ZERO-WAM: IN-CONTEXT WORLD-ACTION MODELING FROM HUMAN VIDEOS FOR OPEN-ENDED TASK GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot cross-task generalization, where a policy must execute manipulation tasks never seen during training, remains a central challenge in robot learning. In large language models, a novel task can be performed simply by specifying it in the context, without any parameter update. This form of in-context learning (ICL) turns generalization into a problem of task specification. To achieve cross-task generalization, we bring this paradigm to robotic manipulation, and argue that the natural task specification for manipulation is a human video: unlike language, it provides rich visual cues about the intended task evolution. We present **Zero-WAM**, a causal video-action model that executes unseen tasks by following in-context human video guidance. To address the scarcity of task-rich paired human-robot data, we propose an automatic pipeline that converts task-sampled robot trajectories into semantically matched human videos, yielding *HumanGen* with 74.2K human-robot ICL pairs across 8.6K tasks. For model training, we further introduce an in-context future chunk prediction (IFP) objective that suppresses shortcuts learned from seen tasks. On seven unseen tasks in RoboTwin 2.0 simulation, Zero-WAM outperforms the strongest video-action baseline by 29.5 percentage points. In real-world evaluations, it generalizes to unseen task configurations involving multi-object scenes, long-horizon manipulation, and fine-grained insertion.

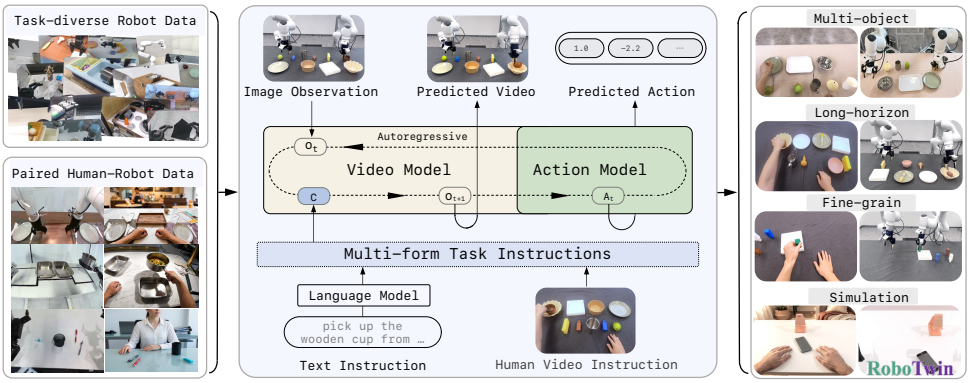


Figure 1: **Overview of Zero-WAM.** The model predicts future robot videos and actions from language or human-video instructions and is evaluated on unseen simulated and real-world tasks.

1 INTRODUCTION

Zero-shot cross-task generalization is essential for general-purpose robotic manipulation: a robot should infer how to act in a task it has not practiced, using only information available at deployment. Inspired by in-context learning (ICL) (Brown et al., 2020; Dong et al., 2024), where a model solves a new problem from input context without parameter updates, we view robotic task generalization as specifying an unseen task through deployment-time context. This perspective suggests a route toward

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.

open-ended task generalization: the policy could infer the intended task from context and translate it into executable robot behavior.

Recent vision-language-action (VLA) models (Kim et al., 2024; O'Neill et al., 2024; Black et al., 2024; Intelligence et al., 2025; Wu et al., 2026a; Yuan et al., 2026a; Zheng et al., 2026a; Bjorck et al., 2025) and video-action models (Li et al., 2026b; Zhang et al., 2026b; Ye et al., 2026b; Xu et al., 2026; Yuan et al., 2026b; Li et al., 2026c) have substantially expanded the capabilities of robot policies, yet both predominantly rely on language as the task interface. Language, however, often underspecifies manipulation tasks because spatial constraints, intermediate states, and temporal structure can be cumbersome to articulate. Human demonstration videos instead provide a natural in-context task specification by directly showing the desired visual state changes and temporal evolution (Jang et al., 2022; Jain et al., 2024). Although they do not provide executable robot actions, they offer a visual reference from which the policy can infer the task and realize it through the robot's own embodiment and dynamics.

Exploiting human videos at scale, however, presents two key challenges. First, large-scale, task-rich human-robot pairs that semantically align human videos with action-labeled robot trajectories are scarce and expensive to collect manually (Sharma et al., 2018; Kareer et al., 2025a; Fang et al., 2024). Second, a policy trained on seen tasks can learn shortcuts that allow it to ignore the in-context video. In particular, the next robot video-action chunk can often be predicted from the robot history and text instruction alone. Consequently, the model may perform well on familiar training tasks without learning to use the human video, and then underuse the video at test time, precisely when it is needed to specify an unseen task.

To address these challenges, we introduce **Zero-WAM**, a causal video-action model for zero-shot robotic task generalization from either language or in-context human-video instructions. Given either task specification, Zero-WAM autoregressively predicts future robot videos and executable actions, conditioning on demonstrated visual state changes when a human video is provided.

To scale training with human-video instructions, we introduce a **in-context human video generation pipeline** that converts task-sampled robot trajectories into semantically matched human videos and pairs them with the corresponding robot trajectories and executable actions. This yields *HumanGen*, with 74.2K human-robot ICL pairs across 8.6K tasks. We also repartition public robot data into more than 5,000 manipulation tasks and sample at the task level, yielding approximately 400K trajectories per epoch (*Task-diverse VA*). This task-balanced curation supplies diverse source trajectories and prevents a small number of tasks from dominating pre-training; Figure 2 summarizes the resulting data and pipeline.

To ensure that Zero-WAM uses the in-context human video rather than relying on shortcuts from robot history and text, we introduce an in-context future chunk prediction (IFP) objective. Standard next-chunk prediction can often be solved using only local robot history, particularly for tasks observed during training. IFP instead supervises multiple strided chunks of future robot video from the current robot-video representation, encouraging the model to encode the longer-term task evolution conveyed by the human video.

We evaluate Zero-WAM on zero-shot cross-task generalization in RoboTwin 2.0 (Chen et al., 2025) simulation and on real-world robotic manipulation. In RoboTwin 2.0, Zero-WAM achieves a 46.95% average success rate across seven unseen tasks, outperforming LingBot-VA (Li et al., 2026b) by an absolute margin of 29.50 percentage points. Real-world experiments further demonstrate human-video-guided generalization to unseen task configurations involving multi-object scenes, long-horizon manipulation, and precision-demand insertion, with Zero-WAM outperforming LingBot-VA across all three task families.

Our contributions are summarized as follows:

- We introduce Zero-WAM, an in-context world action model for zero-shot robotic tasks using language or human-video instructions.
- We build *HumanGen*, 74.2K semantically matched human-robot ICL pairs over 8.6K tasks, using a scalable generation pipeline.
- We demonstrate zero-shot generalization to unseen task configurations in simulation and real-world manipulation without collecting task-specific data or updating model parameters.

