VIDEO PREDICTION POLICY 2: PREDICT BETTER, ACT BETTER

Anonymous authors
Paper under double-blind review

ABSTRACT

World action models (WAMs) have emerged as an important class of generalist robot policies, aiming to transfer video prediction priors to action learning. However, we find that existing WAMs frequently produce incorrect motion predictions in open-ended environment, leading to erroneous actions. We attribute this limitation to two factors: (1) base video models are not optimized for manipulation, and (2) naively incorporating action components into video models can substantially degrade their generalization capabilities. We introduce Video Prediction Policy 2 (VPP2), a WAM that enables strong zero-shot generalization in both video prediction and action generation. First, we curate a large-scale, diverse dataset of manipulation videos to continue pretraining the base video foundation model. We annotate video clips with detailed captions and perform *event-level* video pretraining to promote generalization across open-ended manipulation tasks. Second, we post-train and distill the video model into a single-step visual planner with fixed prediction horizon. Finally, we introduce action module via a mixture-of-transformers (MoT) architecture to learn implicit inverse dynamics model. Experiments demonstrate three key results: (1) VPP2-14B outperforms Cosmos3-64B by 11.0% points in video prediction instruction-following success rate on open-ended tasks; (2) VPP2 surpasses the strongest baseline by 18.5% points in success rate on real-world zero-shot ALOHA manipulation tasks; and (3) following benchmark-specific post-training, VPP2 achieves the highest success rates among evaluated methods on the challenging LIBERO-Pro, LIBERO-OOD, and RoboDojo benchmarks. Videos can be found at `sites.google.com/view/video-prediction-policy-2`.

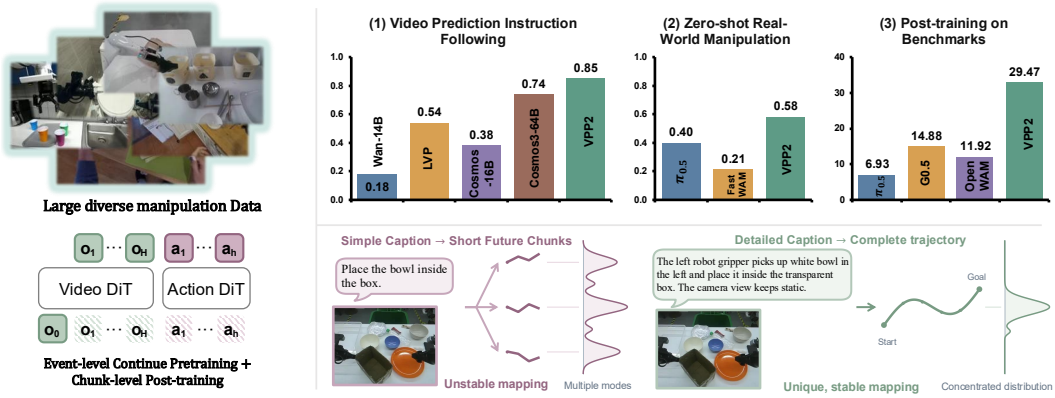


Figure 1: A typical robot policy maps an simple instruction and an observation into an short action chunk. As datasets scale, this mapping becomes increasingly multi-modes and uncertain, leading models to learn spurious short-horizon correlations. VPP2 first establishes consistent semantic-to-trajectory mappings via *next event video prediction* with detailed caption, then post-train and distill model to generate short action chunk.

1 INTRODUCTION

World action models (WAMs) are developing rapidly and have become an important class of generalist robot policies (Hu et al., 2024; Liao et al., 2025; Kim et al., 2026; Yuan et al., 2026; Ma et al., 2026;

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

Li et al., 2026a;b; AgiBot Research Team et al., 2026; Ye et al., 2026). Many WAMs build on video foundation models (Wan et al., 2025; Agarwal et al., 2025; Yang et al., 2025), aiming to transfer their rich priors about physical dynamics to policy learning. This premise is compelling: accurate, instruction-conditioned video predictions can guide action generation through an explicit or implicit inverse dynamics model (Du et al., 2023; Hu et al., 2024). However, prior work (Zhang et al., 2026a) has found that existing WAMs often perform well only on a narrow set of seen tasks, while producing unreliable future motion predictions in unseen, out-of-distribution scenarios. These prediction failures, in turn, lead to erroneous actions. In recent benchmarks that require generalization, robot policies, including WAMs, with near-perfect success rates on standard evaluation suites suffer substantial performance drops under perturbed task configurations and novel skill compositions (Zhou et al., 2025; Li, 2025; Mishra et al., 2026).

We identify two key factors that limit the generalization of existing WAMs to unseen scenarios. First, base video models are primarily optimized for creative content generation and aesthetic quality rather than precise physical dynamics, and therefore often fail to follow manipulation instructions faithfully (Chen et al., 2025; Zhang et al., 2026a). Second, introducing action-specific components or training objectives into pretrained video foundation models can substantially degrade their generalization capabilities (Mishra et al., 2026).

To address these issues, we introduce Video Prediction Policy 2 (VPP2), a WAM that achieves strong zero-shot generalization in both video prediction and action generation. Our first objective is to train a generalizable video foundation that can faithfully follow instruction and make future predictions grounded in the current observation, without unsupported changes to the scene or task. To this end, we annotate each clip with a detailed caption specifying the manipulation process, active end effector, and target object, and perform *event-level* video prediction training. As illustrated in Figure 1, detailed captions and complete event-level trajectory prediction substantially reduce uncertainty in future prediction and encourage consistent mappings from semantic descriptions to visual trajectories, improving generalization across diverse instructions. We then post-train the video model to predict fixed-horizon future chunks and apply consistency distillation (Song et al., 2023) to obtain a single-step visual planner for real-time robot execution. Finally, we introduce an action expert through a mixture-of-transformers (MoT) architecture (Liang et al., 2025) to learn an implicit inverse dynamics model conditioned on the predicted future. During policy execution, we pair VPP2 with a VLM planner (Bai et al., 2025; Bytedance Seed, 2026) that translates open-ended instructions into explicit subtask instructions.

Our experiments demonstrate three key advantages of VPP2. (1) VPP2-14B follows manipulation instructions faithfully, outperforming Cosmos3-64B (Agarwal et al., 2026) by 11% in instruction-following success rate on open-ended tasks. (2) VPP2 exhibits strong zero-shot manipulation capabilities on a real-world ALOHA platform (Zhao et al., 2023), achieving an average success rate of 58.5% across 10 task categories, compared with 40.0% for $\pi_{0.5}$ (Intelligence et al., 2025) and 20.5% for Fast-WAM (Yuan et al., 2026). (3) Following benchmark-specific post-training, VPP2 achieves the highest success rates among evaluated methods on LIBERO-Pro (Zhou et al., 2025) (45.0%, compared with 11.0% for the strongest baseline), LIBERO-OOD (Li, 2025; Mishra et al., 2026) (63.9%), and RoboDojo (Chen et al., 2026) (29.47%, state-of-the-art).

