Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Unified World Models: Coupling Video and Action Diffusion for Pretraining on Large Robotic Datasets

Chuning Zhu[†], Raymond Yu[†], Siyuan Feng[‡], Benjamin Burchfiel[‡], Paarth Shah[‡], and Abhishek Gupta[†]
[†] Paul G. Allen School of Computer Science and Engineering, University of Washington
[‡]Toyota Research Institute

***Abstract*—Imitation learning has emerged as a promising approach towards building generalist robots. However, scaling imitation learning for large robot foundation models remains challenging due to its reliance on high-quality expert demonstrations. Meanwhile, large amounts of video data depicting a wide range of environments and diverse behaviors are readily available. This data provides a rich source of information about real-world dynamics and agent-environment interactions. Leveraging this data directly for imitation learning, however, has proven difficult due to the lack of action annotation required for most contemporary methods. In this work, we present Unified World Models (UWM), a framework that allows for leveraging both video and action data for policy learning. Specifically, a UWM integrates an action diffusion process and a video diffusion process within a unified transformer architecture, where *independent* diffusion timesteps govern each modality. By simply controlling each diffusion timestep, UWM can flexibly represent a policy, a forward dynamics, an inverse dynamics, and a video generator. Through simulated and real-world experiments, we show that: (1) UWM enables effective pretraining on large-scale multitask robot datasets with both dynamics and action predictions, resulting in more generalizable and robust policies than imitation learning, (2) UWM naturally facilitates learning from action-free video data through independent control of modality-specific diffusion timesteps, further improving the performance of finetuned policies. Our results suggest that UWM offers a promising step toward harnessing large, heterogeneous datasets for scalable robot learning, and provides a simple unification between the often disparate paradigms of imitation learning and world modeling. Videos and code are available at https://weirdlabuw.github.io/uwm/**

I. Introduction

Imitation learning provides a simple and scalable way to imbue autonomous robots with complex behaviors using human demonstrations [6, 26, 11, 49]. Imitation learning via supervised learning, often referred to as behavior cloning (BC), has shown remarkable success due to the advent of powerful multimodal generative models such as diffusion [22] or flow-based models [28]. With these methods, acquiring new behaviors amounts to collecting demonstrations and fitting a generative model to the action distributions given observations using a maximum likelihood (or closely related) objective. These models have shown robust and reliable behavior within the training distribution, but can be brittle when tasked beyond this distribution. A natural solution to synthesizing robust and generalizable controllers with imitation learning is to scale up the number of high-quality, on-robot demonstrations collected through robotic teleoperation. While achievable with considerable resources, this data scaling process is expensive and time-consuming. A natural question arises - are prevalent methodologies for imitation learning making maximal use of the available large-scale datasets?

While imitation learning methods learn a mapping from states to optimal actions, they do not explicitly capture temporal dynamics that are naturally present in demonstration trajectories or videos. An alternative paradigm to direct imitation learning that can leverage such dynamics information is that of world modeling; learning approximate models of how the world changes over time. Commonly instantiated as predicting the future observations given current observations (and actions), world models can naturally be trained from large scale robotic datasets [35, 46], but also from alternative sources of data such as uncurated "play" data [13] or even action-free data such as videos. A variety of world modeling techniques, such as video diffusion models [23] or latent state-space modeling [32], have shown impressive results on realistic generation of future frames. However, it is not yet clear how the ability of these world models to capture temporal dynamics can be brought to bear on improving the robustness and generalization of robotic controllers synthesized via imitation learning.

In this work, we propose a new diffusion-based learning framework that unifies imitation learning and world modeling, incorporating knowledge of temporal dynamics gleaned from large robotic datasets into imitation learning policies. Our key insight is to integrate an action diffusion process and an image diffusion process into a single diffusion transformer model conditioned on *independent diffusion timesteps*. Leveraging a connection between diffusion noise at different timesteps of the forward diffusion noising process and partial masking, this allows for flexible sampling from a number of distributions simply by manipulating the diffusion timesteps independently at inference time. For example, to draw a sample from the policy, one can "mask out" the image diffusion process by fixing the diffusion timestep of image denoising to T, thereby marginalizing it. Similarly, one can draw a sample from the forward dynamics model by fixing the action diffusion timestep to 0, inferring next observations given current observations and "clean" actions. The same is also true of inverse dynamics models and unconditional video prediction models into the future. This yields a simple, unified diffusion model that can serve as a policy, dynamics model, video predictor or inverse model depending on the use case (Fig 1).

Concretely, a UWM consists of a coupled score model that predicts action scores and *future* image scores, con-

