

# AffordanceWAM: Affordance-Aware Joint World-Action Modeling for Robot Manipulation

Jiadi You<sup>1,3,†</sup>, Qize Yu<sup>2,†</sup>, Yue Chen<sup>2</sup>, Minghong Cai<sup>4</sup>, Zhide Zhong<sup>3</sup>, Yuran Wang<sup>5</sup>, Bowen Ping<sup>2</sup>, Jiaqi Liang<sup>2</sup>, Zhenhao Shen<sup>2</sup>, Haodong Yan<sup>3</sup>, Yinchuan Li<sup>6</sup>, Ruihai Wu<sup>2</sup>, Xiaojuan Qi<sup>1,\*</sup>, Yingcong Chen<sup>3,\*</sup>

<sup>1</sup>The University of Hong Kong, <sup>2</sup>Peking University, <sup>3</sup>The Hong Kong University of Science and Technology (Guangzhou), <sup>4</sup>The Chinese University of Hong Kong, <sup>5</sup>National University of Singapore, <sup>6</sup>Knowin AI

<sup>†</sup>Equal contribution. <sup>\*</sup>Corresponding authors.

Generalizable robot manipulation requires predicting how a scene will evolve, identifying where interactions are feasible, and determining how to act. Action-labeled robot videos directly supervise control but are costly and limited in diversity, whereas egocentric human videos capture diverse interactions but lack robot actions and differ in embodiment and appearance. We introduce AffordanceWAM, an affordance-aware generative World Action Model that represents object-centric spatiotemporal affordance through Scalar Affordance and Affordance Heatmap, within the generated future World. This representation grounds visual prediction in task-relevant objects and interaction regions for action generation, and provides shared interaction targets across human and robot videos. Built on a pretrained video diffusion Transformer, AffordanceWAM uses separately parameterized World and Action Experts, coupled through Masked Joint Self-Attention, to jointly predict future RGB observations, Scalar Affordance fields, Affordance Heatmaps, and continuous robot actions under a unified flow-matching objective. Human videos supervise all three future-World streams, whereas robot trajectories additionally provide action supervision, enabling transfer without human action labels or retargeting. Experiments on RoboCasa, CALVIN ABC→D, and real-world manipulation demonstrate consistent gains over RGB-only and robot-data-only baselines. Under fixed robot supervision, RoboCasa performance improves monotonically as affordance-annotated human video scales. These results support affordance as an effective interface for both vision-language-action learning and human-to-robot transfer.

**Date:** August 1, 2026

**Version History:** V1: August 1, 2026.

**Project Page:** <https://alexycd2001.github.io/AffordanceWAM>

## 1 Introduction

Egocentric human data offers a source for scaling robot manipulation. Humans perform dexterous manipulation across diverse objects, environments, and tasks, which is difficult to match through teleoperation pipelines [8, 13, 14, 31]. Recent work shows that this experience can improve robot learning by aligning human demonstrations with robot actions [14, 15, 28, 31, 43]. However, these gains depend on alignment in viewpoint, speed, and behavior. Without this bridge, human data may degrade performance.

Recently, World Action Models (WAMs) provide a more direct alternative. Alongside action prediction, they predict future states from a shared backbone [7, 21, 22, 36, 44]. Human and robot videos can therefore supervise object and scene dynamics, which can be transferred across embodiments more readily than actions. By separating how the world evolves from embodiment-specific action execution, WAMs can use large-scale human data more effectively, yet appearance-oriented World representations remain semantically misaligned with action generation.

The question is which World representation supports transfer across embodiments during WAM co-training. Most WAMs reconstruct RGB frames [11, 39] or visual latents [23, 35, 42], modeling appearance, background motion, and viewpoint changes while leaving interaction structure implicit. Affordance [10] makes interaction

structure shared across embodiments explicit by identifying relevant objects and feasible interaction regions and capturing how the task-relevant state evolves [1, 26, 34, 40]. By encoding object-centric interaction semantics rather than embodiment-specific motion, affordance both bridges human and robot data and provides a semantic interface from World prediction to action generation.

We therefore introduce **AffordanceWAM**, a dual-expert WAM designed to share interaction dynamics across embodiments (Figure 1). A pretrained video DiT serves as the World Expert, forecasting future RGB observations and two complementary affordance components: Scalar Affordance and Affordance Heatmap. An Action Expert predicts continuous controls from the layer-wise RGB and Scalar Affordance features of the evolving World. To preserve this separation, Masked Joint Self-Attention lets the Action Expert use these evolving World features, but prevents actions from influencing World prediction and excludes the auxiliary Affordance Heatmap features from the Action input. This reduces an additional RGB-rich pathway and encourages greater use of Scalar Affordance while retaining access to RGB features. This directional coupling leads to asymmetric training: human clips update only the shared World Expert, whereas robot trajectories additionally train the Action Expert. Human and robot data thus meet in interaction space rather than action space.

Experiments on RoboCasa [25] and CALVIN [24] show strong performance across tasks and unseen environments. Crucially, controlled ablations show that affordance is the key bridge through which human interaction dynamics benefit robot learning. Without affordance, human videos lower RoboCasa success from 57.3% to 55.1% and CALVIN sequence length from 3.75 to 3.68; with affordance, the same data raises these metrics from 63.7% to 69.4% and from 3.91 to 4.22. These gains carry over to real-world manipulation. Our contributions are:

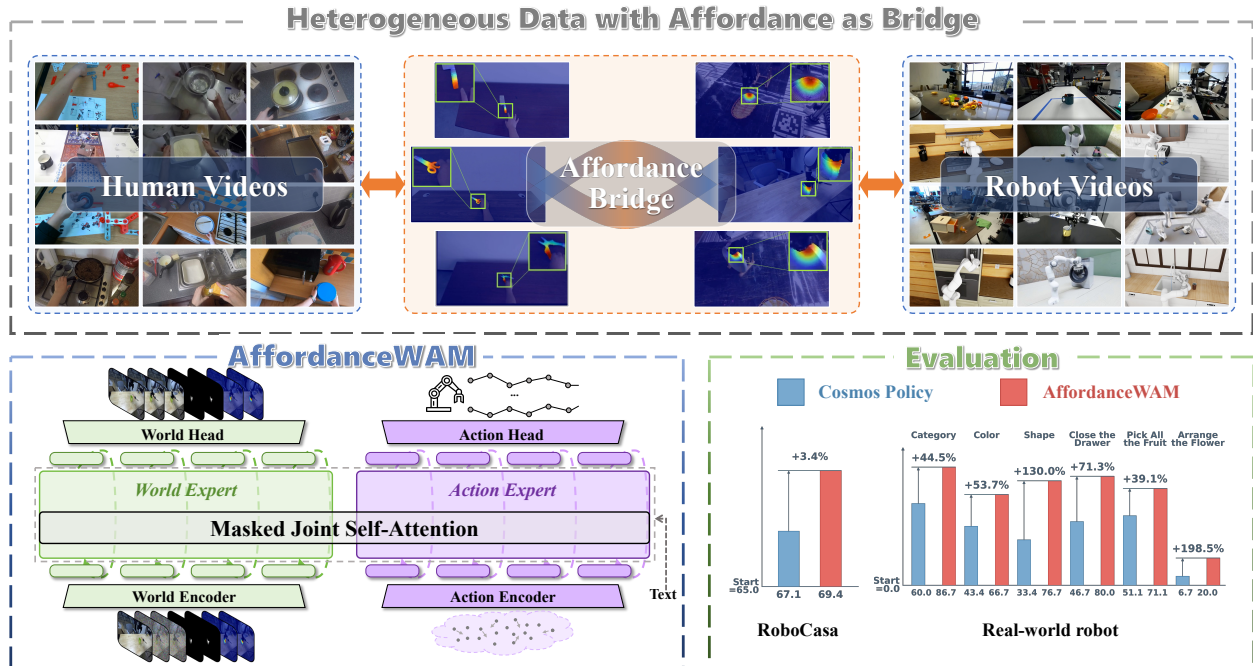
- We introduce AffordanceWAM, a WAM architecture that uses structured affordance forecasting as intermediate supervision to bridge predictive world modeling and robot control enabling generalization across diverse tasks.
- Evidence that affordance forecasting unlocks human-data scale by separating shared interaction dynamics from embodiment-specific control.
- Across simulation and real-world experiments, AffordanceWAM outperforms strong VLAs and WAMs demonstrating generalization and robustness across ablations, qualitative and quantitative analyses.

## 2 Related Work

### 2.1 World Action Models

Video-based world models learn scene dynamics by predicting how visual observations evolve over time. In robot learning, predicted videos can serve as visual plans or provide predictive representations to a downstream policy [9, 12]. These approaches connect visual prediction to control, but the video model and action policy remain separate. World Action Models instead learn future observations and actions within a unified framework. Existing designs include coupled video–action diffusion models [22, 44] and autoregressive models over visual and action tokens [7, 21]. More recent work builds on pretrained video backbones or large-scale causal video–action pretraining [36, 41]. Joint modeling allows visual dynamics to inform action learning and provides a route for using videos without action annotations. Across these formulations, visual prediction may act as a planning space, an auxiliary learning signal, or a jointly generated output.

Despite their architectural differences, these methods usually represent the predicted World with future RGB observations or visual latent tokens. Such representations capture appearance and scene evolution, but they do not explicitly identify the objects and interaction regions that determine a manipulation task. AffordanceWAM augments the generated future World with object-centric spatiotemporal Scalar Affordance and Affordance Heatmaps, and jointly models these streams with future RGB observations and continuous robot actions. It thereby makes interaction structure an explicit prediction target rather than leaving it implicit in visual features.



