WorldScape Policy: Generalizable Robotic Learning via a Foundation World Model

Manifold AI

Abstract. World models offer a promising route to robotic learning by enabling prediction and planning directly in high-dimensional visual space. However, unlike Vision-Language-Action (VLA) policies that reason in language space but often struggle with out-of-distribution visual shifts, turning a powerful generative world model into an executable policy remains challenging. We present **WorldScape Policy**, a world-model-based policy that adapts a pretrained diffusion foundation world model into a generalist robotic planner. By coupling the world model with a lightweight action expert in a unified **Mixture-of-Transformers (MoT)** architecture, our approach overcomes the compounding errors of traditional two-stage inverse dynamics pipelines. The policy jointly models RGB, depth, and continuous action chunks under a unified flow-matching objective, augmented with depth-aware token conditioning to support robust 3D spatial reasoning for contact-rich manipulation. To scale beyond embodiment-specific supervision, we introduce an automated curation and annotation pipeline that converts large-scale egocentric human videos into interleaved instruction-action clips, mixing them with cross-embodiment robot demonstrations. A four-stage training recipe further aligns predictive dynamics with control and supports continual improvement via advantage-conditioned human-in-the-loop post-training. Extensive experiments conducted on a dual-arm physical platform demonstrate that WorldScape Policy achieves unprecedented robustness against severe distribution shifts. Furthermore, our approach exhibits remarkable few-shot adaptability to unseen tasks and embodiments, establishing foundation world models as a transformative and superior paradigm for scalable generalist robot learning.

Date: February 28, 2026
Webpage: https://manifoldai.cn/blogs/WorldScapePolicy.html

1 Introduction

Building generalist robot policies that can follow language instructions and solve long-horizon manipulation tasks remains difficult under real-world constraints. On the one hand, large vision–language–action (VLA) policies can exhibit impressive in-domain performance, but they often struggle with distribution shift (e.g., unseen object instances, backgrounds, lighting, and contact dynamics) and require substantial embodiment-specific data for reliable deployment. On the other hand, model-based approaches promise improved sample efficiency by explicitly modeling dynamics, yet in high-dimensional visual observation spaces they are typically either computationally expensive at inference time (sampling-based planning) or brittle when the learned dynamics are misaligned with downstream control.

This report investigates a third, fundamentally different route: *world-model-based policies* that directly leverage a strong foundation world model as the backbone for action generation. Unlike VLA approaches that perform action reasoning in language space or classical model-based methods that suffer from expensive sampling-based planning, our core hypothesis is that a pretrained embodied world model provides a reusable, visually-grounded prior over physical interaction and scene evolution. Coupling this prior with a lightweight action expert produces a practical policy that (i) generalizes significantly better under severe distribution shifts, and (ii) transfers functional knowledge seamlessly across embodiments through shared predictive representations.

We present **WorldScape Policy**, which adapts the **WorldScape** video foundation world model into a generalist robotic planner. Unlike two-stage pipelines that first predict future states and then learn a separate inverse dynamics model—which often suffer from compounding errors and embodiment-specific retuning—WorldScape Policy adopts a unified **Mixture-of-Transformers (MoT)** architecture with end-to-end optimization for joint "Video–Depth–Action"

modeling. The policy predicts future visual trajectories while simultaneously denoising continuous action chunks, enabling tight video–action alignment and allowing the policy to plan directly in the model's predictive feature space. To support 3D-aware reasoning for contact-rich manipulation, we augment the language / vision reasoning tokens with depth positional embedding and train with depth video diffusion objectives, encouraging the policy to represent geometry consistently across time.

To make this approach scalable, we curate an **embodied data pyramid** that combines (i) large-scale egocentric human videos, (ii) cross-embodiment robot demonstrations, and (iii) high-quality target-robot teleoperation. We further introduce an automated curation pipeline that segments long videos into atomic actions, generates interleaved sub-task instructions, and annotates camera / hand trajectories to align human data with robot action spaces. Training proceeds in four stages: foundation world model pretraining, unified world–action pretraining, target-robot fine-tuning, and advantage-conditioned human-in-the-loop post-training for continual improvement.

We evaluate WorldScape Policy in real-world settings, including randomized environments designed to stress robustness and generalization. Across these evaluations, the results strongly support the superiority of world models for functional control: incorporating predictive dynamics and geometry-aware conditioning yields exceptional robustness under domain randomization, and the multi-stage training recipe enables effective cross-embodiment adaptation, establishing world-model-based policies as a highly resilient alternative to standard VLA paradigms.

In sum, this technical report makes the following contributions:

- **World-model-based policy design:** a unified MoT architecture that jointly models future RGB / depth trajectories and continuous action chunks, overcoming the compounding errors of two-stage pipelines and improving video-action alignment for robust planning.
- **Scalable data curation and annotation pipeline:** an automated pipeline for converting egocentric human videos into instruction–action clips and mixing them with cross-embodiment robot data for generalist pretraining.
- **Practical training recipe:** a four-stage procedure including advantage-conditioned human-in-the-loop post-training to improve closed-loop performance and generalization.
- **Comprehensive evaluation:** experiments are conducted on a real physical embodiment to validate robustness, generalization and few-shot capability.

2 Preliminary

Embodied World Model. Embodied world models (EWMs) are generative models that predict future observations of physical scenes involving robot locomotion and manipulation, and are therefore central to the development of embodied intelligence [1, 2]. Methodologically, they aim to learn general state transition regularities of the physical world in high-dimensional visual space, typically via self-supervised training on interaction data. From the perspective of representation modality, existing EWMs fall into two main categories: video world models [3, 4] and 3D world models [5, 6]. Video world models usually start from general video generation backbones [7, 8] and perform embodied post-training to improve controllability and physical-rule adherence, thereby supporting more reliable action-conditioned prediction. In contrast, 3D embodied world models explicitly construct three-dimensional representations of embodied environments, prioritizing multi-view spatial consistency and geometry-aware prediction under viewpoint changes. In terms of usage, EWMs play two complementary roles. First, they serve as data synthesis engines that generate video–action trajectories for downstream policy learning [9, 10], alleviating the high cost of collecting real-world embodied data. Second, they function as internal simulators for action planning, where robot actions are decoded via forecasting future environment states [11, 12]. This planning-centric pathway has recently gained renewed attention as a route parallel to the VLA-based embodied learning paradigm [13, 14]: while VLA approaches perform action reasoning in the language space, EWMs reason about world states in visual feature space, making them particularly well-suited for embodied learning.