2 Related Works

2.1 Cross-Task Robotic Manipulation

Vision-language-action (VLA) models (Brohan et al., 2022; 2023; Kim et al., 2024; Team et al., 2024; Black et al., 2024; Intelligence et al., 2025; Wu et al., 2026b; Xie et al., 2026a; Yu et al., 2026) scale language-conditioned policies over heterogeneous robot data, but the mismatch between vision-language representations and robot actions still limits cross-task transfer (Zhou et al., 2026a).

World action models (WAMs) for robotics (Li et al., 2026b; Zhang et al., 2026b; Ye et al., 2026b; Li et al., 2026a; Xu et al., 2026; Liu et al., 2026; Su et al., 2026; Yuan et al., 2026b; Li et al., 2026c; Team et al., 2026a; Zhou et al., 2026b; Ma et al., 2026b; Zhang et al., 2026a; Zhao et al., 2026b; Ma et al., 2026a; Feng et al., 2025; Hu et al., 2024; Sadigh and Song; Ye et al., 2026a; Lyu et al., 2026; Du et al., 2023; Liang et al., 2025) jointly predict future visual states and executable robot actions from an instruction and the current observation. By shifting unseen-task generalization from direct action prediction to future-video generation, they can exploit accurate robot actions decoded from predicted future videos (Pai et al., 2025; Mimic Robotics, 2026). The LingBot-VA series (Li et al., 2026b; Zhang et al., 2026b) couples autoregressive causal video and action prediction and studies native video-action pre-training. DreamZero (Ye et al., 2026b) demonstrates zero-shot policy transfer under visual-domain shifts through generated future dynamics.

2.2 Human Video Data for Robotic Manipulation

Human videos have become an important data source for robotic manipulation because they contain diverse interactions that are difficult to collect with robots alone. Prior work uses in-the-wild human videos to learn robot-relevant representations, manipulation priors, or pre-training objectives (Nair et al., 2022; Bahl et al., 2022; Luo et al., 2025; Li et al., 2025; Zheng et al., 2026b; Huang et al., 2026; Wang et al., 2026b; Xie et al., 2026b; Wang et al., 2026a), or reduces the human-robot domain gap through paired, retargeted, jointly trained, or mixed human-robot data (Li et al., 2026a; Zhou et al., 2025b; Kareer et al., 2025a; Cai et al., 2025; Kareer et al., 2025b; Paliwal et al., 2026; Zhou et al., 2025a). A closer line directly conditions policies on human video demonstrations as task prompts (Yu et al., 2018; Jang et al., 2022; Jain et al., 2024; Ye et al., 2025; Patel et al., 2026; Team, 2026; AI, 2026; Dyna Robotics, 2026; Chen et al., 2026; Team et al., 2026b). WAM-TTT (Feng et al., 2026) instead adapts lightweight memory inside a frozen WAM from unlabeled human videos, while RoboTTT (Jiang et al., 2026) absorbs a human-video demonstration into recurrent fast weights for one-shot imitation.

However, in-the-wild videos lack aligned robot actions, task-aligned pairs are costly to scale, and video-conditioned systems often have limited task coverage or require modular alignment or test-time adaptation. Zero-WAM instead generates *HumanGen* from task-sampled robotic video-action data, pairing each human video instruction with a trajectory that retains executable actions. It thereby scales task diversity, semantic alignment, and action grounding, while directly conditioning a frozen policy on a human-video prefix without deployment-time parameter or memory updates.

3 Data Curation

We construct two complementary datasets through task-level sampling. First, we curate *Task-diverse VA* data from public robotic pre-training data (Section 3.1). We then introduce the **in-context human video generation pipeline** to convert task-sampled robot videos into human video instructions (Section 3.2), forming the human-robot ICL pairs in *HumanGen* (Section 3.3). Figure 2 summarizes the data composition and generation pipeline.

3.1 Task-Diverse Video-Action Data

Task-diverse VA data are curated from five public robotic video-action (VA) pre-training datasets: AgiBot (Bu et al., 2025), InternData-A1 (Tian et al., 2026), Open-X-Embodiment (O'Neill et al., 2024), RoboCOIN (Wu et al., 2025), and RoboMIND (Wu et al., 2024), the same sources used by LingBot-VA (Li et al., 2026b). Because their raw distributions overrepresent repeated tele-operation of the same tasks, we group trajectories into tasks defined by manipulation action–object pairs,



Figure 2: Data overview. Top: *Task-diverse VA* data and human-robot ICL data. Bottom: the in-context human video generation pipeline.

obtaining labels from dataset metadata or parsing them from trajectories when needed. We then cap the number of trajectories sampled per task according to each source's intra-task diversity. This task-balanced corpus contains more than 5,000 tasks and approximately 400K robot trajectories per training epoch.

3.2 In-Context Human Video Generation Pipeline

Manually collecting task-diverse human-robot pairs is prohibitively expensive. We instead task-sample robot trajectories with executable actions and use the **in-context human video generation pipeline** to generate semantically matched human manipulation videos.

For each sampled robot video, a VLM (Gemini 3.1 Pro (Google DeepMind, 2026a) or Qwen3.6-Plus (Qwen Team, 2026)) extracts the task name, initial object states, object state changes, and final object states, and produces a prompt for transforming the first robot frame into a human manipulation scene. An image editing model (Nano Banana 2 (Google DeepMind, 2026b) or Qwen-Image-2.0 (Zhao et al., 2026a)) applies the specified variations to generate the initial human observation. During this first-frame editing stage, we diversify visual alignment by varying the background, viewpoint, environment style, object instance, and object placement while preserving the task semantics. The VLM then combines this observation with the extracted object-state information to describe how human hands should complete the task, and a video generation model (Wan 2.7 (Alibaba Cloud, 2026) or Kling AI 3.0 (Kuaishou Technology, 2026)) synthesizes the human manipulation video. Finally, the VLM filters the videos for task-semantic preservation and physical plausibility; qualified videos are paired with their original robot trajectories as ICL samples. The scoring rubric and acceptance criteria are provided in Appendix A.1.

3.3 IN-CONTEXT HUMANGEN DATASET

Generated with the in-context human video generation pipeline, *HumanGen* contains 74.2K pairs of human video instructions and executable robot trajectories over 8.6K tasks, organized into four subsets by robot-trajectory source.

Pre-train ICL (External). This subset uses five public robotic video-action datasets: AgiBot (Bu et al., 2025), InternData-A1 (Tian et al., 2026), Open-X-Embodiment (O'Neill et al., 2024), RoboCOIN (Wu et al., 2025), and RoboMIND (Wu et al., 2024). Its 41,188 pairs over 5,062 tasks span diverse scenes, objects, and more than 45 robot embodiments.