2 Data Process Pipeline

In this section, we describe our data collection and curation pipeline, including how we unify the inputs and outputs across diverse manipulation datasets.

Video Sources. We compile a large-scale, diverse collection of manipulation videos for continued pretraining of an open-source video foundation model (Wan et al., 2025). Our data span three categories: robot manipulation datasets, human activity datasets, and general-purpose video datasets. For general-purpose video datasets, we apply an extensive filtering pipeline similar to that of LVP (Chen et al., 2025) to retain only videos depicting manipulation-related activities. Table 1 summarizes the datasets used for training.

Video Filtering, Segmentation and Captioning. We first remove corrupted trajectories, including those with camera failures or abrupt discontinuities, and segment the remaining trajectories into semantically coherent clips lasting 1–12 seconds. We identify candidate boundaries using heuristic cues associated with natural transitions, such as local minima in human-hand or end-effector velocity

Data Type	Embodiment Type	Data Sources
Robot	Single-arm	OXE, RoboMIND, DROID, RH20T, Molmoact
	Dual-arm	AgibotWorld-Beta, RoboCOIN, RDT, GM100, ABC-130k, Self-collected Aloha
	Mobile & humanoid	InternData-A1, Galaxea Open-World, Self-collectd dexterous hands
Human	Human hands	Open-source: EgoDex, Ego4D, Epic-Kitchen Self-collected egocentric data
General Video	Human hands	Open-vid, Panda-70M

Table 1: We use various types of manipulation datasets and perform different filtering strategy for different datasets to ensure diversity and high-quality.

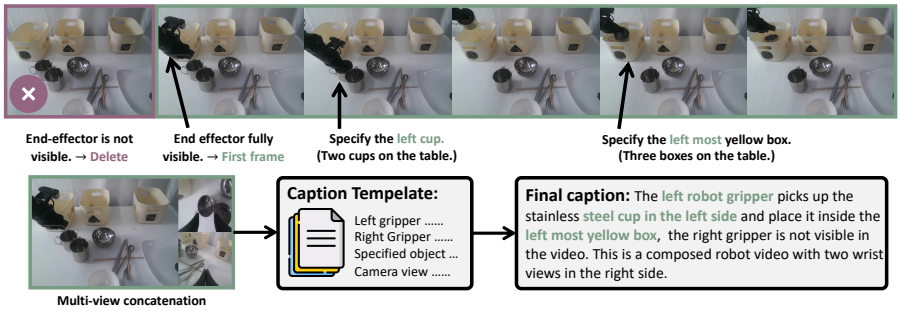


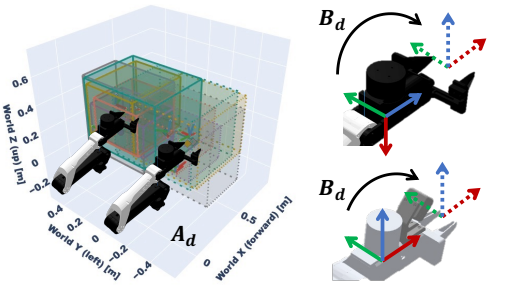
Figure 2: An example of our data process pipeline. The central objective is to reduce uncertainty in future prediction, thereby encouraging the model to learn a consistent mapping from conditioning information to sub-task trajectories. To this end, each caption includes detailed task descriptions, explicit End-Effector Identification, visible target object, and camera-view changes.

and moments when the gripper or fingers open or close. A vision-language model (VLM) then refines the segmentation by merging adjacent clips where appropriate and adjusting their start and end points.

After segmentation, we generate detailed captions that describe the task, explicitly identify the active end effectors, unambiguously specify the target objects, and characterize the camera views and motion. Figure 2 illustrates an example, and Appendix A.1 provides further details.

Unified T-shape Multi-view Input across Datasets. Robotics datasets commonly contain observations from multiple camera views. We arrange them into a T-shaped composite image, with the primary view on the left and two auxiliary views stacked vertically on the right, as illustrated in the bottom-left panel of Figure 2. When fewer than two auxiliary views are available, we fill the missing slots with black placeholder images.

Unified Action Space and Unified End-effector Coordinate Systems. Different robot datasets often adopt inconsistent coordinate systems and camera viewpoints, resulting in conflicting action representations for visually similar motions. For example, a leftward motion may correspond to the positive x-direction in one dataset but the negative y-direction in another, making it difficult for the model to learn the semantic meaning of each action dimension.



(a) Workspace alignment (b) End-effector alignment

Figure 3: (a) Bounding boxes show the aligned workspaces of different datasets. (b) Aligned end-effector coordinate frames.

We focus on egocentric bimanual datasets, including ALOHA-style, humanoid-style, and human egocentric datasets. These datasets account for more than 80% of the total data and share a similar bimanual structure and camera viewpoint. We explicitly align their coordinate systems at two levels: (1) *workspace alignment*, which applies a world-frame transformation so that different robots have comparable end-effector workspaces; and

(2) *end-effector alignment*, which applies a local transformation to standardize end-effector origins and axis orientations.

For an arm dataset d , let $T_d = {}^{W_d}T_{E_d} \in SE(3)$ denote the original end-effector pose, where W_d and E_d are the dataset's world and end-effector frames, respectively. We define the workspace alignment as $A_d = {}^{\bar{W}}T_{W_d}$ and the end-effector alignment as $B_d = {}^{E_d}T_{\bar{E}}$, where $\bar{W}$ and $\bar{E}$ denote the canonical world and end-effector frames. The aligned end-effector pose is given by

$$\widetilde{T}_d = A_d T_d B_d, \tag{1}$$

3 VPP2: A Generalist Policy with Zero-shot Capability

Overview. Our large-scale, diverse, and densely annotated dataset enables us to train a policy with strong generalization capabilities. VPP2 builds upon the pretrained Wan2.1-I2V-14B model (Wan et al., 2025) and introduce a action expert via mixture-of-transformers(MoT) architecture (Liang et al., 2025). Our training objective is twofold: we first adapt the video model to produce generalizable future predictions at real-time inference speed for open-ended manipulation tasks, and then train the action expert conditioned on the KV cache of the video model. To maximize policy generalization, we organize training into the stages summarized in Figure 4.

3.1 Video Prediction Model Training Pipeline

Stage 1: Event-level Video Model Continued Pre-training. We first adapt a pretrained video generation model to the manipulation domain through event-level continued pre-training. Each training example covers a complete manipulation subtask, aligning the subtask description with the corresponding visual evolution from the initial observation to subtask completion.