Video Prediction Policy. Video-based embodied action planning has recently emerged as a distinct paradigm for embodied learning [11, 12, 15]. In this framework, world models are trained in a predominantly self-supervised manner on large-scale video or interaction data to capture general physical dynamics and are subsequently used as predictive simulators for decision making. Existing approaches mainly follow two directions. The first adopts a two-stage inverse

dynamics formulation, where a video world model is trained for future state prediction and then fixed, followed by learning an inverse dynamics model that maps predicted state sequences to actions [11, 12, 16, 17]. Although this decouples dynamics modeling from control inference, it typically requires embodiment-specific adaptation to different robot morphologies and action spaces. The second direction trains a fully action-conditioned world model and performs planning via trajectory optimization by sampling multiple candidate action sequences, rolling them forward to predict future outcomes, and selecting the sequence that maximizes a task reward [18, 19]. Although structurally simple, the sampling-based procedure is computationally expensive and prone to sampling bias in long-horizon settings. More recently, unified architectures have been proposed that jointly optimize world modeling and action prediction within a single framework, typically through coupled video and action branches trained under a multi-task objective [20–22]. By co-training visual dynamics modeling and policy learning, these approaches enhance representation sharing and improve alignment between predicted world states and action generation. Our method follows this unified paradigm and is designed to further strengthen the mutual benefit between video and action learning.

Real-world Reinforcement Learning. Reinforcement learning in real-world robotics imposes stringent constraints on sample efficiency, stability, and scalability, particularly for high-capacity VLA models. Early attempts directly fine-tune pretrained VLAs using PPO-style policy gradients [23, 24], which, while effective in simulation or controlled settings, often become unstable and costly under real-world interaction budgets. To alleviate these issues, subsequent work augments pretrained models through residual policy learning [25], diffusion-space policy optimization [26], sometimes combined with iterative distillation [27]. However, these methods commonly rely on on-policy updates and simplified action distributions, limiting robustness in long-horizon tasks. More recent approaches incorporate value-based objectives, including offline Q-learning [28] from demonstrations [29], preference-driven optimization [30], and value-augmented policy gradients [31] deployed on physical robots. Although improving practical adaptability, such methods still encounter variance and optimization instability when scaling to large generative architectures. To enhance long-horizon learning stability, advantage-aware modeling [32] has emerged as a promising direction, ranging from advantage-weighted regression to explicit state-action advantage estimation. Yet temporal-difference-based advantages remain noisy in complex real-world dynamics. Recent advances [33, 34] therefore propose learning smoother advantage signals from paired observations, sometimes with semantic phase conditioning or discretized optimality supervision, providing more stable and practical training signals for real-world VLA systems.

3 Problem Formulation & Overall Objective

In this paper, we formulate the robotic manipulation task as a conditional action generation problem. Specifically, **WorldScape Policy** π_p jointly predicts video sequences $\mathbf{o}_{s:s+L} \in \mathbb{R}^{L\times N\times H\times W\times 3}$ and corresponding actions $\mathbf{a}_{s:s+L} \in \mathbb{R}^{L\times D}$, conditioned on language instruction i, proprioceptive state $\mathbf{q}_s$ and multi-modal visual observations. These observations include multi-view RGB images $\mathbf{o}_s \in \mathbb{R}^{N\times H\times W\times 3}$ and depth maps $\mathbf{d}_s \in \mathbb{R}^{N\times H\times W\times 1}$. Here, $L > 0$ denotes a fixed time horizon and s is a random index sampled from a trajectory, N is the number of camera views, and D indicates the dimensionality of the robot's action space. We emphasize that the joint prediction of video and action can be decomposed into: (1) autoregressive video prediction π_v and (2) action prediction π_a via an Inverse-Dynamics Model (IDM) conditioned on predictive visual dynamics:

$$\underbrace{\pi_p(\mathbf{o}_{s:s+L}, \mathbf{d}_{s:s+L}, \mathbf{a}_{s:s+L} | \mathbf{o}_s, \mathbf{d}_s, i, \mathbf{q}_s)}_{\text{WorldScape Policy}} = \underbrace{\pi_v(\mathbf{o}_{s:s+L} | \mathbf{o}_s, i, \mathbf{q}_s)\pi_v(\mathbf{d}_{s:s+L} | \mathbf{d}_s, i, \mathbf{q}_s)}_{\text{WorldScape Foundation Model}} \underbrace{\pi_a(\mathbf{a}_{s:s+L} | \mathbf{o}_{s:s+L}, \mathbf{d}_{s:s+L}, \mathbf{q}_s)}_{\text{IDM}} \quad (1)$$

To optimize this objective, instead of employing two separate models, we train a single model end-to-end using the **Mixture-of-Transformers (MoT)** architecture. We posit that this end-to-end design facilitates superior video-action alignment through the deep integration of multiple modalities. Since pretrained video models are already optimized for video prediction on diverse web-scale data, WorldScape Policy only needs to learn to adapt to embodied scenarios and extract coherent actions from the generated visual sequences. We further claim that this paradigm encourages better generalization than the conventional practice of fine-tuning Vision-Language-Action (VLA) models from off-the-shelf VLMs, which often perform action reasoning within the linguistic space. In contrast, our proposed **Embodied World Model (EWM)** explicitly learns temporal dynamics and performs state transitions in the visual feature space, making it particularly well-suited for long-horizon embodied learning tasks.