Pre-train ICL (In-house). This subset contains 30,247 pairs over 3,522 tasks from our in-house robotic datasets.

Simulation ICL. This subset contains 2,500 pairs from all 50 RoboTwin tasks (Chen et al., 2025): 43 for post-training and 7 held out as unseen tasks for zero-shot cross-task evaluation.

Real-world ICL. This subset contains 252 pairs on the bimanual Franka evaluation platform, covering object-to-container placement, three-object sequential manipulation, and two-table-leg insertion. It is used for post-training to adapt the policy to the evaluation robot kinematics.

The external subset emphasizes visual diversity, whereas the in-house and evaluation-domain subsets provide closer human-robot alignment. Detailed subset statistics and comparisons with existing paired human-robot datasets are provided in Appendix A.

4 ZERO-WAM: IN-CONTEXT WORLD ACTION MODELING

In zero-shot cross-task manipulation, the robot must execute a novel task without finetuning. We pursue this ability with a world action model (WAM) pretrained on large-scale human-robot ICL pairs and task-balanced robotic video-action data: at deployment, Zero-WAM uses either a language instruction or a human video demonstration as the task specification to drive execution of unseen tasks.

4.1 PRELIMINARIES

Flow matching for video generation. For video generation, flow matching (Lipman et al., 2022) learns a velocity field along a continuous path between a clean video and Gaussian noise. Given a clean video $\mathbf{x}_0$, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and flow time $t \in [0, 1]$, the noised video and target velocity are

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\boldsymbol{\epsilon}, \qquad \mathbf{v}_t^\star = \boldsymbol{\epsilon} - \mathbf{x}_0. \tag{1}$$

With condition c, the flow-matching loss is

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

Causal video-action modeling. We follow the causal video-action framework of the LingBot-VA series (Li et al., 2026b; Zhang et al., 2026b), in which each control step is a causal prediction problem: given the task condition and history observation, the model predicts the next robot video chunk together with its temporally aligned executable action chunk.

Each robot trajectory is chunked into $\tau = \{(\mathbf{x}^i, \mathbf{a}^i)\}_{i=1}^N$, where $\mathbf{x}^i$ is a video chunk and $\mathbf{a}^i$ is the temporally aligned action chunk. We use chunk-level notation throughout the method. In implementation, each video or action chunk is encoded into modality-specific representations before entering the Transformer, and the attention mask operates over these tokenized representations. We use c to denote the task condition, which can be a language instruction or a same-task human video together with language. At chunk index i, joint prediction of the next video chunk $\mathbf{x}^{i+1}$ and aligned action chunk $\mathbf{a}^{i+1}$ is factorized into video prediction followed by action decoding:

$$\begin{aligned} p_\theta\big(\mathbf{x}^{i+1}, \mathbf{a}^{i+1} \mid \mathbf{x}^{\le i}, \mathbf{a}^{\le i}, c\big) = p_\theta^{\text{vid}}\big(\mathbf{x}^{i+1} \mid \mathbf{x}^{\le i}, \mathbf{a}^{\le i}, c\big) \\ \cdot\, p_\theta^{\text{act}}\big(\mathbf{a}^{i+1} \mid \mathbf{x}^{\le i}, \mathbf{a}^{\le i}, \mathbf{x}^{i+1}, \ell\big), \end{aligned} \tag{3}$$

where p_θ^{vid} predicts the next video chunk, and p_θ^{act} serves as an inverse dynamics model that decodes executable actions from the predicted next robot video chunk. During training, the model is given

the ground-truth trajectory prefix $(\mathbf{x}^{\leq i}, \mathbf{a}^{\leq i})$ and is supervised on the next video-action chunk. For trajectories from *Task-diverse VA* data, the condition is the language instruction, $c = \ell$. For trajectories from *HumanGen* data, the condition is $c = \{\mathbf{h}, \ell\}$, where $\mathbf{h}$ is the same-task human video instruction and ℓ is the language instruction.

Mixture-of-Transformers design. We implement p_θ^{vid} and p_θ^{act} as a video Transformer and an action Transformer under a Mixture-of-Transformers (MoT) design. The two modalities have independent QKV projections, FFNs, and output heads, while their representations interact through shared attention.

4.2 Human Video as Task Specification

For an unseen task, language may underspecify the desired task dynamics. We therefore use a human video $\mathbf{h}$ as an in-context task specification that demonstrates these dynamics directly. During training, each semantically matched *HumanGen* pair provides a generated human video and a robot trajectory with supervised future video-action chunks (Section 3). Since the pair may differ in embodiment, viewpoint, background, and object placement, the model must learn task-level correspondence rather than copy human motion.

During training, we apply a teacher-forcing attention mask over the tokenized representations of human video chunks, robot video chunks, and action chunks. The human video $\mathbf{h}$ is prepended before the robot trajectory and serves as prefix memory for robot video prediction. At chunk index i, the video Transformer predicts the next robot video chunk as

$$p_\theta^{\text{vid}}(\mathbf{x}^{i+1} \mid [\mathbf{h}, \mathbf{x}^{\leq i}], \mathbf{a}^{\leq i}, \ell). \tag{4}$$

For the next robot action chunk prediction, the action Transformer predicts $\mathbf{a}^{i+1}$ from robot-domain history and the next robot video chunk, following the inverse-dynamics factorization in Section 4.1. It does not directly attend to the human video $\mathbf{h}$: the task semantics conveyed by $\mathbf{h}$ are already absorbed into the next robot video chunk, so action decoding reduces to standard inverse dynamics. The action-prediction condition therefore remains unchanged from standard robot video-action training:

$$p_\theta^{\text{act}}(\mathbf{a}^{i+1} \mid \mathbf{x}^{\leq i}, \mathbf{a}^{\leq i}, \mathbf{x}^{i+1}, \ell). \tag{5}$$

During training, $\mathbf{x}^{i+1}$ in the action context is the ground-truth next robot video chunk under teacher forcing; at inference, it is replaced by the video chunk predicted by the video Transformer.

4.3 In-Context Future Chunk Prediction

For in-context samples, the model must infer task evolution from the human video $\mathbf{h}$ and transfer it to the robot domain. Under teacher forcing on seen tasks, however, the immediate next robot video chunk can often be extrapolated from recent robot history $\mathbf{x}^{\leq i}$, allowing the model to reduce its loss while underusing $\mathbf{h}$. This shortcut is especially harmful on unseen tasks, where the ICL signal specifies the novel dynamics. Inspired by the multi-chunk prediction of Next Forcing (Xu et al., 2026), we introduce in-context future chunk prediction (IFP), a training-only auxiliary objective that predicts multiple strided future robot video chunks from the current robot-video representation.