**Figure 1** AffordanceWAM overview. A shared object-focused affordance representation bridges heterogeneous human-robot data with joint world and action modeling, enabling positive transfer to robot manipulation.

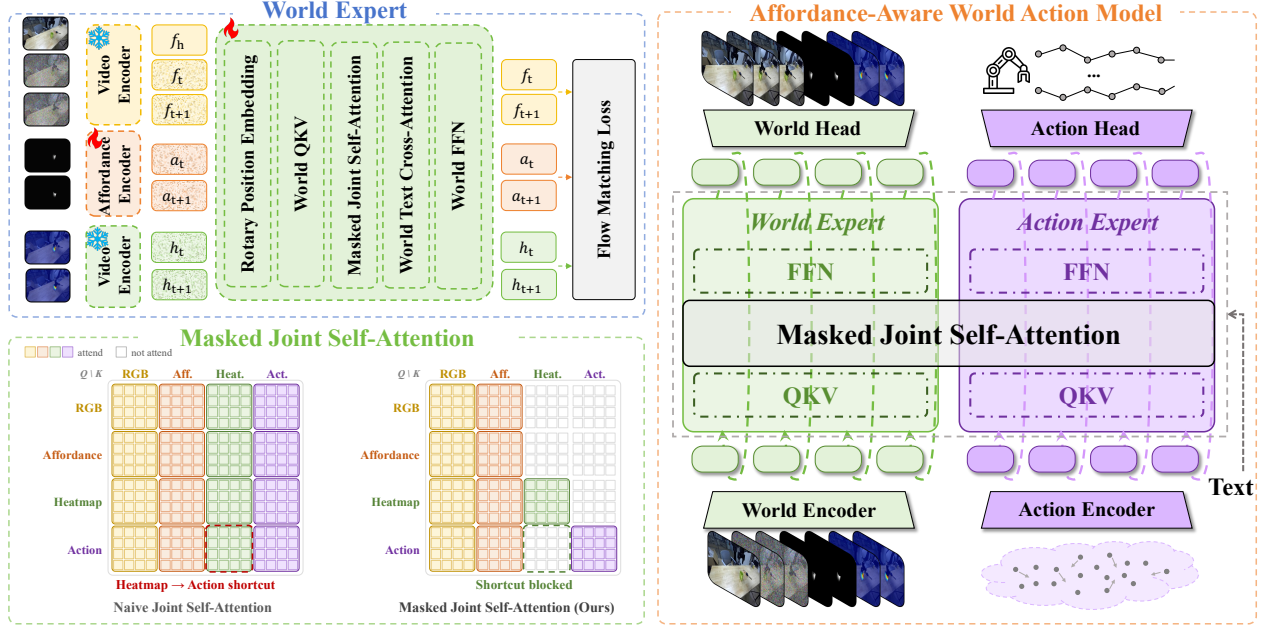
## 2.2 Robot Learning from Human Videos

Human videos have supported robot learning at increasingly direct levels. Earlier methods pretrain visual representations or extract contacts, hand trajectories, and object motion as planning or control interfaces [1, 8, 29, 31, 38]. Recent methods retarget human motion into robot actions and co-train policies across embodiments [13, 14, 43]. EgoScale shows that this route can benefit from large-scale egocentric data, but still relies on retargeted wrist and hand actions followed by aligned human-robot mid-training [43]. More broadly, action-level transfer depends on the quality of human motion recovery and cross-embodiment alignment.

Future-state prediction offers a complementary route. Recently, WAMs couple future prediction with action generation, allowing human videos to shape policy representations through scene dynamics [7, 19, 36, 44]. AffordanceWAM retains future-state prediction but augments it with an affordance representation comprising Scalar Affordance and Affordance Heatmap. Together, they identify task-relevant interaction regions and associate their evolution with visual context.

## 2.3 Affordance Learning for Robotic Manipulation

Affordances describe the interaction possibilities offered by objects [10]. Prior work represents them as contacts or trajectories [1, 40], structured plans [26, 33], or dense fields [34]. Recent affordance-aware VLAs further incorporate these cues through reasoning, feature alignment, or structured forecasting [18, 20, 37], but primarily use them as policy inputs or intermediate predictions. AffordanceWAM instead treats temporally evolving Scalar Affordance and Affordance Heatmap as parts of the generated future World. During joint pretraining, human and robot videos share these prediction targets, while the Action Expert reads layer-wise hidden features from the RGB and Scalar Affordance streams and receives action supervision only from robot trajectories. Affordance therefore serves as a cross-embodiment World representation rather than an embodiment-specific action target.



**Figure 2** AffordanceWAM overview. The World Expert jointly predicts future RGB, Scalar Affordance, and Affordance Heatmaps. The Action Expert reads RGB and Scalar Affordance hidden features through layer-wise Masked Joint Self-Attention to generate continuous actions. The mask excludes Action features from World prediction and Affordance Heatmap features from the Action input, reducing an additional RGB-rich pathway to control.

### 3 Method

AffordanceWAM learns a shared interaction representation from human and robot videos and grounds it in robot control. Its World Expert predicts three future-World streams—RGB, Scalar Affordance, and Affordance Heatmap—while its Action Expert generates actions by accessing the evolving RGB and Scalar Affordance hidden features at each Transformer layer (Fig. 2). We introduce the generative formulation, describe the architecture and World representation, and then present learning from heterogeneous supervision.

#### 3.1 Preliminaries

*Flow matching.* Flow matching learns a velocity field that transports Gaussian noise to data. Given a clean target  $y_0$ , noise  $\epsilon \sim \mathcal{N}(0, I)$ , and flow time  $\tau \in [0, 1]$ , a linear interpolation defines the noisy input and target velocity:

$$y_\tau = (1 - \tau)y_0 + \tau\epsilon, \quad u = \epsilon - y_0. \quad (1)$$

Given conditioning information  $c$ , a velocity predictor  $v_\theta$  with learnable parameters  $\theta$  is trained by

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{y_0, c, \epsilon, \tau} \left[ \frac{1}{n} \|v_\theta(y_\tau, \tau; c) - (\epsilon - y_0)\|_2^2 \right], \quad (2)$$

where  $n$  is the number of scalar target entries and the norm is computed over flattened entries. Generation starts from noise at  $\tau = 1$  and integrates the learned velocity back to the data endpoint at  $\tau = 0$ . We apply this formulation to both future World latents and continuous actions; subscripts 0 and  $\tau$  distinguish clean and noisy states.

*World-action modeling.* Let  $I^- = (I_0, \dots, I_{P-1})$  be a history of  $P$  RGB frames,  $\ell$  a language instruction, and  $s$  the current proprioceptive state. Action modeling predicts a continuous chunk  $U_0 \in \mathbb{R}^{T_A \times d_A}$  from this context, where  $T_A$  and  $d_A$  denote the action horizon and dimension. AffordanceWAM additionally defines the future-World target  $\mathcal{W}^+ = (I^+, F^+, M^+)$  comprising  $T$  future RGB frames  $I^+ = (I_P, \dots, I_{P+T-1})$ , aligned Scalar Affordance  $F^+$ , and Affordance Heatmaps  $M^+$ . The corresponding latent variables, together with the

action chunk, are modeled by

$$p_{\theta}(\mathcal{W}^+, U_0 \mid I^-, \ell, s). \quad (3)$$

The generated future World provides an explicit representation of scene evolution and interaction regions for action generation. We parameterize this joint model through coupled World and Action velocity predictors, with World generation conditioned on RGB history and language and proprioception provided to the Action Expert. During training,  $I^+$ ,  $F^+$ , and  $M^+$  are clean supervision targets. At deployment, these clean targets are unavailable; their latent variables are generated from the conditioning context. Frame indices  $t$  are distinct from flow time  $\tau$ .

### 3.2 AffordanceWAM Architecture

AffordanceWAM consists of a World Expert  $\mathcal{M}_W$  and an Action Expert  $\mathcal{M}_A$ . Both are diffusion Transformers with the same number of layers, but have separate parameters and modality-specific hidden dimensions. Their coupling occurs through joint attention at corresponding layers, allowing the Action Expert to use evolving World hidden features throughout joint denoising.

#### 3.2.1 World Expert.

The World Expert is initialized from a pretrained video diffusion Transformer. Let  $C$  denote the encoded RGB history and  $Z_0^m$  the clean future latent for stream  $m \in \mathcal{S}_W = \{\text{rgb}, \text{afd}, \text{heat}\}$ ; the encoders are defined in Sec. 3.3. The indices afd and heat identify Scalar Affordance and Affordance Heatmap, respectively. At World flow time  $\tau_W$ , collect the three noisy future latents into  $\mathbf{Z}_{W, \tau_W} = (Z_{\tau_W}^{\text{rgb}}, Z_{\tau_W}^{\text{afd}}, Z_{\tau_W}^{\text{heat}})$ . Conditioned on RGB history and language, the Expert predicts their flow velocities:

$$\hat{\mathbf{v}}_W = \mathcal{M}_W(\mathbf{Z}_{W, \tau_W}, \tau_W; C, \ell). \quad (4)$$

Here  $\hat{\mathbf{v}}_W = (\hat{v}_W^{\text{rgb}}, \hat{v}_W^{\text{afd}}, \hat{v}_W^{\text{heat}})$  is the tuple of predicted velocities. For  $m \in \mathcal{S}_W$ ,  $\hat{v}_W^m$  denotes its  $m$ -th stream component and has the same shape as  $Z_0^m$ . The hat denotes a predicted velocity, the subscript identifies the Expert, and the superscript identifies a World stream. History remains clean and is excluded from the velocity targets. The encoded inputs are projected to World tokens and concatenated. Within each block, masked joint attention is followed by language cross-attention and a World-specific feed-forward network. Output projections recover the velocity in each stream’s latent space.

