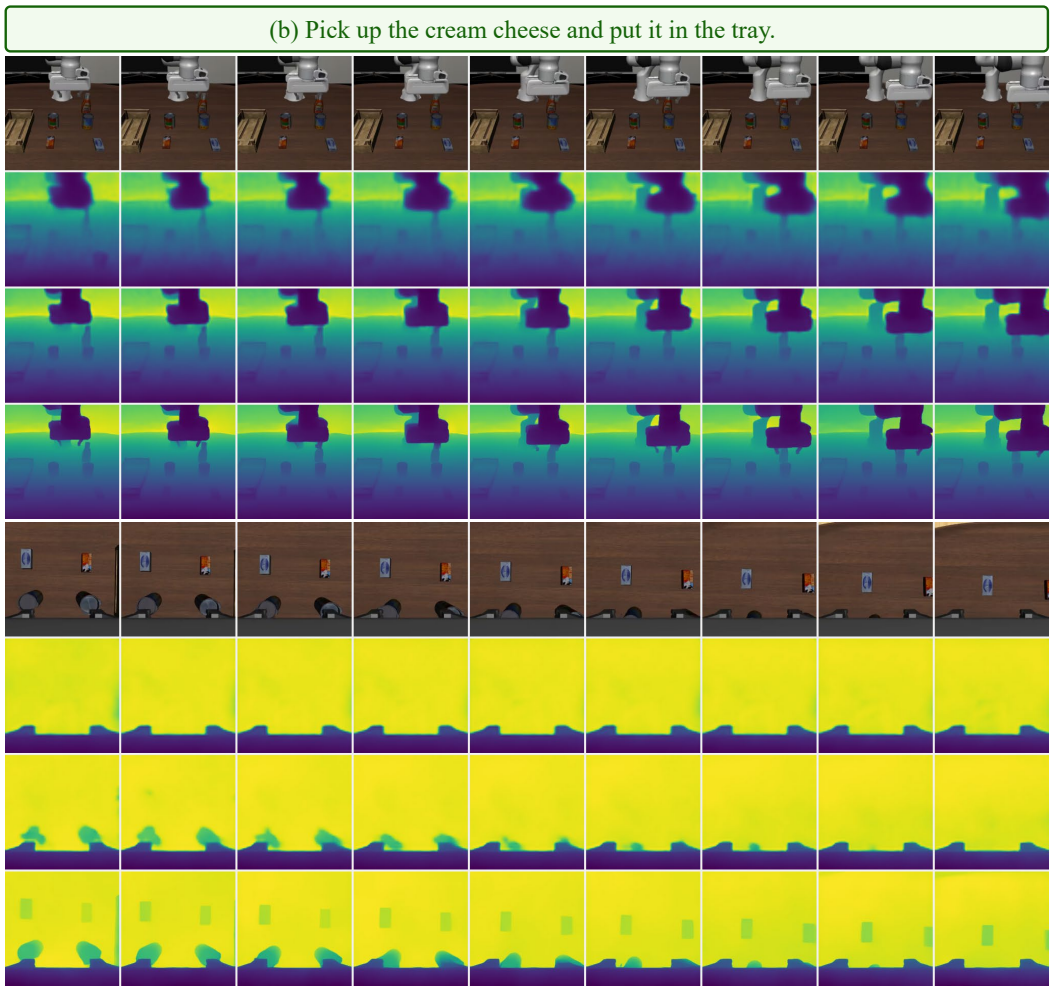


Figure 11: Qualitative visualization of the depth-probing analysis for representation organization. Each example contains a main camera view (upper half) and a wrist camera view (lower half), with columns progressing over nine future frames. Within each view, the rows from top to bottom show the corresponding RGB observations, depth predictions decoded from the unified representation, depth predictions decoded from the predictive geometric representation, and the pseudo-depth references extracted by Depth Anything (Lin et al., 2025). The predictive geometric representation produces depth maps that more faithfully recover object geometry, relative depth, and overall scene structure than the unified representation.

Table 8: Comparison with different predictive representation learning methods under a unified downstream policy framework on the WidowX evaluation.

Methods	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
WoG (Su et al., 2026)	83.3	62.5	37.5	75.0	64.6
FutureVLA (Xu et al., 2026)	79.2	70.8	41.7	**79.2**	67.7
RepWAM (Wang et al., 2026a)	87.5	**75.0**	54.2	58.3	68.8
Ours	**95.8**	70.8	**70.8**	50.0	**71.9**

Comparison with Predictive Representation Learning Methods. Existing predictive representation learning methods differ in both their learned representations and downstream policy architectures, making direct comparisons potentially confounded by differences in the overall training pipeline. To isolate the effect of predictive representation learning, we therefore integrate several

Table 9: Ablation of the pretraining objectives for the predictive geometric representation on the WidowX evaluation.

	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{mask}}$	$\mathcal{L}_{\text{trans}}$	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
(a)				83.3	58.3	54.2	**50.0**	61.5
(b)	✓			87.5	66.7	58.3	**50.0**	65.6
(c)	✓	✓		91.7	**70.8**	66.7	45.8	68.8
(d)	✓	✓	✓	**95.8**	**70.8**	**70.8**	**50.0**	**71.9**

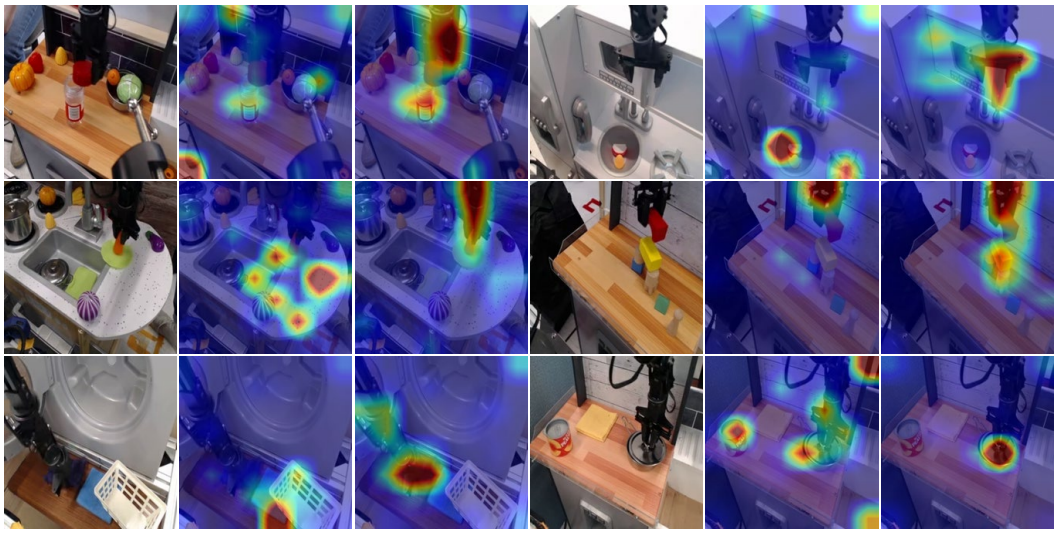


Figure 12: Attention-map comparison between the unguided baseline (middle) and the CPR-guided policy (right). CPR guidance produces more concentrated responses around task-relevant objects and interaction regions.