Let $s \geq 1$ be the temporal stride and K be the number of future chunks to predict. The k-th future target is the robot video chunk $\mathbf{x}^{j_k}$, where the target index j_k is defined as

$$j_k = (i+1) + 1 + (k-1)s, \qquad k = 1, \ldots, K. \tag{6}$$

To predict these targets, we add K IFP modules $\{G_k\}_{k=1}^K$, where G_k is responsible for denoising the k-th strided future video chunk under the flow-matching objective. Each IFP module has the same architecture as a single Transformer layer in the video branch and is initialized from the last video Transformer layer.

IFP operates on the current robot-video representation produced by the main branch. When predicting the current robot video chunk $\mathbf{x}^{i+1}$, we collect robot-video hidden representations from M intermediate layers of the main video Transformer. Let $\{\mathbf{r}_m^{i+1}\}_{m=1}^M$ denote the collected feature representations of $\mathbf{x}_t^{i+1}$, the noised chunk of $\mathbf{x}^{i+1}$ at flow time t. We concatenate these multi-layer representations and project them back to the single-layer hidden dimension with a lightweight MLP projection P_{fuse}:

$$\boldsymbol{\phi}^{i+1} = P_{\text{fuse}}\left(\text{Concat}\left(\{\mathbf{r}_m^{i+1}\}_{m=1}^M\right)\right). \tag{7}$$

Each IFP module G_k predicts its strided future chunk $\mathbf{x}^{j_k}$ conditioned on this fused current representation, the clean robot video/action history, and the language instruction:

$$p^{\text{ifp}}_{\theta,k}\left(\mathbf{x}^{j_k} \mid \boldsymbol{\phi}^{i+1}, \mathbf{x}^{\leq i}, \mathbf{a}^{\leq i}, \ell\right), \qquad k = 1, \ldots, K. \tag{8}$$

The IFP loss applies the flow-matching objective defined above to each strided target, with w_k denoting the loss weight for the k-th future video chunk target:

$$\mathcal{L}_{\text{ifp}} = \sum_{k=1}^{K} w_k \mathcal{L}_{\text{fm}}\left(\mathbf{x}^{j_k}; \boldsymbol{\phi}^{i+1}, \mathbf{x}^{\leq i}, \mathbf{a}^{\leq i}, \ell\right). \tag{9}$$

The IFP modules are not directly conditioned on the human video $\mathbf{h}$; they access its task information only through $\boldsymbol{\phi}^{i+1}$, after the main robot-video Transformer has interacted with the video. Directly conditioning IFP on $\mathbf{h}$ could let the auxiliary branch learn an independent future predictor without requiring the main branch to encode the task information, yet that auxiliary branch is removed at inference. By using only $\boldsymbol{\phi}^{i+1}$, the future loss instead supervises the main robot-video representation. The strided future chunks are predicted in parallel, increasing difficulty without introducing dependence among their predictions during training.

4.4 Training and Inference

Training data mixture. We sample *Task-diverse VA* for diverse robot-domain video-action dynamics and *HumanGen* for human video specifications grounded in executable actions, using sampling weights rather than raw dataset proportions.

Training objective. We optimize three objectives: the next-chunk video flow-matching loss $\mathcal{L}_{\text{fm}}$, an analogous action-space flow-matching loss $\mathcal{L}_a$, and the in-context future chunk prediction loss $\mathcal{L}_{\text{ifp}}$ defined above. The sample-specific objectives are

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{fm}}(\ell) + \lambda_a \mathcal{L}_a(\ell), & \textit{Task-diverse VA} \text{ samples}, \\ \mathcal{L}_{\text{fm}}(\{\mathbf{h}, \ell\}) + \lambda_a \mathcal{L}_a(\ell) + \lambda_{\text{ifp}} \mathcal{L}_{\text{ifp}}, & \textit{HumanGen} \text{ samples}. \end{cases} \tag{10}$$

Inference. At test time, the IFP modules are removed. Language-only mode uses $c = \ell$; ICL mode encodes the human video, caches its tokens as prefix memory, and uses $c = \{\mathbf{h}, \ell\}$ with optional ℓ.

5 Experiments

5.1 Implementation

Following LingBot-VA (Li et al., 2026b), we instantiate Zero-WAM from *Wan-2.2-TI2V-5B* (Wan et al., 2025) and convert it into the MoT causal video-action model described in Section 4. Both streams use RoPE positional encoding (Su et al., 2024); to distinguish human-video and robot-video latents, we offset the height coordinate of human-video RoPE by $\Delta_H = 32$ while leaving robot-video coordinates unchanged. IFP predicts $K = 4$ future robot video chunks with temporal stride $s = 2$ and loss weights $(w_1, \ldots, w_4) = (0.5, 0.25, 0.15, 0.15)$. We pre-train with AdamW (Loshchilov and Hutter, 2017) at a peak learning rate of 1×10^{-4}, sampling *Task-diverse VA* and *HumanGen* at a 1:5 ratio. Pre-training takes 15,360 GPU hours. Additional architecture, training, and inference details are provided in Appendix B.

5.2 Simulation Experiments

Training. We evaluate cross-task generalization on the RoboTwin 2.0 dataset (Chen et al., 2025). For each robot trajectory, we generate a corresponding human video instruction with the in-context human video generation pipeline, forming the Simulation ICL data. We use 43 seen tasks (50 episodes per task) and their ICL data for post-training and reserve 7 tasks for evaluation.[1] The seven evaluation tasks cover unseen-object, articulated-object, bimanual, and long-horizon manipulation;

[1] For *stamp seal* and *move stapler to pad*, we slightly adjust the success criteria for cross-task evaluation. The same criteria are applied to all methods, and we will release the modified task versions.

visualizations are provided in Appendix C. Starting from the pre-training checkpoint, we train with 64 GPUs for 4,000 post-training steps, jointly sampling *Task-diverse VA*, *HumanGen*, and RoboTwin data at a ratio of 2:10:3. Each GPU packs multiple samples up to 160K tokens.

Evaluation. We compare against two video-action baselines under the same cross-task protocol. Since Zero-WAM is initialized from Wan-2.2, we build a direct Wan-based baseline by initializing from *Wan-2.2-TI2V-5B*, adopting the same MoT causal video-action framework; we refer to it as WAN-Action. We also use the released LingBot-VA checkpoint and post-train it on the same seen tasks. For each of three random seeds, we run 100 closed-loop rollouts per unseen task and report the mean and standard deviation of task success.

Table 1: Zero-shot cross-task results on seven unseen RoboTwin tasks. We report task success rates and their macro average over three evaluation seeds.