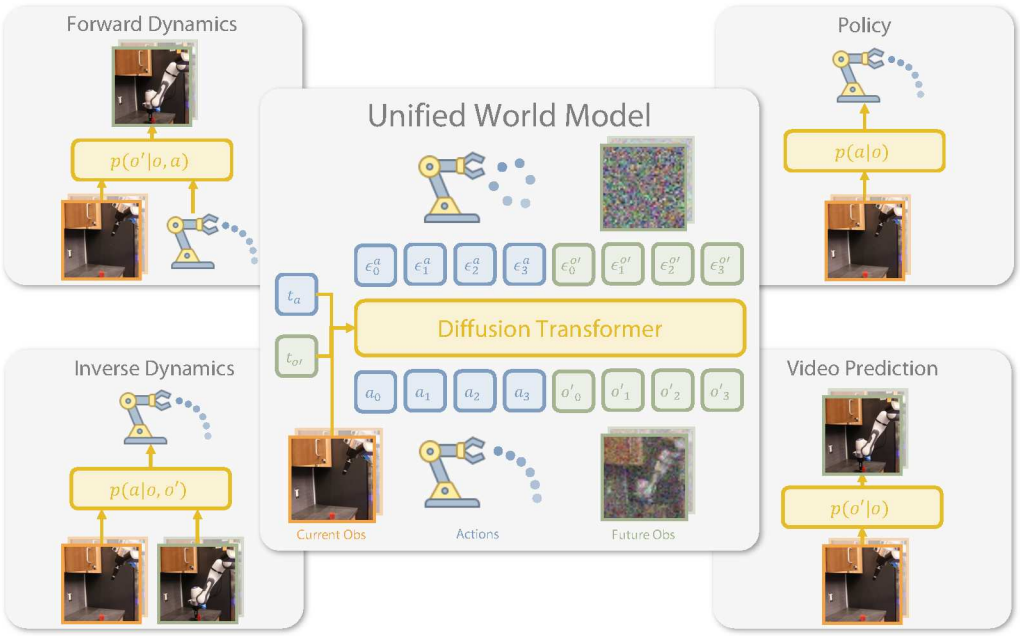

Fig. 1. Unified World Models integrates action and video diffusion in a unified transformer architecture controlled by modality-specific diffusion timesteps. The model can flexibly be trained on large robotics datasets and then flexibly perform a variety of different inferences at test time. Doing so naturally enables improved robustness and generalization for imitation learning.

ditioned on the current image and corresponding diffusion timesteps that are separate for next-observation and action. During training, the timesteps are sampled independently at random, exposing the model to different combinations of image and action noises. During inference, UWMs enable flexible sampling from various distributions by manipulating the diffusion timesteps independently. In particular, a UWM can generate samples from (1) forward dynamics $p(o'|o,a)$, (2) inverse dynamics $p(a|o,o')$ (3) marginal action distribution (policy) $p(a|o)$, (4) marginal image distribution (video generative model) $p(o'|o)$. We show that this learning framework leads to improved policies compared to standard imitation learning since, (1) the unified architecture enables feature sharing between action and pixels, resulting in additional supervision from the same data, (2) the model captures all combinations of marginal and conditional distributions, acquiring an understanding of the causal relationship between actions and images, (3) the model can learn from broader data modalities such as action-free videos.

We demonstrate the effectiveness of UWM through a set of experiments across both simulation and real-world robotic manipulation tasks. We demonstrate that UWM is capable of extracting knowledge from multitask robotic datasets, and further leveraging action-free video data to improve its generalization to out-of-distribution conditions. These models are able to flexibly perform a variety of test-time inference, while retaining strong performance of both policy and dynamics prediction. We show that while conceptually simple, a number of careful design decisions must be made to enable strong performance of the UWM model. Through this investigation of UWM, we take a step towards bridging the gap between policies and world models for robot learning.

II. Preliminaries

Unified world models build on the framework of denoising diffusion models [22], and their application to problems in robotic control [11].

A. Denoising Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [22] are a family of generative models that define a forward noising process and a learned reverse denoising process to generate samples from a complex, multimodal data distribution. Let $p(x_0)$ denote the data distribution from which a number of samples are available. In the *forward* diffusion process, the data $x_0 \sim p(x_0)$ is gradually corrupted by iteratively adding Gaussian noise over T steps through a Markov chain according to a variance schedule $\{\beta_t\}_{t=1}^T$. Concretely, the forward process is defined as

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t \mid \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t \mathbf{I}\right), \quad t = 1, \ldots, T.$$

After T steps, x_T is nearly an isotropic Gaussian [22]. The corresponding *reverse* process aims to map x_T back to a clean sample x_0 from the data distribution by iteratively denoising. While the exact reverse conditional $q(x_{t-1} \mid x_t)$ is generally intractable, one learns a parametrized approximation,

$$p(x_{t-1} \mid x_t) = \mathcal{N}\big(\mu(x_t, t),\, \Sigma(x_t, t)\big),$$

In practical settings, the variance $\Sigma(x_t, t)$ is set to a simple time varying constant $\sigma_t^2 I$. As shown in prior work [3], the optimal mean under MLE is:

$$\mu(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mathbb{E}\left[\epsilon_x | x_t\right]\right)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and ϵ_x is the Gaussian noise injected into x_t. To approximate this conditional expectation $\mathbb{E}[\epsilon_x | x_t]$, DDPMs are trained using a variant of *denoising score matching*, which approximates this using a neural network s_θ

$$\min_\theta \mathbb{E}_{x_0,\, t,\, \epsilon}\Big[\| s_\theta(x_t,\, t) - \epsilon \|_2^2\Big],$$

where $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ and $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. Intuitively, this "score function" predicts the noise added at each step using a simple regression objective. This learned noise prediction network $s_\theta(x_t,\, t)$ can then directly parameterize the reverse diffusion process as

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} s_\theta(x_t,\, t)), \sigma_t^2 I\big)$$