representative approaches into the same downstream VLA framework and evaluate them under a unified setting on the WidowX benchmark. As shown in Tab. 8, our method achieves the highest average success rate of 71.9%, compared with 64.6% for WoG (Su et al., 2026), 67.7% for FutureVLA (Xu et al., 2026), and 68.8% for RepWAM (Wang et al., 2026a). Since these methods are evaluated with the same downstream policy framework, the comparison provides a more controlled assessment of the predictive representations used to guide policy learning. The result suggests that the benefit of predictive guidance depends not only on exposing the policy to future information, but also on how that information is organized for downstream control. Our method provides structured predictive geometric information together with action-aware information extracted from geometric transitions, allowing these complementary signals to be transferred through dedicated supervision pathways.

Effect of Predictive Representation Pretraining Objectives. We further examine the contribution of the three objectives used to learn the predictive geometric representation. Starting from the downstream baseline, introducing geometry reconstruction $\mathcal{L}_{\text{rec}}$ together with the action-supervised motor branch improves the average success rate from 61.5% to 65.6%, indicating that preserving geometric structure already provides useful guidance for downstream control. Adding masked geometric state prediction $\mathcal{L}_{\text{mask}}$ further increases the performance to 68.8%, suggesting that contextual prediction across temporal states helps the representation capture richer temporal dependencies. Finally, incorporating geometric transition prediction $\mathcal{L}_{\text{trans}}$ yields the best result of 71.9%. This additional gain indicates that explicitly modeling transitions between consecutive geometric states provides further information about scene evolution that is useful for downstream action generation. Overall, the progressive improvements support the complementary roles of structural, contextual, and transition-level supervision in shaping the predictive geometric representation.

Table 10: Effect of the input used to learn the motor representation on the WidowX evaluation.

Motor Input	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
Predictive States $\mathbf{G}$	87.5	66.7	66.7	**54.2**	68.8
Geometric Transitions $\Delta\mathbf{G}$	**95.8**	**70.8**	**70.8**	50.0	**71.9**

Table 11: Effect of the motor representation alignment depth on the WidowX evaluation. The alignment is applied to intermediate features from different layers.

Alignment Layer	6	12	18	24	30	36
SR (%)	69.8	**71.9**	67.7	64.6	65.6	61.5

Table 12: Effect of the foundation encoder used during predictive representation pretraining on the WidowX evaluation.

Methods	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
DINO	87.5	**75.0**	54.2	**62.5**	69.8
VGGT	**95.8**	70.8	**70.8**	50.0	**71.9**

Qualitative Analysis of Visual Attention. We also visualize the attention maps of the baseline and CPR-guided policies under the same observations and language instructions. As shown in Fig. 12, the baseline exhibits relatively diffuse attention, whereas the CPR-guided model more consistently concentrates on task-relevant objects, target regions, and surrounding spatial context. In particular, stronger localization is observed around manipulated objects and regions involved in the ongoing interaction. These qualitative observations are consistent with the intended roles of the two guidance signals: predictive geometric supervision encourages structured scene understanding, while motor guidance emphasizes information associated with action-aware transitions. Although attention maps do not directly measure policy reasoning, they provide qualitative evidence that CPR can encourage more task-focused visual representations.

Effect of Geometric Transitions for Motor Representation Learning. We next examine why the motor representation is learned from geometric transitions rather than directly from predictive geometric states. We construct a variant in which the motion queries attend directly to $\mathbf{G}$ instead of $\Delta\mathbf{G}$ while keeping the remaining training procedure unchanged. As shown in Tab. 10, directly using the predictive geometric states yields an average success rate of 68.8%, whereas using geometric transitions increases the performance to 71.9%, an absolute gain of 3.1 percentage points. This result supports the use of transitions as the input to motor representation learning. A possible interpretation is that consecutive predictive states share substantial information describing scene structure, whereas their difference emphasizes temporal variation. Using $\Delta\mathbf{G}$ therefore provides a more action-oriented signal from which action supervision can identify information associated with motor commands.

Effect of the Motor Representation Alignment Layer. We investigate where motor guidance should be introduced within the downstream VLA backbone. Specifically, we apply the same motor representation alignment objective to intermediate features from different transformer layers while keeping the remaining training configuration unchanged. As reported in Tab. 11, performance reaches its maximum at layer 12, with an average success rate of 71.9%. The performance decreases when the alignment point is moved toward deeper layers. This suggests that motor guidance is most effective at an intermediate stage of the policy representation, where the features have incorporated sufficient contextual information while retaining flexibility to influence subsequent action generation. Based on this result, we use layer 12 for motor representation alignment in the remaining experiments.

Effect of the Foundation Encoder. We further examine the dependence of representation pre-training on the chosen foundation encoder. We compare DINO(Oquab et al., 2023), which provides generic visual representations, with VGGT(Wang et al., 2025), which provides stronger geometric structure. As shown in Tab. 12, both encoders lead to competitive downstream performance, achieving 69.8% and 71.9% average success rates, respectively. VGGT provides the stronger overall result and a notable advantage on *Stack Cube*, which requires reasoning about relative object geometry. At the same time, DINO remains competitive and performs better on *Put Carrot* and *Put Eggplant*. These results suggest that the proposed representation learning framework can utilize different pre-trained feature spaces while benefiting from a foundation model with stronger geometric structure.

Table 13: Ablation study of the geometry-guidance loss weight λ_{geo} on the WidowX evaluation.

λ_{geo}	0.1	0.3	0.5	0.7	0.9
SR (%)	69.8	**71.9**	70.8	68.8	65.6

Effect of the Geometry-Guidance Weight. We investigate the effect of the geometry-guidance loss weight λ_{geo} in the downstream post-training objective on the WidowX evaluation. As shown in Tab. 13, the average success rate initially improves as the weight increases from 0.1 to 0.3, reaching the best performance of 71.9%. Further increasing λ_{geo} leads to a gradual performance degradation. These results suggest that the predictive geometric supervision is most effective when its contribution is appropriately balanced with the native action-generation objective. A relatively small geometry-guidance weight may provide insufficient supervision for transferring the structured predictive geometric information, whereas overly strong geometric supervision may place excessive emphasis on representation matching and reduce the relative influence of the action-generation objective. We therefore set $\lambda_{\text{geo}} = 0.3$ in the remaining experiments.