#### 3.2.2 Action Expert.

The Action Expert projects a noisy action chunk  $U_{\tau_A}$  into action tokens and conditions on  $\ell$ ,  $s$ , and layer-wise World hidden features. Let  $\mathcal{H}_W$  denote the collection of RGB and Scalar Affordance hidden features at corresponding World layers. It contains clean historical RGB features and evolving future RGB and Scalar Affordance features, while excluding the Affordance Heatmap stream. The action velocity is

$$\hat{v}_A = \mathcal{M}_A(U_{\tau_A}, \tau_A, \ell, s; \mathcal{H}_W) \in \mathbb{R}^{T_A \times d_A}. \quad (5)$$

On the joint-generation path, the features in  $\mathcal{H}_W$  evolve with the World flow state and are exchanged within the same model evaluation. The Action head predicts a velocity in continuous action space; integrating this velocity yields the action chunk.

#### 3.2.3 Masked Joint Self-Attention.

The Experts use independent query, key, and value projections to interact in a common attention space at each layer. As shown in Fig. 2, RGB and Scalar Affordance tokens attend to both streams; Affordance Heatmap tokens attend to RGB, Scalar Affordance, and themselves; and Action tokens attend to RGB, Scalar Affordance, and themselves. This mask blocks Action-to-World information flow and prevents Affordance Heatmap features from reaching Action tokens, including through intermediate RGB or Scalar Affordance features. Affordance Heatmap supervision can still shape the shared World features through training gradients, while RGB and Scalar Affordance provide the World interface for action generation. Excluding Affordance Heatmap features reduces an additional source of RGB-dominated conditioning and is intended to mitigate RGB shortcuts, encouraging the Action Expert to make greater use of Scalar Affordance. Direct access to RGB features remains available.

### 3.3 Affordance-Aware World Representation

We represent *affordance* with two complementary components: **Scalar Affordance** identifies task-relevant interaction regions, and **Affordance Heatmap** associates those regions with their visual context. Together with RGB scene prediction, they form the three future-World streams. Throughout this paper, affordance refers to both components unless a specific stream is named. Their encoded future latents share a temporal grouping, enabling attention to relate interaction structure to visual dynamics. The exact frame and latent-slot configuration is given in Sec. B.

#### 3.3.1 RGB Future Representation.

A frozen causal video encoder  $\mathcal{E}_{\text{vid}}$  compresses the observed and future RGB sequence into history  $C$  and future target  $Z_0^{\text{rgb}}$ :

$$[C; Z_0^{\text{rgb}}] = \mathcal{E}_{\text{vid}}([I^-; I^+]). \quad (6)$$

Brackets denote temporal concatenation. Causality ensures that  $C$  depends only on observed frames, while the future slots represent appearance, object motion, and scene-state changes. Their temporal grouping defines the alignment used by the other two streams.

#### 3.3.2 Scalar Affordance Representation.

For each future frame  $t$ , a scalar field  $F_t \in [0, 1]^{H \times W}$  encodes the task-relevant object and its interaction region, where  $H$  and  $W$  are the image height and width. The field specifies *where* to interact, without prescribing embodiment-specific motor commands. Human and robot videos use the same field semantics, providing a common supervision interface across embodiments; target construction is detailed in Sec. A.

A trainable Scalar Affordance codec, comprising encoder  $\mathcal{E}_{\text{afd}}$  and decoder  $\mathcal{D}_{\text{afd}}$ , encodes the future fields  $F^+ = (F_P, \dots, F_{P+T-1})$  as the temporally aligned latent  $Z_0^{\text{afd}} = \mathcal{E}_{\text{afd}}(F^+)$ . Its reconstruction  $F_{\text{rec}}^+ = \mathcal{D}_{\text{afd}}(Z_0^{\text{afd}})$  is supervised by a frame-level MSE to preserve spatial affordance information:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{F^+} \left[ \frac{1}{THW} \sum_{t=P}^{P+T-1} \|F_{\text{rec},t} - F_t\|_F^2 \right], \quad (7)$$

where  $F_{\text{rec},t}$  is the reconstructed field for frame  $t$  and  $\|\cdot\|_F$  is the Frobenius norm. This loss supervises the encode–decode reconstruction of the clean fields. Future Scalar Affordance generation uses latent flow matching.

#### 3.3.3 Affordance Heatmap Representation.

A scalar field omits much of the appearance needed to associate an interaction region with its object. We therefore construct a complementary RGB-space target by rendering the field on the corresponding frame with a fixed function  $\mathcal{R}$ :

$$M_t = \mathcal{R}(I_t, F_t), \quad Z_0^{\text{heat}} = \Pi_{\text{fut}} \mathcal{E}_{\text{vid}}([I^-; M^+]), \quad (8)$$

where  $M^+ = (M_P, \dots, M_{P+T-1})$ , brackets denote temporal concatenation, and  $\Pi_{\text{fut}}$  retains only the future latent slots. Using the original RGB history and the same frozen video encoder preserves the temporal correspondence with  $Z_0^{\text{rgb}}$ . Affordance Heatmap prediction supplies a joint visual and interaction target, encouraging RGB and Scalar Affordance features to retain the information needed to reconstruct their association. The attention mask in Fig. 2 prevents the Action Expert from using Affordance Heatmap features as an additional conditioning stream.

### 3.4 Data Sources and Training Strategy

We first jointly pre-train on affordance-annotated human and robot interactions, then adapt the Action Expert to the target robot using vision–language–action demonstrations without affordance labels. Human samples contain  $(I^-, I^+, F^+, \ell)$ ; robot pre-training samples additionally provide  $(s, U_0)$ . Affordance Heatmaps are derived from RGB and Scalar Affordance targets for both sources. Dataset composition, loss weights, and optimization settings are provided in Secs. A and C.

### 3.4.1 Unified Flow-Matching Objective.

We instantiate Eq. 2 for each prediction branch, using its velocity output and the conditioning information specified in Sec. 3.2. Following the path in Eq. 1, all future-World streams share flow time  $\tau_W$  and receive independent Gaussian noise  $\epsilon_W^m$ , while actions use their own noise  $\epsilon_A$  and flow time  $\tau_A$ . The corresponding target velocities are  $u_W^m = \epsilon_W^m - Z_0^m$  and  $u_A = \epsilon_A - U_0$ . The branch-specific losses are

$$\begin{aligned}\mathcal{L}_m &= \mathbb{E} \left[ \frac{1}{n_m} \|\hat{v}_W^m - u_W^m\|_2^2 \right], \quad m \in \mathcal{S}_W, \\ \mathcal{L}_A &= \mathbb{E} \left[ \frac{1}{n_A} \|\hat{v}_A - u_A\|_2^2 \right].\end{aligned}\tag{9}$$

Here  $\hat{v}_W^m$  is the  $m$ -th component of the joint World output in Eq. 4, evaluated at  $\tau_W$ , and  $\hat{v}_A$  is the Action output in Eq. 5, evaluated at  $\tau_A$  with the corresponding World features. The normalization  $n_m$  counts the scalar entries in  $Z_0^m$ , and  $n_A = T_A d_A$ . The expectations cover training samples, Gaussian noise, and flow times, whose sampling scheme is summarized in Sec. D. History  $C$  remains clean and is excluded from the loss. Both Scalar Affordance and Affordance Heatmap generation are supervised by latent velocity MSE. We combine the World prediction losses with frame-level codec reconstruction:

$$\mathcal{L}_W = \mathcal{L}_{\text{rgb}} + \lambda_{\text{afd}} \mathcal{L}_{\text{afd}} + \lambda_{\text{heat}} \mathcal{L}_{\text{heat}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}.\tag{10}$$

Here the nonnegative coefficients  $\lambda_{\text{afd}}$ ,  $\lambda_{\text{heat}}$ , and  $\lambda_{\text{rec}}$  balance the auxiliary objectives relative to RGB prediction.

### 3.4.2 Stage I: Heterogeneous Joint Pre-training.

Human videos activate only the World Expert and supervise all three future-World streams. Robot trajectories activate both Experts and additionally supervise action generation through the layer-wise coupling. Their respective objectives are

$$\mathcal{L}_{\text{human}} = \mathcal{L}_W, \quad \mathcal{L}_{\text{robot}} = \mathcal{L}_W + \lambda_A \mathcal{L}_A.\tag{11}$$

No Action tokens or pseudo-action targets are constructed for human videos. Human data thus improves the World representation through shared RGB, Scalar Affordance, and Affordance Heatmap supervision, while robot data grounds that representation in continuous control.

Training begins with World-only warmup, followed by Action pre-training with a gradual introduction of action supervision. Robot Action training mixes two paths: *joint training*, which conditions action denoising on noisy World features, and *fixed-World training*, which conditions it on a completed World. The curriculum progressively favors joint training and increases timestep synchronization between the two Experts. Within the fixed-World path, detached clean targets are gradually replaced by the model’s own predictions. This eases early optimization while preparing the Action Expert for both joint generation and final refinement.

### 3.4.3 Stage II: Task-Specific Post-training.

We adapt to the target robot and task distribution using robot vision–language–action demonstrations without affordance annotations. The complete World model is frozen, and only the Action Expert is optimized with  $\mathcal{L}_A$ . Each batch retains both training paths: noisy World features at randomly sampled timesteps maintain joint-generation ability, while completed predicted Worlds train final Action refinement. Gradients stop at the World interface in both paths. This preserves the learned World representation while adapting control without restricting the Action Expert to fixed final-World conditions. The effect of predicted-World exposure during post-training is reported in Table 11.