Given this reverse diffusion model, samples can approximately be drawn from the data distribution using a simple "denoising procedure". Starting with a sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ drawn from Gaussian noise, new samples are iteratively drawn from $p_\theta(x_{t-1} \mid x_t)$ until a "clean sample" x_0 is obtained. This procedure allows for the representation of complex multimodal distributions where performing MLE tractably is challenging. Extensions have been applied to robotic control as well [11].

a) Conditional Generation with Diffusion Models: While the above-mentioned generative modeling process is unconditional, diffusion models are naturally extended to conditional settings. Consider a setting where multivariate data $(x_0, z_0) \sim p(x_0, z_0)$ is available, and the conditional distribution $p(x_0|z_0)$ must be modeled. In these settings, a bulk of the machinery from above can be reused, simply with an additional conditioning variable. The forward process remains identical, while the reverse process is modified as

$$p(x_{t-1} \mid x_t, z_0) = \mathcal{N}\big(\frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \mathbb{E}[\epsilon_x | x_t, z_0]), \sigma_t^2 I\big)$$

In this case, the expectation $\mathbb{E}[\epsilon_x | x_t, z_0]$ can be approximated using a *conditional* noise prediction network that is also trained with denoising score matching:

$$\min_\theta \mathbb{E}_{(x_0,z_0)\sim p(x_0,z_0)\, t,\, \epsilon}\Big[\| s_\theta(x_t, z_0,\, t) - \epsilon \|_2^2\Big],$$

III. Method

In this section, we introduce Unified World Models as a way to incorporate temporal dynamics into diffusion-based action prediction models, proving a bridge between the often disparate worlds of imitation learning and world modeling.

A. Problem Setup

We build on typical sequential decision making settings, assuming access to a dataset of (observation, action, next-observation) pairs $\mathcal{D}_{\text{e}} = \{(o_i, a_i, o_i')\}_{i=1}^N$ provided by an expert demonstrator. For the sake of exposition, we will assume that the environment is Markovian in observations o. In addition to this action-labeled dataset, we may also be provided an action-free dataset $\mathcal{D}_{\text{af}} = \{(o_i, o_{i+1})\}_{i=1}^M$. The question becomes - how can we extract the most learning signal out of these datasets for synthesizing robot controllers?

In this context, several different models may be desired - (1) a policy $p(a|o)$ (often referred to as $\pi(a|o)$), that samples optimal actions to execute at a particular observation, (2) a dynamics model $p(o'|o, a)$, that samples future observations, given a current observation and action, (3) an inverse model, $p(a|o, o')$ that predicts what distribution of actions can transition between a current observation and a desired next observation, (4) a video-prediction model $p(o'|o)$ that predicts marginal future observations given current ones. While these models have each seen use in different contexts, they are largely considered to be disparate fields of study. In this work, we show that these are many sides of the same dice; they can be unified into a single model to benefit each other.

B. Unified World Models via Coupled Video-Action Diffusion

The core idea of a UWM is to develop a single diffusion model that can be trained on samples from the joint distribution of data $p(o, a, o')$ and used to flexibly perform inference for the policy $p(a|o)$, the dynamics model $p(o'|o, a)$, the inverse model $p(a|o, o')$ and a video prediction model $p(o'|o)$, with simple modifications to test-time inference.

A unified world model instantiates a joint diffusion model that integrates next observation o' and action prediction a into a single diffusion model conditioned on current observation o. This can naturally be done by parameterizing a joint noise prediction network $s_\theta(o, a_t, o_t',\, t)$ that approximates a conditional expectation over both action and next observation noise $\mathbb{E}[\epsilon_a, \epsilon_{o'} | o, a_t, o_t',]$, with $\epsilon_a, \epsilon_{o'}$ referring to noise on actions and next observations, and t referring to the *coupled* timestep of the joint diffusion process.[1] However, training such a joint noise prediction network [20] $s_\theta(o, a_t, o_t',\, t)$ does not accomplish flexible inference since it can only sample from the joint distribution of (a, o').

For flexible inference, we can leverage a connection between diffusion time-steps and masking - *noising input tokens by setting the inference timestep for diffusion appropriately can naturally induce a form of partial masking.* Timesteps closer to T (fully noised) indicate full masking, while timesteps closer to 0 (unnoised) indicate no masking. Based on this key insight, UWM modifies the joint diffusion process mentioned above and *decouples* the timesteps between that of the diffusion processes of next observation prediction $t_{o'}$ and that of action prediction t_a in a joint noise prediction network s_θ. This separation of time steps allows for independent control of $t_{o'}$ and t_a during training and inference, which gives rise to flexible inference capabilities.

A UWM models a coupled noise prediction network $s_\theta(o, a_{t_a}, o'_{t_{o'}}, t_a, t_{o'})$ that approximates a conditional expectation over noise $\mathbb{E}\left[\epsilon_a, \epsilon_{o'} | o, a_{t_a}, o'_{t_{o'}}\right]$, with $\epsilon_a, \epsilon_{o'}$ referring to

[1] It is important to note that throughout this section o will refer to the current observation, o_t' will refer to timestep t in the diffusion process for the **next** observation, and a_t will refer to timestep t in the diffusion process for action.