Model	Place obj. on scale	Stamp seal	Open microwave	Move stapler to pad	Place bread in basket	Place empty cup	Stack three blocks	**Average**
WAN-Action	3.00±2.16	7.33±1.25	2.26±1.60	10.67±1.70	15.26±2.55	38.33±2.05	0.00±0.00	10.98±1.07
LingBot-VA	6.17±4.87	3.67±2.49	29.33±10.66	23.33±8.22	17.33±6.18	42.33±7.85	0.00±0.00	17.45±1.40
Zero-WAM *w/o pre-train*	22.00±1.63	53.33±2.36	25.90±11.03	74.33±0.47	32.63±2.83	46.33±0.47	0.00±0.00	36.36±1.81
Zero-WAM	**24.67**±2.05	**47.00**±4.55	**59.00**±2.83	**69.14**±2.93	**35.00**±3.74	**84.87**±0.18	**9.00**±2.16	**46.95**±0.72

Results. As shown in Table 1, Zero-WAM achieves an average success rate of 46.95%, outperforming LingBot-VA (17.45%) and WAN-Action (10.98%) by 29.50 and 35.97 percentage points, respectively. It outperforms both baselines across all seven unseen tasks, with particularly large gains on *open microwave* and *stamp seal*, which require inferring unseen articulated-object and relocation dynamics. On the hardest long-horizon task, *stack blocks three*, Zero-WAM is the only method to achieve non-zero success. These results demonstrate strong cross-task transfer enabled by human video instructions and task-balanced robotic pre-training.

5.3 Real-World Experiments

We evaluate Zero-WAM on a bimanual Franka robot across three task families (Table 2). For each family, a small set of seen-task robot demonstrations adapts the policy to the robot kinematics, while the evaluated task configurations are held out. At test time, Zero-WAM uses only a human video instruction, whereas LingBot-VA receives a detailed textual description.

Tasks. In *object-to-container placement*, each test configuration contains at least one object or container absent from the seen-task robot demonstrations. In *three-object sequential manipulation*, the human video specifies an arbitrary order for manipulating three objects, including unseen objects and containers. In *two-table-leg insertion*, it specifies which colored leg should enter each target hole, requiring two precise insertions. Detailed task settings are provided in Appendix D.

Results. Zero-WAM outperforms LingBot-VA across all three task families. The largest gain occurs in three-object sequential manipulation (33.3% vs. 10.0%), showing stronger transfer of long-horizon order specifications. It also achieves 16.7% on unseen table-leg insertion configurations, whereas LingBot-VA obtains zero success, demonstrating the benefit of human video for fine-grained task specification. Detailed protocols and qualitative results are provided in Appendix D.

5.4 Ablation Studies

All ablations use the same 43 seen and 7 unseen RoboTwin tasks as the main simulation experiment and evaluate human-video ICL, IFP, and task-balanced robotic pre-training.

Effect of in-context human video prompts. To isolate ICL from large-scale pre-training, Zero-WAM w/o pre-train and WAN-Action share the Wan-2.2 initialization and train only on the 43 seen RoboTwin tasks, using human-video and language instructions, respectively; LingBot-VA uses language while retaining robotic video-action pre-training. As shown in Table 1, Zero-WAM w/o pre-train achieves a 36.36% average success rate with human-video ICL, outperforming WAN-Action (10.98%) and LingBot-VA (17.45%). However, all three variants obtain zero success on *stack blocks*

Table 2: Real-world unseen-configuration evaluation. We report task success rates over 30 real-robot trials. LingBot-VA is evaluated with language instructions, while Zero-WAM uses human video instructions.

Task	Train demos	LingBot-VA	**Zero-WAM**
Object-to-container placement	120	43.3	**53.3**
Three-object sequential manipulation	96	10.0	**33.3**
Two-table-leg insertion	36	0.0	**16.7**

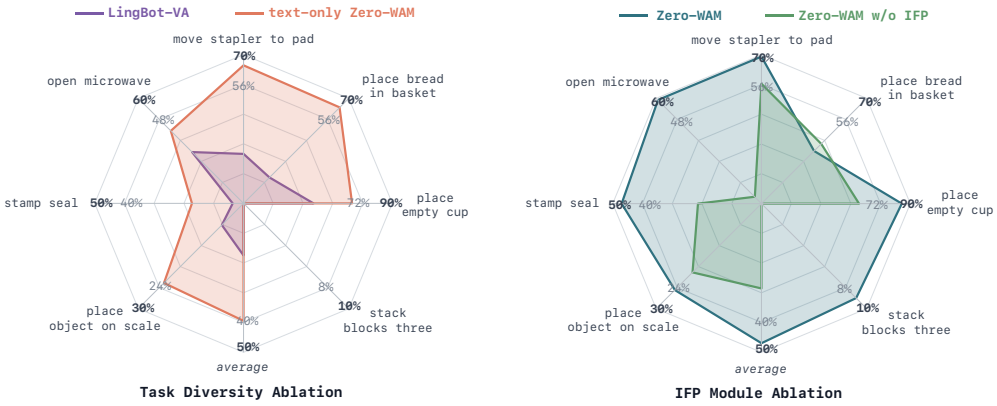


Figure 3: Left: task-balanced robotic data ablation, comparing LingBot-VA with a text-only Zero-WAM variant. Right: IFP module ablation, comparing full Zero-WAM with Zero-WAM w/o IFP. Radar axes indicate per-task success rates and the macro average.

three, whereas the full Zero-WAM succeeds, indicating that task-diverse ICL pre-training is needed for complex unseen tasks.

Effect of task-balanced robotic data. To isolate task-level sampling from human-video conditioning, we construct a text-only Zero-WAM variant that retains the ICL samples but masks their human-video condition, reducing them to text-conditioned robotic video-action samples. We compare against LingBot-VA, whose pre-training corpus covers most of the same raw robotic data. Figure 3 (left) shows that the text-only variant achieves a 39.44% seven-task average, outperforming LingBot-VA by 21.99 percentage points. Its gains on *move stapler to pad*, *place bread in basket*, and *place empty cup* show that task-balanced robotic pre-training improves cross-task transfer even without human-video conditioning.

Effect of in-context future chunk prediction. When trained with human-video instructions, the model may predict the next robot video from language and robot history without sufficiently using the video prompt. IFP counters this shortcut by supervising strided future chunks from the current robot-video representation, encouraging the main branch to encode task evolution from the human video. Figure 3 (right) shows that IFP raises the seven-task average from 28.55% to 46.95%, with large gains on *open microwave* and *stamp seal*. On the unseen long-horizon *stack blocks three* task, IFP breaks the zero-success barrier, improving success from 0.00% to 9.00%. These results show that IFP effectively elicits the human-video-following capability learned from ICL data.