Comparison of Guidance Paradigms. To further examine the design of CPR, we compare our approach with two representative guidance paradigms under the same downstream VLA framework and training setting. The *input-centric* variant directly provides current-frame VGGT features to the policy, allowing it to access explicit geometric information from the observed scene. The *prediction-centric* variant instead uses future VGGT features as a predictive target for guidance. In contrast, our method learns a predictive geometric representation from dense geometric features and further extracts action-aware information from geometric transitions.

As shown in Tab. 14, the input-centric and prediction-centric variants achieve average success rates of 63.5% and 65.6%, respectively, while CPR reaches 71.9%. The consistent improvement suggests that explicitly incorporating current geometry or future geometric prediction alone does not fully capture the complementary information needed for action generation. By organizing predictive geometry together with action-aware information extracted from geometric transitions, CPR provides a more effective interface for downstream policy learning.

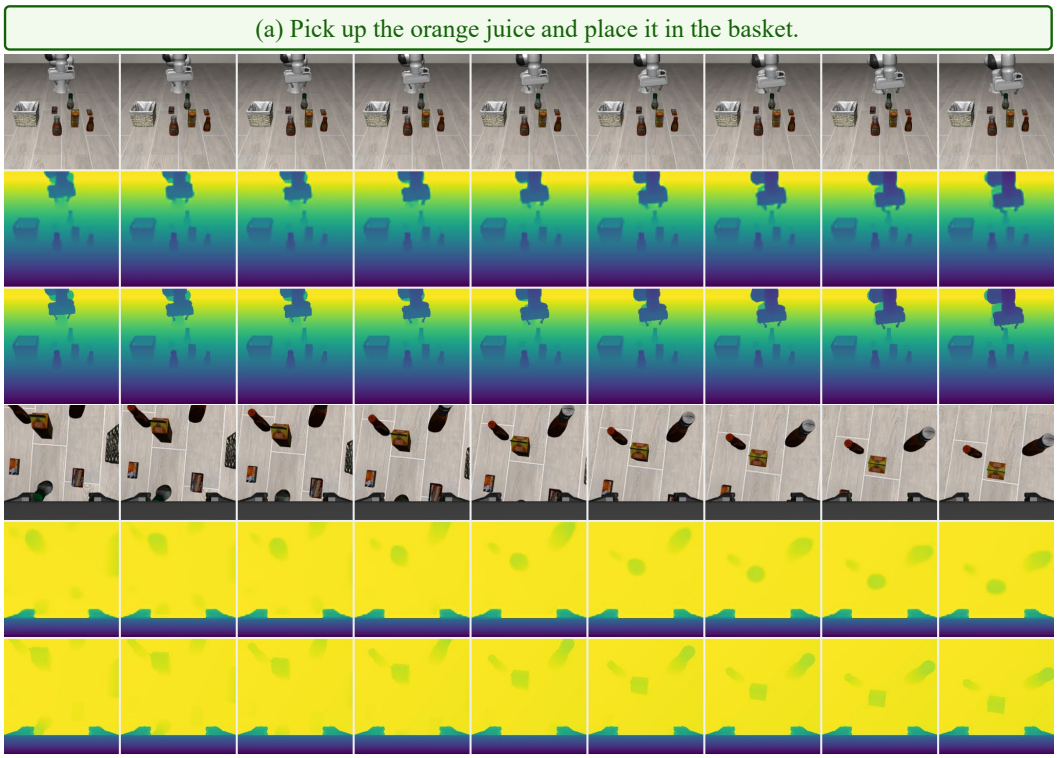
Qualitative Analysis of Predictive Geometry. We further qualitatively examine whether the geometry-query pathway transfers meaningful predictive geometric information into the downstream policy. Following the protocol described in the main paper, we freeze the post-trained policy and extract its predicted geometry-query features from current-time inputs. A lightweight DPT-based depth probe (Ranftl et al., 2021) is then trained on these frozen features, while keeping the VLA backbone and the geometry-query representations fixed. Since ground-truth depth annotations are unavailable for the evaluated observations, we use pseudo-depth maps generated by Depth Anything (Lin et al., 2025) as supervision for the probe.

As illustrated in Fig. 13, we provide qualitative results on both LIBERO and real-world manipulation scenes. For the LIBERO examples, the geometry-query features recover major spatial structures across both the main and wrist views, including object boundaries, relative depth relationships, and scene layout. Similar behavior is observed in the real-world examples across the main view and the two wrist views, suggesting that the learned geometry-query pathway preserves meaningful spatial structure under different embodiments and visual configurations. Importantly, these geometric features are inferred solely from current-time inputs, while their predictive supervision during post-training is constructed from future-informed representations. The qualitative results therefore

Table 14: Controlled comparison of different geometry-guidance paradigms under the same downstream VLA framework on the WidowX evaluation. The input-centric variant uses current-frame VGGT features as policy input, while the prediction-centric variant predicts future VGGT features as auxiliary guidance.

Guidance Paradigm	Put Spoon	Put Carrot	Stack Cube	Put Eggplant	Avg.
Input-centric	83.3	66.7	58.3	45.8	63.5
Prediction-centric	87.5	62.5	62.5	**50.0**	65.6
Ours	**95.8**	**70.8**	**70.8**	**50.0**	**71.9**

provide evidence that the geometry-query pathway transfers predictive geometric information into the downstream policy, rather than merely encoding the current visual appearance.