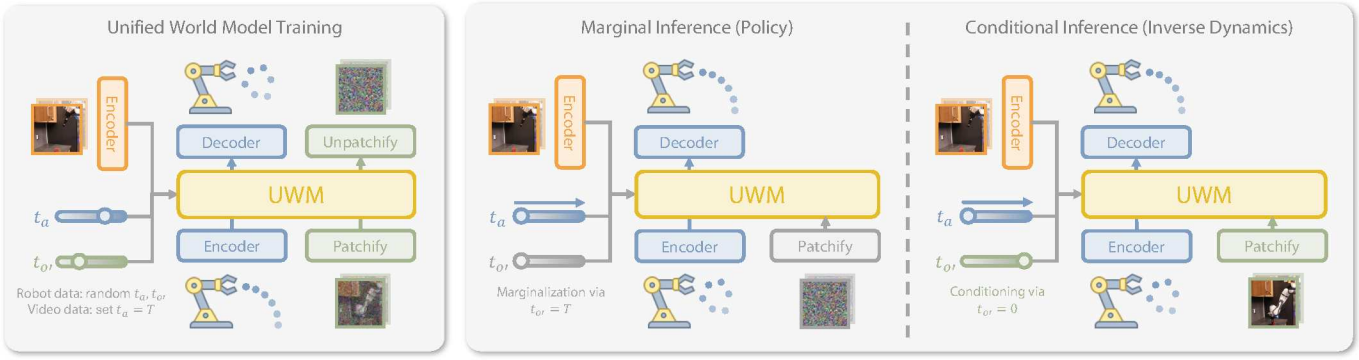

Fig. 2. Unified World Model training and policy inference pipeline. The left panel shows UWM pretraining on robot trajectories with actions and co-training on action-free videos by masking out actions using diffusion timesteps. The right panel illustrates marginal and conditional inference modes, corresponding to the policy and the inverse dynamics.

noise on next observations and actions, and t_a and $t_{o'}$ referring to the *decoupled* steps of the diffusion process with respect to actions and next-observations respectively. The ability to set diffusion timesteps independently allows for marginalization and conditioning of different variables. Fixing the timestep for either t_a or $t_{o'}$ to T marginalizes the corresponding variable a, o', while setting the timestep to 0 performs conditioning. By setting timestep $t_{o'} = T$, the joint model is approximating the expectation $\mathbb{E}\left[\epsilon_a, \epsilon_{o'}, | o, a_{t_a}, o'_T\right]$. Since o'_T is approximately an isotropic Gaussian, this reduces to $\mathbb{E}\left[\epsilon_a | o, a_{t_a}\right]$, which represents a policy $p(a|o)$, thereby performing marginalization. Similarly, setting the timestep $t_{o'} = 0$ reduces the approximated distribution to $\mathbb{E}\left[\epsilon_a | o, a_{t_a}, o'\right]$ which corresponds to an inverse model $p(a|o, o')$, thereby performing conditioning. Simply setting combinations of t_a and $t_{o'}$ allows for flexible inference of policies, dynamics models, inverse models, and video prediction from the same model!

This suggests a recipe for unified world modeling using a simple modification to the standard denoising objective [22]. To train a joint noise prediction diffusion model $(\epsilon^\theta_a, \epsilon^\theta_{o'}) = s_\theta(o'_{t_{o'}}, a_{t_a}, o, t_a, t_{o'})$, we independently sample action timestep t_a and next observation timestep $t_{o'}$, draw noisy action and next-observation samples from their respective distributions, and train the coupled conditional score model conditioned on the current observation with a standard denoising objective across both next-observations and actions:

$$\ell(\theta) = \mathbb{E}_{\substack{(o,a,o')\sim\mathcal{D} \\ t_a, t_{o'}\sim\mathcal{U}(0,T) \\ \epsilon_a, \epsilon_{o'}\sim\mathcal{N}(0,1)}} \left[w_a \parallel \epsilon^\theta_a - \epsilon_a \parallel^2_2 + w_{o'} \parallel \epsilon^\theta_{o'} - \epsilon_{o'} \parallel^2_2 \right], \tag{1}$$

$$\begin{aligned} \textit{where} \quad \epsilon^\theta_a, \epsilon^\theta_{o'} &= s_\theta(o, a_{t_a}, o'_{t_{o'}}, t_a, t_{o'}), \\ a_{t_a} &= \sqrt{\bar{\alpha}_{t_a}} a + \sqrt{1 - \bar{\alpha}_{t_a}} \epsilon_a, \\ o'_{t_{o'}} &= \sqrt{\bar{\alpha}_{t_{o'}}} o' + \sqrt{1 - \bar{\alpha}_{t_{o'}}} \epsilon_{o'}. \end{aligned}$$

where w_a and $w_{o'}$ are weights chosen to trade off between the action prediction and next-observation prediction objectives. Intuitively, this training paradigm exposes the model to all combinations of noise levels of the modalities, which enables flexible sampling at inference. We can use the trained model to flexibly draw samples from various distributions by simply controlling the time-steps t_a and $t_{o'}$ during inference, as follows:

1) **Policy** To sample from the policy $p(a|o)$, we must *marginalize* out the next observation o'. To do so, we can simply set $t_{o'} = T$ and perform denoising steps to synthesize an action a conditioned on current observation o. We perform the reverse diffusion process on actions going from $t_a = T, \ldots, 1$, with frozen next observation timestep (at full masking) $t_{o'} = T$, and $a_T \sim \mathcal{N}(0, I)$, $o'_T \sim \mathcal{N}(0, I)$:

$$\begin{aligned} a_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(a_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} s_\theta(o, a_t, o'_T, t, T) \right) \\ + \sigma_t \delta_t, \qquad \delta_t \sim \mathcal{N}(0, I) \end{aligned} \tag{2}$$

While this is simply performing action diffusion, it benefits from the temporal dynamics grounding obtained from the other modes.

