SIP-WAM: ENHANCING WORLD ACTION MODELS WITH INTERACTION-AWARE SPATIOTEMPORAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

World Action Models (WAMs) have demonstrated strong potential for robot policy learning by jointly modeling future visual observations and actions. However, existing RGB-centric WAMs remain limited in learning interaction-aware spatiotemporal representations. Temporally sparse visual prediction provides insufficient supervision for fine-grained interaction dynamics, whereas temporally dense RGB prediction incurs substantial computational overhead. Meanwhile, full-frame RGB objectives do not explicitly prioritize task-relevant interaction regions, limiting interaction-aware spatial grounding. To address these limitations, we introduce SIP-WAM, a framework that enhances WAMs with Interaction-aware Spatiotemporal Prediction. Specifically, alongside the WAM backbone, an Interaction DiT jointly predicts the current spatial locations and temporally dense future trajectories of a sparse set of points uniformly sampled on interaction participants, providing efficient and control-relevant interaction supervision for both spatial grounding and fine-grained temporal dynamics. To avoid additional inference overhead, we couple the two branches through layerwise asymmetric joint attention, enabling the Interaction DiT to guide representation learning during training and be removed at inference. Extensive experiments on RoboCasa, LIBERO-Plus, and real-world single-arm and bimanual manipulation demonstrate strong policy performance and generalization, validating the effectiveness of interaction-aware spatiotemporal supervision for world action modeling.

1 INTRODUCTION

In contrast to conventional Vision-Language-Action (VLA) policies (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024), which are optimized primarily for direct action prediction from visual observations and language instructions, World Action Models (WAMs) extend robot policy learning to a joint generative formulation of future visual observations and robot actions (Kim et al., 2026b; Ye et al., 2026b;a; Li et al., 2026a). By incorporating future visual prediction into policy optimization, this formulation encourages the acquisition of predictive representations that capture the physical dynamics underlying manipulation, thereby supporting more robust and generalizable robot manipulation (Kim et al., 2026b; Ye et al., 2026b; Zhang et al., 2026).

However, existing WAMs that rely primarily on future RGB prediction face limitations in learning interaction-aware spatiotemporal representations. On the one hand, most current WAMs employ temporally sparse visual prediction, with future RGB targets sampled at a lower temporal frequency than that of the action sequence, providing limited supervision for the intermediate motion of control-relevant entities (Kim et al., 2026b; Yuan et al., 2026; Ye et al., 2026a). Consequently, *fine-grained temporal interaction dynamics* remain insufficiently constrained. Increasing the temporal density of RGB prediction (Zhen et al., 2026) better constrains intermediate evolution, while incurring substantial computational overhead, with model capacity also devoted to predicting static, control-irrelevant visual content (Li et al., 2026b). On the other hand, full-frame RGB objectives require WAMs to model task-relevant visual content alongside nuisance factors such as background and illumination (Guan et al., 2026). Such objectives do not directly prioritize regions involved in the manipulation task, thereby limiting *interaction-aware spatial grounding*.

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.

Point-trajectory prediction offers a compact alternative for providing temporally dense motion guidance. However, representative formulations forecast trajectories from supplied query points or predefined image grids (Wen et al., 2023; Yang et al., 2025; Guan et al., 2026). Besides lacking direct spatial grounding supervision, most of these approaches do not specifically focus on the motion of task-specified interaction participants, resulting in limited emphasis on the instructed interaction. Some methods construct object-centric query points using pretrained detection models (Xu et al., 2024). However, object localization remains externally supplied, and current point coordinates serve as conditions rather than prediction targets, leaving direct supervision for interaction-aware spatial grounding outside the trajectory prediction objective.

To address these limitations, we introduce **SIP-WAM**, a framework that enhances WAMs with Interaction-aware Spatiotemporal Prediction. We formulate a joint prediction objective that addresses both spatial interaction grounding and temporal interaction dynamics modeling. As shown in Fig. 1, SIP-WAM jointly predicts the current spatial locations and temporally dense future trajectories of sparse points on the interaction participants, namely the task-relevant entities involved in the manipulation. This joint prediction objective enables SIP-WAM to promote interaction-aware spatial grounding while efficiently constraining fine-grained temporal interaction dynamics, encouraging more control-relevant spatiotemporal representations. Specifically, SIP-WAM introduces an Interaction DiT that predicts these current-to-future coordinate sequences through flow matching, with group-wise Hungarian matching aligning predictions with ground truth within each participant group. The Interaction DiT and Video-Action DiT are coupled through layerwise asymmetric joint attention in a MoT-style architecture (Liang et al., 2024), while the Video-Action DiT retains the original WAM objective. This design allows our proposed supervision to update the WAM backbone during joint training without introducing a forward dependency on the Interaction DiT. The Interaction DiT can therefore be removed at inference with no additional computational overhead.

We evaluate SIP-WAM on the RoboCasa (Nasiriany et al., 2024) and LIBERO-Plus (Fei et al., 2025) simulation benchmarks and on real-world single-arm and bimanual manipulation tasks. SIP-WAM achieves strong performance across these settings, outperforming advanced VLA and WAM methods, thereby demonstrating the effectiveness of interaction-aware spatiotemporal prediction for robot policy learning. Notably, real-world evaluations show that SIP-WAM more effectively leverages human manipulation videos for cross-embodiment transfer than the strong RGB-only Cosmos Policy (Kim et al., 2026b) baseline. In summary, our contributions are as follows:

- We introduce SIP-WAM, a framework that enhances WAMs with Interaction-aware Spatiotemporal Prediction, providing explicit supervision for interaction-aware spatial grounding and fine-grained temporal interaction dynamics.
- We develop an Interaction DiT coupled with the WAM backbone through layerwise asymmetric joint attention, enabling interaction-aware spatiotemporal supervision during joint training without additional inference overhead.
- We demonstrate consistent improvements across the RoboCasa and LIBERO-Plus simulation benchmarks and real-world manipulation tasks, together with more effective cross-embodiment transfer from human manipulation videos.

2 Related Works

2.1 World Action Models

World Action Models (WAMs) incorporate future-observation prediction into robot policy learning, coupling visual dynamics modeling with robot action generation. Earlier approaches follow a two-stage paradigm, first generating video plans and subsequently deriving robot actions using inverse dynamics models (Du et al., 2023; Zhou et al., 2024) or policies conditioned on generated visual goals or videos (Du et al., 2024; Bharadhwaj et al., 2024a). More recent approaches jointly model future observations and robot actions within a unified or closely coupled framework (Guo et al., 2024; Li et al., 2025; Zhu et al., 2025; Liang et al., 2025; Kim et al., 2026b; Li et al., 2026a; Ye et al., 2026b). Beyond full-frame RGB prediction, SIP-WAM enhances WAMs with Interaction-aware Spatiotemporal Prediction while preserving the original WAM objective and introducing no additional inference overhead.