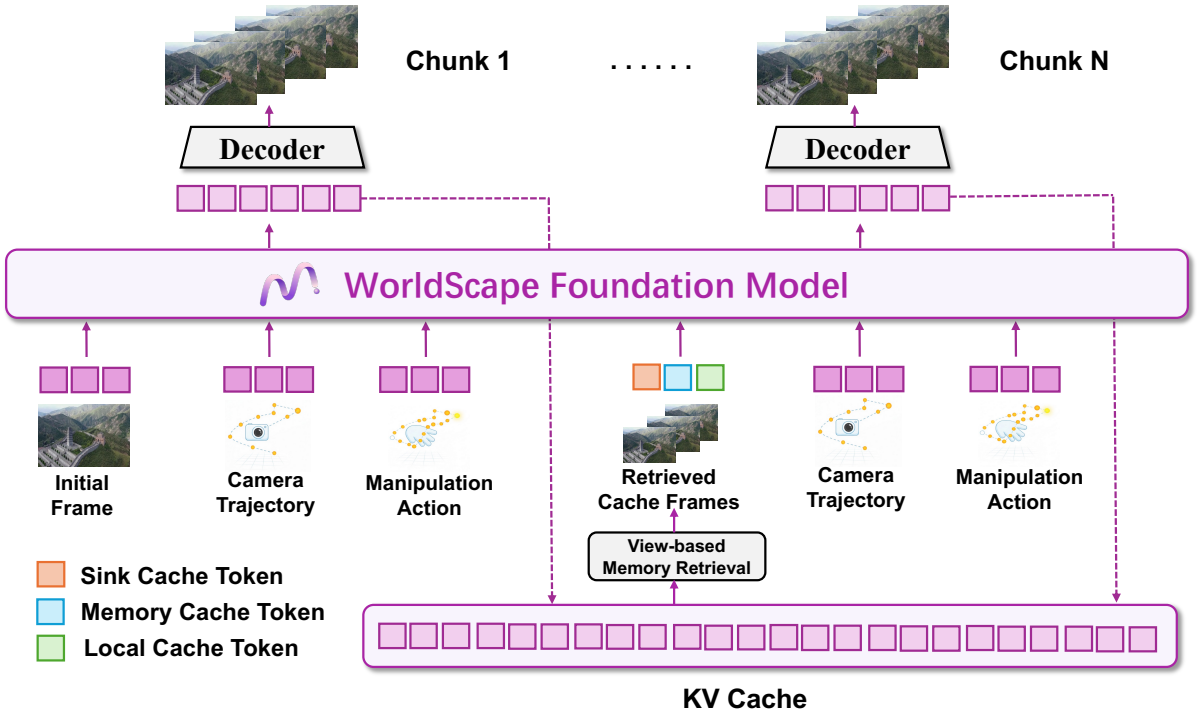


Figure 1: WorldScape foundation model overview. WorldScape employs a chunk-wise autoregressive video generation framework. In this design, each video segment is denoised sequentially, conditioned on camera trajectory, manipulation action, and the cached key–value states from previously generated chunks.

4 WorldScape Policy: Adapting Embodied World Model for Robotic Planning

4.1 WorldScape Foundation Model

We adopt a foundation world model, WorldScape, as the core of WorldScape Policy. WorldScape Policy is obtained by fine-tuning this foundation model under a multi-task objective, without modifying its core diffusion transformer structure. In this section, we detail the key architectural components of the foundation world model.

Autoregressive Diffusion Architecture. WorldScape is built on an autoregressive diffusion-based video transformer for action-conditioned future prediction, as shown in Figure 1. Given current observations and control inputs, the model generates temporally evolving future states in a chunk-level autoregressive manner. Each generation step conditions on: (i) visual latents of the current observation, (ii) structured control signals, and (iii) cached representations from previously generated chunks. The backbone adopts a diffusion transformer with causal attention over chunked latent tokens. All modalities are modeled within a unified denoising framework.

Unified Interaction Conditioning. The foundation model integrates structured control inputs within a unified token-level conditioning mechanism in the diffusion transformer. It supports two categories of interaction signals: locomotion control, represented by camera trajectory parameters, and manipulation control, represented by articulated pose sequences. Locomotion signals are parameterized using per-frame Plücker embeddings $\mathbf{P}_i \in \mathbb{R}^{6\times h\times w}$ and processed by a lightweight convolutional adapter. The resulting spatially aligned control features are injected into the transformer hidden states via residual addition. Manipulation signals are represented as pose sequences converted into a control video. This control video is concatenated with the visual latent along the channel dimension before denoising, allowing joint processing of perception and action signals within the same diffusion step. Inside each World–Action DiT block, image, depth, and action latents are treated as separate token streams. These tokens interact through a shared joint-attention module, where cross-modal dependencies are modeled explicitly. Adaptive layer normalization (AdaLN) conditions the transformer layers on time embeddings and control embeddings, ensuring consistent integration of action signals across denoising stages. Both locomotion and manipulation controls are therefore modeled within the same diffusion backbone

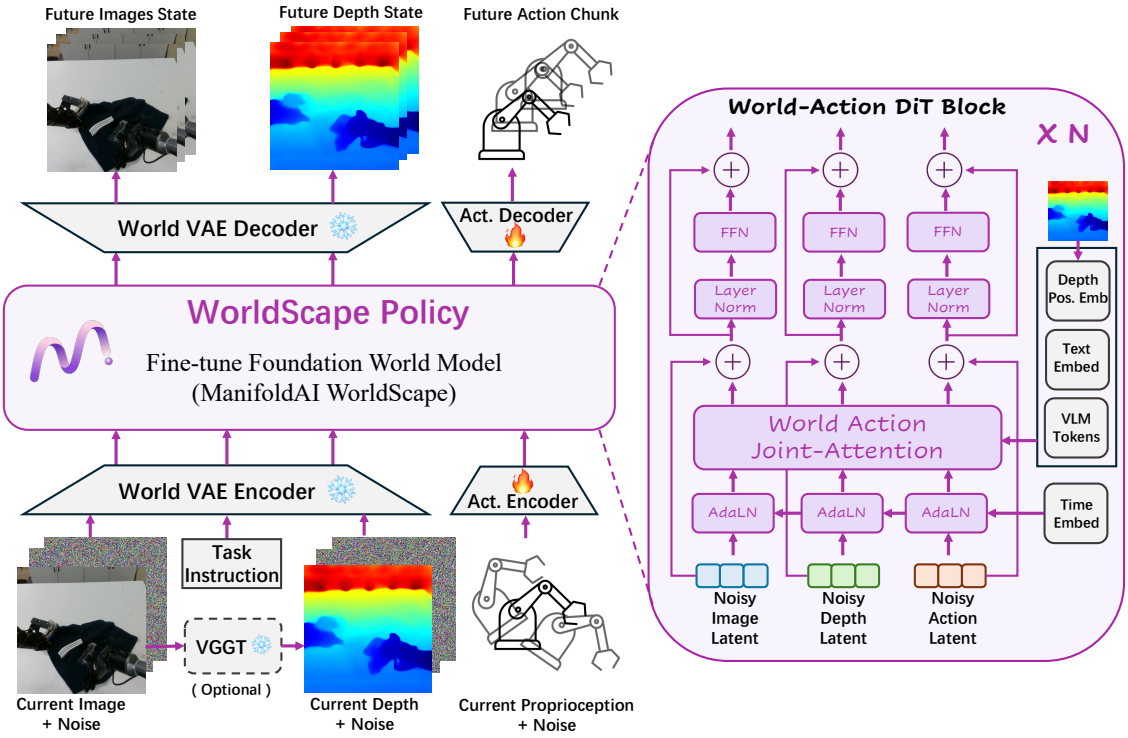


Figure 2: WorldScape Policy framework overview. WorldScape Policy serves as a generalist robotic planner which is finetuned from the ManifoldAI WorldScape foundation model. It handles multimodal inputs and predicts both future states and executable robot actions within a unified model. All modalities are jointly modeled through diffusion training paradigm under a multi-task objective.

at the token level, without introducing modality-specific branches. During pretraining, the transformer learns a joint distribution over control tokens and visual state tokens, establishing shared video–action representations in latent space. This learned video-action correspondence is preserved during fine-tuning and provides an initialized action-conditioned generative prior when extending the model to predict robot action chunks in WorldScape Policy.