2) **Video Prediction Model** To sample from the video prediction model $p(o'|o)$, we must *marginalize* out the action a. We can simply set $t_a = T$ and perform denoising steps to synthesize the next observation o' conditioned on current observation o. We perform the reverse diffusion process on next observations going from $t_{o'} = T, \ldots, 1$, with frozen action timestep (at full masking) $t_{o'} = T$, and $a_T \sim \mathcal{N}(0, I)$, $o'_T \sim \mathcal{N}(0, I)$:

$$\begin{aligned} o'_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(o'_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} s_\theta(o, a_T, o'_t, T, t) \right) \\ + \sigma_t \delta_t, \qquad \delta_t \sim \mathcal{N}(0, I) \end{aligned} \tag{3}$$

3) **Forward Dynamics** In order to use the UWM as a forward dynamics model $p(o'|o, a)$, we must *condition* on a particular executed action a. To do so, we can simply set $t_a = 0$ and denoise in order to synthesize the next observation o' conditioned on current observation o and the particular action a. We perform the reverse diffusion process on next observations going from $t_{o'} = T, \ldots, 1$,

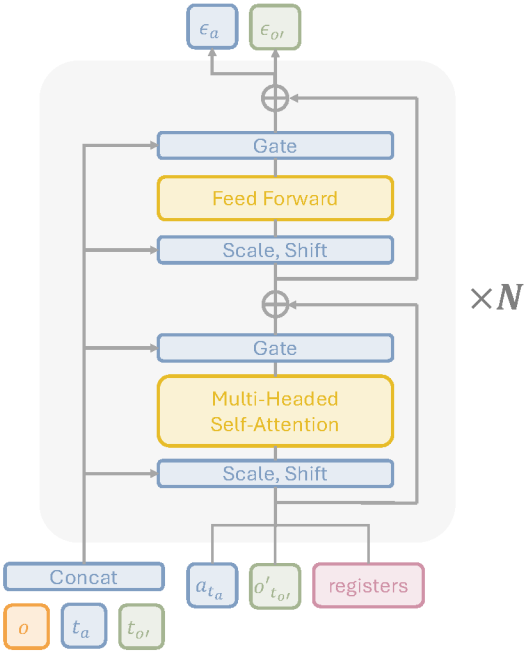

Fig. 3. A single Unified World Model (UWM) block consists of a transformer block with observations and diffusion timesteps conditioning via adaptive layer norm. In addition, we add randomly initialized register tokens which allows for better multi-modal feature sharing.

with frozen action timestep (at no masking) $t_a = 0$, and $o'_T \sim \mathcal{N}(0, I)$:

$$o'_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(o'_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} s_\theta(o, a, o'_t, 0, t)\right) + \sigma_t \delta_t, \qquad \delta_t \sim \mathcal{N}(0, I) \tag{4}$$

4) **Inverse Dynamics** Lastly, to sample from the inverse dynamics model $p(a|o, o')$, we must *condition* on a particular next observation o'. To do so, we can simply set $t_{o'} = 0$ and perform denoising steps to synthesize the action a conditioned on current observation o and the particular next observation o'. Specifically, we perform the reverse diffusion process on next observations going from $t_a = T, \ldots, 1$, with frozen next observation timestep (at no masking) $t_{o'} = 0$, and $a_T \sim \mathcal{N}(0, I)$:

$$a_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(a_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} s_\theta(o, a_t, o', t, 0)\right) + \sigma_t \delta_t, \qquad \delta_t \sim \mathcal{N}(0, I) \tag{5}$$

This simple modification to the standard diffusion training paradigm allows a single model to be trained, benefiting from feature sharing between different models of action and future observation prediction. This model can then be flexibly used for inference with just a simple of choice of time-step, making it a versatile, general-purpose decision-making model.

C. Architecture

We model the UWM as a diffusion transformer shown in Fig. 2 and 3. The model predicts actions and observation noises ϵ_a and $\epsilon_{o'}$ given current observations o, noisy actions a_{t_a}, noisy observations $o'_{t_{o'}}$, action timestep t_a, and observation timestep t_o. The actions are action chunks of length h_a. The current observations o and next observations o' are frame-stacked observations of length h_o from n_c camera views.

To condition the model on current observations, we encode each frame from each camera view using a ResNet-18 [21] encoder to obtain an n_{embd} dimension feature. the features are concatenated to form an embedding of size $n_c \cdot h_o \cdot n_{\text{embd}}$. The diffusion timesteps are encoded using a shared sinusoidal timestep encoder from [22], resulting in two timestep embeddings. These timestep embeddings are concatenated with the image features, and the combined features are used to condition the transformer via Adaptive Layer Normalization (AdaLN) [33].