Given a demonstration $\tau = (o_0, o_1, \ldots, o_T)$, we uniformly sample N future frames across the entire event:

$$\mathbf{y}_{\text{event}} = \left(o_0, o_{\lfloor T/N \rfloor}, o_{\lfloor 2T/N \rfloor}, \ldots, o_T\right), \tag{2}$$

where T is the final frame index and N excludes the initial conditioning frame o_0. Let $\mathbf{x}_1 = \mathcal{E}(\mathbf{y}_{\text{event}})$ denote the corresponding video latent representation, where $\mathcal{E}$ is the video encoder. We construct the interpolated latent as

$$\mathbf{x}_s = (1-s)\boldsymbol{\epsilon} + s\mathbf{x}_1, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad s \in [0,1], \tag{3}$$

where s denotes flow time, with $s = 0$ corresponding to noise and $s = 1$ to data. The model is optimized with the flow matching objective (Lipman et al., 2023)

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{(\mathbf{x}_1, c),\, \boldsymbol{\epsilon},\, s}\left[\left\| v_\theta(\mathbf{x}_s, s, c) - (\mathbf{x}_1 - \boldsymbol{\epsilon})\right\|_2^2\right], \tag{4}$$

where v_θ predicts the flow velocity and c includes the subtask instruction and initial observation o_0.

In experiments, we predict 16 sparse frames at a resolution of 416×240 and continue pre-training Wan2.1-I2V-14B for $30{,}000$ optimization steps with batch size $1{,}024$ and learning rate 1e-5. Our experiments show that the resulting model achieves better instruction following on manipulation tasks than the base Wan2.1-I2V-14B model (Wan et al., 2025). Despite having only 14B parameters, VPP2 also outperforms the 64B Cosmos3-Super-Image2Video model (Agarwal et al., 2026) on our manipulation instruction-following evaluation.

Stage 2: Chunk-level Video Model Post-training. Event-level continued pre-training enables the model generalize the best in manipulation tasks. However, sampling a fixed number of frames from events of different durations produces variable temporal intervals between predicted frames, complicating their alignment with downstream actions. We therefore further post-train the model to predict a fixed-duration future chunk, establishing a consistent temporal scale for action learning.

Let H denote the prediction horizon in seconds and f the dataset frames per seconds (fps). The nominal interval between sampled frames is $\delta = fH/N$, measured in dataset frames. For a chunk starting at frame t, the target video is

$$\mathbf{y}^{(t)}_{\text{chunk}} = \left(o_t, o_{t+\lfloor \delta \rfloor}, o_{t+\lfloor 2\delta \rfloor}, \ldots, o_{t+\lfloor N\delta \rfloor}\right), \tag{5}$$

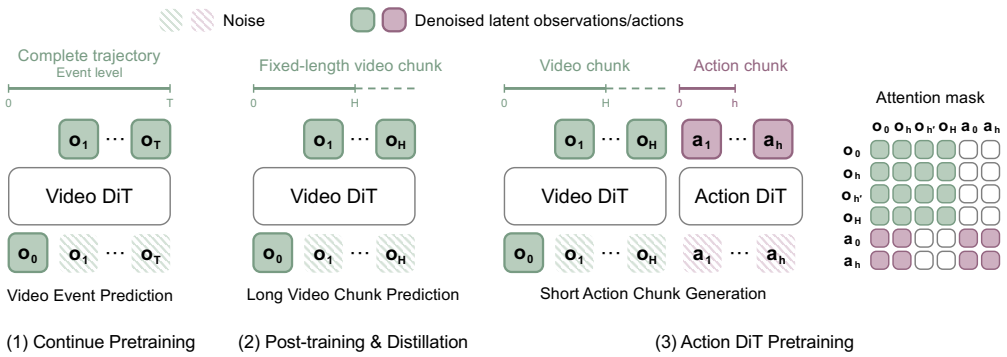


Figure 4: The VPP2 training pipeline is designed to maximize generalization. Stage 1 uses large-scale, event-level video pretraining to learn a generalizable video model for manipulation. Stage 2 post-trains and distills this model to predict long-horizon video chunks spanning 8 seconds in a single forward pass, taking approximately 0.1 seconds. Finally, Stage 3 trains an action expert to generate 2-second action chunks conditioned on the one-step video latents.

with conditioning inputs c containing the observation o_t and the corresponding subtask instruction. Indices $t + \lfloor fH \rfloor$ is bounded to max length T and we optimize the same flow matching objective in Eq. equation 4, with $\mathbf{x}_1 = \mathcal{E}(\mathbf{y}^{(t)}_{\text{chunk}})$.

Stage 2: Consistency Distillation. We distill the chunk-level video model for single-step generation (Song et al., 2023) by enforcing consistent terminal predictions at adjacent flow times:

$$\mathcal{L}_{\text{CD}} = \mathbb{E}\left[\lambda(s, s') \left\| F_\theta(\mathbf{x}_s, s, c) - \text{stopgrad}\left(F_{\bar{\theta}}(\widehat{\mathbf{x}}_{s'}, s', c)\right)\right\|_2^2\right], \tag{6}$$

where $0 \le s < s' \le 1$, $\widehat{\mathbf{x}}_{s'}$ is obtained by integrating the frozen teacher flow from $(\mathbf{x}_s, s)$ to s', and $\bar{\theta}$ denotes the EMA student parameters. The student satisfies $F_\theta(\mathbf{x}, 1, c) = \mathbf{x}$. To emphasize single-step generation, we sample $s' = 1$ with probability 0.5; otherwise, we sample s' from the remaining flow times. At inference, the video latent is generated in a single forward pass as $F_\theta(\boldsymbol{\epsilon}, 0, c)$.

3.2 Action Modeling

Stage 3: Action Pretraining across Datasets. After training a generalizable video model, we start training action on our processed bimanual datasets with unified workspace and eef coordinate systems. The action expert is a 0.9B-parameter diffusion transformer (DiT) (Peebles & Xie, 2023) with standard MoT architecture. Since the action expert is newly initialized, we freeze the base parameters of the video DiT during the early stages of action training and adapt the video backbone using only LoRA (Hu et al., 2022). This strategy helps preserve the pretrained video representations while limiting disruption from action-training gradients.

Inference Latency. We can generate video very fast since we reduce the video input to $17 \times 416 \times 240$ in post-training and distill the video model to enable one-step generation as described in Sec. 3.1. With bfloat16 precision and torch.compile, video model part takes approximately 0.12 seconds. The action expert then generates an action chunk conditioned on the video latents using five denoising steps, taking approximately 0.1 seconds. The total latency per action chunk is therefore approximately 0.22 seconds.

3.3 VLM for High-level Planning.

Sub-task Planning. To tackle long-horizon tasks and handle ambiguous human instructions, we adopt a hierarchical architecture in which a VLM generates subtask plans (Shi et al., 2025). The VLM handles high-level reasoning, semantic understanding, and memory, allowing VPP2 to focus on mapping explicit instructions to trajectories. In practice, the VLM can be deployed locally or accessed through an API to a frontier model.