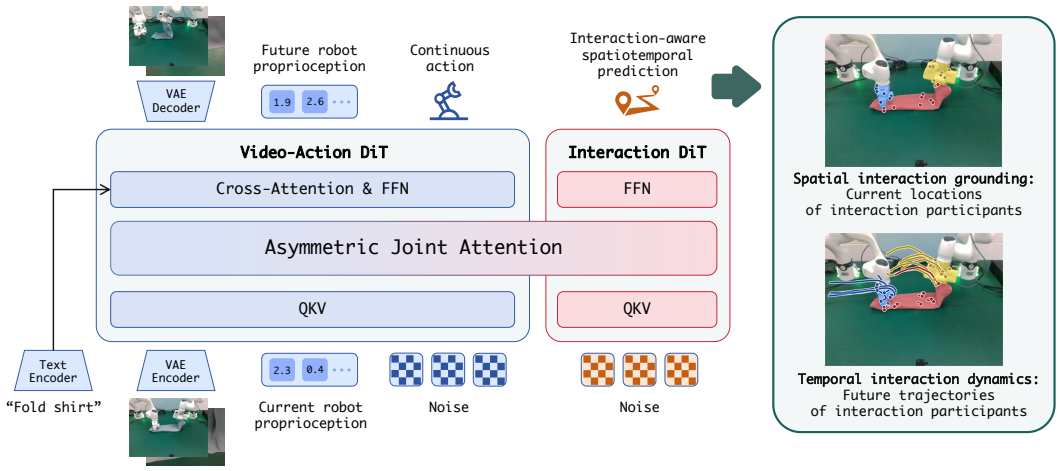

Figure 1: **Overview of SIP-WAM.** SIP-WAM introduces an Interaction DiT alongside the WAM backbone to jointly predict the current spatial locations and temporally dense future trajectories of sparse points uniformly sampled on the interaction participants. The Interaction DiT and Video-Action DiT are coupled layerwise through asymmetric joint attention, allowing the Interaction DiT to be removed at inference without additional computational overhead.

2.2 Trajectory Modeling for Robot Control

A growing body of work leverages point trajectories as a compact and temporally dense motion representation for robot control. Some approaches condition action policies on predicted trajectories (Wen et al., 2023; Yang et al., 2025; Xu et al., 2024; Hung et al., 2026). Another line of work translates predicted trajectories into explicit motion guidance or plans for robot execution (Yuan et al., 2024; Bharadhwaj et al., 2024b; Gao et al., 2025). Recent work integrates trajectory prediction directly into policy learning, jointly optimizing trajectory and action prediction (Li et al., 2026b; Kim et al., 2026a; Guan et al., 2026). Beyond future trajectory prediction, SIP-WAM also predicts current point coordinates and focuses on task-specified interaction participants, providing explicit supervision for interaction-aware spatial grounding and fine-grained temporal interaction dynamics.

3 Method

As illustrated in Fig. 1, SIP-WAM introduces an Interaction DiT alongside the WAM backbone for Interaction-aware Spatiotemporal Prediction. Within a MoT-style two-branch architecture, the Interaction DiT predicts the current spatial locations and temporally dense future trajectories of sparse points uniformly sampled on the interaction participants and is coupled layerwise with the Video-Action DiT through asymmetric joint attention. We jointly optimize the original WAM objective and our proposed interaction-aware spatiotemporal prediction objective, with group-wise Hungarian matching applied to the latter. The asymmetric information flow allows the Interaction DiT to be removed at inference without incurring additional computational overhead.

3.1 Preliminaries

World Action Models. We adopt Cosmos Policy (Kim et al., 2026b) as the WAM backbone of SIP-WAM. Cosmos Policy builds upon the pretrained Cosmos-Predict2 video model (Ali et al., 2025) and adapts it for robot policy learning. At control step t, the model receives multi-view RGB observations $\mathbf{I}_t = \{\mathbf{I}_t^{(n)}\}_{n=1}^{N_{\text{cam}}}$, the robot proprioception $\mathbf{q}_t$, and the language instruction ℓ. For an action prediction horizon of H steps, we denote the current state and the future state at the end of the horizon as $\mathbf{s}_t = (\mathbf{I}_t, \mathbf{q}_t)$ and $\mathbf{s}_{t+H} = (\mathbf{I}_{t+H}, \mathbf{q}_{t+H})$, respectively. Conditioned on $(\mathbf{s}_t, \ell)$, the WAM jointly models an action chunk, the resulting future state, and its value estimate:

$$p_\theta(\mathbf{a}_{t:t+H-1}, \mathbf{s}_{t+H}, V(\mathbf{s}_{t+H}) \mid \mathbf{s}_t, \ell)\,. \tag{1}$$

Cosmos Policy employs latent frame injection to organize VAE-encoded RGB observations together with robot proprioception, action chunks, and value estimates into a unified latent sequence. A

shared Video-Action DiT jointly denoises the action, future state, and value conditioned on the clean current state and the language instruction.

Flow Matching. Flow matching (Lipman et al., 2022; Liu et al., 2022) learns a time-dependent velocity field that defines an ODE connecting the data distribution and a simple noise distribution. Given a clean sample $\mathbf{x}_0 \sim p_{\text{data}}$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\tau \in [0, 1]$, we define the interpolated sample and target velocity as

$$\mathbf{x}_\tau = (1 - \tau)\mathbf{x}_0 + \tau\boldsymbol{\epsilon}, \qquad \mathbf{v}_\tau = \frac{\partial \mathbf{x}_\tau}{\partial \tau} = \boldsymbol{\epsilon} - \mathbf{x}_0. \tag{2}$$

Conditioned on $\mathbf{c}$, the velocity network $\mathbf{v}_\theta$ is trained via the following flow-matching objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}, \tau}\left[\|\mathbf{v}_\theta(\mathbf{x}_\tau, \tau; \mathbf{c}) - \mathbf{v}_\tau\|_2^2\right]. \tag{3}$$

SIP-WAM applies this formulation to both the Video-Action DiT and the Interaction DiT.

3.2 Interaction-Aware Spatiotemporal Prediction

We formulate **Interaction-aware Spatiotemporal Prediction** as jointly predicting the current spatial locations and future motion of interaction participants during an instructed manipulation. Intuitively, these two components specify where interaction occurs and how the interaction unfolds. In our formulation, the interaction participants comprise the manipulation subject and target, corresponding to the robot gripper and the object to be manipulated, respectively. Specifically, SIP-WAM predicts the current coordinates of sparse points uniformly sampled on each interaction participant, together with the temporally dense future trajectories of these sampled points.

Formally, let $\mathcal{E} = \{\text{sub}, \text{tar}\}$ denote the manipulation subject and target. We define the interaction-aware spatiotemporal prediction target as

$$\mathcal{Y}_t^{\text{int}} = \left\{ \underbrace{(\mathbf{P}_t^e, \mathbf{M}_t^e)}_{\text{spatial interaction grounding}}, \underbrace{(\mathbf{P}_{t+1:t+H}^e, \mathbf{M}_{t+1:t+H}^e)}_{\text{temporal interaction dynamics}} \right\}_{e \in \mathcal{E}}. \tag{4}$$

Here, $\mathbf{P}_t^e \in [0, 1]^{N_e \times 2}$ contains the normalized current coordinates of N_e sparse points sampled on participant e, providing explicit supervision for interaction-aware spatial grounding, while $\mathbf{P}_{t+1:t+H}^e \in [0, 1]^{N_e \times H \times 2}$ contains their corresponding coordinates over the future H steps, forming temporally dense point trajectories over the same horizon as the action chunk, constraining fine-grained temporal interaction dynamics. The variables $\mathbf{M}_t^e \in \{0, 1\}^{N_e}$ and $\mathbf{M}_{t+1:t+H}^e \in \{0, 1\}^{N_e \times H}$ denote the corresponding point visibility indicators.