The context of the diffusion transformer consists of action embeddings and image embeddings. The action embeddings are obtained by encoding the action chunk per-timestep using a shallow MLP. For image diffusion, we adopt the latent diffusion paradigm [36] and encode full-size $(224, 224, 3)$ images into $(28, 28, 4)$ latent images using a frozen SDXL VAE [34]. We then patchify the latent images using a spatiotemporal patchifier of size $(4, 4, 2)$. These image patch embeddings are then concatenated with the action embeddings and passed into the transformer backbone. The image noising and denoising processes are performed in the latent space, and the final image sample is decoded using the same VAE to generate full-size images.

Empirically, we found that adding redundant tokens that are eventually discarded (i.e. registers [14]) helps with model performance. We hypothesize that this is because images and actions are distinct modalities that can benefit from having an intermediary medium to exchange information. However, since all output embeddings of the diffusion transformer are meaningful noise predictions, there is no room for such communication. The registers can store information from either modality, which can then be retrieved in subsequent transformer layers. We demonstrate the effectiveness of registers in our ablation experiments in Section IV-D.

D. Training Paradigms

In this work, we evaluate the effectiveness of UWM as a pretraining method for learning the dynamics information from large multitask robotic datasets. To train a UWM, we sample sequences of observations and actions from the dataset, construct (o, a, o') tuples, sample random diffusion timesteps $t_a, t_{o'} \sim \mathcal{U}(0, T)$ and optimize the denoising score matching objective in Eq. 1.

Moreover, UWM naturally enables co-training on action-free video data by using diffusion timesteps for masking. Given action-free video samples, instead of sampling the action timestep randomly, we fix the action timestep to T and fill in the missing actions with random noise $\epsilon_a \in \mathcal{N}(0, 1)$ and optimize the same loss in Eq. 1. We validate the effectiveness of co-training on videos in our experiments in Section. IV-B.

Fig. 4. Visualization of datasets used for pretraining and finetuning. The pretraining/cotraining dataset consists of diverse tasks performed by Franka robots in various environments to ensure broad generalization capabilities. The finetuning datasets include five tasks, each designed to evaluate task-specific performance under controlled conditions.

IV. Experiments

In this section, we examine the following research questions: (1) can UWM effectively learn from large robotic datasets as a pretraining paradigm? (2) can UWM further benefit from additional video data without action labels in a co-training paradigm? (3) what are the key design choices that contribute to UWM's performance. We answer these questions through a number of real robot experiments with a Franka robot using the DROID [25] manipulation platform, as well as simulated experiments in the LIBERO [29] benchmark.

A. Baselines

We compare UWM to the following baselines throughout our experiments. Detailed descriptions of each baseline are deferred to Appendix A.

1) **Diffusion Policy (DP)** [11] is a behavior cloning method that fits a conditional diffusion model to a dataset of expert observation-action data. While the original framework is evaluated in single-task settings, we extend it to the pretraining-finetuning setting by fitting a model to the behavior distribution of a multitask dataset and then finetuning it to the downstream task demonstrations. We compare to DP as a baseline to validate the effectiveness of the additional supervisory signals in UWM. To minimize the discrepancy from UWM, we adopt a diffusion transformer backbone similar to [15] instead of the original UNet architecture [10]. Unlike the encoder-decoder architecture used in [15], however, we only employ the decoder side of the architecture.
2) **PAD** [19] is a joint video-action diffusion model that learns a joint distribution of actions and future observations conditioned on current observations. The key conceptual difference between PAD and UWM is the decoupling of timesteps between actions and next-observations. In addition, PAD conditions the model on the current observations by concatenating the clean latents of the current observations to the noisy latents of the next observations along the channel dimension, similar to Stable Video Diffusion [7]. This is different from the AdaLN condition in UWM. PAD supports co-training on videos by masking the action tokens with a learned mask token.
3) **GR1** [45] is a video-action transformer model that predicts actions and future image observations conditioned on current image observations. Contrary to other baselines, GR1 does not model a distribution over data using a diffusion generation process. Instead, it directly regresses the actions and images by minimizing a least squares loss. We compare to GR1 to validate the effectiveness of diffusion as a pretraining objective relative to supervised regression. GR1 supports co-training on videos by masking the action tokens with a learned mask token.

B. Real Robot Experiments

1) Setup: To evaluate UWM and baselines as pretraining methods, we leverage the DROID dataset [25] as a source of pretraining data. The DROID dataset is a diverse dataset consisting of robot trajectories collected across various institutions and operators, covering a large variety of tasks, camera positions and backgrounds in natural settings. We curate a pretraining dataset by sampling a subset of 2000 trajectories from the DROID dataset based on location (Fig 4, top row). For methods that support co-training on videos (e.g. GR-1, PAD, and UWM), we want to additionally evaluate their capability of learning from action-free videos. To this end, we curate another 2000 trajectories from the rest of the dataset to use as videos (Fig 4, middle row).