6 Conclusion

Zero-shot cross-task generalization is essential for robots to perform new manipulation tasks without task-specific robot data or deployment-time adaptation. Our results highlight the value of combining scalable human-video task specifications with task-balanced robotic pre-training while explicitly training the policy to use the in-context demonstration. Zero-WAM instantiates this idea in a causal video-action model that accepts either language or human-video instructions. Results in RoboTwin and real-world experiments demonstrate transfer to unseen task configurations, suggesting a practical path toward more generalizable robot manipulation.

AI USE STATEMENT

We used generative AI tools to improve the grammar and clarity of the manuscript. As part of the proposed data-generation pipeline, we also use multimodal understanding and generation models to synthesize training data. Our use of these models follows their official usage policies and guidelines. The authors reviewed all AI-assisted writing and generated data and take responsibility for the final content of this work.

ETHICS STATEMENT

The synthetic data used in this work are derived from publicly available external datasets and internally collected data. All source data were obtained and used lawfully and in accordance with their applicable licenses, terms, and permissions. The data do not contain private or personally sensitive information, and their use does not introduce privacy concerns.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the model architecture, data-generation procedure, training setup, and key hyperparameters in the paper and appendix. Upon acceptance, we will release the datasets, model weights, and training configurations needed to reproduce our results.

REFERENCES

Skild AI. Introducing s1: In-context learning for robotics. August 2026. URL https://skild.ai/blogs/s1.

Alibaba Cloud. Wan2.7: Text-to-video api reference. Alibaba Cloud Model Studio documentation, https://www.alibabacloud.com/help/en/model-studio/text-to-video-api-reference, July 2026.

Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Xiongyi Cai, Ri-Zhao Qiu, Geng Chen, Lai Wei, Isabella Liu, Tianshu Huang, Xuxin Cheng, and Xiaolong Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

Guangyan Chen, Meiling Wang, Te Cui, Zichen Zhou, Qi Shao, Shalfun Li, Hang Su, Roy Gan, Hao Wang, Mengyin Fu, et al. Robots acquire manipulation skills in seconds from a single human video. *arXiv preprint arXiv:2607.20033*, 2026.

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

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

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

Dyna Robotics. Dyna-2: A 1-million-hour scaling law for world-action models. August 2026. URL `https://dyna.co/dyna-2`.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024.

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

Yusen Feng, Bingchen Han, Jiangran Lyu, Kai Liu, Yixin Zheng, Yuxuan Wan, Weiheng Liu, Sun Han, Ruiqin Li, Yulong Zhang, et al. Wam-ttt: Steering world-action models by watching human play at test time. *arXiv preprint arXiv:2607.06988*, 2026.

Google DeepMind. Gemini 3.1 Pro. Model card, `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026a.

Google DeepMind. Gemini 3.1 Flash Image. Model card, `https://deepmind.google/models/model-cards/gemini-3-1-flash-image/`, February 2026b.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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.

Wei-Jin Huang, Yue-Yi Zhang, Yi-Lin Wei, Zhi-Wei Xia, Juantao Tan, Yuan-Ming Li, Zhilin Zhao, and Wei-Shi Zheng. Learning whole-body human-humanoid interaction from human-human demonstrations. *arXiv preprint arXiv:2601.09518*, 2026.

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.

Vidhi Jain, Maria Attarian, Nikhil J Joshi, Ayzaan Wahid, Danny Driess, Quan Vuong, Pannag R Sanketi, Pierre Sermanet, Stefan Welker, Christine Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.

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

Yunfan Jiang, Yevgen Chebotar, Ruijie Zheng, Fengyuan Hu, Yunhao Ge, Jimmy Wu, Tianyuan Dai, Scott Reed, Li Fei-Fei, Yuke Zhu, et al. Robottt: Context scaling for robot policies. *arXiv preprint arXiv:2607.15275*, 2026.

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 *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025a.

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. *arXiv preprint arXiv:2512.22414*, 2025b.

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

Kuaishou Technology. Kling AI launches 3.0 model, ushering in an era where everyone can be a director. Official press release, https://ir.kuaishou.com/node/11216/pdf, February 2026.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. Egowam: World action models beyond pixels with in-the-wild egocentric human data. In *Robot World Models*, 2026a.

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

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 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*, 2026c.

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

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

Mingyu Liu, Jiuhe Shu, Hui Chen, Zeju Li, Canyu Zhao, Jiange Yang, Shenyuan Gao, Hao Chen, and Chunhua Shen. Stamo: Unsupervised learning of generalizable robot motion from compact state representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35014–35024, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

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

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

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*, 2026b.

Mimic Robotics. Introducing FLUX-mimic: Scaling Video-Action Models for General Purpose Dexterity. Official blog, https://www.mimicrobotics.com/blog/introducing-flux-mimic, 2026.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

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

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as i do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

Austin Patel, Ben Pekarek, Joel Enrique Castro Hernandez, and Shuran Song. Behavior prompting policy: Demonstrations as prompts for manipulation. *arXiv preprint arXiv:2606.30457*, 2026.

Qwen Team. Qwen3.6-Plus: Towards real world agents. Official blog, https://qwen.ai/blog?id=qwen3.6, April 2026.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Shuang Li Yihuai Gao Dorsa Sadigh and Shuran Song. Unified video action model.

Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In *Conference on robot learning*, pp. 906–915. PMLR, 2018.

Modi Shi, Shijia Peng, Jin Chen, Haoran Jiang, Tianyu Li, Di Huang, Ping Luo, Hongyang Li, and Li Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. *arXiv preprint arXiv:2602.10106*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Generalist Team. Gen-1.5: Embodied foundation models are one-shot learners. *Generalist AI Blog*, 2026. https://generalistai.com/blog/gen-1.5.

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

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Rhoda AI Team et al. Causal video models are data-efficient robot policy learners. *Rhoda AI Blog*, 2, 2026b.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 976–985, 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.

Ye Wang, Pei Lin, Xiong-Hui Chen, Haoqi Yuan, Zhixuan Liang, Yiyang Huang, Anzhe Chen, Zixing Lei, Jie Zhang, Tao Zhang, et al. Ego2robot: Scalable robot data synthesis from egocentric human data. *arXiv preprint arXiv:2608.02580*, 2026a.

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. *arXiv preprint arXiv:2605.24934*, 2026b.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

Wei Wu, Fangjing Wang, Fan Lu, He Sun, Shi Liu, Yunnan Wang, Yibin Yan, Yong Wang, Shuailei Ma, Xinyang Wang, et al. From foundation to application: Improving vla models in practice. *arXiv preprint arXiv:2607.06403*, 2026a.

Xiao-Ming Wu, Bin Fan, Kang Liao, Jian-Jian Jiang, Runze Yang, Yihang Luo, Zhonghua Wu, Wei-Shi Zheng, and Chen Change Loy. Vlanext: Recipes for building strong vla models. *arXiv preprint arXiv:2602.18532*, 2026b.