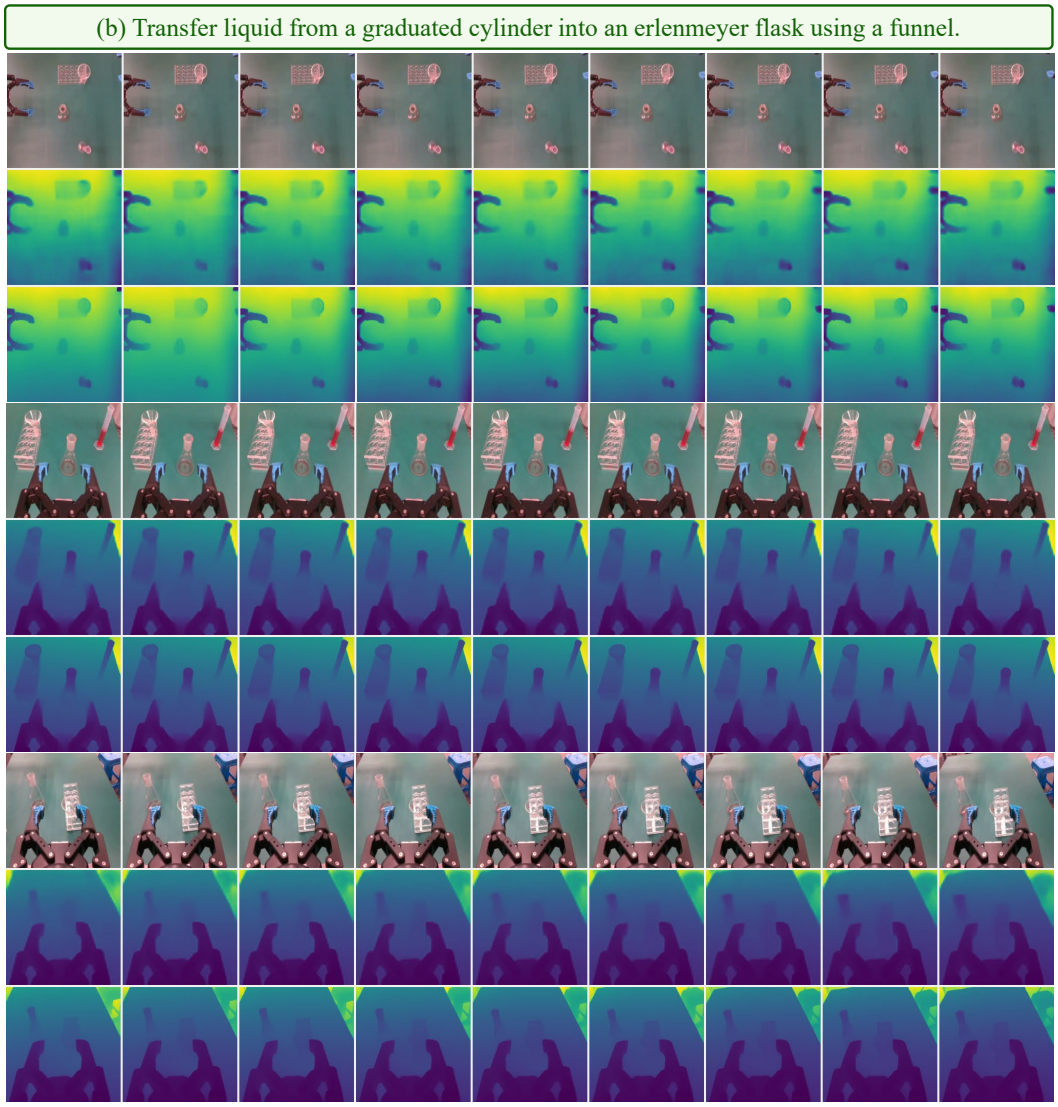


Figure 13: Qualitative visualization of predictive geometry transferred through the geometry-query pathway. The first panel shows examples from LIBERO, where each example contains a main camera view (upper half) and a wrist-camera view (lower half); within each view, the rows from top to bottom show the RGB observation, the depth prediction decoded from the predicted geometry-query features, and the pseudo-depth reference generated by Depth Anything. The second panel shows real-world manipulation examples with three camera views, corresponding to the main view, the left-arm wrist view, and the right-arm wrist view. For each view, the rows from top to bottom follow the same order: RGB observation, predicted depth, and pseudo-depth reference. The results illustrate that the geometry-query pathway can recover meaningful object geometry and relative spatial structure across both simulated and real-world settings.

B Training Details

B.1 Datasets

For predictive representation pretraining, we construct a mixed training corpus that combines robot manipulation datasets with additional general-purpose data. The robot manipulation data are designed to provide diverse interaction trajectories with explicit visuomotor structure and action supervision. We organize these data into two broad groups. The first group consists primarily of conventional robot manipulation datasets, including Open X-Embodiment (OXE)(Collaboration, 2023), RoboTwin(Chen et al., 2025), and LIBERO(exclude LIBERO-90)(Liu et al., 2024). These

datasets provide diverse robotic interaction trajectories and action-labeled manipulation behaviors, which expose the representation model to a broad range of object interactions, robot motions, and action-conditioned scene changes. The second group consists of ego data, which provide substantially broader interaction diversity and complementary visuomotor observations beyond conventional fixed-base robot manipulation datasets(Jin et al., 2026; Grauman et al., 2022). In addition, we incorporate several general-purpose datasets(Goyal et al., 2017) to further increase the diversity of visual scenes and improve the model's ability to capture geometric structure across a broader range of environments. We use the mixed corpus throughout representation pretraining rather than restricting individual training stages to a specific data source. This combination provides both manipulation-oriented action supervision and diverse visual-geometric observations, allowing the predictive representations to learn scene structure and its evolution from a broader distribution of interaction data.

Table 15: Our training data mixture used during the pretraining phase.

Dataset	Ratio
UTAustin Mutex (Shah et al., 2023)	2.3
DLR EDAN Shared Control (Quere et al., 2020)	0.7
Taco Play (Rosete-Beas et al., 2023; Mees et al., 2022)	2.3
Austin Sirius Dataset (Liu et al., 2025)	1.5
Roboturk (Mandlekar et al., 2019)	1.6
Austin Sailor Dataset (Nasiriany et al., 2022)	1.7
Berkeley Fanuc Manipulation (Zhu et al., 2023a)	1.9
UCSD Kitchen Dataset (Yan et al., 2023)	0.6
LIBERO (Liu et al., 2024)	3.4
Viola (Zhu et al., 2023b)	0.8
Austin Buds Dataset (Zhu et al., 2022)	0.2
Bridge (Walke et al., 2023; Ebert et al., 2021b)	11.6
FMB Dataset (Luo et al., 2025)	1.3
Furniture Bench Dataset (Heo et al., 2025)	2.3
NYU Franka Play Dataset (Cui et al., 2022)	1.8
Robotwin (Chen et al., 2025)	7.3
Berkeley Cable Routing (Luo et al., 2024)	1.0
BC-Z (Jang et al., 2022)	5.8
Berkeley Autolab UR5 (Chen et al., 2023)	2.3
Stanford Hydra Dataset (Belkhale et al., 2023)	4.3
Fractal (Brohan et al., 2022)	10.7
Jaco Play (Dass et al.)	2.1
DobbE (Shafiullah et al., 2023)	0.5
Ego4D (Grauman et al., 2022)	12.3
UMI (Jin et al., 2026)	10.8
Something-Something V2 (Goyal et al., 2017)	8.9

B.2 Training Protocol

Pretraining Configurations. We train the predictive representation model using a progressive multi-stage training protocol. Rather than introducing all objectives simultaneously, we gradually add the supervision objectives according to the role they play in shaping the complementary representations. This curriculum first establishes structurally informative geometric representations, then encourages temporal contextual modeling and transition prediction, and finally introduces action supervision to extract the motor representation.