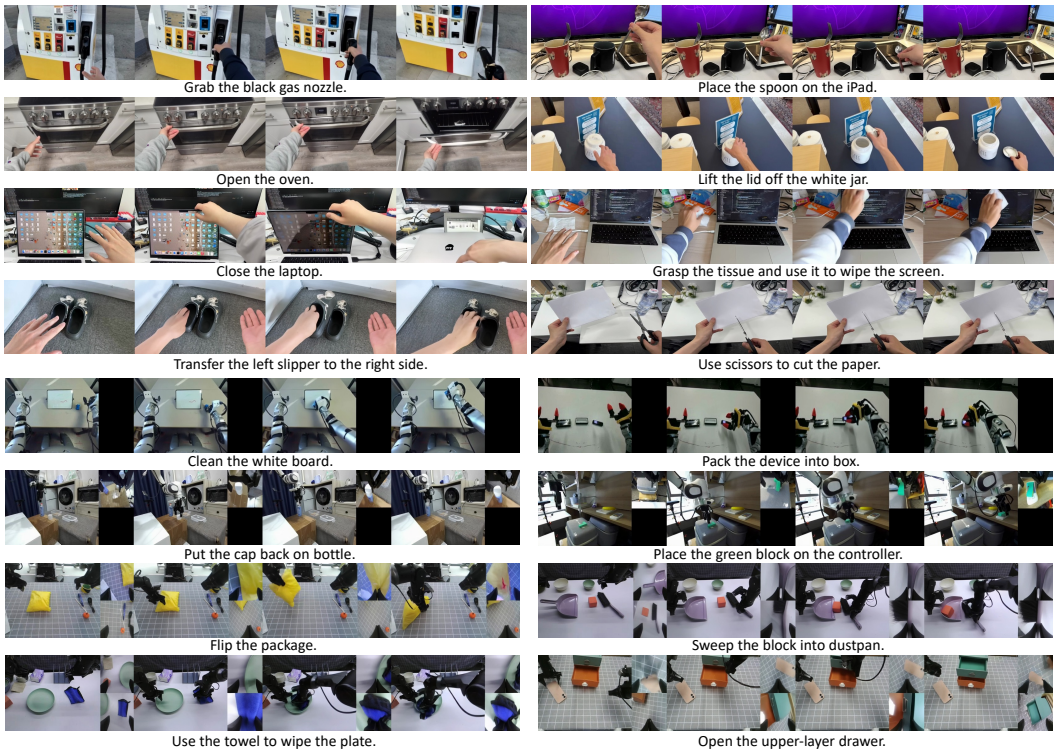


Figure 5: VPP2 video predictions for human-hand and robot manipulation with 30 steps. Trained on large-scale, diverse manipulation datasets, VPP2 generalizes across human hands and a wide range of robot embodiments. For robot manipulation, VPP2 jointly predicts one to three camera views arranged in a T-shaped composite. Due to space constraints, single-step video predictions after distillation are shown in Appendix Figure 10.

Prompt Enhancement. Another function of VLM planning is to generate detailed subtask similar to template prompt during training caption. Also, model video generation models also find that detailed caption is essential for generation quality (Wan et al., 2025; Yang et al., 2025).

4 Experiments

In this section, we conduct experiments to answer the following questions: (1) How does the motion quality of VPP2's video predictions compare with that of state-of-the-art video foundation models on manipulation tasks? (2) How broad are VPP2's zero-shot manipulation capabilities? (3) How well does VPP2 perform after domain-specific post-training?

4.1 Video Prediction Quality Analyses

Video Prediction for Human and Robot Manipulation. Figure 5 visualizes predictions for human-hand manipulation and diverse robot embodiments from pretrained VPP2 model. To assess generalization, we randomly sample test cases from prior work (Chen et al., 2025; Zhang et al., 2026a) and take photographs in office environments and robot workspaces, pairing them with freely chosen manipulation instructions. Trained on data spanning more than 20 robot embodiments, VPP2 also generalizes to unseen embodiments and tasks, as shown in Figure 5. VPP2 also supports joint prediction across one to three camera views, with the primary view placed on the left and up to two auxiliary views stacked on the right.

Quantitative Comparisons. We compare VPP2 with four video foundation models and one ablation variant: (1) Wan-2.1-14B (Wan et al., 2025), our base model; (2) LVP (Chen et al., 2025), which also adapts Wan-2.1-14B to manipulation tasks; (3) Cosmos3-Nano-16B, which is extensively trained on robot manipulation data; (4) Cosmos3-Super-Image2Video, the strongest model in the Cosmos 3

Method	Human	Robot
Wan2.1-14B	0.32	0.04
LVP	0.62	0.46
Cosmos3-16B	0.44	0.32
Cosmos3-64B	0.70	0.78
VPP2-Fixed-Step	0.48	0.42
VPP2 (ours)	**0.80**	**0.90**

Table 2: Instruction following success rates.

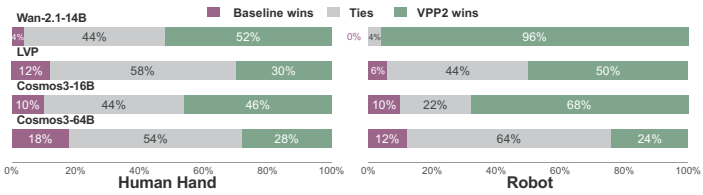
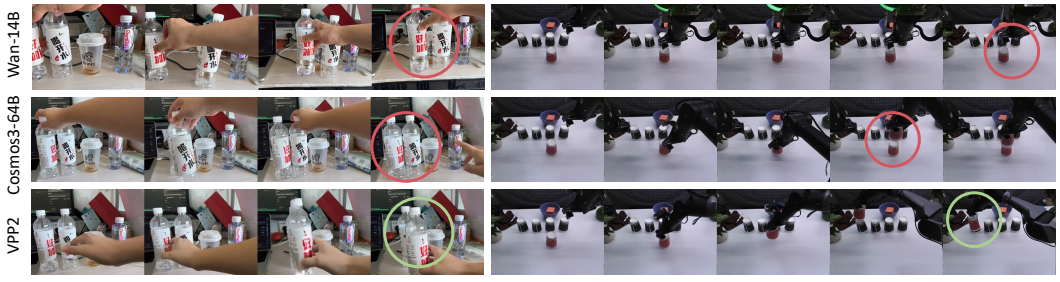


Figure 6: Win rates of the pretrained VPP2 model against different baselines on video prediction tasks.



(a) Grasp leftmost bottle and place it in front of the short cup. **(b) Stack red paper cup on leftmost black paper cup.**

Figure 7: Comparisons on instruction following capability between VPP2, Wan-14B, Cosmos3-64B. VPP2 demonstrates better instruction following on complex tasks requiring spatial understanding.

family (Agarwal et al., 2026); and (5) VPP2-Fixed-Step, an ablation of our method that predicts a fixed-duration future chunk rather than a complete event, potentially misaligning the predicted future with the instruction describing the full event.

We randomly collect 50 image–instruction pairs each for human-hand and robot manipulation and generate video predictions for every pair using each model. We first use a GPT-based evaluator to assess instruction-following success rates, reported in Table 2. Human evaluators then compare pairs of predictions to determine which better follows the instruction, yielding VPP2's pairwise win rates against each baseline in Figure 6. VPP2's advantages are most pronounced in complex manipulation scenarios requiring spatial understanding, as illustrated in Figure 7. We hypothesize that training with detailed captions encourages more accurate mappings from instructions to trajectories, contributing to these improvements.