Hengyi Xie, Chenfei Yao, Xianjin Wu, Yingying Zhu, Dingkang Liang, Xiang Bai, and Han Ding. Turbovla: Real-time vision-language-action model at 32 hz on an rtx 4090 with $<$ 1 GB VRAM. *arXiv preprint arXiv:2607.27205*, 2026a.

Sicheng Xie, Haidong Cao, Zejia Weng, Zhen Xing, Haoran Chen, Shiwei Shen, Jiaqi Leng, Zuxuan Wu, and Yu-Gang Jiang. Human2robot: Learning robot actions from paired human-robot videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11078–11086, 2026b.

Gangwei Xu, Qihang Zhang, Jiaming Zhou, Xing Zhu, Yujun Shen, Xin Yang, and Yinghao Xu. Next forcing: Causal world modeling with multi-chunk prediction. *arXiv preprint arXiv:2606.11187*, 2026.

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

Ke Ye, Jiaming Zhou, Yuanfeng Qiu, Jiayi Liu, Shihui Zhou, Kun-Yu Lin, and Junwei Liang. From watch to imagine: Steering long-horizon manipulation via human demonstration and future envisionment. *arXiv preprint arXiv:2509.22205*, 2025.

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

Ryan Yu, Pushi Zhang, Starrick Liu, Brae Liu, Miracle Kang, Shalfun Li, Lights Shi, Ellie Ma, Ping Yang, Chris Pan, et al. Wall-oss-0.5 technical report. *arXiv preprint arXiv:2605.30877*, 2026.

Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026a.

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*, 2026b.

Mengqi Zhang, Sahil Khose, Simar Kareer, Yuchen Song, Unnat Jain, and Judy Hoffman. Deva: Decoupled video-action model with physical guidance for robot policy learning. *arXiv preprint arXiv:2607.24159*, 2026a.

Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, et al. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026b.

Bing Zhao, Chenfei Wu, Deqing Li, Hao Meng, Jiahao Li, Jie Zhang, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kuan Cao, et al. Qwen-image-2.0 technical report. *arXiv preprint arXiv:2605.10730*, 2026a.

Weiheng Zhao, Haoyi Jiang, Xin Shi, Liu Liu, Fan Huang, Zhizhong Su, Wei Sui, and Xinggang Wang. Faster-wam: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026b.

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*, volume 2026, pp. 60580–60606, 2026a.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026b.

Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: Learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025a.

Jiaming Zhou, Teli Ma, Kun-Yu Lin, Zifan Wang, Ronghe Qiu, and Junwei Liang. Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22551–22561. IEEE, 2025b.

Jiaming Zhou, Ke Ye, Jiayi Liu, Teli Ma, Zifan Wang, Ronghe Qiu, Kun-Yu Lin, Zhilin Zhao, and Junwei Liang. Exploring the limits of vision-language-action manipulation in cross-task generalization. *Advances in Neural Information Processing Systems*, 38:139899–139927, 2026a.

Pengfei Zhou, Shengcong Chen, Di Chen, Jiaxu Wang, Rongjun Jin, Bingwen Zhu, Yike Pan, Songen Gu, Kuanning Wang, Shufeng Nan, et al. τ_0-wm: A unified video-action world model for robotic manipulation. *arXiv preprint arXiv:2606.01027*, 2026b.

A Dataset Details and Comparisons

A.1 VLM-Based Quality Filtering

For each generated human video, the VLM assigns scores along two dimensions relative to the sampled source robot video: task-semantic consistency and physical plausibility. Task-semantic consistency measures whether the two videos perform the same task, including the manipulated objects, action goal, and key object-state changes. Physical plausibility measures whether the generated human motion, object interactions, and scene evolution are physically coherent. Both dimensions use the three-level rubric in Table 3.

Table 3: VLM quality-filtering rubric for generated human videos.

Dimension	Score	Criterion
Task-semantic consistency	5 (perfect)	The task, manipulated objects, action goal, and key state changes match the source robot video.
	3 (acceptable)	The overall task is recognizable, but nonessential object or state details differ.
	1 (bad)	The task is different or ambiguous, or a key object, goal, or state change is missing or incorrect.
Physical plausibility	5 (perfect)	Human motion, object interactions, and scene evolution are coherent, with no meaningful physical artifacts.
	3 (acceptable)	Minor artifacts or temporal inconsistencies are present, but the task execution remains physically plausible.
	1 (bad)	Implausible motion, interactions, or severe generation artifacts invalidate the task execution.

We retain a generated video only when its task-semantic consistency score is 5 and its physical plausibility score is either 3 or 5. All other videos are discarded. Each retained video is paired with its source robot trajectory to construct a human-robot ICL sample.

A.2 Detailed HumanGen Composition

Following the same task-diverse sampling principle used to curate *Task-diverse VA* data, source robot videos are sampled by task rather than by raw trajectory frequency, so that the ICL data are not dominated by repeated executions of a small set of tasks.

Pre-train ICL (External). This subset is built from the same five public robotic video-action datasets used for *Task-diverse VA* data: AgiBot (Bu et al., 2025), InternData-A1 (Tian et al., 2026), Open-X-Embodiment (O'Neill et al., 2024), RoboCOIN (Wu et al., 2025), and RoboMIND (Wu et al., 2024). These datasets cover diverse robot embodiments, scenes, and objects, including more than 45 robot embodiments. We sample robot trajectories by task from each dataset and convert them into semantically matched human videos using the in-context human video generation pipeline. Specifically, we sample 3,354 tasks and 6,660 trajectories from AgiBot, 261 tasks and 12,515 trajectories from InternData-A1, 438 tasks and 9,290 trajectories from Open-X-Embodiment, 512 tasks and 6,513 trajectories from RoboCOIN, and 497 tasks and 6,210 trajectories from RoboMIND. In total, Pre-train ICL (External) contains 5,062 tasks and 41,188 human-robot ICL pairs. For this subset, we intentionally diversify the visual alignment between the human and robot videos: we enforce changes in the scene environment, tabletop background, object instances, and object placement, and balance first-person and third-person viewpoints. Because task semantics are preserved across these visual variations, the model is pushed to learn human-robot task correspondences that are invariant to changes in scenes, objects, and viewpoints, improving its robustness to visually diverse ICL prompts.

Pre-train ICL (In-house). To further increase task coverage, we sample robot trajectories by task from our in-house robotic datasets. This subset covers multiple robot embodiments, such as bimanual Franka and Galaxea R1 Pro, and contains 3,522 tasks and 30,247 human-robot ICL pairs. When using the in-context human video generation pipeline to produce the corresponding human videos for this subset, we reduce the ratio of third-person viewpoints and decrease the frequency of scene, object,

and object-placement changes. This produces more visually aligned ICL samples while maintaining task diversity.