3D Spatial Consistency Training. To incorporate geometric constraints into training, we adopt a spatial-aware multi-task objective that combines diffusion supervision with 3D consistency regularization. Let $\mathbf{z}_t$ denote the noisy latent at timestep t, and $\mathbf{v}_\theta(\mathbf{z}_t, t)$ denote the predicted velocity field. The diffusion objective is

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\mathbf{z}_0}\left[\|\mathbf{v}_\theta(\mathbf{z}_t, t) - \mathbf{v}_t\|_2^2\right], \tag{2}$$

where $\mathbf{v}_t$ is the target velocity from the forward process. To enforce 3D consistency, the denoised latent $\hat{\mathbf{z}}_0$ is used to reconstruct depth and RGB signals under K camera viewpoints. The geometric loss is defined as

$$\mathcal{L}_{\text{3D}} = \frac{1}{K}\sum_{k=1}^{K}\left(\|\hat{\mathbf{D}}_k - \mathbf{D}_k\|_2^2 + \lambda\|\hat{\mathbf{I}}_k - \mathbf{I}_k\|_2^2\right), \tag{3}$$

where $(\mathbf{D}_k, \mathbf{I}_k)$ denote ground-truth depth and RGB observations under viewpoint k. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \alpha\mathcal{L}_{\text{3D}}. \tag{4}$$

Hierarchical Memory Mechanism. WorldScape maintains long-horizon consistency through a hierarchical KV cache design embedded in the autoregressive transformer. At each denoising step t, the attention context is composed of three memory partitions: the anchor memory stores persistent scene tokens, the global memory pool retains selected historical states, and the local window preserves recent temporal context. Each cached entry is indexed by its associated camera

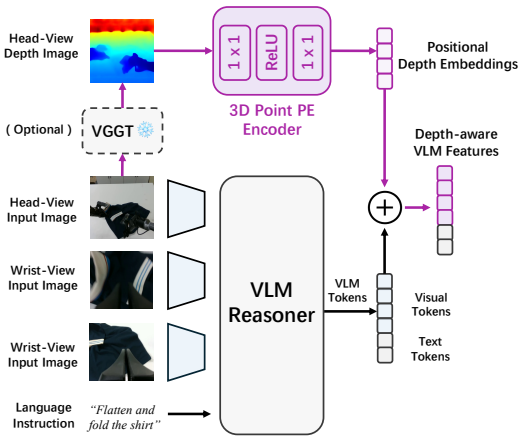


Figure 3: Details of Depth Positional Embedding.

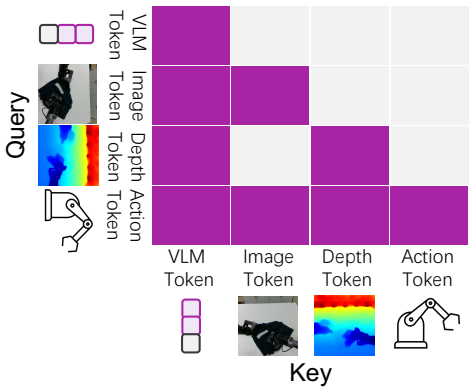


Figure 4: Example of the WorldScape Policy attention mask.

pose, represented by orientation and translation parameters. Memory retrieval is performed using a geometry-aware similarity score that combines directional alignment and spatial proximity between the current query pose and candidate memory poses. Only entries with high geometric relevance are attended during cross-token attention. To control redundancy, newly generated KV entries are evaluated against the global pool using a similarity-based gating criterion. Entries with high reconstructibility from existing memory are discarded, while novel viewpoints are inserted into the global pool subject to a fixed capacity constraint. When the memory budget is exceeded, a diversity-preserving pruning step is applied. A subset of entries is selected based on pose-space dispersion, encouraging coverage over both translation and rotation dimensions. This hierarchical memory structure allows the model to maintain spatial continuity across revisited regions while ensuring bounded cache growth during long autoregressive rollouts.

Adaptation to WorldScape Policy. WorldScape Policy is obtained by fine-tuning this foundation model under a joint objective that includes future state prediction and action prediction. The diffusion transformer backbone, interaction conditioning scheme, 3D spatial consistency training paradigm, and hierarchical memory mechanism remain unchanged. The policy therefore inherits the same generative dynamics and geometric inductive biases as the foundation model while extending it to world-action modeling.

4.2 Architecture

As shown in Figure 2, we resort to a unified framework which follows the MoT (Mixture-of-Transformers) architecture aiming to perform joint "Video-Depth-Action" diffusion training under a multi-task objective. Specifically, we first employ an off-the-shelf 3D foundation model (e.g., VGGT [35]) to infer the corresponding depth clues under the current observations. Then our WorldScape Policy converts the multimodal inputs into the latent space through a pretrained world VAE encoder and an additional action encoder, which are further processed with a sequence of World-Action DiT blocks to conduct joint interaction among different token streams conditioned on diverse conditions, including VLM tokens and textual embeddings. Besides, to equip the reasoning conditions with depth awareness, we inject the positional depth embeddings [36–38] into the VLM tokens through position-wise depth encoder and element-wise addition, as shown in Figure 3. Concurrently, the robot's proprioceptive sequences, specifically initial states and noisy action chunks, are fed into the action expert for action denoising. We employ Flow Matching [39] for both visual generation and continuous action modeling, which facilitates comprehensive robotic control, ensuring high-precision execution across complex tasks and diverse embodiments. The world-action joint attention layer follows the attention mask pattern as shown in Figure 4. Finally, the denoised multimodal latents are decoded into structural outputs for future state prediction and robotic planning.

To examine the versatility of our embodied world model as a robotic planner, our WorldScape Policy combines three Transformer-based experts for different usages through a cross-modal **World-Action Joint Attention** mechanism. For reasoning, we leverage the capabilities of off-the-shelf large language models (e.g., Qwen3-VL-2B [40] and PaliGemma-2B [41]). The WorldScape foundation model is built upon the Wan2.1-1.3B backbone, while the action expert utilizes a lightweight structure consisting of several DiT blocks.