1) Stage 1: Geometry Reconstruction. We first train the representation model using only the geometry reconstruction objective $\mathcal{L}_{\text{rec}}$. At this stage, the mixed robot manipulation and general-purpose data are used to establish an initial predictive geometric representation. Since the representation is obtained through hierarchical tokenization and temporal compression, the reconstruction objective provides a direct signal for preserving the geometric information contained in the pretrained foundation features.

2) Stage 2: Masked Geometric Representation Prediction. After the representation acquires basic geometric structure, we introduce the masked geometric state prediction objective $\mathcal{L}_{\text{mask}}$. Masked temporal slots are inferred from the available contextual states, encouraging the model to capture dependencies across the interaction sequence. This stage improves the ability of the predictive geometric representation to infer missing geometric information from temporal context.

3) Stage 3: Geometric Transition Prediction. We next introduce the geometric transition prediction objective $\mathcal{L}_{\text{trans}}$. In this stage, the representation model is explicitly trained to predict changes between consecutive predictive geometric states. This encourages the learned representation not only to describe scene structure but also to capture how the geometric configuration evolves over time.

4) Stage 4: Motor Representation Learning. Finally, we introduce the action prediction objective $\mathcal{L}_{\text{act}}^{\text{pre}}$ using conventional robot manipulation data that provide temporally aligned action supervision. Motion queries attend to the geometric transitions $\Delta\mathbf{G}$, and the resulting motor representation is supervised with the corresponding action chunks. This final stage extracts action-aware information from the learned geometric transitions and completes the learning of the complementary predictive geometric and motor representations.

Unless otherwise specified, each stage is initialized from the parameters obtained in the preceding stage. We set $\lambda_{\text{mask}} = 0.5$, $\lambda_{\text{trans}} = 0.5$, and $\lambda_{\text{vc}} = 0.06$. This progressive training strategy allows the model to first establish a structurally informative geometric representation before introducing increasingly task-oriented temporal and action supervision. We perform representation pretraining on 32 NVIDIA H200 GPUs with a global batch size of 512. The base learning rate is set to 1×10^{-5} with learning-rate decay, and the model is trained for a total of 120K optimization steps. The pretrained 3D foundation encoder remains frozen throughout representation pretraining.

Post-training Configurations. For downstream post-training, the pretrained predictive representation model is kept frozen. For each training clip, it is used to construct the two privileged targets: the predictive geometric representation $\mathbf{G}$ and the motor representation $\mathbf{Z}$. We initialize the downstream VLA visual-language backbone from Qwen3-VL-4B-Instruct (Yang et al., 2025). The native action-generation objective of the corresponding VLA is retained throughout post-training, while the representation-guidance objectives are jointly optimized with it. We use a linear warm-up schedule for the first 5,000 optimization steps. To facilitate stable adaptation, we further apply a cosine decay schedule to the coefficient of the latent motor-alignment loss, λ_{align}, gradually reducing the strength of latent alignment supervision as training progresses. Importantly, the pretrained representation model is not used by the downstream policy at inference time. After post-training, the policy operates solely on its standard current-time inputs, with the predictive and motor representations serving only as privileged supervision during training.

C Model Details

C.1 Hierarchical Geometric Tokenization

We use a hierarchical tokenization process to convert dense multi-view video features extracted by VGGT (Wang et al., 2025) into a compact and temporally organized geometric token sequence. Given a multi-frame, multi-view input clip, each RGB image is first resized and padded following the input preprocessing procedure of VGGT, resulting in a resolution of 518×518. The processed images are then independently fed into a frozen VGGT encoder to obtain dense geometric features for each frame-view pair. Since the number of available camera views varies across the pretraining datasets, we use a maximum of three input views at each time step. Samples with fewer than three views are zero-padded to the maximum view number, and the corresponding padded views are masked throughout the subsequent attention operations. After view-level compression, we further aggregate frame-level geometric states along the temporal dimension. Inspired by hierarchical temporal compression designs (Wan et al., 2025), we retain the first frame $\mathbf{H}_0$ as an initial-observation anchor and group every four subsequent frames into a temporal slot. Specifically, for each temporal group, a set of learnable temporal queries performs cross-attention over the frame-level geometric features within the corresponding group. The frame features serve as keys and values, while the learnable temporal queries serve as queries. Learnable frame-position embeddings are added before temporal aggregation to preserve the relative temporal ordering of frames within each group.

We further construct the temporal attention mask to match the grouping structure of the temporal slots. Frame features belonging to the same temporal group are mutually visible to the corresponding temporal queries, whereas features from different groups are prevented from attending to each other during this local aggregation stage. Within each temporal group, padded frame positions are additionally masked so that invalid frames do not affect the resulting representation. The first frame is treated as a dedicated temporal anchor and is retained without being merged with subsequent frames.

Through this hierarchical view- and time-wise compression, dense multi-view VGGT features are progressively transformed into a compact tokenized geometric sequence S, where the first temporal slot corresponds to the initial observation and each subsequent slot summarizes a group of four consecutive frames. This hierarchical design substantially reduces the number of geometric tokens while preserving both view-specific information and temporal ordering. More importantly, the resulting sequence explicitly organizes geometric information at two levels: view aggregation captures a compact representation of the scene from multiple camera perspectives, while temporal aggregation provides a structured representation for modeling the evolution of scene geometry over the interaction sequence.

C.2 Details of FastWAM

FastWAM (Yuan et al., 2026) differs from conventional VLA models (Bjorck et al., 2025; Kim et al., 2025; Black et al., 2024), in that it jointly models future visual observations and robot actions through a generative world-action architecture. We therefore introduce additional interfaces to incorporate the two complementary predictive representations into the original FastWAM architecture while preserving its native world- and action-generation mechanisms.