### 3.4.4 Inference.

Given RGB history, a language instruction, and current proprioception, we initialize future RGB, Scalar Affordance, Affordance Heatmap, and action latent variables from independent Gaussian noise. The two Experts jointly generate future-World latents and actions using their respective flow schedulers, integrating from noise to data while exchanging features at each layer. History RGB latents are re-clamped after each

World update, and Action tokens read only RGB and Scalar Affordance hidden features throughout generation. After joint generation, we hold the completed predicted World fixed and refine the generated action chunk through a short Action-only denoising pass. Joint generation produces the main prediction; refinement provides final correction under the completed World. These two inference phases correspond to the noisy-World and fixed predicted-World training paths, respectively. The refined action chunk is used for control. Sampling settings are provided in Sec. D.

## 4 Experiments

We evaluate AffordanceWAM on simulation benchmarks and real-world manipulation tasks. Our experiments are organized to address three core questions:

- **Q1 (Affordance as a V-L-A Interaction Interface):** Can object-centric affordance ground language-conditioned intent in visual interaction regions and guide action generation, thereby bridging vision, language, and action?
- **Q2 (Affordance as a Human-Robot Bridge):** Does a shared affordance representation bridge human and robot interactions, enabling action-free human videos to improve robot control beyond the benefits of heterogeneous data or auxiliary supervision alone?
- **Q3 (Scalable and Effective Action Learning):** Does the benefit of affordance-mediated human experience grow with data scale and translate into stronger manipulation performance across manipulation tasks?

### 4.1 Experimental Setup

*Simulation benchmarks.* We consider RoboCasa [25] and CALVIN ABC→D [24]. RoboCasa evaluates task-specific manipulation in diverse household environments; AffordanceWAM uses 300 demonstrations per task for downstream adaptation. CALVIN ABC→D evaluates cross-environment generalization by training on environments A, B, and C and testing on the unseen environment D. Following the standard protocol, we report the success rates of completing at least one to five consecutive tasks and the average completed sequence length. Unless otherwise noted, AffordanceWAM results are averaged over multiple runs with different random seeds.

*Real-world setup.* We further compare AffordanceWAM against Cosmos Policy in the real world. The evaluation contains five basic pick-and-place tasks and three complex tasks. Both methods are fine-tuned using the same 50 trajectories per task and evaluated under identical conditions over 15 trials/task. This setting evaluates whether the advantages of AffordanceWAM persist under equal-data real-world adaptation.

### 4.2 Simulation Benchmark Results

*RoboCasa.* Table 1 reports the average success rate on RoboCasa. AffordanceWAM achieves the highest performance among the compared methods, reaching 69.4% and exceeding Cosmos Policy by 2.3 percentage points. The result establishes the overall competitiveness of affordance-aware joint world-action modeling across diverse household manipulation tasks.

| Method                      | Average SR (%) |
|-----------------------------|----------------|
| GR00T-N1 [27]               | 49.6           |
| UVA [22]                    | 50.0           |
| UWM [44]                    | 60.8           |
| $\pi_0$ [4]                 | 62.5           |
| GR00T-N1.5 [2]              | 64.1           |
| Cosmos Policy [17]          | <u>67.1</u>    |
| <b>AffordanceWAM (Ours)</b> | <b>69.4</b>    |

**Table 1** Results on RoboCasa.

*CALVIN ABC→D.* Table 2 evaluates cross-environment generalization on CALVIN ABC→D. AffordanceWAM achieves the best results across all reported metrics, reaching an average sequence length of 4.22 compared with 4.01 for the strongest included baseline. It also improves the success rates for completing one to five consecutive tasks to 96.1%, 92.8%, 85.5%, 77.8%, and 69.5%, respectively. Together with the RoboCasa results, these improvements demonstrate that the affordance-aware world-action formulation remains effective across both task-specific adaptation and unseen-environment evaluation.

| Method               | 1/5         | 2/5         | 3/5         | 4/5         | 5/5         | Avg. Len.   |
|----------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| SuSIE [3]            | 87.0        | 69.0        | 49.0        | 38.0        | 26.0        | 2.69        |
| GR-1 [32]            | 85.4        | 71.2        | 59.6        | 49.7        | 40.1        | 3.06        |
| OpenVLA [16]         | 91.3        | 77.8        | 62.0        | 52.1        | 43.5        | 3.27        |
| CLOVER [5]           | <u>96.0</u> | 83.5        | 70.8        | 57.5        | 45.4        | 3.53        |
| UniVLA [6]           | 95.5        | 85.8        | 75.4        | 66.9        | 56.5        | 3.80        |
| $\pi_0$ [4]          | 93.8        | 85.0        | 76.7        | 68.6        | 60.1        | 3.84        |
| Seer [30]            | 94.4        | 87.2        | 79.9        | 72.2        | 64.3        | 3.98        |
| VPP [12]             | 95.3        | <u>88.2</u> | <u>80.3</u> | <u>72.9</u> | <u>64.5</u> | <u>4.01</u> |
| <b>AffordanceWAM</b> | <b>96.1</b> | <b>92.8</b> | <b>85.5</b> | <b>77.8</b> | <b>69.5</b> | <b>4.22</b> |

**Table 2** Results on CALVIN ABC→D. Avg. Len. denotes the average sequence length.

### 4.3 Ablation Studies

*Affordance as a Vision-Language-Action interaction interface.* We first examine whether incorporating affordance into the generated future World improves action learning. Here affordance comprises both Scalar Affordance and Affordance Heatmap; w/o Afd. removes both streams and their supervision, leaving RGB as the only World target. Under robot-only training, affordance supervision raises RoboCasa success rate from 57.3% to 63.7% and CALVIN average sequence length from 3.75 to 3.91. When human and robot videos are jointly used, the improvement becomes substantially larger: RoboCasa success rate increases from 55.1% to 69.4%, while CALVIN average sequence length increases from 3.68 to 4.22.

| Setting          | Human | Afd. | H→A    | RoboCasa    |             |             |             |             | CALVIN      |             |
|------------------|-------|------|--------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                  |       |      |        | SR          | 1/5         | 2/5         | 3/5         | 4/5         | 5/5         | Avg. Len.   |
| <b>Full</b>      | Yes   | Yes  | Masked | <b>69.4</b> | <b>96.1</b> | <b>92.8</b> | <b>85.5</b> | <b>77.8</b> | <b>69.5</b> | <b>4.22</b> |
| w/o Human        | No    | Yes  | Masked | 63.7        | 93.6        | 86.9        | 77.1        | 70.6        | 62.4        | 3.91        |
| w/o Human & Afd. | No    | No   | Masked | 57.3        | 92.8        | 84.6        | 74.7        | 66.3        | 56.2        | 3.75        |
| w/o Afd.         | Yes   | No   | Masked | 55.1        | 92.1        | 84.2        | 73.4        | 64.7        | 53.8        | 3.68        |
| w/o H→A Mask     | Yes   | Yes  | Open   | 63.5        | 93.4        | 85.8        | 76.7        | 69.4        | 59.2        | 3.85        |

**Table 3** Ablations on RoboCasa and CALVIN ABC→D. Afd. comprises both Scalar Affordance and Affordance Heatmap; w/o Afd. removes both streams. H→A Mask indicates Affordance Heatmap-to-action attention is masked.

All four controlled settings are evaluated over three independently trained seeds. Seed-wise results and sample standard deviations are reported in Tables 8 and 9.

The two affordance components are generated under language conditioning: Scalar Affordance identifies task-relevant interaction regions, while Affordance Heatmap associates them with visual context through auxiliary World supervision. The Action Expert reads RGB and Scalar Affordance hidden features to generate continuous controls. The consistent control gains therefore support the combined affordance representation as an effective interaction interface: language specifies task intent, affordance grounds that intent in the visual world, and action generation determines how the grounded interaction is executed.

*Affordance as a Human-Robot Bridge.* As shown in Table 3, the  $2 \times 2$  factorial ablation isolates the interaction between human-video pretraining and affordance supervision. With affordance, adding human videos improves RoboCasa success from 63.7% to 69.4% and CALVIN average sequence length from 3.91 to 4.22. Without

affordance, the same human videos instead reduce the two metrics from 57.3% to 55.1% and from 3.75 to 3.68. The opposite trends across both benchmarks show that additional human data alone is insufficient; its benefit depends on the representation through which it enters the policy. This interaction amounts to 7.9 percentage points on RoboCasa and 0.38 average sequence length on CALVIN. Without a shared interaction target, human videos introduce substantial appearance and embodiment variation without clarifying which changes are relevant to control. Affordance supervision instead expresses human and robot interactions through shared Scalar Affordance and Affordance Heatmap targets, allowing both domains to supervise task-relevant World dynamics. Since human clips never activate the Action Expert, their control gains must arise through the learned World representation and its coupling to action generation. The sign reversal therefore provides direct evidence that affordance is the mechanism that makes action-free human experience useful for robot learning.

*Complementary roles of Scalar Affordance and Affordance Heatmap.* We additionally remove either the Scalar Affordance stream or the Affordance Heatmap stream, separately from w/o Afd., which removes both. Qualitative predictions reveal two distinct failure modes. Without Scalar Affordance, the RGB-dominated Affordance Heatmap permits the World Expert to model visual appearance while largely discarding interaction-specific structure. Without Affordance Heatmap, the predicted Scalar Affordance increasingly drifts from its corresponding object regions, indicating weaker consistency between visual dynamics and interaction prediction. Removing Affordance Heatmap also reduces RoboCasa success from 69.4% to 65.4% and CALVIN average sequence length from 4.22 to 3.99 (Table 11).