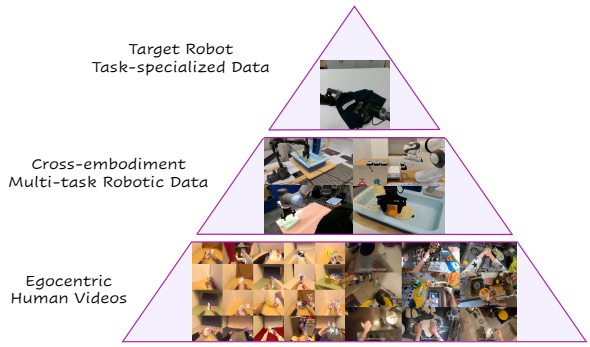


Figure 5: Illustration of the **Embodied Data Pyramid** formulated for **WorldScape Policy** training. It categorizes the diverse training corpus into three hierarchical levels, from Level 1 at the base to Level 3 at the top. While the data scale progressively decreases from bottom to top, the data quality and task relevance significantly increase.

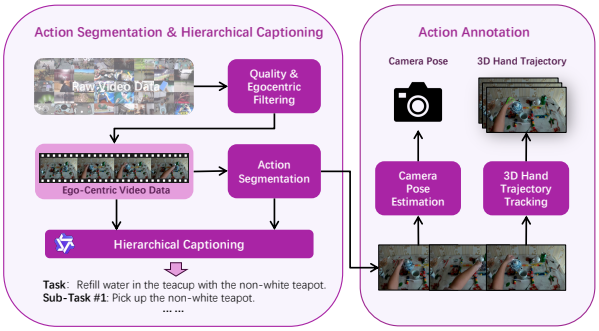


Figure 6: Illustration of the **Automated Curation** pipeline for **Interleaved** action human data. This pipeline automatically processes raw egocentric videos to produce high-quality structured data with hierarchical task descriptions (high-level task instructions and detailed step-by-step sub-task captions) alongside precise action trajectory annotations.

5 WorldScape Policy Data

To develop a robust WorldScape Policy model, we curate a large-scale dataset that encompasses diverse human-centric activities and multi-embodiment robot demonstrations with rich context annotations and action labels. Our approach aims to achieve both task and embodiment generalization by leveraging a heterogeneous data mixture from diverse sources covering a wide range of activities. The core of our strategy involves transforming raw and unstructured egocentric human videos from various source formats into a structured and unified format with detailed context and action annotations, while augmenting this with cross-embodiment robot data to ensure broad task coverage and target-specific precision. In total, 13.4K hours of human and cross-embodiment videos are collected and curated for WorldScape Policy model training.

5.1 Data Composition

As shown in Figure 5, we construct an embodied data pyramid for WorldScape Policy training, which includes three hierarchical levels of data with varying quantities and qualities tailored to different training stages. The pyramid structure reflects our training strategy: Level 1 (base) contains the largest volume of diverse egocentric human data for foundational world model pretraining; Level 2 (middle) incorporates cross-embodiment robot demonstrations for general action representation learning; Level 3 (top) consists of high-quality target-robot teleoperation data for embodiment-specific fine-tuning. Our data is sourced from three primary domains to ensure both task and platform diversity.

Egocentric Human Data. We utilize large-scale egocentric human video datasets[42–50], which provide a rich collection of tasks ranging from atomic skill executions to complex long-horizon activities, covering diverse everyday human behaviors. This data forms the foundation of our data pyramid (Level 1) and is critical for enabling our WorldScape model to learn world physics, object affordances, and achieve improved generalizability across a broad spectrum of manipulation tasks. The natural diversity in human demonstrations exposes the model to a wide variety of object interactions, environmental contexts, and manipulation strategies that are difficult to capture through robot-only data collection.

Cross-embodiment Robot Data. To cultivate general manipulation capabilities in our WorldScape Policy, we incorporate data from a wide variety of robotic platforms[51–53], forming Level 2 of our data pyramid. This cross-embodiment mixture enables the model to learn robot-agnostic features and transfer functional knowledge across different kinematic structures and morphologies. By training on diverse robot embodiments, the policy learns to abstract away embodiment-specific details and focus on the underlying manipulation principles that generalize across platforms.

Target-Robot Data. We collect high-quality tele-operated demonstrations specifically for our target robot platform,

constituting Level 3 of the data pyramid. This subset consists of diverse tasks ranging from pick-and-place operations to articulated object manipulation, ensuring that the model achieves high performance and precision on the specific hardware deployed. This embodiment-specific data is essential for fine-tuning the policy to the target robot's kinematic constraints, workspace limitations, and control characteristics.

5.2 Curation Pipeline for Interleaved Action Human Data

The primary challenge in utilizing human video for WorldScape model training is the lack of unified action labels and structured context annotations. We address this challenge through an automated curation pipeline that converts raw egocentric videos into structured data with action trajectory annotations and interleaved context captions, as shown in Figure 6.

Action segmentation & hierarchical context captioning. We first perform multi-stage filtering on the collected raw video data. Specifically, we apply quality filtering to remove low-quality videos (e.g., over-exposed, dark, or blurry frames) and data that does not conform to our target domain (such as vertical screen aspect ratios). Concurrently, we conduct egocentric filtering to retain only data from egocentric perspectives by prompting a Qwen3-VL model[54], thereby ensuring that only high-quality egocentric videos are retained to avoid noisy data.

Next, we employ Qwen3-VL[54] to generate comprehensive context captions for entire video sequences, capturing the overall task instruction and scene description at the video level. Subsequently, we segment the videos into short, atomic clips corresponding to single actions by performing temporal segmentation through thresholding on ORB keypoint matching ratios between consecutive frames. This segmentation strategy identifies action boundaries by detecting significant changes in visual features, ensuring that each clip represents a semantically coherent manipulation primitive. For each atomic clip, we again employ Qwen3-VL to generate specific sub-task instructions (e.g., "grasping the handle of the drawer"), providing the necessary semantic alignment for instruction-following and enabling hierarchical task decomposition.