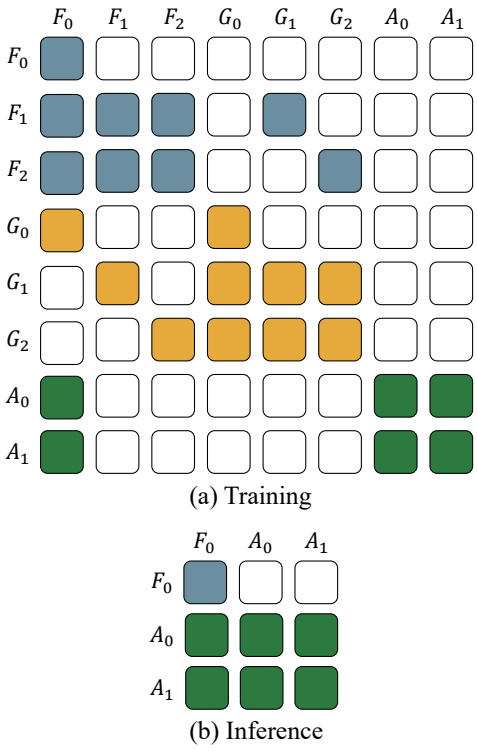


Figure 14: Attention mask of FastWAM. F denotes video frames, G indicates predictive geometric representation and A represents actions to be denoised.

Specifically, FastWAM consists of a visual world modeling branch implemented with a WAN(Wan et al., 2025) and an action-generation branch. To incorporate the predictive geometric representation, we introduce an additional geometry branch with an architecture and initialization strategy analogous to the original action branch. The geometry branch is trained to model the predictive geometric representations provided by the pretrained representation model, and operates in parallel with the original world and action modeling branches. This enables FastWAM to receive explicit supervision on the structured evolution of scene geometry in addition to its original visual world-prediction objective. The motor representation is incorporated through the action-generation pathway. Rather than introducing an additional generative branch for the motor representation, we use the pretrained motor representation as a latent guidance target for intermediate features within the action branch. Specifically, the corresponding intermediate action features are aligned with the motor representation using the same representation-level alignment objective as in the VLA setting. This provides action-aware geometric information directly to the branch responsible for generating robot actions, while leaving the native action modeling objective of FastWAM unchanged.

Because the predictive geometric representation is organized into temporal slots, an additional temporal alignment mechanism is required to establish correspondence between the geometric representation and the visual tokens in FastWAM. We therefore introduce a structured attention mask that explicitly couples temporally corresponding video and geometry slots. As illustrated in Fig. 14, each

video slot is allowed to attend to its corresponding geometry slot, enabling the visual representation at a given temporal position to access the geometric information associated with the same stage of the interaction. Conversely, each geometry slot is also allowed to attend to the corresponding video slot. This bidirectional local visibility establishes an explicit correspondence between visual and geometric representations while preventing a geometry slot from indiscriminately attending to temporally unrelated video tokens. The resulting attention pattern therefore encourages localized video-geometry interaction: Such slot-level correspondence allows the model to jointly reason about visual appearance and geometric structure at matched temporal positions, while preserving the temporal organization of the predictive representation. The mask is applied in addition to the native attention structure of FastWAM and does not require additional geometric observations at inference time. During inference, the additional geometry branch and the pretrained representation model are not required to receive externally provided geometric observations. The downstream FastWAM operates using its standard current-time visual input, from which the learned model generates the future world evolution and corresponding actions. Thus, as in the VLA setting, the predictive geometric and motor representations are used to provide privileged supervision during training, while no extra geometry input is required during deployment.

D Limitations and Future Work

Our current framework primarily focuses on visual geometry and its temporal evolution to construct predictive and action-aware representations. An interesting direction for future work is to further incorporate tactile information into the same representation-learning framework. Tactile observations provide complementary cues about contact states, interaction forces, and fine-grained object configurations that may not be fully observable from vision alone. Integrating visual geometry with tactile feedback could therefore provide a richer representation of scene evolution, particularly for contact-rich manipulation and interactions involving occlusion or subtle physical state changes. More broadly, exploring multimodal predictive representations that jointly capture visual, geometric, and tactile evolution may further strengthen the connection between world modeling and action generation.

E Simulation and Real-World Evaluation Details

E.1 Simulation Evaluation Details

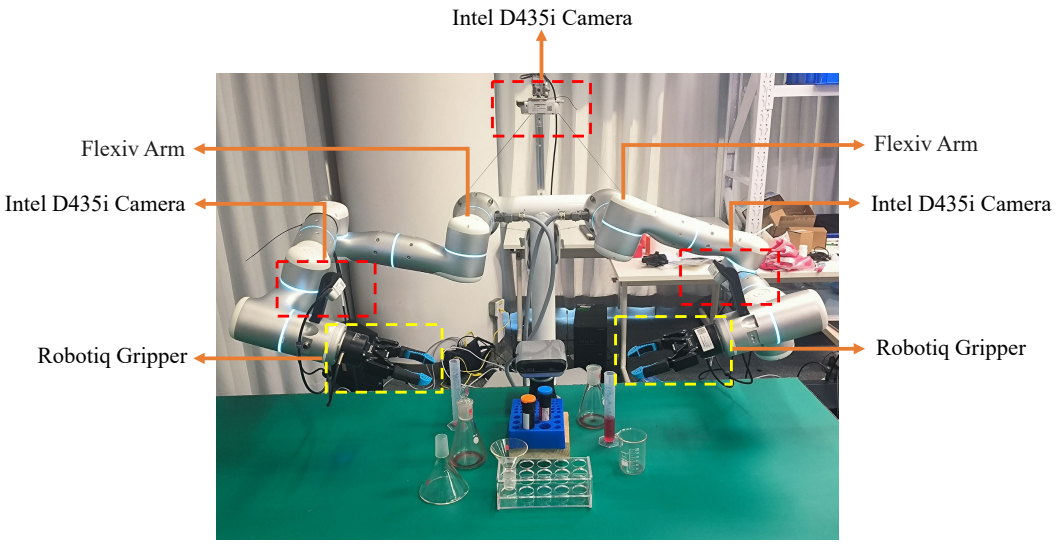
We provide detailed descriptions of the simulation benchmarks employed in our experiments.

SimplerEnv Benchmark. To assess cross-embodiment generalization and investigate the correlation between simulation performance and real-world performance, we conduct experiments on the SimplerEnv benchmark (Li et al., 2024b) using the WidowX robot.