The two streams therefore provide complementary inductive biases. Scalar Affordance explicitly represents task-relevant interaction regions, whereas Affordance Heatmap encourages these regions to remain spatially aligned with the predicted visual world. Their distinct failure modes suggest that the benefit arises from the structured roles of the two representations, rather than treating additional prediction streams as interchangeable capacity. Qualitative examples are provided in Figure 4.

*Scaling human interaction experience.* We vary the fraction of affordance-annotated human data while keeping robot-data exposure, training updates, and all other settings fixed. As shown in Table 4, RoboCasa success increases monotonically but non-uniformly: the first 60% of human data yields only a 1.5-point gain, whereas scaling from 60% to 100% contributes a further 4.2 points, accounting for approximately 74% of the total improvement. Seed-wise values and sample standard deviations for each data fraction are provided in Table 10.

This threshold-like pattern suggests that limited human data provides insufficient interaction coverage to overcome the substantial visual and embodiment gap. Once the human dataset reaches broader coverage of objects and interaction regions, additional examples can reinforce the shared affordance structure instead of merely adding appearance variation. Since robot action supervision and optimization updates remain fixed throughout this comparison, the improvement reflects more effective use of the same action-labeled robot dataset. The result supports affordance-mediated human interaction experience as a practical route toward more data-efficient robot action learning.

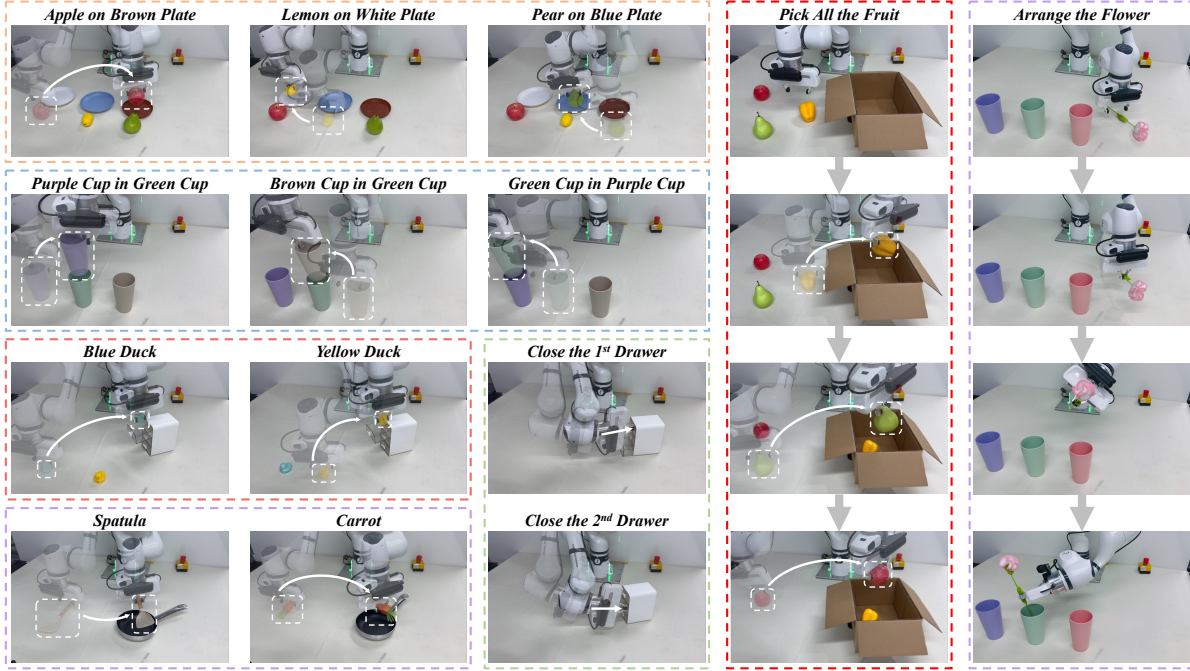
| Human data (%)  | 0    | 20   | 40   | 60   | 80   | 100         |
|-----------------|------|------|------|------|------|-------------|
| RoboCasa SR (%) | 63.7 | 63.9 | 64.3 | 65.2 | 67.1 | <b>69.4</b> |

**Table 4** Scaling affordance-annotated human video on RoboCasa. Robot-data exposure and training updates are fixed.

*Effect of Affordance Heatmap-to-action masking.* Affordance Heatmap retains substantial RGB appearance. Allowing the Action Expert to read its features introduces an additional RGB-rich pathway that may favor appearance-based shortcuts. Masking this pathway reduces the additional RGB signal and is intended to encourage greater use of Scalar Affordance. It mitigates this potential shortcut while preserving direct access to RGB features; it does not require actions to depend exclusively on Scalar Affordance. Opening the Affordance Heatmap-to-action attention path reduces RoboCasa success from 69.4% to 63.5% and CALVIN average sequence length from 4.22 to 3.85 (Table 11). These results support keeping Affordance Heatmap as auxiliary World supervision and using RGB and Scalar Affordance as the direct World interface for control, consistent with the intended reduction of RGB shortcuts.

#### 4.4 Real-World Evaluation

Figure 3 provides an overview of our real-world task suite and representative robot rollouts, comprising five simple language-conditioned tasks that test category-, color-, and shape-based object and target selection, as well as three complex tasks: *Close the Drawer*, *Pick All the Fruits*, and *Arrange the Flower*.



**Figure 3** Overview of the real-world manipulation tasks. The task suite contains simple language-conditioned manipulation tasks involving fruit–plate placement, cup insertion, object selection, and tool/object placement (left), together with three complex tasks: *Close the Drawer*, *Pick All the Fruit*, and *Arrange the Flower* (right).

| (a) Basic real-world tasks   |                  |                    |             |             |                    |             |
|------------------------------|------------------|--------------------|-------------|-------------|--------------------|-------------|
| Method                       | Category         | Color              |             | Shape       |                    | Avg.        |
|                              | Fruit to Plate   | Cup                | Duck        | Carrot      | Spatula            |             |
| Cosmos Policy [17]           | 60.0             | 40.0               | 46.7        | 40.0        | 26.7               | 42.7        |
| <b>AffordanceWAM (Ours)</b>  | <b>86.7</b>      | <b>60.0</b>        | <b>73.3</b> | <b>86.7</b> | <b>66.7</b>        | <b>74.7</b> |
| (b) Complex real-world tasks |                  |                    |             |             |                    |             |
| Method                       | Close the Drawer | Pick All the Fruit |             |             | Arrange the Flower |             |
|                              | Success          | ≥1 Fruit           | ≥2 Fruits   | ≥3 Fruits   | Success            |             |
| Cosmos Policy [17]           | 46.7             | 73.3               | 53.3        | 26.7        | 6.7                |             |
| <b>AffordanceWAM (Ours)</b>  | <b>80.0</b>      | <b>86.7</b>        | <b>73.3</b> | <b>53.3</b> | <b>20.0</b>        |             |

**Table 5** Success rates (%) on real-world manipulation tasks. Both methods use 50 fine-tuning trajectories per task and are evaluated over 15 trials under identical conditions. Panel (a) reports results on basic tasks and their macro-average. Panel (b) reports results on *Close the Drawer* and *Arrange the Flower*, together with stage-wise success rates on *Pick All the Fruit*.

Under the same adaptation budget, AffordanceWAM outperforms Cosmos Policy on all simple tasks and raises the macro-average success rate from 42.7% to 74.7% (Table 5). It also improves drawer-closing success from 46.7% to 80.0%, improves all three fruit-placement thresholds—including the three-fruit success rate from 26.7% to 53.3%—and raises flower-arrangement success from 6.7% to 20.0%. These results extend the simulation findings to physical manipulation and show that the affordance-aware interaction representation

remains beneficial under equal-data real-world adaptation.

## 5 Conclusion

In this work, we present AffordanceWAM to bridge the gap between human and robot embodiments. Rather than aligning human motion with robot actions or modeling the future with RGB imagery alone, AffordanceWAM forecasts Scalar Affordance and Affordance Heatmap as a shared interaction representation within the future World across embodiments. A directionally coupled dual-expert architecture learns this World from human and robot data while grounding action generation exclusively in robot trajectories. AffordanceWAM achieves strong results on RoboCasa, CALVIN, and real-world tasks. Controlled ablations further suggest that the benefits of human video for robot control are pronounced when affordance serves as the shared predictive representation across embodiments.