Action annotation. To ensure robust WorldScape Policy action training on human video data, we provide reliable camera pose and action trajectory annotations for human videos to align the action space between egocentric human and robotic demonstrations. For camera extrinsic parameters, we adopt VIPE[55] to estimate camera extrinsic, which enables us to normalize the visual input and facilitate the translation of human observations into a robot-centric coordinate frame. For the extraction of action labels, we track hand keypoints and wrist poses in egocentric human videos[56, 57], and translate them to global coordinates using the previously acquired camera extrinsic. This coordinate transformation enables us to align action coordinates of human and robotic data into a unified system, thereby facilitating faster and more robust WorldScape Policy training.

6 WorldScape Policy Training

WorldScape Policy is trained in four structured stages to progressively integrate physical interaction priors from diverse datasets into a policy transferable to a target robot.

Stage1: World Model Pretraining.

The first stage trains the WorldScape foundation model to learn action-conditioned visual dynamics. Initialized from a pretrained video diffusion backbone, the model is optimized to predict future visual states under structured control inputs. Using video–trajectory pairs, it models viewpoint-dependent state transitions under locomotion signals in a closed-loop generative setting. To incorporate geometric structure into the learned dynamics, multi-view depth supervision is jointly optimized with the diffusion objective through a spatial consistency loss. The model is further adapted to manipulation signals via LoRA-based fine-tuning, integrating articulated pose conditioning into the same diffusion framework. The resulting WorldScape foundation model forms the visual dynamics foundation for subsequent policy adaptation.

Stage2: Unified World Model Action Pretraining.

The second stage aims to learn a generalized action representation conditioned on predictive visual dynamics. To translate physics-informed priors into effective robotic control, we jointly pretrain the entire WorldScape Policy on the curated video-action trajectory dataset. This includes annotated hand trajectories extracted from egocentric human

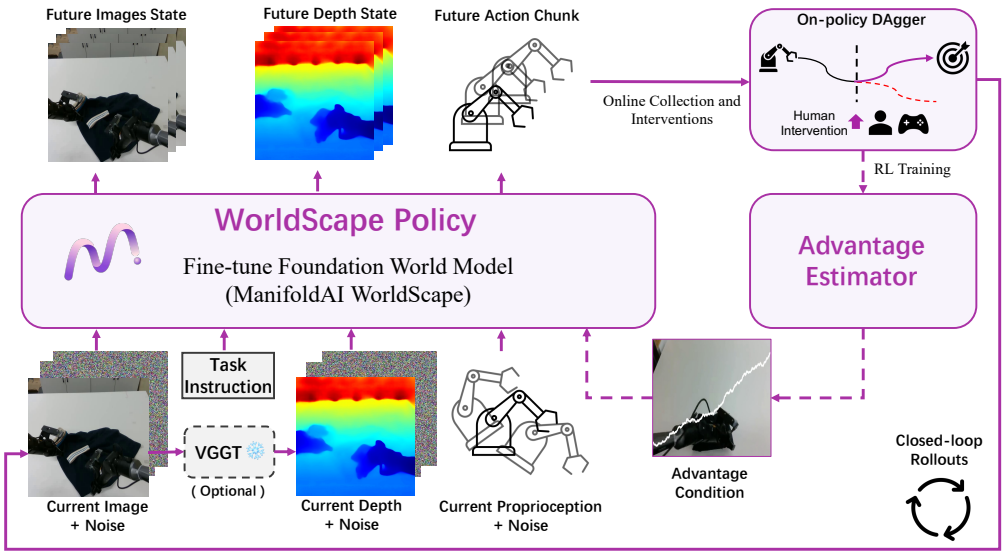


Figure 7: Illustration of Human-In-The-Loop experience learning process.

videos and cross-embodiment robotic demonstrations, establishing a robust alignment between visual state transitions and continuous action spaces.

Stage3: Unified World Model Action Finetuning.

This stage adapts the generalized action priors to the specialized kinematics and control frequencies of the target robot. Prior to on-device deployment, we fine-tune the WorldScape Policy exclusively on high-quality, task-specific teleoperation data collected from the target embodiment, ensuring precise execution and embodiment-specific morphological alignment.

Stage4: Advantage Conditioned Training.

To continually enhance on-policy performance and robustness against real-world distribution shifts, we introduce an additional post-training stage incorporating a Human-In-The-Loop (HITL) learning paradigm, as illustrated in Figure 7. This is achieved through iterative on-policy data collection via DAgger [58], coupled with real-time value prediction. This closed-loop interaction facilitates advantage-conditioned policy optimization, allowing the model to effectively correct execution errors and refine its long-horizon planning capabilities.

7 Experiments

We conduct extensive evaluation to assess the planning capability of our WorldScape Policy in real-world environments.

7.1 Baselines & Settings

We evaluate two mainstream policies used for robotic manipulation, namely VLA-based policy and video-based policy.

VLA-based Policy. We compare our method with several state-of-the-art Vision-Language-Action (VLA) models. $\pi_{0.5}$ [59] is a flow-matching based VLA model that directly outputs continuous actions. `X-VLA` [60] is an open-source VLA model built upon large vision-language models with soft prompts, demonstrating strong generalization capabilities. `GigaBrain-0` [61] is a large-scale VLA model trained on extensive robotic datasets for general-purpose manipulation.

Video-based Policy. We also compare with recent video-based policies that leverage video generation for robotic planning. `Motus` [21] is a representative video-based policy that learns to predict future video frames and actions simultaneously, providing strong spatial-temporal reasoning.

Table 1: Real-world experiment results of robotic manipulation tasks. Each experiment was conducted 10 times for success rate calculation. **Note that other comparing methods yield a 0% success rate under few-shot constraint across all evaluated tasks**.

Real-world Task	$\pi_{0.5}$	X-VLA	Motus	GigaBrain-0	Ours	Ours (20 shots)
Pick Shirt From Basket	80%	70%	40%	10%	90%	60%
Flatten Shirt	50%	20%	0%	0%	60%	30%
Fold Shirt	60%	40%	0%	20%	90%	50%
Get Water From Dispenser	60%	40%	40%	10%	60%	40%
Pour Water	60%	70%	10%	10%	90%	70%
Clean Up Trash On Table	60%	60%	50%	10%	70%	40%
Wipe Table Clean	50%	20%	20%	10%	60%	40%