To construct the training targets, we use SAM 3 (Carion et al., 2026) to segment the robot gripper and the instruction-specified manipulation target in real-world demonstrations. For simulation data, we replay the demonstrations and render the corresponding segmentation masks directly from the simulator. We then uniformly sample a fixed number of points within each mask and track their future trajectories and visibility over the prediction horizon using CoTracker3 (Karaev et al., 2025).

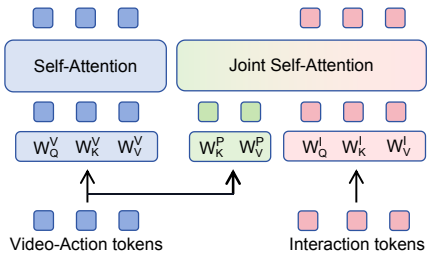

Figure 2: **Asymmetric joint attention.** W_K^P and W_V^P project Video-Action tokens to match the dimension of Interaction tokens for joint self-attention.

Unlike conventional trajectory formulations that forecast future motion from supplied query points or predefined image grids, SIP-WAM treats both current and future point coordinates as prediction targets. The model must therefore infer the current spatial positions of the task-specified interaction participants and predict how they evolve over the prediction horizon. Jointly modeling the manipulation subject and target further renders the spatiotemporal prediction interaction-aware. This unified objective equips the WAM with the ability to perform interaction-aware spatial grounding and model fine-grained temporal interaction dynamics efficiently, encouraging the backbone to learn spatiotemporal representations better aligned with robot control.

Asymmetric Joint Attention. Mixture-of-Transformers (MoT) (Liang et al., 2024) enables cross-modal information exchange through joint attention while retaining modality-specific Transformer parameters. Inspired by MoT, SIP-WAM maintains separate Video-Action and Interaction DiT branches and performs asymmetric joint attention at corresponding layers. At layer l, the attention operations of the two branches are

$$\begin{aligned} \mathbf{A}_l^{\mathrm{VA}} &= \mathrm{Attn}\big(\mathbf{Q}_l^{\mathrm{VA}}, \mathbf{K}_l^{\mathrm{VA}}, \mathbf{V}_l^{\mathrm{VA}}\big), \\ \mathbf{A}_l^{\mathrm{I}} &= \mathrm{Attn}\big(\mathbf{Q}_l^{\mathrm{I}}, \big[\mathbf{K}_l^{\mathrm{VA}\to\mathrm{I}}; \mathbf{K}_l^{\mathrm{I}}\big], \big[\mathbf{V}_l^{\mathrm{VA}\to\mathrm{I}}; \mathbf{V}_l^{\mathrm{I}}\big]\big). \end{aligned} \tag{5}$$

Here, $[\cdot;\cdot]$ denotes concatenation along the sequence length dimension, while $\mathbf{K}_l^{\mathrm{VA}\to\mathrm{I}}$ and $\mathbf{V}_l^{\mathrm{VA}\to\mathrm{I}}$ are obtained by projecting Video-Action features into keys and values with the Interaction DiT's attention dimension. The Video-Action branch attends only to its own tokens, whereas Interaction queries attend to both branches. Moreover, the two branches retain distinct attention dimensions, allowing the Interaction DiT to operate at a substantially smaller dimension. This asymmetric information flow enables interaction-aware spatiotemporal supervision to update the WAM backbone during joint training without introducing a forward dependency on the Interaction DiT, which can therefore be removed at inference without additional overhead.

3.3 Training Objective

We use group-wise Hungarian matching (Kuhn, 1955; Carion et al., 2020) to provide permutation-invariant supervision over sets of current-to-future point sequences, without imposing a fixed correspondence between predictions and ground truth. Each matching group consists of points sampled on the same interaction participant, and matching is performed independently within each group.

Let $\widehat{\mathbf{P}}^{g}_{i,t:t+H}$ and $\mathbf{P}^{g}_{j,t:t+H}$ denote the predicted clean coordinate sequence for prediction i and the ground-truth coordinate sequence for point j in group g, respectively. We obtain the optimal assignment as

$$\widehat{\pi}^{g} = \arg\min_{\pi\in\Pi_{N_g}} \sum_{i=1}^{N_g} \mathcal{C}\Big(\widehat{\mathbf{P}}^{g}_{i,t:t+H}, \mathbf{P}^{g}_{\pi(i),t:t+H}\Big), \tag{6}$$

where Π_{N_g} denotes the set of permutations over the N_g points in group g, and $\mathcal{C}$ evaluates coordinate discrepancies over the entire current-to-future sequence. The same assignment aligns the coordinate sequences and their associated visibility labels. The matched ground-truth coordinates are then used to recompute the flow-matching targets.

The interaction-aware spatiotemporal prediction objective and the overall training objective are

$$\mathcal{L}_{\mathrm{int}} = \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}}^{\mathrm{int}} + \lambda_{\mathrm{vis}}\mathcal{L}_{\mathrm{vis}}, \qquad \mathcal{L}_{\text{SIP-WAM}} = \mathcal{L}_{\mathrm{WAM}} + \lambda_{\mathrm{int}}\mathcal{L}_{\mathrm{int}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{FM}}^{\mathrm{int}}$ is the flow-matching loss for the Hungarian-matched current-to-future coordinate sequences, $\mathcal{L}_{\mathrm{vis}}$ is the binary cross-entropy loss for point visibility, and $\mathcal{L}_{\mathrm{WAM}}$ denotes the original WAM denoising objective. The coefficients λ_{FM}, λ_{vis}, and λ_{int} control the relative weights of the loss terms. The Video-Action DiT and Interaction DiT are optimized jointly.

4 Experiments

4.1 Simulation Experimental Setup

Benchmarks. We evaluate SIP-WAM on two simulation benchmarks. RoboCasa (Nasiriany et al., 2024) consists of 24 kitchen manipulation tasks, including pick-and-place, opening and closing doors and drawers, turning knobs, and pressing buttons, with 50 human-teleoperated demonstrations per task for training. Following Cosmos Policy (Kim et al., 2026b), we perform 50 evaluation rollouts per task. These rollouts are evenly distributed across five scenes with different layouts and visual styles, with 10 rollouts per scene. Two of the five scenes use styles not encountered during training. LIBERO-Plus (Fei et al., 2025) builds upon the original LIBERO benchmark to evaluate policy robustness across 10,030 task variants spanning seven perturbation categories: object layouts, camera viewpoints, robot initial states, language instructions, light conditions, background textures, and sensor noise. We train on the original LIBERO demonstrations and evaluate directly on LIBERO-Plus without additional fine-tuning. We report task success rate as the primary evaluation metric on both benchmarks.