4.2 Policy Performance Analysis

Zero-shot Performance on Real-world Aloha Robot. After pretraining a strong video foundation model for manipulation, we post-train and distill it into a fast, single-step generator to support action learning. With its generalizable video predictions, VPP2 demonstrates strong zero-shot capabilities across open-ended manipulation tasks. We deploy VPP2 directly on a robot embodiment seen during training, without task-specific fine-tuning. For comparison, we train (1) $\pi_{0.5}$ on the same action data and (2) Wan-14B version of Fast-WAM which only see the representation of first frames. We evaluate models on 10 different categories of zero-shot tasks and present results in Figure 8. VPP2 achieves an average success rate of 58.5%, compared with 40.0% for $\pi_{0.5}$ and 20.5% for Fast-WAM, and performs best in 9 of the 10 categories.

LIBERO-ID, LIBERO-Pro, and LIBERO-OOD Benchmarks. All models are trained exclusively on the four standard LIBERO suites (Liu et al., 2023) and are evaluated under three complementary settings. LIBERO-ID measures standard in-distribution manipulation performance. To assess generalization beyond the training distribution, we further evaluate on LIBERO-Pro (Zhou et al., 2025) and LIBERO-OOD (Li, 2025) without additional training. Following HarnessVLA (Zhang et al., 2026c), we evaluate the Position and Task perturbations of LIBERO-Pro, which test generalization to changes in object positions and task specifications, respectively. We further follow (Mishra et al., 2026) to evaluate compositional generalization on LIBERO-OOD, which recombines familiar objects, layouts, and goals into unseen task configurations along spatial, object, and goal dimensions. Per-suite LIBERO-ID results are provided in Appendix C.1.

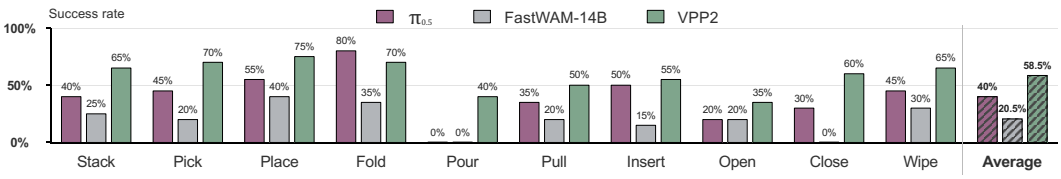


Figure 8: Zero-shot success rates on the real-world ALOHA across 10 randomly selected task categories. For a fair comparison, we finetune $\pi_{0.5}$ and FastWAM on all ALOHA datasets exposed in VPP2's training data and use the 14B variant of FastWAM.

	LIBERO-ID	LIBERO-Pro			LIBERO-OOD			
Method	**Overall**	**Position**	**Task**	**Overall**	**Spatial**	**Object**	**Goal**	**Overall**
π_0 (Black et al., 2024)	94.2	0.5	0.0	0.3	0.7	0.3	4.3	1.7
$\pi_{0.5}$ (Intelligence et al., 2025)	96.9	20.8	1.3	11.0	36.7	2.3	41.7	26.8
MolmoAct (Lee et al., 2025)	86.6	1.5	1.5	1.5	–	–	–	–
X-VLA (Zheng et al., 2026)	98.1	0.8	6.8	3.8	–	–	–	–
AtomVLA (Sun et al., 2026)	97.0	7.3	5.3	6.3	–	–	–	–
Cosmos-Policy (Kim et al., 2026)	98.5	–	–	–	30.7	0.3	0.7	10.5
Fast-WAM (Yuan et al., 2026)	97.6	–	–	–	12.7	0.0	15.7	9.4
DiT4DiT (Ma et al., 2026)	98.6	–	–	–	9.0	0.0	10.3	6.4
Temporal Ratio (Mishra et al., 2026)	94.0	–	–	–	58.6	40.0	80.3	59.4
VPP2 (Ours)	**98.8**	**42.3**	**47.8**	**45.0**	**59.3**	**44.5**	**87.7**	**63.9**

Table 3: Post-training on LIBERO-ID, LIBERO-Pro, and LIBERO-OOD benchmarks. All models are trained exclusively on the four standard LIBERO suites (Liu et al., 2023) and are evaluated under three complementary settings.

As shown in Table 3, while most baselines exceed 94% on LIBERO-ID, this in-distribution performance does not transfer to generalization. On LIBERO-Pro, VLA baselines drop to at most 11.0% overall, whereas VPP2 reaches 45.0%, with the largest gain on the Task perturbation (47.8% vs. 6.8%), indicating that VPP2 follows the given instruction rather than replaying memorized trajectories. On LIBERO-OOD, video-based policies such as Cosmos-Policy, Fast-WAM, and DiT4DiT reach at most 10.5%, while VPP2 achieves 63.9%, also surpassing Temporal Ratio (Mishra et al., 2026), which specifically targets this generalization gap. We attribute these gains to next-event video prediction with detailed captions, which establishes consistent mappings from instructions to future trajectories.

RoboDojo Benchmark. RoboDojo (Chen et al., 2026) is a unified sim-and-real benchmark for evaluating generalist robot manipulation policies. Its simulation benchmark contains 42 bimanual tasks covering five capability dimensions: generalization, memory, precision, long-horizon execution, and open-vocabulary instruction following. We follow the official training and evaluation protocol and report both the success rate and the average score, which captures partial task progress.

As shown in Table 4, VPP2 achieves an average score of 35.51 and a success rate of 29.47%, achieving state-of-the-art performance.[1] It outperforms representative VLAs such as $\pi_{0.5}$ (Intelligence et al., 2025), X-VLA (Zheng et al., 2026), and Xiaomi-Robotics-1 (Guo et al., 2026), world action models such as Fast-WAM (Yuan et al., 2026) and OpenWAM-α (Wang et al., 2026), and the frontier foundation model GPT-6-Astra (Zhang et al., 2026b) (28.97 / 22.48%). Per-dimension results are provided in Appendix C.2.

VPP2 with Subtask Planning. The VPP2 results in Table 4 use full-task instructions without subtask planning. We further investigate VLM-based subtask planning on five selected RoboDojo task groups covering object classification, block stacking, block swapping, mahjong, and tic-tac-toe. To train the subtask-conditioned policy, we segment demonstrations and pair each segment with a subtask instruction. At test time, a vision-language model (VLM) uses visual observations and execution feedback to select the next subtask for VPP2 to execute.

Figure 9 summarizes the results across the five selected task groups. The evaluated planning configurations achieve an average success rate of 57.6%, compared with 27.6% for the baseline. The largest gains occur in language-conditioned block stacking (10% to 52%) and tic-tac-toe (0% to 38%).