7.2 Implementation Details

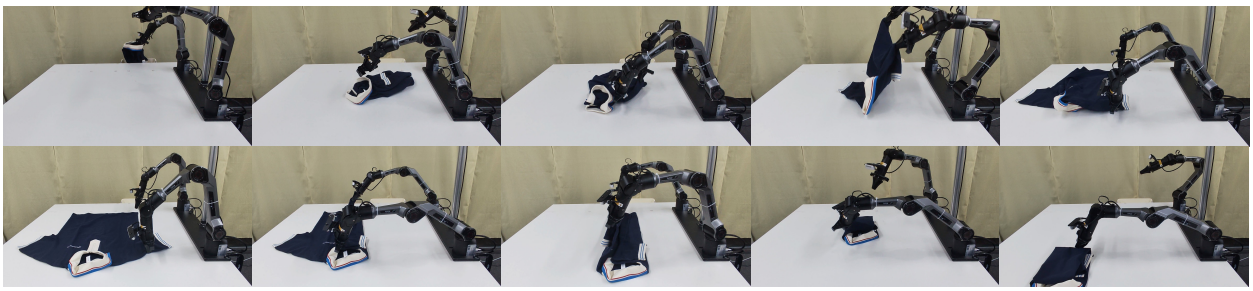
Training details for different stages. We adopt a multi-stage training strategy to progressively equip the model with world knowledge and action capabilities. In **Stage 1 (World Model Pretraining)**, we train the visual dynamics model using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 48×10. The model is trained on large-scale egocentric human videos to predict future frames conditioned on visual observations. In **Stage 2 (Unified World Model Action Pretraining)**, we incorporate the action expert and jointly train the video and action branches. We use a learning rate of 5×10^{-5} and a batch size of 32×10., training on a mixture of human video-action trajectories and cross-embodiment robot data. The loss weights for the video and action objectives are balanced to ensure stable co-training. In **Stage 3 (Unified World Model Action Finetuning)**, we fine-tune the entire policy on target-robot data. The learning rate is reduced to 1×10^{-5} with a batch size of 48. This stage focuses on adapting the general action priors to the specific kinematics and control frequencies of the target embodiment. In **Stage 4 (Advantage Conditioned Training)**, we perform on-policy fine-tuning using DAgger [58] and advantage-conditioned optimization. We collect online interaction data and update the policy with a learning rate of 5×10^{-6}, enabling the model to correct its own execution errors and improve robustness in real-world scenarios.

Action Representation. We adopt a combined formulation of joint angles and 6D end-effector poses for action representation. The policy systematically controls the robotic arm by predicted end-pose based action chunk.

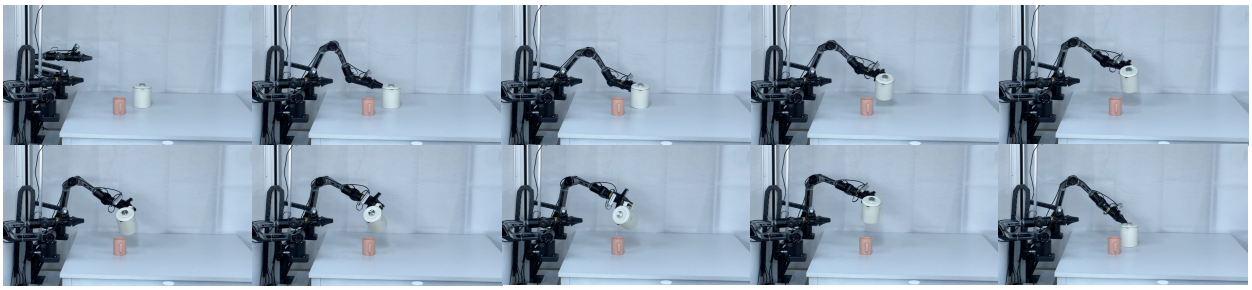
7.3 Evaluation in Real-World Environments

Real-world Task Setup. We evaluate WorldScape Policy on the dual-arm PIPER platform across 7 long-horizon dexterous manipulation tasks: *Pick Shirt From Basket*, *Flatten Shirt*, *Fold Shirt*, *Get Water From Dispenser*, *Pour Water*, *Clean Up Trash On Table*, and *Wipe Table Clean*. These tasks require complex spatial-temporal reasoning, precise bimanual coordination, and robust physical interaction in unstructured environments.

Main Results & Analysis. ❶ ***Seen Task Evaluation.*** Experimental results are shown in Table 1. As demonstrated, WorldScape Policy significantly outperforms existing VLA-based and video-based baselines across a variety of complex, long-horizon manipulation tasks. For instance, in contact-rich tasks such as *Flatten Shirt* and *Fold Shirt*, our method achieves a much higher success rate compared to $\pi_{0.5}$ and Motus. This superior performance is largely attributed to the versatile WorldScape foundation model and the unified MoT architecture, which effectively aligns predictive visual dynamics with continuous action generation, enabling precise and robust execution in the physical world. Figure 8 illustrates the real-world deployment of WorldScape Policy on the dual-arm PIPER platform, showcasing its capability to handle dexterous manipulation tasks. Furthermore, Figure 9 visualizes the joint generation of future video frames and corresponding actions during the `Folding Clothes` task, highlighting the model's strong spatial-temporal reasoning and planning abilities. ❷ ***Few-shot Evaluation.*** To evaluate the sample efficiency and generalization capability of our approach, we conduct few-shot adaptation experiments on unseen tasks as shown in Table 1. Thanks to the rich physical priors acquired during the world model pretraining and cross-embodiment co-training stages, WorldScape Policy can rapidly adapt to novel tasks with minimal target-domain demonstrations, significantly reducing the data collection burden for real-world deployment. Notably, previous approaches struggle to generalize to the few-shot regime,



(a) Flattening and folding clothes throughout the entire process



(b) Picking up the water kettle and pouring water into the cup

Figure 8: Real-world deployment of **WorldScape Policy** on PIPER arms for long-horizon dexterous task evaluation.

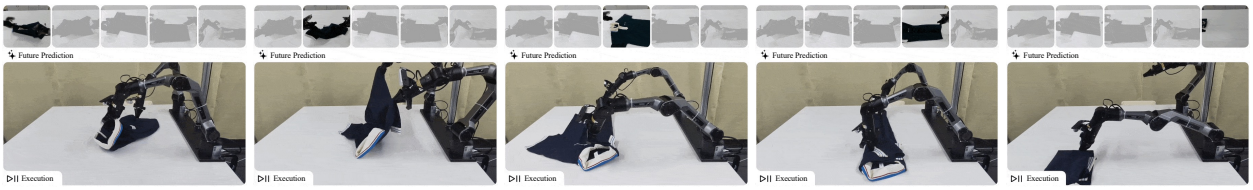


Figure 9: Illustration of both video and action generation results during the real-world execution process of `Folding Clothes` task.

resulting in a zero success rate across all evaluated tasks. ❸ ***Training Efficiency Evaluation.*** Benefiting from the embodied foundation model - WorldScape with general physics-informed priors, our WorldScape Policy demonstrates a significantly faster training convergence speed compared to previous methods. By comparing the required training epochs during the post-training stage, our method achieves a reasonably high task execution success rate with only a small number of fine-tuning steps (5K steps as default). In contrast, other baseline methods require substantially longer fine-tuning steps (50K steps as default) to achieve comparable success rates. This highlights the efficiency of our approach in adapting to new tasks and environments.