Table 1: **RoboCasa simulation results.** We report average success rates across 24 tasks. Cosmos Policy* denotes the Cosmos Policy baseline further fine-tuned on the same replayed dataset as SIP-WAM. Best results are highlighted in **bold**.

Method	Demos / task	Avg. SR (%)
GR00T-N1 (Bjorck et al., 2025)	300	49.6
π_0 (Black et al., 2024)	300	62.5
GR00T-N1.5 (Bjorck et al., 2025)	300	64.1
GR00T-N1.5 + HAMLET (Koo et al., 2025)	300	66.4
UVA (Li et al., 2025)	50	50.0
UWM (Zhu et al., 2025)	1000	60.8
Video Policy (Liang et al., 2025)	300	66.0
FLARE (Zheng et al., 2025)	300	66.4
Cosmos Policy (Kim et al., 2026b)	50	67.1
Cosmos Policy*(Baseline)	50	69.5
SIP-WAM (ours)	**50**	**72.3**

Table 2: **OOD evaluation on LIBERO-Plus.** We report success rates (%) across seven perturbation categories and the overall success rate (Avg.). Best results are highlighted in **bold**.

Method	Camera	Robot	Lang.	Light	BG	Noise	Layout	Avg.
UniVLA (Bu et al., 2025)	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0 (Black et al., 2024)	13.8	6.0	58.8	85.0	81.4	79.0	68.9	53.6
OpenVLA-OFT (Kim et al., 2025)	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
FoMoVLA (Li et al., 2026b)	64.0	62.2	**94.0**	94.1	**96.2**	82.2	79.6	80.5
Fast-WAM (Yuan et al., 2026)	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
GE-Act (Liao et al., 2026)	60.7	**77.0**	77.4	95.8	86.0	**90.9**	80.2	80.3
Cosmos Policy (Kim et al., 2026b)	71.8	53.4	88.9	**98.5**	86.1	87.8	84.1	80.6
SIP-WAM (ours)	**73.6**	54.9	89.8	**98.5**	86.5	89.1	**84.4**	**81.5**

Implementation Details. We initialize the Video-Action DiT of SIP-WAM from the pretrained Cosmos Policy (Kim et al., 2026b) and jointly optimize it with the Interaction DiT. The Interaction DiT matches the Video-Action DiT in depth with 28 Transformer blocks, but uses smaller hidden and feed-forward dimensions of 128 and 512, respectively. For each interaction participant, we uniformly sample eight points and jointly predict their current locations and future trajectories. We replay the RoboCasa Human-50 and LIBERO demonstrations following the Cosmos Policy data preprocessing pipeline to construct the training targets for interaction-aware spatiotemporal prediction described in Section 3.2. For RoboCasa, the regenerated RGB observations differ from those used to train Cosmos Policy, while actions and other non-RGB data remain unchanged. For a fair comparison, we therefore fine-tune the RGB-only Cosmos Policy baseline on the same replayed dataset and denote the resulting model as Cosmos Policy*. For LIBERO, deterministic replay yields essentially the same demonstrations as those used to train Cosmos Policy. Additional model architecture and training details are provided in Appendix A.1.

4.2 Simulation Results

RoboCasa. As shown in Table 1, SIP-WAM achieves strong performance on RoboCasa, attaining an average success rate of 72.3% across 24 kitchen manipulation tasks. Notably, this performance is achieved using only 50 demonstrations per task, surpassing several methods trained with 300–1,000 demonstrations per task, including GR00T-N1.5, FLARE, and UWM. This comparison suggests favorable data efficiency for SIP-WAM. Compared with the original Cosmos Policy, SIP-WAM improves the average success rate

Table 3: **Future visual prediction quality on RoboCasa.**

Method	LPIPS ↓	FID ↓
RGB-only WAM	0.129	23.34
SIP-WAM	**0.128**	**22.92**

Table 4: **Ablation studies on RoboCasa.** We report average success rates (%) across 24 tasks. Experiments in (a)–(c) are carried out without Hungarian matching.

(a) Spatiotemporal prediction target.

Target	Avg.
Spatial only	71.1
Temporal only	71.6
Spatial + Temporal	**72.1**

(b) Trajectory point sampling.

Method	Avg.
Grid-style	70.8
ATM-style	70.7
Ours	**72.1**

(c) Interaction DiT architecture.

Architecture	Avg.
Sparse-layer fusion	70.8
All-layer fusion	69.9
Ours	**72.1**

(d) Hungarian matching.

Variant	Avg.
w/o Hungarian	72.1
w/ Hungarian	**72.3**

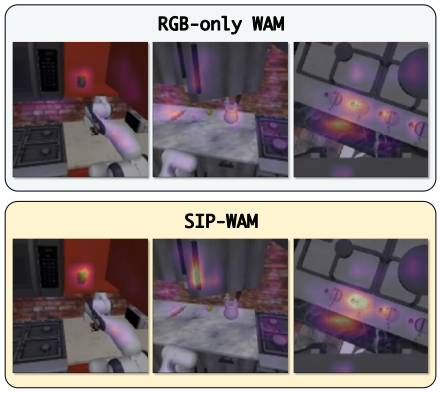

Figure 3: **Attention visualization.** RGB-only WAM (top) and SIP-WAM (bottom) on the same observation.

from 67.1% to 72.3%, a gain of 5.2 percentage points. Compared with Cosmos Policy*, which is further fine-tuned on the same replayed dataset for a fair comparison, SIP-WAM achieves a 2.8-percentage-point improvement, suggesting that our proposed interaction-aware spatiotemporal supervision translates into stronger policy performance. Moreover, these policy gains are achieved without additional inference overhead. More detailed results are provided in Appendix A.2.

Beyond policy performance, we visualize the action-to-video attention maps of the RGB-only WAM (Cosmos Policy*) and SIP-WAM in Fig. 3. We find that SIP-WAM focuses more on interaction-relevant regions than the RGB-only baseline, with reduced attention to the surrounding background. We further evaluate future visual prediction on an offline RoboCasa simulation test set containing 6,000 samples across 24 tasks (250 per task). None of these samples were seen by either model during training. As shown in Table 3, SIP-WAM achieves comparable future visual prediction quality to the Cosmos Policy* baseline.

LIBERO-Plus. SIP-WAM demonstrates strong out-of-distribution generalization on LIBERO-Plus. As shown in Table 2, SIP-WAM achieves an overall success rate of 81.5%, outperforming advanced VLA and WAM methods. It also surpasses FoMoVLA (Li et al., 2026b) (80.5%), which incorporates 2D point-trajectory supervision. Compared with the RGB-only Cosmos Policy baseline, SIP-WAM improves the overall success rate from 80.6% to 81.5%, achieving gains in six of the seven perturbation categories while maintaining the same performance under lighting variations. In particular, SIP-WAM achieves notable improvements over Cosmos Policy under camera-viewpoint changes, robot-initial-state perturbations, and sensor noise, with gains of 1.8, 1.5, and 1.3 percentage points, respectively. These consistent gains across diverse perturbations highlight the improved robustness of SIP-WAM under distribution shifts.