[1]Baseline results are taken from the official RoboDojo simulation leaderboard as of September 2026.

	π_0	$\pi_{0.5}$	X-VLA	Fast-WAM	OpenWAM-α	Xiaomi-Robotics-1	GPT-6-Astra	**VPP2 (Ours)**
Avg. Score	3.48	11.44	10.13	3.48	17.18	20.07	28.97	**35.51**
Success Rate (%)	1.53	6.93	6.52	2.03	11.92	13.93	22.48	**29.47**

Table 4: Post-training results on the RoboDojo simulation benchmark. Average score captures partial task progress, and success rate measures binary task completion; both are averaged over the five capability dimensions. Baseline results are taken from the official leaderboard.

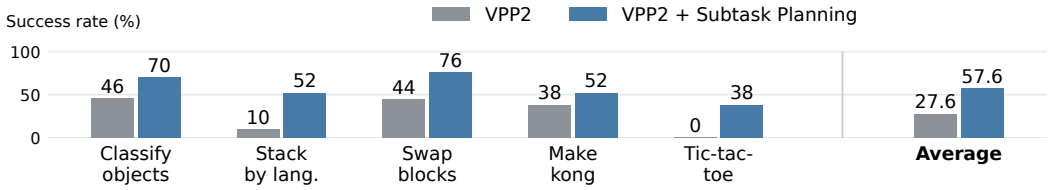


Figure 9: Success rates with and without VLM-based subtask planning (50 trials per task). Average is the five-task mean; the baseline checkpoint matches Table 4.

5 Related Works

Video Foundation Model. Video generation has advanced rapidly through latent diffusion models (Blattmann et al., 2023b;a) and large-scale diffusion transformer models (Yang et al., 2025; Wan et al., 2025; Agarwal et al., 2025; 2026). Large-scale bidirectional video models provide strong general-purpose video priors; even recent autoregressive generators build on pretrained bidirectional models, using them for initialization or as teachers before causal adaptation and distillation (Yin et al., 2025; Huang et al., 2025). This motivates our strategy of first learning a generalizable video model and then adapting it for efficient prediction. Beside training strategy, prompt enhancement is an important component in video foundation models (Yang et al., 2025; Wan et al., 2025). Similarly, we use a VLM to produce detailed, manipulation-specific descriptions that reduce ambiguity in these mappings. After establishing this manipulation-specific video prior, we post-train the model for fixed-horizon prediction and distill it for single-step generation to support efficient action learning.

World Action Model. World action models (WAMs) leverage video prediction to facilitate robot policy learning. Early approaches first generate future frames and then infer actions through inverse dynamics (Du et al., 2023; Black et al., 2023; Bharadhwaj et al., 2024; Liang et al., 2024; Feng et al., 2025), but iterative video generation can incur substantial latency for closed-loop control. More recent approaches condition actions on features from video models or jointly model videos and actions (Hu et al., 2024; Guo et al., 2024; Li et al., 2025; Zhu et al., 2025; Liao et al., 2025; Zheng et al., 2025; Ma et al., 2026; Kim et al., 2026; Yan et al., 2026b;a; Ye et al., 2026; Li et al., 2026a). Recent work also explores event-grounded world-action learning (Li et al., 2026b) and large-scale manipulation-specific pretraining (AgiBot Research Team et al., 2026). Nevertheless, preserving the generalization of pretrained video models during action learning remains challenging (Mishra et al., 2026). VPP2 emphasizes instruction-aligned future prediction, combining event-level pretraining with detailed captions, fixed-horizon post-training, and single-step distillation to provide an efficient video backbone for action learning.

6 Conclusion

We presented Video Prediction Policy 2 (VPP2), a world We presented Video Prediction Policy 2 (VPP2), a world action model that achieves strong zero-shot generalization in both video prediction and action generation. We first continue pretraining the base video model to produce generalizable, instruction-following predictions, then learn a policy while preserving these predictive capabilities. Experiments show that VPP2 follows manipulation instructions more faithfully than substantially larger video foundation models, exhibits strong zero-shot manipulation capabilities on a real-world ALOHA platform, and achieves the highest success rates among evaluated methods on the challenging LIBERO-Pro, LIBERO-OOD, and RoboDojo benchmarks. We hope these findings encourage future WAM research to look beyond action modeling and place greater emphasis on learning generalizable video predictions and preserving this generalization during action learning.

AI USE STATEMENT

We used generative AI tools to assist with language polishing, improve the clarity and readability of the manuscript, and draft portions of the paper. In the research process, these tools were used to suggest experimental parameter settings and assist with searches for relevant literature and technical information. We also used AI tools to annotate human and robotic manipulation datasets at scale for model training. The authors evaluated the suggestions and made the final decisions regarding experimental design and parameter selection. All AI-assisted text was reviewed and revised by the authors, and retrieved information was checked against original sources before use. The authors take full responsibility for the final content of the paper, including its data, claims, results, and references.

REPRODUCIBILITY STATEMENT

We describe the model architecture, training objectives, and evaluation protocols in the main text, with additional data processing details provided in the appendix. Detailed training configurations, code, and model checkpoints are available on our anonymous project website.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

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

AgiBot Research Team, Renhang Liu, Wenzhi Zhao, Zhuo Yang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Guanghui Ren, Youlun Peng, Rongjun Jin, et al. GE-Act 2.0: Pretraining and scaling a world-action model for robotic manipulation. *arXiv preprint arXiv:2609.05588*, 2026. URL https://arxiv.org/abs/2609.05588.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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*, 2024.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Caiyi Zhang, Peihao Li, Kiwhan Song, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025.

Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Haoran Lu, Weijie Wan, Baijun Chen, Songling Liu, Haowen Yan, et al. Robodojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv preprint arXiv:2607.04434*, 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, 2023.

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

Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, Nan Sun, et al. Xiaomi-robotics-1: Scaling vision-language-action models with over 100k hours of real-world trajectories. *arXiv preprint arXiv:2607.15330*, 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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. *arXiv preprint arXiv:2412.14803*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5576. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 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, and Jinwei Gu. Cosmos policy: Fine-tuning video models for visuomotor control and planning. *arXiv preprint arXiv:2601.16163*, 2026.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a. doi: 10.48550/arXiv.2601.21998. URL https://arxiv.org/abs/2601.21998.

Quanyi Li. Vlas are confined yet capable of generalizing to novel instructions. *arXiv preprint arXiv:2505.03500*, 2025.

Shalfun Li, Victor Yao, Charles Yang, Truth Qu, Regis Cheng, Ryan Yu, Howard Lu, Newton Von, Vincent Chen, Yohann Tang, et al. WALL-WM: Carving world action modeling at the event joints. *arXiv preprint arXiv:2606.01955*, 2026b. doi: 10.48550/arXiv.2606.01955. URL https://arxiv.org/abs/2606.01955.