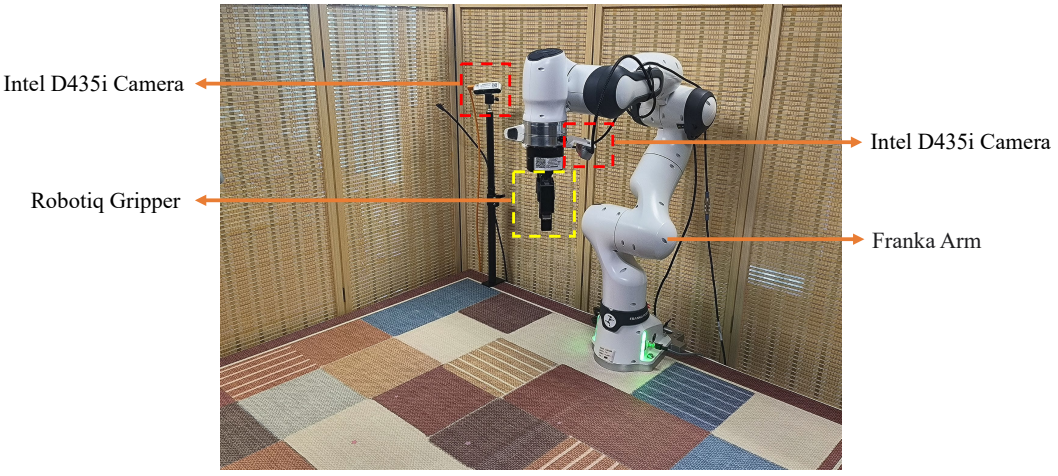
For the WidowX robot, we use the *Visual Matching* setting to ensure consistency with the data distribution of the Bridge dataset(Ebert et al., 2021a). We evaluate the robot on four manipulation tasks: *Put Carrot on Plate* (*Put Carrot*), *Put Spoon on Towel* (*Put Spoon*), *Stack Green on Yellow* (*Stack Cube*), and *Put Eggplant in Basket* (*Put Eggplant*). All experiments on SimplerEnv strictly follow the standard evaluation protocol established by CogAct (Li et al., 2024a).

LIBERO Benchmark. We additionally evaluate our method on the LIBERO benchmark (Liu et al., 2024) to examine different dimensions of temporal reasoning and knowledge transfer. In particular, we consider four task suites: *Spatial*, which evaluates spatial reasoning under different object configurations; *Object*, which focuses on object-centric manipulation involving diverse object categories; *Goal*, which evaluates goal-conditioned instruction following in static environments; and *Long*, which targets long-horizon planning over multi-stage manipulation tasks. Each suite consists of 10 tasks, and each task is provided with 50 human-teleoperated demonstrations.

LIBERO-Plus Benchmark. LIBERO-Plus (Fei et al., 2025) evaluates the robustness of policies trained on LIBERO under controlled perturbations at test time. It introduces variations across seven dimensions: *Camera Viewpoints*, *Robot Initial States*, *Language Instructions*, *Light Conditions*, *Background Textures*, *Sensor Noise*, and *Object Layouts*. These perturbations are categorized into



(a) Flexiv platform



(b) Franka platform

Figure 15: Bimanual and single-arm manipulation platforms.

five difficulty levels, resulting in a total of 10,030 evaluation instances. We directly evaluate the checkpoint trained on LIBERO without performing any adaptation to the LIBERO-Plus perturbations. This evaluation protocol is designed to examine whether the learned policy representations remain effective when the observation distribution, robot initialization, or instruction formulation differs from those encountered during training.

E.2 Real-World Robot Evaluation Details

Real-World Robot Setup. We conduct real-world experiments on both bimanual and single-arm manipulation platforms to evaluate the proposed CPR guidance under different embodiments and interaction requirements, as shown in Fig.15. For the bimanual setting, we use dual Flexiv robotic arms, each equipped with a Robotiq 2F-85 gripper. Visual observations are captured using three Intel RealSense D435i RGB-D cameras, including two wrist-mounted cameras and one head-mounted camera. The combination of third-person and wrist-view observations provides complementary global and local views of the manipulation scene, which is particularly important for bimanual tasks involving object coordination and fine-grained manipulation.



Figure 16: Visualization of four real-robot manipulation tasks.

For the single-arm setting, we use a Franka Research 3 robotic platform, equipped with a 7-DoF manipulator and joint-level torque and force sensing. Visual observations are obtained from an Intel RealSense D435i RGB-D camera mounted in an eye-on-base configuration, providing a global view of the manipulation workspace. We additionally mount a second RealSense camera on the robot wrist to provide close-up observations of the manipulated objects and end-effector interaction. The robot is equipped with a Robotiq 2F-85 gripper for end-effector manipulation. These multi-view configurations provide both scene-level spatial context and fine-grained object-interaction cues for downstream action generation.

Baseline Models. We compare our method with representative VLA baselines using different pre-trained visual-language backbones. **OpenVLA**(Kim et al., 2024) is initialized from the publicly released checkpoint pretrained on the OXE(Collaboration, 2023) dataset. For our real-world experi-

ments, we extend the original input interface to jointly process third-person and wrist-view images. We also evaluate π_0(Black et al., 2024), initialized from the official JAX checkpoint pretrained on the DROID(Khazatsky et al., 2024) dataset. Unless otherwise specified, the same downstream training protocol is used when incorporating CPR into the corresponding baseline model.

Real-World Task Suites. As shown in Fig.16, we evaluate four real-world manipulation tasks covering bimanual coordination, geometric perception, contact-sensitive manipulation, and long-horizon task execution. The first three tasks are performed with the bimanual Flexiv setup, while the final task is conducted using the single-arm Franka platform.

- **Remove a Stopper.** The robot removes a stopper from an erlenmeyer flask and places it at a designated holder. The task first requires the left arm to grasp and stabilize the flask, after which the right arm grasps the stopper and places it accurately into the corresponding circular opening on the storage rack. Successful execution therefore requires precise relative positioning between the two arms and the manipulated objects, as well as coordinated bimanual actions during the stopper removal and placement stages.
- **Pick up Tubes into Box.** The robot sequentially transfers two tubes into a blue box using both arms. The left arm first grasps the brown tube and places it into the box, followed by the right arm grasping the blue tube and placing it into the corresponding location. The task requires accurate perception of object locations and target geometry, together with appropriate end-effector positioning for grasping and insertion. Since the two manipulation stages involve different objects and hand-object configurations, the policy must maintain consistent spatial grounding throughout the sequential execution.
- **Transfer Liquid into an Erlenmeyer.** The robot transfers liquid from a graduated cylinder into an erlenmeyer flask using a funnel. The right arm first grasps the funnel and inserts it into the flask with sufficient spatial precision. It then grasps the liquid-filled graduated cylinder and carefully tilts it to pour the liquid through the funnel. After the transfer, the right arm stabilizes the flask while the left arm removes the funnel and returns it to a designated holder on the rack. This task combines fine-grained geometric alignment, coordinated bimanual manipulation, sequential object-role changes, and long-horizon execution, making the relative spatial configuration of the objects and tools important throughout the interaction.
- **Toast Breads and Make a Burger.** This task is performed on the single-arm Franka platform and involves a longer sequence of manipulation stages. The robot first places two slices of bread into the toaster, then transfers one slice onto a plate. It subsequently grasps a meat patty from the rack and stacks it onto the bread, before placing the remaining bread slice on top to complete the burger. The task requires maintaining the correct object order and intermediate states across multiple manipulation stages, while accurately positioning each object for grasping, placement, and stacking.