4.3 Ablation Studies

To assess the key design choices of SIP-WAM, we perform controlled ablations on RoboCasa. We examine the proposed prediction objective and key architectural choices through the following four questions.

Q1: Is joint modeling of spatial interaction grounding and temporal interaction dynamics important for policy learning? We compare three prediction settings: *Spatial only*, which predicts only the current spatial locations of the interaction participants; *Temporal only*, which forecasts the corresponding future trajectories from supplied current query points; and *Spatial + Temporal*, which jointly predicts both. As shown in Table 4(a), spatial-only and temporal-only prediction both improve over the RGB-only Cosmos Policy* baseline (69.5%), achieving average success rates of 71.1% and 71.6%, respectively. Joint prediction further improves performance to 72.1%, yielding the highest success rate among the three settings. These results suggest that interaction-aware spatial grounding and the modeling of fine-grained temporal interaction dynamics provide complementary benefits for policy learning.

Q2: Is interaction-aware spatiotemporal prediction more effective than conventional trajectory supervision? We compare our approach with two representative trajectory supervision schemes. For *Grid-style*, following FoMoVLA (Li et al., 2026b), we uniformly sample an 8×8 grid of points on the current RGB observation and predict their future trajectories. For *ATM-style*, we select moving query points following the strategy of ATM (Wen et al., 2023) and forecast their subsequent trajectories. As shown in Table 4(b), SIP-WAM achieves an average success rate of 72.1%, outperforming Grid-style (70.8%) and ATM-style (70.7%). These results suggest that explicit spatial grounding and a focus on task-specified interaction participants offer benefits beyond conventional trajectory supervision alone.

Q3: How does the Interaction DiT architecture affect policy performance? We compare three Interaction DiT architectures: *sparse-layer fusion*, which fuses features from one early, one middle, and one late Video-Action DiT layer to condition a lightweight DiT; *all-layer fusion*, which fuses features from all 28 backbone layers, following Faster-WAM (Ma et al., 2026); and *ours*, which couples the Interaction DiT layerwise with the Video-Action DiT through asymmetric joint attention. Our architecture yields gains of 1.3 and 2.2 percentage points over sparse-layer and all-layer fusion, respectively, reaching an average success rate of 72.1% (Table 4(c)). These results suggest that our architecture facilitates more effective cross-modal feature interaction, helping the WAM learn spatiotemporal representations better suited to policy learning.

Q4: Does Hungarian matching improve policy performance? We evaluate whether training with Hungarian matching benefits policy learning. Without Hungarian matching, predictions are paired with ground-truth point sequences in a fixed order. Hungarian matching instead finds an optimal one-to-one assignment within each participant group, relaxing the fixed-order constraint. In our formulation, the matching cost is computed over each point's complete current-to-future coordinate sequence, as described in Section 3.3. Table 4(d) shows that Hungarian matching improves the average success rate from 72.1% to 72.3%, suggesting that the more flexible assignment enabled by Hungarian matching offers a modest benefit for policy learning.

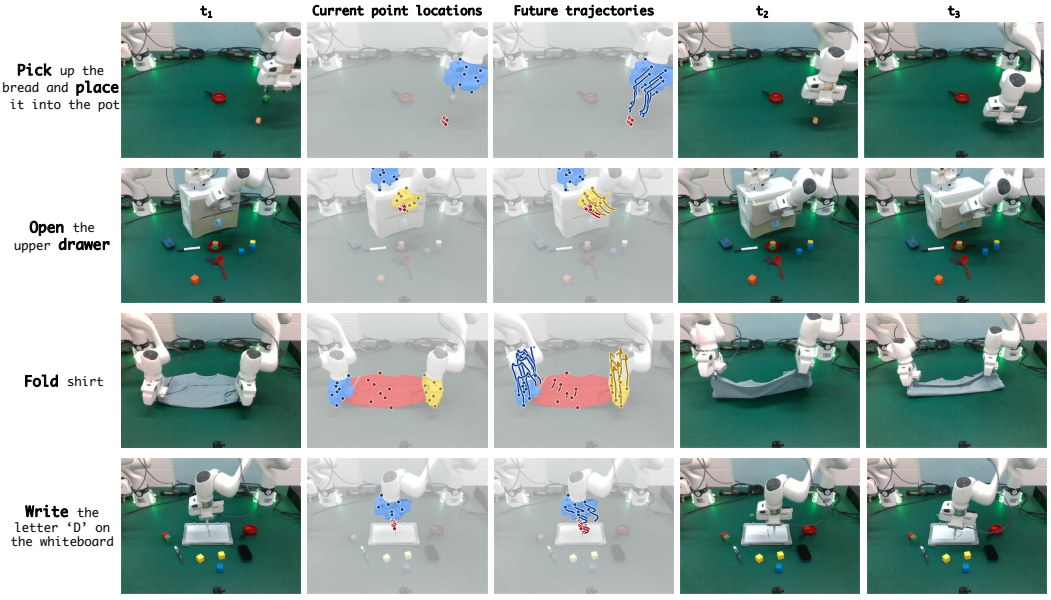

Figure 4: **Overview of real-world experiments.** We evaluate four tasks covering both single-arm and bimanual manipulation. We also visualize the current point locations and future trajectories used for interaction-aware spatiotemporal supervision.

4.4 Real-World Experiments

Experimental setup. We conduct real-world experiments on a dual-arm platform comprising two Franka Research 3 (FR3) robot arms, supporting both single-arm and bimanual manipulation. As

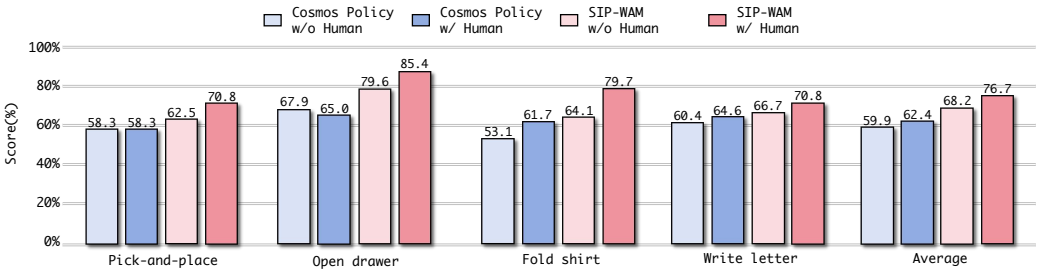

Figure 5: **Real-world experimental results.** We compare SIP-WAM and Cosmos Policy across four tasks under robot-only training and human-video co-training. Scores represent the average task completion percentage.

illustrated in Fig. 4, we evaluate four tasks covering single-arm pick-and-place, bimanual articulated-object interaction, bimanual long-horizon deformable-object manipulation, and single-arm letter writing. We collect 80 teleoperated robot demonstrations per task for training. To assess the benefits of human–robot co-training, we additionally collect 80 human manipulation videos per task. We compare SIP-WAM with Cosmos Policy (Kim et al., 2026b) under two training settings: robot demonstrations alone, and co-training on the robot demonstrations and human videos. During co-training, action supervision is applied only to robot data for both methods. SIP-WAM additionally applies interaction-aware spatiotemporal supervision to both human and robot data. Each method is evaluated over 24 trials per task under the same setup. We report the task completion score, computed as the average task completion percentage across evaluation trials.