## References

- [1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023. doi: 10.1109/CVPR52729.2023.01324. [https://openaccess.thecvf.com/content/CVPR2023/html/Bahl\\_Affordances\\_From\\_Human\\_Videos\\_as\\_a\\_Versatile\\_Representation\\_for\\_Robotics\\_CVPR\\_2023\\_paper.html](https://openaccess.thecvf.com/content/CVPR2023/html/Bahl_Affordances_From_Human_Videos_as_a_Versatile_Representation_for_Robotics_CVPR_2023_paper.html).
- [2] Johan Bjorck, Valts Blukis, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Jim Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Xiaowei Jiang, Kaushil Kundalia, Jan Kautz, Zhiqi Li, Kevin Lin, Zongyu Lin, Loic Magne, Yunze Man, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Shihao Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Seonghyeon Ye, Zhiding Yu, Yizhou Zhao, Zhe Zhang, Ruijie Zheng, and Yuke Zhu. GR00T N1.5: An improved open foundation model for generalist humanoid robots. NVIDIA Research, <https://research.nvidia.com/labs/gear/gr00t-n15/>, June 2025. Accessed: 2026-07-29.
- [3] Kevin Black, Mitsuhiro Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. In *International Conference on Learning Representations*, 2024. <https://openreview.net/forum?id=c0chJTSbci>.
- [4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky.  $\pi_0$ : A vision-language-action flow model for general robot control. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010. <https://www.roboticsproceedings.org/rss21/p010.html>.
- [5] Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4411. [https://proceedings.neurips.cc/paper\\_files/paper/2024/hash/fad8962279154544ed69bb63eb14d677-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2024/hash/fad8962279154544ed69bb63eb14d677-Abstract-Conference.html).
- [6] 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. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.014. <https://www.roboticsproceedings.org/rss21/p014.html>.
- [7] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, Deli Zhao, and Hao Chen. WorldVLA: Towards autoregressive action world model, 2025. <https://arxiv.org/abs/2506.21539>.
- [8] Hanzhi Chen, Boyang Sun, Anran Zhang, Marc Pollefeys, and Stefan Leutenegger. VidBot: Learning generalizable 3D actions from in-the-wild 2D human videos for zero-shot robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27661–27672, 2025. doi: 10.1109/CVPR52734.2025.02576. [https://openaccess.thecvf.com/content/CVPR2025/html/Chen\\_VidBot\\_Learning\\_Generalizable\\_3D\\_Actions\\_from\\_In-the-Wild\\_2D\\_Human\\_Videos\\_CVPR\\_2025\\_paper.html](https://openaccess.thecvf.com/content/CVPR2025/html/Chen_VidBot_Learning_Generalizable_3D_Actions_from_In-the-Wild_2D_Human_Videos_CVPR_2025_paper.html).

- [9] 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. In *Advances in Neural Information Processing Systems*, volume 36, pages 9156–9172, 2023. doi: 10.52202/075280-0403. [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html).
- [10] James J. Gibson. The theory of affordances. In Robert Shaw and John Bransford, editors, *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.
- [11] 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, 2026. <https://arxiv.org/abs/2604.26694>.
- [12] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24328–24346. PMLR, 2025. <https://proceedings.mlr.press/v267/hu25g.html>.
- [13] Vidhi Jain, Maria Attarian, Nikhil J. Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R. Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, Igor Gilitschenski, Yonatan Bisk, and Debidatta Dwibedi. Vid2Robot: End-to-end video-conditioned policy learning with cross-attention transformers. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.052. <https://www.roboticsproceedings.org/rss20/p052.html>.
- [14] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. EgoMimic: Scaling imitation learning via egocentric video. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 13226–13233. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127989. <https://doi.org/10.1109/ICRA55743.2025.11127989>.
- [15] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human-to-robot transfer in vision-language-action models. In *Proceedings of Robotics: Science and Systems*, 2026. <https://roboticsconference.org/program/papers/72/>.
- [16] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 2679–2713. PMLR, 2025. <https://proceedings.mlr.press/v270/kim25c.html>.
- [17] Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026. <https://openreview.net/forum?id=wPEIStHxYH>.
- [18] Weijie Kong, Zhian Su, Wei Yu, and Huixu Dong. AffordVLA: Injecting affordance representations into vision-language-action models via implicit feature alignment, 2026. <https://arxiv.org/abs/2605.17517>.
- [19] Baoyu Li, Xinchun Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data, 2026. <https://arxiv.org/abs/2607.08436>.
- [20] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, Yan Peng, and Feifei Feng. CoA-VLA: Improving vision-language-action models via visual-text chain-of-affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9759–9769, 2025. doi: 10.1109/ICCV51701.2025.00910. [https://openaccess.thecvf.com/content/ICCV2025/html/Li\\_CoA-VLA\\_Improving\\_Vision-Language-Action\\_Models\\_via\\_Visual-Text\\_Chain-of-Affordance\\_ICCV\\_2025\\_paper.html](https://openaccess.thecvf.com/content/ICCV2025/html/Li_CoA-VLA_Improving_Vision-Language-Action_Models_via_Visual-Text_Chain-of-Affordance_ICCV_2025_paper.html).
- [21] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. In *Proceedings of Robotics: Science and Systems*, 2026. <https://roboticsconference.org/program/papers/16/>.
- [22] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.074. <https://www.roboticsproceedings.org/rss21/p074.html>.

- [23] Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable VLA pretraining in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35047–35058, 2026. [https://openaccess.thecvf.com/content/CVPR2026/html/Luo\\_Joint-Aligned\\_Latent\\_Action\\_Towards\\_Scalable\\_VLA\\_Pretraining\\_in\\_the\\_Wild\\_CVPR\\_2026\\_paper.html](https://openaccess.thecvf.com/content/CVPR2026/html/Luo_Joint-Aligned_Latent_Action_Towards_Scalable_VLA_Pretraining_in_the_Wild_CVPR_2026_paper.html).
- [24] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. doi: 10.1109/LRA.2022.3180108. <https://doi.org/10.1109/LRA.2022.3180108>.
- [25] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of household tasks for generalist robots. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.050. <https://www.roboticsproceedings.org/rss20/p050.html>.
- [26] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. RT-Affordance: Affordances are versatile intermediate representations for robot manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 8249–8257. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127525. <https://doi.org/10.1109/ICRA55743.2025.11127525>.
- [27] NVIDIA, Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Jim Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. GR00T N1: An open foundation model for generalist humanoid robots, March 2025. <https://arxiv.org/abs/2503.14734>.
- [28] Ryan Punamiya, Dhruv Patel, Patcharapong Aphiwetsa, Pranav Kuppili, Lawrence Y. Zhu, Simar Kareer, Judy Hoffman, and Danfei Xu. EgoBridge: Domain adaptation for generalizable imitation from egocentric human data. In *Human to Robot: Workshop on Sensorizing, Modeling, and Learning from Humans*, 2025.
- [29] Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. HRP: Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.068. <https://www.roboticsproceedings.org/rss20/p068.html>.
- [30] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, 2025. <https://openreview.net/forum?id=meRCKuUpmc>.
- [31] Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. HumanEgo: Zero-shot robot learning from minutes of human egocentric videos, 2026. <https://arxiv.org/abs/2605.24934>.
- [32] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024. <https://openreview.net/forum?id=NxoFmGgWC9>.
- [33] Rongtao Xu, Jian Zhang, Minghao Guo, Youpeng Wen, Haoting Yang, Min Lin, Jianzheng Huang, Zhe Li, Kaidong Zhang, Liqiong Wang, Yuxuan Kuang, Meng Cao, Feng Zheng, and Xiaodan Liang. A0: An affordance-aware hierarchical model for general robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13491–13501, 2025. [https://openaccess.thecvf.com/content/ICCV2025/html/Xu\\_A0\\_An\\_Affordance-Aware\\_Hierarchical\\_Model\\_for\\_General\\_Robotic\\_Manipulation\\_ICCV\\_2025\\_paper.html](https://openaccess.thecvf.com/content/ICCV2025/html/Xu_A0_An_Affordance-Aware_Hierarchical_Model_for_General_Robotic_Manipulation_ICCV_2025_paper.html).
- [34] Siyu Xu, Zijian Wang, Yunke Wang, Chenghao Xia, Tao Huang, and Chang Xu. Affordance field intervention: Enabling VLAs to escape memory traps in robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37206–37215, 2026. [https://openaccess.thecvf.com/content/CVPR2026/html/Xu\\_Affordance\\_Field\\_Intervention\\_Enabling\\_VLAs\\_to\\_Escape\\_Memory\\_Traps\\_in\\_CVPR\\_2026\\_paper.html](https://openaccess.thecvf.com/content/CVPR2026/html/Xu_Affordance_Field_Intervention_Enabling_VLAs_to_Escape_Memory_Traps_in_CVPR_2026_paper.html).
- [35] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025. [https://proceedings.iclr.cc/paper\\_files/paper/2025/hash/45d74e190008c7bff2845ffc8e3facd3-Abstract-Conference.html](https://proceedings.iclr.cc/paper_files/paper/2025/hash/45d74e190008c7bff2845ffc8e3facd3-Abstract-Conference.html).

- [36] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies, 2026. <https://arxiv.org/abs/2602.15922>.
- [37] Qize Yu, Jiadi You, Yuran Wang, Jiaqi Liang, Bowen Ping, Yang Tian, Yue Chen, Minghong Cai, Zeying Gong, Ruihai Wu, Yinchuan Li, Junwei Liang, and Yingcong Chen. AffordanceVLA: A vision-language-action model empowering action generation through affordance-aware understanding, 2026. <https://arxiv.org/abs/2606.06155>.
- [38] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 1541–1566. PMLR, 2025. <https://proceedings.mlr.press/v270/yuan25a.html>.
- [39] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination?, 2026. <https://arxiv.org/abs/2603.16666>.
- [40] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction in robotics. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 4005–4020. PMLR, 2025. <https://proceedings.mlr.press/v270/yuan25c.html>.
- [41] Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, Jiaming Zhou, Yishu Shen, Yudong Jin, Fangyi Xu, Shuailei Ma, Jiaqi Liao, Guanxing Lu, Zifan Shi, Yongkun Wen, Yujie Zhao, Weixuan Tang, Xinyang Wang, Chaojian Li, Jiapeng Zhu, Ka Leong Cheng, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Native video-action pretraining for generalizable robot control, 2026. <https://arxiv.org/abs/2607.08639>.
- [42] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loïc Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. FLARE: Robot learning with implicit world modeling. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 3952–3971. PMLR, 2025. <https://proceedings.mlr.press/v305/zheng25a.html>.
- [43] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. EgoScale: Scaling dexterous manipulation with diverse egocentric human data, 2026. <https://arxiv.org/abs/2602.16710>.
- [44] 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. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.015. <https://www.roboticsproceedings.org/rss21/p015.html>.