Real-World Data Collection. For downstream post-training, we collect 100 demonstrations for each of the three bimanual tasks and 60 demonstrations for the single-arm task. The collected trajectories are used to train the downstream policies.

F Details of Compared Methods

We compare our approach with representative methods from three groups: standard policy baselines, input-centric methods, and prediction-centric methods. This categorization distinguishes how geometric or predictive information is introduced into policy learning: standard policies rely primarily on the model's native visual-language-action inputs, input-centric methods explicitly augment the current observation with geometric or spatial information, while prediction-centric methods introduce additional predictive or auxiliary supervision to encourage the policy to learn future, temporal, or geometric structure.

Standard Policy Baselines. These methods generate robot actions from visual observations, language instructions, and, when applicable, proprioceptive states, without introducing an explicit future-representation or auxiliary geometric prediction target.

- **OpenVLA-OFT**(Kim et al., 2025). OpenVLA-OFT (Kim et al., 2025) is an optimized post-training recipe for OpenVLA(Kim et al., 2024) that studies action representations, decoding strategies, and fine-tuning objectives. Its policy uses continuous action representations together with action chunking, parallel decoding, and an ℓ_1 regression objective for efficient multi-step action prediction. These design choices provide a strong and efficient baseline for adapting pretrained VLA models to downstream manipulation tasks.
- π_0(Black et al., 2024). π_0 is a general-purpose VLA that augments a pretrained vision-language model with a dedicated action expert. The action expert generates continuous action chunks using conditional flow matching, allowing the model to produce smooth, high-frequency robot trajectories across multiple embodiments. The model is pretrained on diverse multi-robot manipulation data and subsequently adapted to downstream tasks through post-training.

Input-Centric Methods. Input-centric methods enhance the policy primarily by providing additional geometric or spatial information derived from the current observation. Rather than constructing a predictive target from future observations, these approaches enrich the policy's current-time spatial representation through depth, point clouds, geometric embeddings, or spatially structured tokens.

- **GeoVLA**(Sun et al., 2025). GeoVLA augments a VLA with explicit 3D geometric information derived from depth observations. Depth maps are converted into point clouds and processed by a dedicated Point Embedding Network to obtain 3D geometric embeddings. These embeddings are combined with the vision-language representation and processed by a 3D-enhanced action expert, providing the policy with explicit spatial information for action generation.
- **3D-CAVLA** (Bhat et al., 2025). 3D-CAVLA enhances VLA spatial awareness by incorporating depth perception and task-oriented visual context. The framework combines depth information with chain-of-thought reasoning and region-of-interest detection to focus the policy on task-relevant objects and spatial relationships. Its goal is to improve scene-context understanding and generalization to unseen manipulation tasks.
- **SpatialVLA** (Qu et al., 2025). SpatialVLA explicitly introduces spatial representations into a generalist robot policy through Ego3D Position Encoding, which injects 3D positional information into the visual representation. It further uses Adaptive Action Grids to represent spatial robot movements in a discretized and robot-adaptive action space, facilitating spatially grounded manipulation and transfer across robot setups.
- **FALCON** (Zhang et al., 2025b). FALCON injects spatially rich tokens into the action-generation pathway of a VLA through an Embodied Spatial Model and a Spatial-Enhanced Action Head. The spatial branch can use RGB input alone and can additionally incorporate geometric modalities such as depth or camera pose when available. By consuming spatial tokens directly at the action head rather than concatenating them into the vision-language backbone, FALCON aims to strengthen 3D spatial understanding while preserving the semantic representation of the VLA.

Prediction-Centric Methods. Prediction-centric methods introduce additional predictive or temporally structured supervision during policy learning. Rather than only enriching the current observation, these approaches encourage the policy to capture future scene evolution, depth structure, robot motion, or other predictive information that can support action generation.

- **4D-WAM** (Yang et al., 2026). 4D-WAM introduces spatiotemporal supervision into World Action Models using 3D trajectory fields. It employs motion alignment to encourage consistency of feature changes across adjacent frames and destination alignment to guide the model toward long-horizon target information. These objectives provide local motion and long-range destination supervision through representation alignment, without requiring additional geometric inputs at inference time.
- **ELAN4D** (He et al., 2026). ELAN4D provides embodiment-centric 4D supervision to VLA policies through future robot keypoint tracks. The tracks are derived from proprioceptive states using forward kinematics and describe the 3D displacement of robot joints

and end-effectors over time. A lightweight auxiliary track decoder injects this predictive spatiotemporal supervision into the action-generation pathway during training, while the auxiliary branch can be removed at inference time.

- **QDepth-VLA** (Li et al., 2025b). QDepth-VLA introduces depth prediction as an auxiliary supervision task for VLA training. A dedicated depth expert predicts quantized latent depth tokens obtained from a VQ-VAE representation of depth maps, encouraging the policy to learn depth-aware features that capture geometric structure relevant to manipulation. Unlike methods that directly provide depth as an additional input, QDepth-VLA uses depth prediction to shape the learned representation through auxiliary supervision.
- **WoG** (Su et al., 2026). World Guidance (WoG) maps future observations into compact condition representations and uses them to guide action generation. The VLA is trained to jointly predict the compressed future condition and the corresponding action trajectory, introducing future-world modeling supervision in a compact condition space rather than requiring full future-image reconstruction. This design aims to retain predictive information that is useful for action generation while keeping the prediction target compact.
- **WorldVLA** (Cen et al., 2025). WorldVLA integrates visual world modeling and action generation within a unified autoregressive framework. The world model predicts future visual observations conditioned on the current scene and actions, while the action model predicts subsequent actions from visual observations. The two components are jointly trained so that future-world modeling and action generation provide mutual supervision. WorldVLA further introduces selective attention masking to reduce error propagation from previously generated actions during autoregressive action prediction.

G AI USE STATEMENT

In this work, we used LLM-based assistants to polish the wording of text. We have not used generative Al tools to produce research ideas or experimental designs. No generative Al tool had access to non-public data. We have reviewed all AI-assisted work: all Al-assisted text was checked, corrected, and approved by the authors, and all Al-assisted code was verified for correctness, with every reported result produced by the authors' own runs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative Al.