Results. Figure 5 presents the real-world experimental results. With robot demonstrations alone, SIP-WAM achieves an average task completion score of 68.2% across the four tasks, surpassing Cosmos Policy at 59.9%. This improvement highlights the effectiveness of interaction-aware spatiotemporal prediction for real-world single-arm and bimanual manipulation.

With human-video co-training, SIP-WAM and Cosmos Policy achieve average scores of 76.7% and 62.4%, respectively. Cosmos Policy improves on two tasks, remains unchanged on one, and declines on the remaining task, whereas SIP-WAM improves on all four. Notably, SIP-WAM benefits more from the additional human videos, obtaining a larger average gain over robot-only training than Cosmos Policy, suggesting that interaction-aware spatiotemporal supervision enables more effective cross-embodiment transfer from human manipulation videos.

5 Limitations

SIP-WAM relies on additional off-the-shelf segmentation and tracking models (SAM 3 and Co-Tracker3) to construct training targets for interaction-aware spatiotemporal prediction. Moreover, although SIP-WAM builds on a pretrained backbone, we study this prediction objective only during fine-tuning on simulation benchmarks and a limited set of real-world demonstrations, rather than during large-scale WAM pretraining. We believe extending this approach to large-scale WAM pretraining is a promising direction, but its effectiveness remains to be validated.

6 Conclusion

In this work, we introduce SIP-WAM, a novel framework that enhances WAMs via Interaction-Aware Spatiotemporal Prediction. By incorporating an Interaction DiT that jointly predicts the current spatial locations and temporally dense future trajectories of the interaction participants, SIP-WAM provides supervision for interaction-aware spatial grounding and fine-grained temporal interaction dynamics. Layerwise asymmetric joint attention couples the Interaction and Video-Action DiTs during training, allowing the Interaction DiT to be removed at inference without additional computational overhead. Results on RoboCasa, LIBERO-Plus, and real-world manipulation tasks demonstrate strong policy performance and generalization, while real-world co-training experiments further suggest more effective cross-embodiment transfer from human manipulation videos.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation and language editing, including improving the grammar, clarity, readability, and wording of the manuscript. These tools also assisted with identifying and summarizing relevant literature and provided feedback on the methodology.

The authors retained responsibility for all research decisions and critically reviewed and revised the AI-assisted content to ensure consistency with the methods and experimental results reported in this work. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training objectives, and training-target construction in the main paper. Model hyperparameters and simulation training details are reported in Section A.1. Real-world data collection, robot-only training, and human-video co-training are detailed in Sections B.1 and B.2. The real-world evaluation protocol and task-specific scoring criteria are specified in Section B.3. We will publicly release our code, model checkpoints, and the replayed RoboCasa dataset to facilitate reproducibility.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024a.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision*, pp. 306–324. Springer, 2024b.

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. π_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, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, Leslie Kaelbling, et al. Video language planning. In *International Conference on Learning Representations*, volume 2024, pp. 31138–31155, 2024.

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.

Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Cai Zhehao, and Lin Shao. Flip: Flow-centric generative planning as general-purpose manipulation world model. In *International Conference on Learning Representations*, volume 2025, pp. 21927–21948, 2025.

Jiarui Guan, Wenshuai Zhao, Yue Pei, Ziliang Chen, Arno Solin, and Juho Kannala. Point tracking improves world action models. *arXiv preprint arXiv:2605.23856*, 2026.

Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. *Advances in Neural Information Processing Systems*, 37:112386–112410, 2024.

Adam Hung, Bardienus Pieter Duisterhof, and Jeffrey Ichnowski. 3pointr: 3d point tracks for learning manipulation from unconstrained human videos. *arXiv preprint arXiv:2603.08485*, 2026.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, October 2025.

Jisoo Kim, Jungbin Cho, Sanghyeok Chu, Ananya Bal, Jinhyung Kim, Gunhee Lee, Sihaeng Lee, Seung Hwan Kim, Bohyung Han, Hyunmin Lee, et al. Pri4r: Learning world dynamics for vision-language-action models with privileged 4d representation. *arXiv preprint arXiv:2603.01549*, 2026a.

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.

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

Myungkyu Koo, Daewon Choi, Taeyoung Kim, Kyungmin Lee, Changyeon Kim, Younggyo Seo, and Jinwoo Shin. Hamlet: Switch your vision-language-action model into a history-aware policy. *arXiv preprint arXiv:2510.00695*, 2025.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

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*, 2026a.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025.

Wei Li, Peijin Jia, Yuan Ma, Xuefeng Jiang, Titong Jiang, Sheng Sun, Yujian Li, Xin Wen, Han Hong, Zhikang Liu, et al. Fomovla: Bridging visual foresight and motion guidance for vision-language-action models. *arXiv preprint arXiv:2607.14739*, 2026b.

Junbang Liang, Pavel Tokmakov, Ruoshi Liu, Sruthi Sudhakar, Paarth Shah, Rares Ambrus, and Carl Vondrick. Video generators are robot policies. *arXiv preprint arXiv:2508.00795*, 2025.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 88446–88463, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Liheng Ma, Rui Heng Yang, Zhanguang Zhang, Mateo Clemente, Ziwen Hu, Tongtong Cao, and Yingxue Zhang. Faster-wam: Do world action models need deep action modules? *arXiv preprint arXiv:2608.02365*, 2026.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024.

Jiange Yang, Haoyi Zhu, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. Tra-moe: Learning trajectory prediction model from multiple domains for adaptive policy conditioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6960–6970. IEEE, 2025.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

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*, 2026b.

Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024.

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.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, Rui Heng Yang, Yintao Ma, Amir Rasouli, Sajjad Pakdamansavoji, Yangzheng Wu, Lingfeng Zhang, Tongtong Cao, et al. Do world action models generalize better than vlas? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026.

Haoyu Zhen, Zixian Gao, Qiao Sun, Yilin Zhao, Yuncong Yang, Yilun Du, Pengsheng Guo, Tsun-Hsuan Wang, Yi-Ling Qiao, and Chuang Gan. Action images: End-to-end policy learning via multiview video generation. *arXiv preprint arXiv:2604.06168*, 2026.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, et al. Flare: Robot learning with implicit world modeling. In *Conference on Robot Learning*, pp. 3952–3971. PMLR, 2025.

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

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025.

A Simulation Experimental Details

A.1 Training Details

RoboCasa. We train on the Human-50 dataset, using 50 human-teleoperated demonstrations per task across 24 kitchen manipulation tasks. Following the Cosmos Policy data preprocessing pipeline (Kim et al., 2026b), we replay the demonstrations to obtain simulator-rendered segmentation masks and construct the training targets for interaction-aware spatiotemporal prediction described in Section 3.2. Replay regenerates the RGB observations while preserving the original actions and other non-RGB data. For a fair comparison, we additionally fine-tune the RGB-only Cosmos Policy baseline on the same replayed dataset as SIP-WAM and denote the resulting model as Cosmos Policy*.