# Appendix

This supplement provides data and implementation details, training and sampling settings, seed-wise results, and additional quantitative and qualitative analyses. Notation and model objectives follow Sec. 3.

## A Data Preparation

### A.1 Sources and Splits

Human pre-training videos come from Ego4D-FHO, EPIC-KITCHENS VISOR, HOI4D, H2O, and MEC-CANO. Robot trajectories come from RoboInter (DROID and RH20T), BridgeData V2, InternData-A1, and RoboCasa365. Table 6 summarizes the retained scale, sampling ratio, and downstream data budgets. The stratified validation holdout is used for training-time validation and checkpoint selection.

| Scope             | Scale or configuration                         |
|-------------------|------------------------------------------------|
| Stage-I human     | More than 300 hours                            |
| Stage-I robot     | Approximately 0.4 million trajectories         |
| Stage-I sampling  | Human:robot = 1:1                              |
| Validation        | Fixed 2,000 clips plus a stratified 1% holdout |
| RoboCasa Stage II | $24 \times 300 = 7,200$ demonstrations         |
| CALVIN Stage II   | Complete ABC training split                    |
| Real Stage II     | $8 \times 50 = 400$ trajectories               |

**Table 6** Data composition for heterogeneous pre-training, validation, and task-specific post-training.

### A.2 Affordance Target Construction

We construct targets from the source datasets’ existing object and interaction annotations. The following procedure implements the shared field semantics defined in Sec. 3.3.

1. **Clip and annotation alignment.** Sample clips using the temporal and spatial configuration in Table 7, and align object and interaction-region annotations with the sampled timestamps.
2. **Label normalization.** Map source-specific annotations to the common task-relevant object and interaction-region labels.
3. **Field rasterization.** Construct  $F_t$  using a Gaussian kernel and per-frame peak normalization. Interpolate missing annotations over at most two frames; under occlusion, retain only visible support.
4. **Affordance Heatmap targets.** Apply the fixed renderer in Eq. 8 to each aligned RGB-Scalar Affordance pair.

### A.3 Filtering and Quality Control

Human clips are filtered for excessive camera shake using motion derived from camera parameters and for invalid Affordance Heatmaps under the central-9/16 region criterion. Retained clips undergo random manual review. Robot trajectories are checked for temporal alignment, valid states and actions, and valid interaction annotations. These checks precede the retained counts in Table 6.

## B Model Implementation

### B.1 Backbone and Expert Dimensions

Table 7 specifies the pretrained backbone, Expert dimensions, and prediction interface. Native hidden widths and joint-attention dimensions are listed separately, following the architecture in Sec. 3.2.

| Component            | Configuration                                                 |
|----------------------|---------------------------------------------------------------|
| World initialization | Wan2.2-TI2V-5B                                                |
| World Expert         | 30 layers; hidden dimension 3,072                             |
| Action Expert        | 30 layers; hidden dimension 1,024; FFN dimension 4,096        |
| Joint attention      | 24 heads $\times$ 128 dimensions = 3,072                      |
| Training clip        | 17 frames: 5 observed and 12 future; 10 FPS; $256 \times 320$ |
| Deployment history   | 5 observed RGB frames                                         |
| Action output        | $12 \times 7$ continuous values                               |

**Table 7** Model dimensions and input/output configuration.

## B.2 Temporal Latent Layout

The configuration corresponds to  $P = 5$ ,  $T = T_A = 12$ , and  $d_A = 7$  in the Method. The causal video encoder produces five temporal slots: one for frame 0, one for frames 1–4, and three for the future groups 5–8, 9–12, and 13–16. The first two slots form  $C$ ; the remaining three form  $Z_0^{\text{rgb}}$ . The Scalar Affordance codec produces three slots with the same future grouping, and the Affordance Heatmap encoder output is restricted to these three future slots. Each slot contains spatial latent features, not a single Transformer token.

## C Training Configuration

Stage I calibrates the Scalar Affordance codec and warms up the World Expert before introducing Action training. World-only warmup uses 5,000 updates for the 8-slot configuration and 7,000 for the 11-slot configuration. During Action pre-training, the action-loss weight increases gradually from zero to one. Fixed-World conditioning is emphasized early, then reduced as joint noisy-World/Action training becomes dominant. Within the joint path, synchronized timesteps become more frequent; within the fixed-World path, clean target Worlds are gradually replaced entirely by predicted Worlds. This curriculum prepares joint generation and final refinement together.

The final loss weights are  $\lambda_{\text{afd}} = 0.2$ ,  $\lambda_{\text{heat}} = 0.5$ ,  $\lambda_{\text{rec}} = 0.1$ , and  $\lambda_A = 1$ , with unit weight on RGB prediction. Here afd denotes the Scalar Affordance branch and heat denotes the Affordance Heatmap branch. We use AdamW and BF16 training on 64 NVIDIA H200 GPUs.

Stage II freezes the complete World path and trains only the Action Expert for 10,000 updates at a learning rate of  $10^{-5}$  for each target benchmark. Within each batch, 75% of samples use frozen World features at randomly sampled noise timesteps to preserve joint-generation ability; the remaining 25% use completed predicted Worlds to train final Action refinement. Both paths stop gradients at the World interface.

## D Joint Generation and Action Refinement

Joint training uses noisy World and Action latents along the flow path in Eq. 1. World timesteps are sampled from  $\mathcal{U}(0, 1)$ ; Action timesteps are either synchronized with World or sampled independently. Fixed-World training holds the completed World condition fixed while denoising actions. During pre-training this condition transitions from detached clean targets to model predictions; Stage II uses completed predicted Worlds for its fixed-World samples. In both paths, Action reads RGB and Scalar Affordance features, with Affordance Heatmap features excluded by the attention mask.

All main AffordanceWAM policy results use the same inference procedure: **35 World steps and 35 Action steps of joint generation, followed by 8 steps of Action refinement with strength 0.25**. Refinement operates on the jointly generated action chunk while keeping the completed predicted World fixed. Thus, noisy-World training supports main generation, and fixed predicted-World training supports final correction.

## E Evaluation Protocols and Seed-Wise Results

### E.1 Evaluation Units and Aggregation

Controlled RoboCasa and CALVIN experiments use three independently trained seeds. Each RoboCasa seed is evaluated on 24 tasks with 50 episodes per task, giving 1,200 episodes per seed and 3,600 per setting. Each CALVIN seed is evaluated on 1,000 five-task sequences, giving 3,000 per setting. We report the mean and sample standard deviation across training seeds; the seed is the unit of run-to-run variability.

Tables 8 and 9 expand the aggregate results in Table 3. In both tables, w/o Afd. removes both Scalar Affordance and Affordance Heatmap, leaving an RGB-only World target; w/o Human & Afd. additionally removes human pre-training data. Seed identifiers are aligned across the controlled settings.

| Setting          | Seed 1 | Seed 2 | Seed 3 | Mean $\pm$ std                   |
|------------------|--------|--------|--------|----------------------------------|
| Full             | 69.9   | 69.4   | 68.9   | <b>69.4 <math>\pm</math> 0.5</b> |
| w/o Human        | 63.8   | 63.2   | 64.1   | 63.7 $\pm$ 0.5                   |
| w/o Human & Afd. | 57.4   | 56.8   | 57.7   | 57.3 $\pm$ 0.5                   |
| w/o Afd.         | 55.2   | 54.5   | 55.6   | 55.1 $\pm$ 0.6                   |

**Table 8** Seed-wise controlled ablations on RoboCasa. Values are average task success rates in percent. Each seed contains 1,200 evaluation episodes; the final column reports the mean and sample standard deviation over three independently trained seeds.