8 Conclusion and Future Work

In this paper, we present WorldScape Policy, a novel world-model-based robotic policy that adapts a pretrained video foundation model into a generalist action planner. By employing a unified Mixture-of-Transformers (MoT) architecture, our approach jointly models future visual trajectories, depth information, and continuous action chunks, effectively overcoming the compounding errors typical of two-stage inverse dynamics pipelines. To scale the learning process, we introduce an automated data curation pipeline that leverages large-scale egocentric human videos and cross-embodiment robotic demonstrations. Furthermore, our four-stage training recipe, culminating in advantage-conditioned human-in-the-loop post-training, ensures robust alignment between predictive dynamics and physical control. Extensive evaluations conducted on a real-world dual-arm robot demonstrate that WorldScape Policy achieves superior robustness against severe distribution shifts and exhibits strong generalization capabilities across diverse manipulation tasks.

Future Work. While WorldScape Policy demonstrates the immense potential of foundation world models for robotic control, several avenues remain for future exploration. First, scaling up the model parameters and the diversity of the pretraining data could further enhance zero-shot generalization to entirely unseen environments and object categories. Second, integrating more fine-grained tactile sensing and audio modalities into the unified diffusion framework could improve performance in highly delicate, contact-rich manipulation tasks. Finally, exploring more efficient inference techniques for the diffusion-based action generation process will be crucial for deploying these high-capacity models on resource-constrained edge devices with higher control frequencies.

References

[1] Yu Shang, Yinzhou Tang, Xin Zhang, Shengyuan Wang, Yuwei Yan, Honglin Zhang, Zhiheng Zheng, Jie Zhao, Jie Feng, Chen Gao, et al. A survey of embodied world models. 2025.

[2] Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025.

[3] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

[4] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

[5] Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Zhengkai Jiang, Yue Hu, Yue Liao, Peng Gao, Hongsheng Li, Maoqing Yao, et al. Enerverse: Envisioning embodied future space for robotics manipulation. *arXiv preprint arXiv:2501.01895*, 2025.

[6] Hongyan Zhi, Peihao Chen, Siyuan Zhou, Yubo Dong, Quanxi Wu, Lei Han, and Mingkui Tan. 3dflowaction: Learning cross-embodiment manipulation from 3d flow world model. *arXiv preprint arXiv:2506.06199*, 2025.

[7] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[8] 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.

[9] Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. Roboscape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

[10] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[11] 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.

[12] Yu Shang, Yangcheng Yu, Xin Zhang, Xin Jin, Haisheng Su, Wei Wu, and Yong Li. Mowm: Mixture-of-world-models for embodied planning via latent-to-pixel feature modulation. *arXiv preprint arXiv:2509.21797*, 2025.

[13] 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.

[14] 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.

[15] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

[16] 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.

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

[18] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[19] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

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

[21] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

[22] Yucheng Hu, Jianke Zhang, Yuanfei Luo, Yanjiang Guo, Xiaoyu Chen, Xinshu Sun, Kun Feng, Qingzhou Lu, Sheng Chen, Yangang Zhang, et al. Bagelvla: Enhancing long-horizon manipulation via interleaved vision-language-action generation. *arXiv preprint arXiv:2602.09849*, 2026.

[23] Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025.

[24] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. Simplevla-rl: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025.

[25] Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15665–15672. IEEE, 2025.

[26] Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. *arXiv preprint arXiv:2506.15799*, 2025.

[27] Charles Xu, Qiyang Li, Jianlan Luo, and Sergey Levine. Rldg: Robotic generalist policy distillation via reinforcement learning. *arXiv preprint arXiv:2412.09858*, 2024.

[28] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[29] Dongchi Huang, Zhirui Fang, Tianle Zhang, Yihang Li, Lin Zhao, and Chunhe Xia. Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning. *arXiv preprint arXiv:2508.02219*, 2025.

[30] Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. *arXiv preprint arXiv:2411.19309*, 2024.

[31] Seyed Kamyar Seyed Ghasemipour, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, and Igor Mordatch. Self-improving embodied foundation models. *arXiv preprint arXiv:2509.15155*, 2025.

[32] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[33] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, et al. $\pi_{0.6}$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.

[34] Checheng Yu, Chonghao Sima, Gangcheng Jiang, Hai Zhang, Haoguang Mai, Hongyang Li, Huijie Wang, Jin Chen, Kaiyang Wu, Li Chen, et al. χ_0: Resource-aware robust manipulation via taming distributional inconsistencies. *arXiv preprint arXiv:2602.09021*, 2026.

[35] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[36] Haisheng Su, Junjie Zhang, Feixiang Song, Sanping Zhou, Wei Wu, Junchi Yan, and Nanning Zheng. Freqpde: Rethinking positional depth embedding for multi-view 3d object detection transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28145–28155, 2025.

[37] Haisheng Su, Wei Wu, Feixiang Song, Junjie Zhang, Zhenjie Yang, and Junchi Yan. Drivemamba: Task-centric scalable state space model for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2602.13301*, 2026.

[38] Changyong Shu, Jiajun Deng, Fisher Yu, and Yifan Liu. 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3589, 2023.

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

[40] 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.

[41] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[42] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[43] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

[44] Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation, 2025. URL https://arxiv.org/abs/2503.11423.

[45] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[46] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20270–20281, October 2023.

[47] Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024.

[48] Build AI. Egocentric-10k. 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-10K.

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[50] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

[51] AgiBot World Colosseum contributors. Agibot world colosseum. https://github.com/OpenDriveLab/AgiBot-World, 2024.

[52] 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. In *Robotics: Science and Systems (RSS) 2025*. Robotics: Science and Systems Foundation, 2025. URL https://www.roboticsproceedings.org/rss21/p152.pdf.

[53] InternData-A1 contributors. Interndata-a1. https://github.com/InternRobotics/InternManip, 2025.

[54] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[55] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025.

[56] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[57] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025.

[58] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083. IEEE, 2019.

[59] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. 2025. URL https://arxiv.org/abs/2504.16054.

[60] 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. *arXiv preprint arXiv:2510.10274*, 2025.

[61] GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025.