Each training sample contains the language instruction, robot proprioception, and RGB observations from the left, right, and wrist cameras. Images are resized to 224×224 pixels and augmented with random resized cropping, rotation, and color jitter. The same geometric transformations are applied to the ground-truth current-to-future coordinate sequences of points sampled on the interaction participants, maintaining alignment with the images. Actions and proprioceptive states are normalized using training-set statistics. We use an action horizon of $H = 32$ and train the Interaction DiT to jointly predict the current point locations and future trajectories over the subsequent 32 time steps.

We follow the Cosmos Policy training protocol (Kim et al., 2026b) and jointly optimize the Video-Action DiT and Interaction DiT using AdamW under the objective in Equation 7, retaining the original WAM prediction objectives. Table 5 summarizes the optimization settings and model architecture details.

LIBERO-Plus. For evaluation on LIBERO-Plus, we train on the original LIBERO demonstrations and construct training targets for interaction-aware spatiotemporal prediction using the same replay-based procedure as for RoboCasa. Deterministic replay preserves the demonstrations used to train Cosmos Policy. We follow the same overall training procedure as for RoboCasa, with dataset-specific learning rates and training steps reported in Table 5. Following Cosmos Policy, we use an action horizon of $H = 16$ and the corresponding 16-step future trajectories. The resulting model is evaluated directly on LIBERO-Plus without additional fine-tuning.

Table 5: **Simulation training hyperparameters and model architecture details.**

Optimization	**RoboCasa**	**LIBERO-Plus**
Video-Action DiT learning rate	2×10^{-5}	1.5×10^{-5}
Interaction DiT learning rate	4×10^{-5}	3×10^{-5}
Video-Action DiT weight decay	0.1	0.1
Interaction DiT weight decay	0.01	0.01
Effective batch size	768	1,920
Training steps	4500	2500
Warm-up steps	100	100
Gradient clipping norm	5.0	5.0
Model architecture	**Video-Action DiT**	**Interaction DiT**
Transformer blocks	28	28
Hidden dimension	2048	128
Attention dimension (all heads)	2048	128
Feed-forward dimension	8192	512
Attention heads	16	4

A.2 Detailed RoboCasa Results

Multi-Seed Evaluation. To assess the consistency of the performance gains, we evaluate SIP-WAM and the Cosmos Policy* baseline across three random evaluation seeds. For each seed, both

Table 6: **RoboCasa results across evaluation seeds.** We report average success rates (%) across 24 tasks, with 50 rollouts per task for each seed. Seed 2 corresponds to the main-paper results. Δ denotes the improvement over Cosmos Policy* in percentage points.

Method	Seed 1	Seed 2	Seed 3
SIP-WAM	**71.1**	**72.3**	**70.7**
Cosmos Policy*	68.1	69.5	67.3
Δ	+3.0	+2.8	+3.4

methods follow the same evaluation protocol, with 50 rollouts per task across the 24 RoboCasa tasks. As shown in Table 6, SIP-WAM outperforms Cosmos Policy* by 3.0, 2.8, and 3.4 percentage points, respectively. These results support the stability of SIP-WAM's performance gains. Seed 2 corresponds to the primary evaluation setting reported in the main paper.

Task-level performance analysis. Table 7 presents the detailed RoboCasa results using evaluation Seed 2, with 50 rollouts per task. SIP-WAM matches or exceeds Cosmos Policy* on most tasks. Grouping tasks by manipulation type, we observe an average gain of 2.5 percentage points over Cosmos Policy* across the eight pick-and-place tasks and 4.7 percentage points across the six articulated-object manipulation tasks involving door and drawer opening and closing. These gains support the effectiveness of interaction-aware spatiotemporal prediction for tasks requiring precise spatial alignment and motion coordination.

Table 7: **Task-level performance comparison on RoboCasa.** We report success rates (%) over 50 rollouts per task using evaluation Seed 2. Δ denotes the difference between SIP-WAM and Cosmos Policy* in percentage points. Best results, including ties, are highlighted in **bold**.

Task	**SIP-WAM**	Cosmos Policy*	Δ
PnPSinkToCounter	**68.0**	66.0	+2.0
PnPMicrowaveToCounter	**52.0**	40.0	+12.0
PnPCounterToSink	**82.0**	72.0	+10.0
PnPCounterToMicrowave	**42.0**	38.0	+4.0
PnPStoveToCounter	66.0	**72.0**	-6.0
PnPCabToCounter	42.0	**44.0**	-2.0
PnPCounterToCab	**64.0**	**64.0**	0.0
PnPCounterToStove	**50.0**	**50.0**	0.0
CoffeeSetupMug	26.0	**30.0**	-4.0
CoffeeServeMug	**60.0**	58.0	+2.0
OpenSingleDoor	**84.0**	80.0	+4.0
CloseSingleDoor	**98.0**	96.0	+2.0
OpenDoubleDoor	**100.0**	96.0	+4.0
CloseDoubleDoor	**90.0**	78.0	+12.0
OpenDrawer	**88.0**	82.0	+6.0
CloseDrawer	**100.0**	**100.0**	0.0
TurnOnSinkFaucet	**96.0**	84.0	+12.0
TurnOffSinkFaucet	**94.0**	84.0	+10.0
TurnSinkSpout	82.0	**84.0**	-2.0
TurnOnStove	54.0	**60.0**	-6.0
TurnOffStove	**28.0**	**28.0**	0.0
CoffeePressButton	**100.0**	94.0	+6.0
TurnOnMicrowave	72.0	**76.0**	-4.0
TurnOffMicrowave	**96.0**	92.0	+4.0
Average	**72.3**	69.5	+2.8

B Real-World Experimental Details

B.1 Experimental Setup and Data Collection

Our real-world platform consists of two 7-DoF Franka Research 3 (FR3) robot arms equipped with two-finger grippers, supporting both single-arm and bimanual manipulation. Three Intel RealSense D435i cameras capture the workspace: one fixed third-person camera and one wrist-mounted camera on each arm. We use only RGB images from these cameras, stored at a resolution of 480×640 and resized to 224×224 for policy input. Figure 6 shows the platform and camera configuration.

For each arm, the policy predicts a 7-dimensional action comprising relative end-effector translation, relative end-effector rotation, and a gripper command. The commands for both arms form a 14-dimensional action vector. The combined proprioceptive state of both arms is represented by a 16-dimensional vector, comprising the end-effector position, quaternion orientation, and gripper width of each arm.

We collect 80 teleoperated robot demonstrations per task using 3D SpaceMouse across four tasks: pick-and-place, opening drawers, folding shirts, and letter writing. Each demonstration contains synchronized multi-view RGB observations, robot proprioception, actions, and a corresponding language instruction. We additionally collect 80 human manipulation videos per task, recording direct human execution of the same tasks without robot action or proprioceptive annotations. Examples are shown in Fig. 7. Both SIP-WAM and Cosmos Policy use the same robot demonstrations and, in the co-training setting, the same additional human videos. For SIP-WAM, we construct interaction-aware spatiotemporal prediction targets for both data sources following Section 3.2.