| Setting          | Run            | 1/5 (%)                          | 2/5 (%)                          | 3/5 (%)                          | 4/5 (%)                          | 5/5 (%)                          | Avg. Len.                         |
|------------------|----------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|-----------------------------------|
| Full             | Seed 1         | 96.4                             | 92.8                             | 85.5                             | 78.6                             | 69.5                             | 4.23                              |
|                  | Seed 2         | 95.8                             | 93.4                             | 84.8                             | 77.0                             | 70.5                             | 4.22                              |
|                  | Seed 3         | 96.1                             | 92.2                             | 86.2                             | 77.8                             | 68.5                             | 4.21                              |
|                  | Mean $\pm$ std | <b>96.1 <math>\pm</math> 0.3</b> | <b>92.8 <math>\pm</math> 0.6</b> | <b>85.5 <math>\pm</math> 0.7</b> | <b>77.8 <math>\pm</math> 0.8</b> | <b>69.5 <math>\pm</math> 1.0</b> | <b>4.22 <math>\pm</math> 0.01</b> |
| w/o Human        | Seed 1         | 93.6                             | 87.5                             | 76.4                             | 70.6                             | 63.1                             | 3.91                              |
|                  | Seed 2         | 94.1                             | 86.3                             | 77.8                             | 69.9                             | 62.4                             | 3.91                              |
|                  | Seed 3         | 93.1                             | 86.9                             | 77.1                             | 71.3                             | 61.7                             | 3.90                              |
|                  | Mean $\pm$ std | 93.6 $\pm$ 0.5                   | 86.9 $\pm$ 0.6                   | 77.1 $\pm$ 0.7                   | 70.6 $\pm$ 0.7                   | 62.4 $\pm$ 0.7                   | 3.91 $\pm$ 0.01                   |
| w/o Human & Afd. | Seed 1         | 93.3                             | 84.6                             | 75.4                             | 66.3                             | 56.2                             | 3.76                              |
|                  | Seed 2         | 92.3                             | 85.2                             | 74.7                             | 65.6                             | 56.9                             | 3.75                              |
|                  | Seed 3         | 92.8                             | 84.0                             | 74.0                             | 67.0                             | 55.5                             | 3.73                              |
|                  | Mean $\pm$ std | 92.8 $\pm$ 0.5                   | 84.6 $\pm$ 0.6                   | 74.7 $\pm$ 0.7                   | 66.3 $\pm$ 0.7                   | 56.2 $\pm$ 0.7                   | 3.75 $\pm$ 0.01                   |
| w/o Afd.         | Seed 1         | 92.6                             | 84.2                             | 73.4                             | 64.0                             | 54.5                             | 3.69                              |
|                  | Seed 2         | 92.1                             | 83.6                             | 74.1                             | 65.4                             | 53.1                             | 3.68                              |
|                  | Seed 3         | 91.6                             | 84.8                             | 72.7                             | 64.7                             | 53.8                             | 3.68                              |
|                  | Mean $\pm$ std | 92.1 $\pm$ 0.5                   | 84.2 $\pm$ 0.6                   | 73.4 $\pm$ 0.7                   | 64.7 $\pm$ 0.7                   | 53.8 $\pm$ 0.7                   | 3.68 $\pm$ 0.01                   |

**Table 9** Seed-wise controlled ablations on CALVIN ABC $\rightarrow$ D. Each seed contains 1,000 five-task evaluation sequences. Values after  $\pm$  are sample standard deviations over three independently trained seeds.

### E.2 Human-Data Scaling

Table 10 provides the three individual runs and sample standard deviations underlying Table 4. The experimental controls follow Sec. 4.3.

| Human data (%) | 0    | 20   | 40   | 60   | 80   | 100         |
|----------------|------|------|------|------|------|-------------|
| Seed 1         | 63.8 | 64.4 | 64.3 | 65.7 | 67.1 | 69.9        |
| Seed 2         | 63.2 | 63.4 | 63.8 | 64.7 | 67.7 | 69.4        |
| Seed 3         | 64.1 | 63.9 | 64.8 | 65.2 | 66.5 | 68.9        |
| Mean           | 63.7 | 63.9 | 64.3 | 65.2 | 67.1 | <b>69.4</b> |
| Std            | 0.5  | 0.5  | 0.5  | 0.5  | 0.6  | 0.5         |

**Table 10** Seed-wise RoboCasa success rates for the human-data scaling experiment. The standard deviation is computed across three training seeds.

## F Additional Quantitative Evaluations

### F.1 Additional Architectural Ablations

Table 11 extends the controls in Table 3 with alternative fusion designs, removal of predicted-World post-training, and removal of the Affordance Heatmap stream. Data and update budgets are held fixed. The fusion and masking controls also match token count and trainable parameters.

| Variant                           | RoboCasa SR       | CALVIN Len.        |
|-----------------------------------|-------------------|--------------------|
| Masked Joint Self-Attn.           | <b>69.4 ± 0.5</b> | <b>4.22 ± 0.01</b> |
| Late fusion                       | 65.9 ± 0.6        | 4.00 ± 0.01        |
| Independent World→Action          | 64.5 ± 0.6        | 3.90 ± 0.01        |
| Open Affordance Heatmap→Action    | 63.5 ± 0.6        | 3.85 ± 0.01        |
| w/o predicted-World post-training | 66.6 ± 0.6        | 4.06 ± 0.01        |
| w/o Affordance Heatmap stream     | 65.4 ± 0.6        | 3.99 ± 0.01        |

**Table 11** Architectural ablations under the same data and update budget. Fusion and masking controls additionally match token count and trainable parameters; the stream-removal row is a targeted component ablation.

*Variant definitions.* In *late fusion*, the Experts communicate only through a final World-to-Action fusion module. The *independent pipeline* first completes World prediction, then supplies it as a fixed condition to a separately executed Action Expert. *Open Affordance Heatmap→Action* removes the block on Affordance Heatmap keys for Action queries. The *predicted-World post-training* ablation removes Stage II exposure to World predictions; the *Affordance Heatmap* ablation removes that auxiliary stream. The post-training ablation concerns training and is separate from the final inference-time Action refinement in Sec. D.

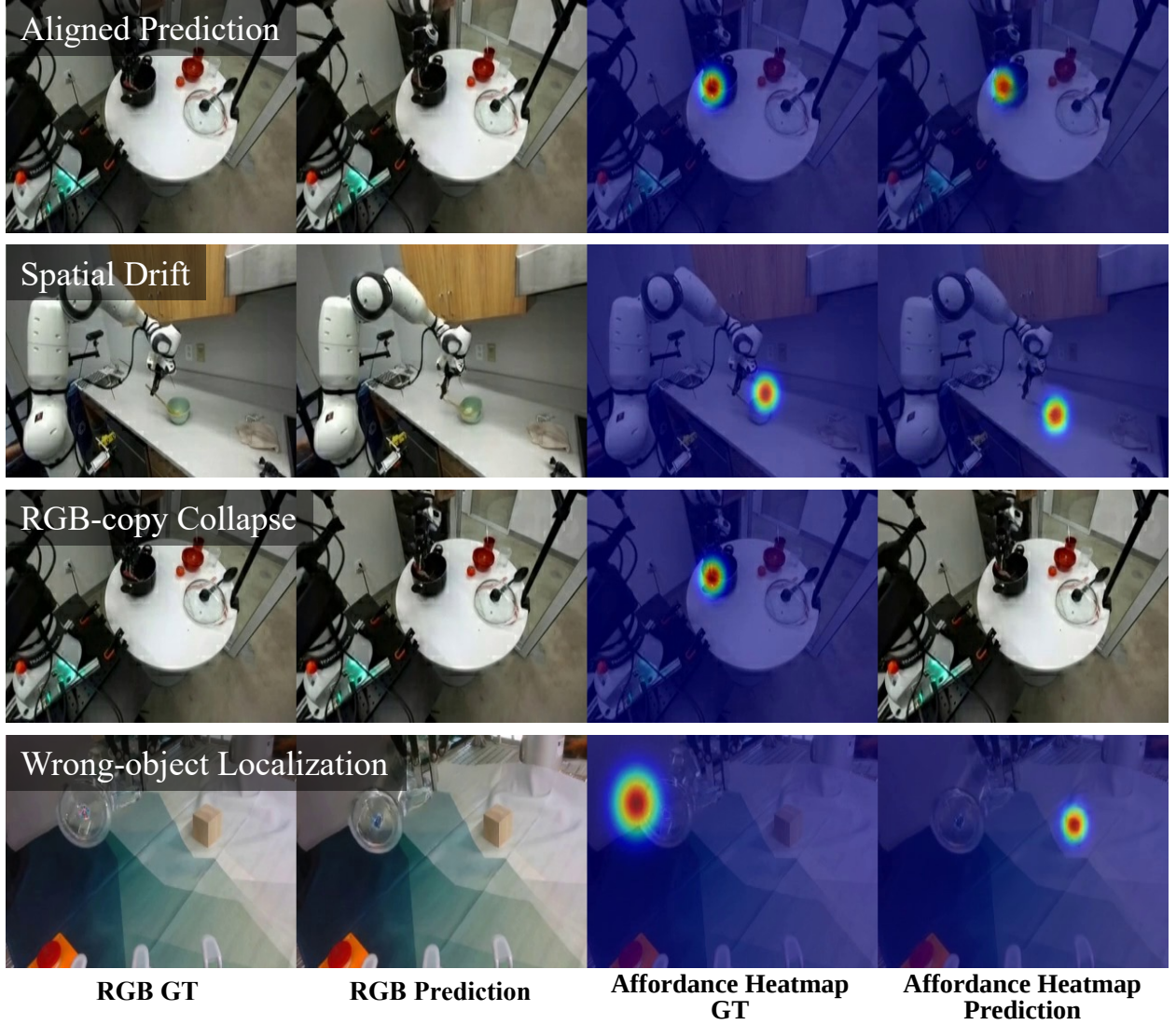
## G Qualitative Analysis

### G.1 Qualitative Predictions and Failure Modes

Figure 4 compares generated RGB observations and Affordance Heatmaps with their clean targets. The aligned example localizes the interaction region on the intended object. The remaining rows illustrate three failure types:

- **Spatial drift:** a predicted interaction region is displaced from the corresponding object despite plausible RGB prediction.
- **RGB-copy collapse:** appearance structure dominates the Affordance Heatmap prediction, weakening interaction-specific localization.
- **Wrong-object localization:** a compact prediction selects an object inconsistent with the instruction.

Further diagnostic categories are *correct World*, *failed control*, where plausible visual predictions are followed by contact, kinematic, or execution errors, and *long-horizon degradation*, where localization quality worsens across successive future groups. These categories separate representation errors from downstream control failures.



**Figure 4** Qualitative RGB and Affordance Heatmap predictions. From left to right: RGB GT, RGB Prediction, Affordance Heatmap GT, and Affordance Heatmap Prediction, where GT denotes ground truth. From top to bottom: aligned prediction, spatial drift, RGB-copy collapse, and wrong-object localization.