To evaluate the efficacy of the pretrained models, we construct five different real-world tasks (shown in Fig 5) using the portable manipulation platform proposed in DROID [25]. The tasks involve different kinds of robotic manipulation:

- **Stack-Bowls** aims to train robotic controllers to place the pink bowl into the blue bowl across various positions.
- **Block-Cabinet** aims to open the cabinet and grasp a small red block from the table to place it in the cabinet.
- **Paper-Towel** involves precisely grasping a paper towel from a cabinet and placing it upright on a wooden stand on the table.
- **Hang-Towel (deformable object)** involves grasping a towel by the corner and hanging it on a hook attached to the cabinet.
- **Rice-Cooker (long horizon)** involves pouring a cup of rice into the inner bin of a rice cooker, and placing the inner bin on the rice cooker.

Each of these tasks involves positional and visual generalization, and requires reasonably precise robotic manipulation. We curate the finetuning datasets by teleoperating the robot and collecting a dataset of expert trajectories.

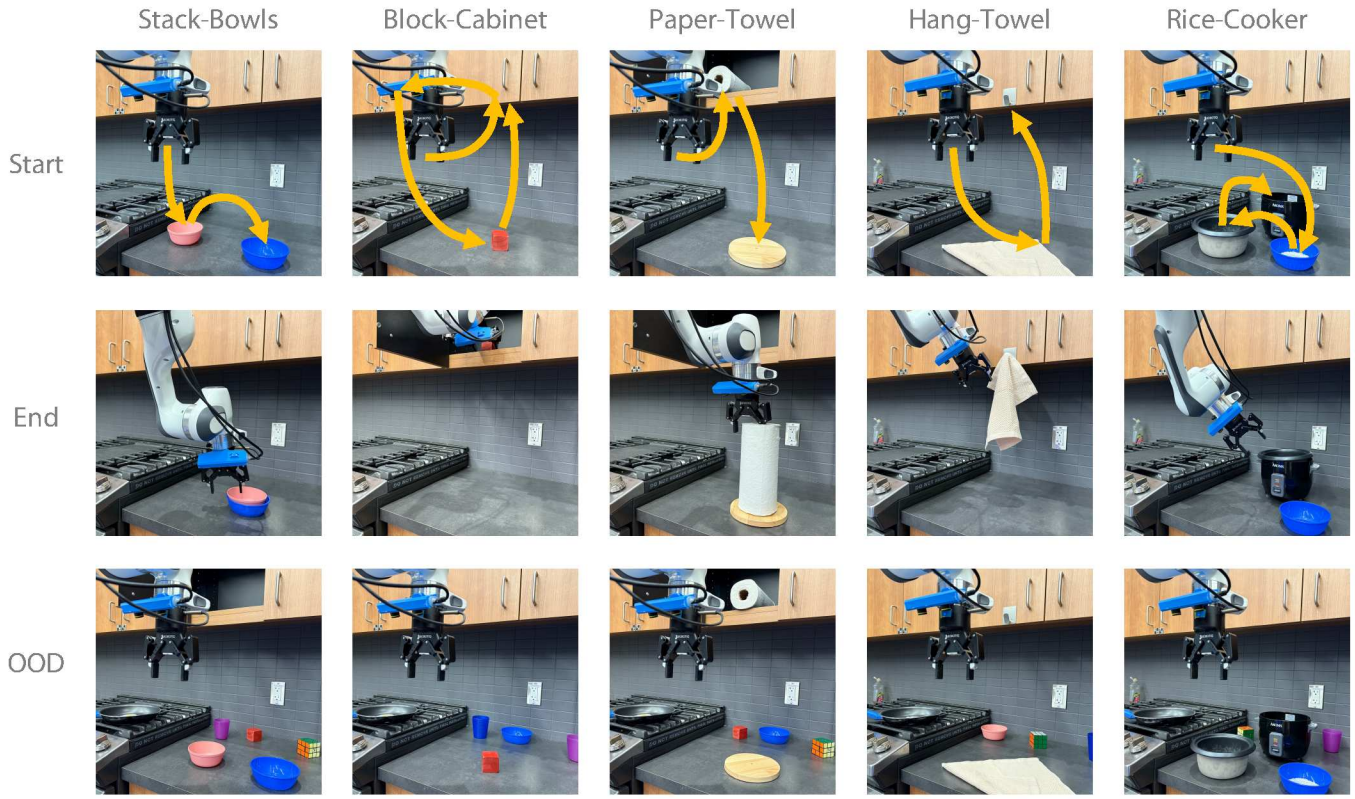

Fig. 5. Real-world setup for real robot tasks: Stack-Bowls, Block-Cabinet, Paper-Towel, Hang-Towel, and Rice-Cooker. The first row illustrates the initial configurations for each task, while the second row demonstrates successful task completions. The third row highlights the out-of-distribution (OOD) configurations designed to evaluate the robustness of each method.

We pretrain all methods on the pretraining / co-training datasets for 100K steps and then finetune to the above-mentioned evaluation tasks (task-specific parameters shown in Table. VII.) Specifically, for cotraining experiments, we mix up the robot and video datasets and sample batches uniformly from the mixture dataset, where each batch may contain action-labeled and action-free data. We then apply the method-specific masking techniques and optimize the cotraining loss. For each task, we evaluate in scenarios approximately similar to those encountered during data collection (referred to as in-distribution), and we also construct an out-of-distribution evaluation setting by introducing distractions that are unseen in the finetuning dataset, as shown in Fig 5. To ensure statistically significant evaluation, we test each task on a fixed set of randomly chosen initialization positions. We provide details for the task-specific setups in Appendix C1.