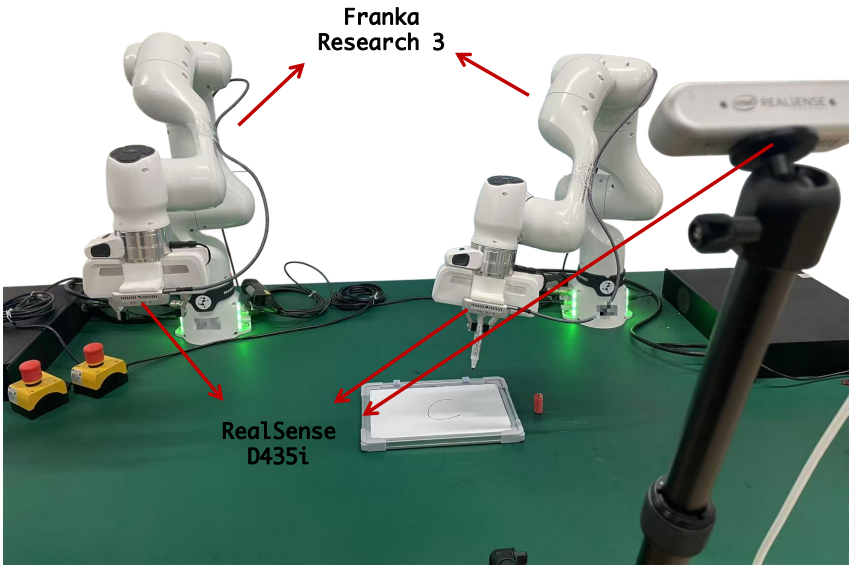

Figure 6: **Real-world experimental setup.** Our dual-arm platform comprises two Franka robot arms and three Intel RealSense D435i cameras: one fixed third-person camera and one wrist-mounted camera on each arm.

B.2 Training Details

Robot-only training. We follow the Cosmos Policy training formulation, retaining the original WAM prediction objectives. For both Cosmos Policy and SIP-WAM, we initialize the Video-Action DiT from the same pretrained Cosmos Predict2 checkpoint (Ali et al., 2025). For SIP-WAM, we jointly optimize the Video-Action DiT and Interaction DiT using AdamW under the objective in Equation 7. We use an action horizon of $H = 32$ and follow the image augmentation and nor-

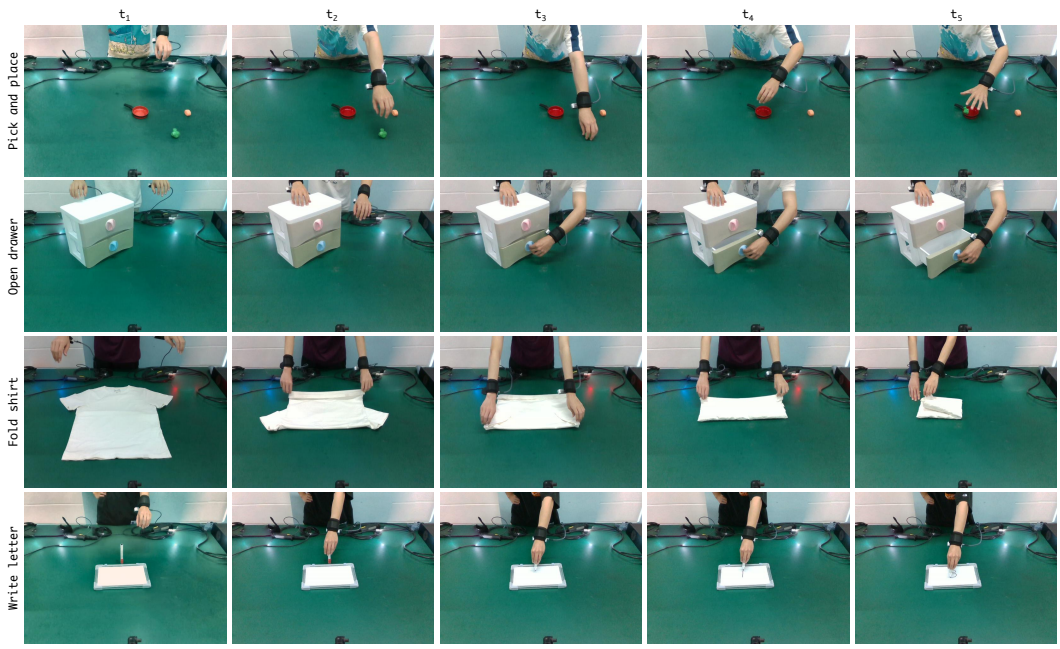

Figure 7: **Examples of human manipulation videos.** We show sequences from the four tasks used for human-video co-training. Each row corresponds to one task, with time progressing from left to right.

malization procedures described in Section A.1, with normalization statistics computed from the real-world robot demonstrations.

Human-video co-training. We co-train both methods on robot demonstrations and human manipulation videos. Robot samples retain the original WAM prediction objectives. For human samples, both methods learn future visual prediction conditioned on the current RGB observations and language instruction. Unlike robot demonstrations, human manipulation videos provide visual observations of task execution without corresponding robot action, proprioception, or value labels.

SIP-WAM additionally applies interaction-aware spatiotemporal supervision to both data sources, jointly predicting the current locations and future trajectories of sparse points sampled on the interaction participants.

B.3 Task Definitions and Evaluation Protocol

Each method is evaluated over 24 trials per task under both robot-only training and human-video co-training, using the same experimental setup. Each trial receives a score between 0 and 100 based on the completed stages, and we report the average task completion score across evaluation trials. The task-specific scoring criteria are detailed below.

- **Pick-and-place.** The robot must grasp the object specified by the language instruction and place it into a pot. We award 50 points for successfully grasping the specified object and an additional 50 points for placing that object into the pot.
- **Open drawer.** The language instruction specifies whether to open the upper or lower drawer. The left arm presses down on the top of the drawer unit, while the right arm grasps the corresponding handle and pulls the instructed drawer open. We award 30 points for the left gripper contacting and pressing down on the top of the drawer, 30 points for the right gripper grasping the handle of the instructed drawer, and 40 points for successfully opening that drawer without jamming.
- **Fold shirt.** We evaluate this task using two different shirts. Each of the following eight stages contributes 12.5 points: (1) grasping the lower edge of the garment; (2) making the

first fold in half; (3) taking hold of both sleeves; (4) folding the sleeves toward the center; (5) regrasping the lower edge of the folded garment; (6) making a second fold in half; (7) grasping the right edge with the right gripper; and (8) completing a final fold in half. For stages (1)–(6), we award 12.5 points when both arms complete their respective parts of the stage, 6.25 points when only one arm does so, and 0 points otherwise. Stages (7) and (8) each receive 12.5 points upon completion. The maximum score is 100 points.

- **Write letter.** The robot must grasp a pen and write the letter specified by the language instruction. We evaluate four uppercase letters: A, B, C, and D. We award 50 points for successfully grasping the pen and an additional 50 points for correctly writing the instructed letter.