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

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=OutjGuJnNk.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Teli Ma, Jia Zheng, Zifan Wang, Chunli Jiang, Andy Cui, Junwei Liang, and Shuo Yang. Dit4dit: Jointly modeling video dynamics and actions for generalizable robot control. *arXiv preprint arXiv:2603.10448*, 2026.

Utkarsh A Mishra, Yongxin Chen, Danfei Xu, Yang Liu, Xi Chen, and Jiayuan Mao. Understanding and mitigating the video-action generalization gap via temporal ratio. *arXiv preprint arXiv:2607.08127*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Xiaoquan Sun, Zetian Xu, Chen Cao, Zonghe Liu, Yihan Sun, Jingrui Pang, Ruijian Zhang, Zhen Yang, Kang Pang, Dingxin He, et al. Atomvla: Scalable post-training for robotic manipulation via predictive latent world models. *arXiv preprint arXiv:2603.08519*, 2026.

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

Yuran Wang, Siqiao Huang, Mingleyang Li, Chenhao Zhang, Jiaqi Liang, Weiyang Jin, Yue Chen, Xuemin Chi, Donghao Zhou, Qize Yu, et al. Openwam: An open, modular exploration towards systematic world-action model pretraining. *arXiv preprint arXiv:2609.07398*, 2026.

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

Haodong Yan, Zhide Zhong, Jiaguan Zhu, Junjie He, Weilin Yuan, Wenxuan Song, Xin Gong, Yingjie Cai, Guanyi Zhao, Xu Yan, et al. S-vam: Shortcut video-action model by self-distilling geometric and semantic foresight. *arXiv preprint arXiv:2603.16195*, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026. doi: 10.48550/arXiv.2602.15922. URL https://arxiv.org/abs/2602.15922.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.html.

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.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-robotworld technical report: Unifying embodied world modeling through language-conditioned video generation. *arXiv preprint arXiv:2606.17030*, 2026a.

Wenbo Zhang, Kaixuan Wang, Yutao Ouyang, Xiaoyu Huang, Liyang Li, Kailun Su, Weiyang Jin, Wenhao Chai, Haotian Liang, Zhiyang Dou, et al. An unexpected robot policy: Early evaluations of gpt-6 astra on robodojo and beyond. *arXiv preprint arXiv:2609.24170*, 2026b.

Yixian Zhang, Huanming Zhang, Feng Gao, Xiao Li, Zhihao Liu, Chunyang Zhu, Jiaxing Qiu, Yuchen Yan, Jiyuan Liu, Wenhao Tang, et al. Harness vla: Steering frozen vlas into reliable manipulation primitives via memory-guided agents. *arXiv preprint arXiv:2607.08448*, 2026c.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, 2026.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

Xueyang Zhou, Yangming Xu, Guiyao Tie, Yongchao Chen, Guowen Zhang, Duanfeng Chu, Pan Zhou, and Lichao Sun. Libero-pro: Towards robust and fair evaluation of vision-language-action models beyond memorization. *arXiv preprint arXiv:2510.03827*, 2025.

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 Dataset Process Details

A.1 Video Captioning.

After segmentation, we use a VLM to generate detailed captions for each clip according to a predefined template. Our captioning pipeline aims to reduce uncertainty in future prediction and encourage consistent mappings from semantic descriptions to visual trajectories. Specifically, we address four major sources of uncertainty:

(1) Detailed Task Description. We describe how the end effector completes the task, including the sequence and manner of manipulation.

(2) Explicit End-Effector Identification. The active end effector must be visible in the first frame; otherwise, we trim the clip to begin when it first becomes visible. When multiple end effectors are involved, we describe the motion of each one explicitly.

(3) Unambiguous Target-Object Specification. When multiple identical or similar objects are present, we identify the target using distinctive attributes or spatial relationships. The target object must also be visible in the first frame.

(4) Camera-View Description. We use optical flow to remove egocentric clips with excessive viewpoint changes. For clips with moderate camera motion, a VLM describes the viewpoint changes and the camera views included in the video, and we incorporate this information into the caption.

A.2 Unified Action Space

We focus on egocentric bimanual datasets, including ALOHA-style, humanoid-style, and human egocentric datasets. These datasets account for more than 80% of the total data and share a similar bimanual structure and camera viewpoint. We explicitly align their coordinate systems at two levels: (1) *workspace alignment*, which applies a world-frame transformation so that different robots have comparable end-effector workspaces; and (2) *end-effector alignment*, which applies a local transformation to standardize end-effector origins and axis orientations.

For an arm dataset d , let $T_d = {}^{W_d}T_{E_d} \in SE(3)$ denote the original end-effector pose, where W_d and E_d are the dataset's world and end-effector frames, respectively. We define the workspace alignment as $A_d = {}^{\bar{W}}T_{W_d}$ and the end-effector alignment as $B_d = {}^{E_d}T_{\bar{E}}$, where $\bar{W}$ and $\bar{E}$ denote the canonical world and end-effector frames. The aligned end-effector pose is given by