Simulation ICL. To support cross-task evaluation in simulation, we generate ICL data for 50 RoboTwin tasks (Chen et al., 2025). This subset contains 2,500 ICL samples, with 50 samples for each task. Among them, 43 tasks are used for post-training (2,150 training samples), while the remaining 7 tasks are reserved as unseen tasks for zero-shot cross-task evaluation.

Real-world ICL. Real-world ICL contains human-robot pairs collected on the bimanual Franka evaluation embodiment. It includes 252 human-robot ICL pairs across the three real-world task families used in our evaluation: 120 pairs from 30 object-to-container placement training task combinations, 96 pairs from 16 three-object sequential manipulation training task combinations, and 36 pairs for two-table-leg insertion. This subset is used for post-training so that the model can fit the evaluation robot kinematics for cross-task evaluation.

A.3 Comparison with Existing Datasets

Table 4: Comparison with existing task-level paired human-robot datasets. "Multi-source" indicates whether the dataset is constructed from multiple data sources. "Embod." is the number of robot embodiments. "Human view" specifies the viewpoint of human videos. "Human acquisition" indicates whether the human videos require manual collection or are automatically generated. "Align. div." indicates whether human-robot pairs include diverse degrees of visual alignment, such as changes in background, viewpoint, environment style, object instance, or object placement. "Samples" is the number of human video instructions, and "Tasks" is the number of task categories.

Dataset	Multi-source	Embod.	Human view	Human acquisition	Align. div.	Samples	**Tasks**
MIME (Sharma et al., 2018)	✗	1	third	manual	✗	8.3K	20
EgoMimic (Kareer et al., 2025a)	✗	1	ego	manual	✗	2.1K	3
BC-Z (Jang et al., 2022)	✗	1	ego	manual	✓	18.7K	100
EgoHumanoid (Shi et al., 2026)	✗	1	ego	manual	✗	1.2K	4
RH20T (Fang et al., 2024)	✗	4	ego & third	manual	✗	110K	147
EgoScale (Zheng et al., 2026b)	✗	1	ego	manual	✗	10.3K	344
***HumanGen* (ours)**	✓	>45	**ego & third**	**auto-generated**	✓	**74.2K**	**8.6K**

Table 4 compares *HumanGen* with existing task-level paired human-robot datasets. The statistics are taken from their official papers. Unlike prior datasets that rely on manual human data collection, *HumanGen* uses the in-context human video generation pipeline to automatically generate semantically matched human videos from robot trajectories, enabling scalable construction of human-robot ICL pairs. *HumanGen* also combines multiple data sources (public, in-house, simulation, and real-world), covers substantially more robot embodiments, contains human data from both ego and third-person views, and contains paired data with diverse degrees of visual alignment. Most importantly, *HumanGen* contains substantially broader task coverage, which is a key factor for zero-shot cross-task generalization.

B Additional Implementation Details

Architecture. The video Transformer follows the Wan-2.2 backbone with hidden dimension $d_v = 3072$ and 30 Transformer layers. The action Transformer has hidden dimension $d_a = 3072$, is initialized from the video branch, and is connected to the video Transformer through the MoT architecture. We use the Wan-2.2 VAE to encode human and robot videos into latent representations and T5 (Raffel et al., 2020) to encode language instructions. For IFP feature fusion, we collect hidden states from $M = 4$ layers of the 30-layer video Transformer, specifically layers $\{4, 12, 20, 30\}$.

Training. We use the flow-matching video loss, action loss, and IFP loss defined in Section 4. AdamW uses weight decay 0.01. For non-ICL samples, we drop the language instruction with probability 0.1. For ICL samples, we drop the human-video latent with probability 0.1 and increase language-instruction dropout from the Wan-2.2 default of 0.1 to 0.4. This modality dropout reduces reliance on language and encourages the use of human-video instructions. The robot video chunk size is randomly sampled from 1 to 4 during pre-training. Following LingBot-VA, each GPU packs

samples of different token lengths into one sequence and uses attention masks to isolate them, improving throughput while limiting each GPU to 160K tokens.

Inference. In language-only mode, the model uses only the text instruction, without the ICL human video, and the video CFG scale is 5. In ICL mode, it uses only the human video, disables language, and applies ICL classifier-free guidance with scale 5 (Ho and Salimans, 2022). In both modes, the inference chunk size is 2 and the action CFG scale is 1.0.

C Additional Simulation Experiment Details

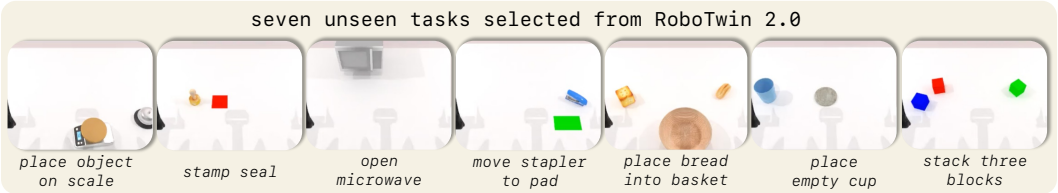


Figure 4: Seven unseen evaluation tasks in RoboTwin 2.0. These tasks involve unseen-object manipulation, articulated-object manipulation, bimanual manipulation, and long-horizon manipulation.

D Additional Real-World Experiment Details

D.1 Task Protocols

Object-to-container placement. The robot picks an object and places it into a target container using either single-arm manipulation or bimanual coordination. We collect 120 seen-task demonstrations from 30 object-container training combinations to adapt the policy to the Franka embodiment. We hold out part of the object and container set, so each test configuration contains at least one object or container absent from the seen-task robot demonstrations.

Three-object sequential manipulation. This task evaluates whether the policy can follow a human video that specifies an arbitrary order for manipulating three objects. Each episode selects three objects from a scene containing multiple objects and containers, and the human video shows their manipulation in a random order. We collect 96 demonstrations from 16 training combinations and corresponding human video instructions. At test time, the video specifies an arbitrary order that may include unseen objects and containers.

Two-table-leg insertion. This task tests fine-grained human-video task specification by requiring two table legs to be inserted into selected holes on a tabletop base. The human video specifies which colored leg should be inserted into each target hole. We collect 36 demonstrations for this precise insertion setting, and the robot must complete both insertions at test time. The absolute success rate is limited by the difficulty of precise physical insertion, but Zero-WAM succeeds on unseen insertion configurations while LingBot-VA obtains zero success.

D.2 Qualitative Results

We further evaluate generalization to a new task involving an unseen object by directly deploying the checkpoint trained on two-table-leg insertion on an unseen light-bulb insertion task, without additional training or fine-tuning. As shown in the fourth panel of Figure 5, Zero-WAM follows the human video instruction and successfully transfers the learned insertion skill to this novel object.



Figure 5: Qualitative real-world evaluations of Zero-WAM with human video instructions.