$$\widetilde{T}_d = A_d T_d B_d, \tag{7}$$

B More Video Prediction Results

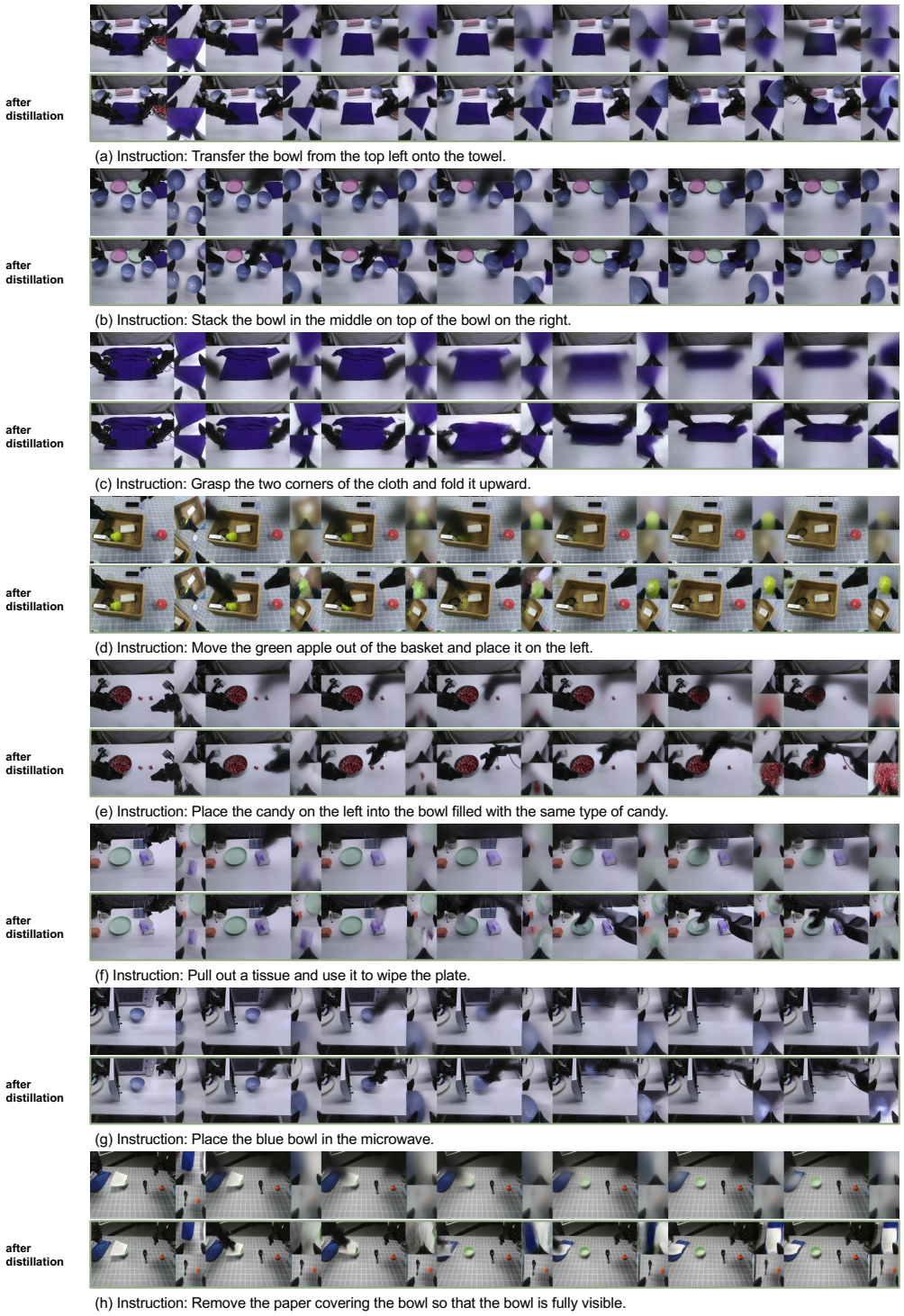


Figure 10: Additional video prediction results on open-ended tasks. We compare **single-step** video predictions before and after distillation. Distillation enables high-quality video prediction with just one sampling step.

C Detailed Benchmark Results

C.1 Detailed LIBERO Results

Table 5 reports per-suite success rates on the four standard LIBERO suites (Liu et al., 2023), which correspond to the LIBERO-ID results in Table 3. Baseline results are taken from the original papers and from Zhang et al. (2026c) and Mishra et al. (2026).

Method	Spatial	Object	Goal	Long	Avg.
π_0 (Black et al., 2024)	96.8	98.8	95.8	85.2	94.2
$\pi_{0.5}$ (Intelligence et al., 2025)	98.8	98.2	98.0	92.4	96.9
MolmoAct (Lee et al., 2025)	87.0	95.4	87.6	77.2	86.6
X-VLA (Zheng et al., 2026)	98.2	98.6	97.8	97.6	98.1
AtomVLA (Sun et al., 2026)	96.4	99.6	97.6	94.4	97.0
Cosmos-Policy (Kim et al., 2026)	98.1	**100.0**	98.2	97.6	98.5
Fast-WAM (Yuan et al., 2026)	98.2	**100.0**	97.0	95.2	97.6
DiT4DiT (Ma et al., 2026)	98.4	99.6	**98.6**	97.6	98.6
Temporal Ratio (Mishra et al., 2026)	96.3	99.6	97.6	82.6	94.0
VPP2 (Ours)	**99.2**	99.8	98.2	**98.0**	**98.8**

Table 5: Per-suite success rates (%) on the standard LIBERO benchmark. Best results in each column are in bold.

C.2 Detailed RoboDojo Results

RoboDojo (Chen et al., 2026) evaluates generalist manipulation policies on 42 bimanual simulation tasks organized into five capability dimensions. *Generalization* tests robustness to unseen backgrounds, lighting, clutter, and target objects, and is evaluated under both standard and randomized settings. *Precision* requires fine-grained target localization and contact-rich control. *Long-Horizon* requires completing all sub-steps of multi-step tasks. *Memory* contains tasks whose correct actions depend on information observed earlier in the episode. *Open* evaluates unseen task specifications whose required skills appear in the training data under different contexts. We follow the official training and evaluation protocol. The average score captures partial task progress, and the success rate measures binary task completion.

Tables 6 and 7 report the per-dimension average score and success rate, respectively, complementing Table 4. Baseline results are taken from the official RoboDojo simulation leaderboard as of September 2026.

	Generalization						
Method	Std.	Rand.	Precision	Long-Horizon	Memory	Open	Avg.
π_0 (Black et al., 2024)	7.18	0.71	3.56	6.19	3.47	0.25	3.48
$\pi_{0.5}$ (Intelligence et al., 2025)	20.93	5.82	12.40	23.54	5.89	1.98	11.44
X-VLA (Zheng et al., 2026)	17.90	3.04	18.32	16.53	4.76	0.55	10.13
Fast-WAM (Yuan et al., 2026)	4.33	0.34	1.96	9.14	3.55	0.42	3.48
OpenWAM-α (Wang et al., 2026)	33.16	8.26	18.45	34.93	10.41	1.41	17.18
Xiaomi-Robotics-1 (Guo et al., 2026)	35.65	11.44	26.69	38.39	7.81	3.94	20.07
GPT-6-Astra (Zhang et al., 2026b)	35.32	31.40	12.65	21.45	43.04	34.36	28.97
VPP2 (Ours)	32.34		39.29	41.01	59.58	5.34	**35.51**

Table 6: Per-dimension average score on the RoboDojo simulation benchmark. Generalization is evaluated under standard (Std.) and randomized (Rand.) settings. Avg. is the mean over the five capability dimensions, where the generalization score is the mean of the Std. and Rand. settings. For VPP2, we report the generalization result pooled over both settings (25 episodes each, following the official protocol), which equals their mean. Baseline results are taken from the official leaderboard.

Method	Generalization		Precision	Long-Horizon	Memory	Open	Avg.
	Std.	Rand.					
π_0 (Black et al., 2024)	4.89	0.22	0.75	2.00	2.11	0.25	1.53
$\pi_{0.5}$ (Intelligence et al., 2025)	14.89	1.44	5.50	14.67	4.67	1.67	6.93
X-VLA (Zheng et al., 2026)	12.22	1.33	12.00	9.75	3.56	0.50	6.52
Fast-WAM (Yuan et al., 2026)	2.11	0.11	0.00	5.17	3.44	0.42	2.03
OpenWAM-α (Wang et al., 2026)	25.56	4.11	9.25	25.33	9.11	1.08	11.92
Xiaomi-Robotics-1 (Guo et al., 2026)	28.00	6.00	18.83	23.67	6.56	3.58	13.93
GPT-6-Astra (Zhang et al., 2026b)	32.67	28.33	4.00	8.25	38.67	31.00	22.48
VPP2 (Ours)	25.00		30.50	28.00	59.33	4.50	**29.47**

Table 7: Per-dimension success rate (%) on the RoboDojo simulation benchmark, computed in the same way as Table 6.