2) Discussion: We report the results on the real robot experiments across three tasks in Table I and the average performance in Figure 6. For each method and each task, we provide results in the in-distribution (ID) and out-of-distribution (OOD) scenarios. Furthermore, for methods that support co-training on videos, we additionally report the finetuning results of co-trained model (separated by "/").

We first examine the pretraining results on the in-distribution setting. This set of experiments reflect the models' ability to accurately capture the expert policy's distribution.

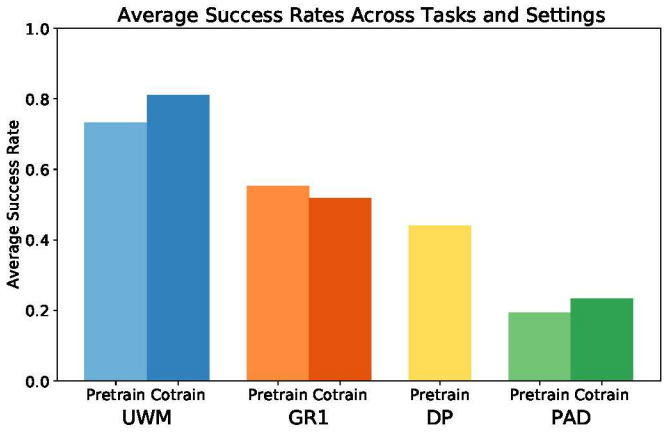

Fig. 6. Average success rates across all real robot tasks and in-distribution and out-of-distribution settings. UWM exhibits strong performance and can further improve by co-training from action-free videos.

We find that UWM achieves the highest success rates across all five tasks among the methods, surpassing the best baseline by as much as 20%. This demonstrates the strength of coupled action-video diffusion in absorbing rich dynamic information from multitask datasets. In particular, since the model is trained to capture all possible conditional and marginal distributions, it is instilled with an understanding of the causal re-

TABLE I
EVALUATION RESULTS ACROSS REAL ROBOT TASKS (PRETRAIN / COTRAIN)

(Pretrain / Cotrain)	Stack-Bowls ID	Stack-Bowls OOD	Block-Cabinet ID	Block-Cabinet OOD	Paper-Towel ID	Paper-Towel OOD	Hang-Towel ID	Hang-Towel OOD	Rice-Cooker ID
UWM (Ours)	**0.86 / 0.92**	**0.76 / 0.84**	**0.76 / 0.84**	**0.60 / 0.72**	**0.78 / 0.86**	**0.78 / 0.84**	**0.82 / 0.86**	**0.64 / 0.76**	**0.60 / 0.65**
DP	0.48 / –	0.36 / –	0.60 / –	0.26 / –	0.52 / –	0.48 / –	0.64 / –	0.28 / –	0.35 / –
PAD	0.08 / 0.20	0.08 / 0.12	0.00 / 0.00	0.00 / 0.00	0.42 / 0.42	0.34 / 0.44	0.52 / 0.54	0.30 / 0.38	0.00 / 0.00
GR1	0.66 / 0.62	0.48 / 0.38	0.66 / 0.74	0.44 / 0.64	0.60 / 0.46	0.60 / 0.46	0.66 / 0.66	0.48 / 0.44	0.40 / 0.25

lationship between actions and image observations, explaining its superior performance compared to joint prediction models such as GR1 and PAD. We observe that GR1 consistently outputs the second best results, establishing a strong baseline performance for deterministic regressive models. On the other hand, diffusion policies fail to leverage the rich and dynamic pixel information in the pretraining datasets, being inefficient at learning from diverse multitask trajectories. Finally, due to a combination of conditioning methods and joint image-action diffusion, PAD achieves the lowest success across the board. We attribute its low performance largely to the conditioning via concatenation. Compared to AdaLN in UWM and DP which takes in image features preprocessed by an encoder, PAD takes in raw pixels, thus needing to incorporate the feature extraction in the same transformer model. This limits its performance at accurately capturing the conditional action distribution without expanding model capacity.

We then examine the OOD scenarios. This set of experiments tests the models' robustness to distribution shifts. We find that all models experience performance drops in the presence of visual distractions. This is especially pronounced in Stack-Bowls, Block-Cabinet, and Hang-Towel. In the Paper-Towel task, the models seem unaffected by the visual distractions, potentially due to the task not requiring the models to pay attention to the table top when grasping the paper towel. Despite a slight performance drop compared to the ID setting, we find UWM to outperform the baselines, showcasing strong robustness under distribution shifts.

Finally, we test the methods' potential to scale with videos by cotraining with action-free videos. Results are reported after the / in each entry of Table I. We find UWM to consistently improve performance when exposed to additional videos during pretraining. This suggests using diffusion time steps for masking as an effective strategy for co-training on multimodal data. While GR1 is able to learn from videos by masking the actions with a learnable token, we found mixed results of the cotrained model. In Stack-Bowls, Paper-Towel, and Rice-Cooker, the cotrained GR1 model is worse than the pretrained model, which implies that incorporating videos dilutes the action learning signal. While PAD showcases weak positive transfer as a result of cotraining, its baseline performance is suboptimal. In Table. IV, we perform evaluations in a larger set of OOD scenarios and found video cotraining to provide significant gains in those settings.

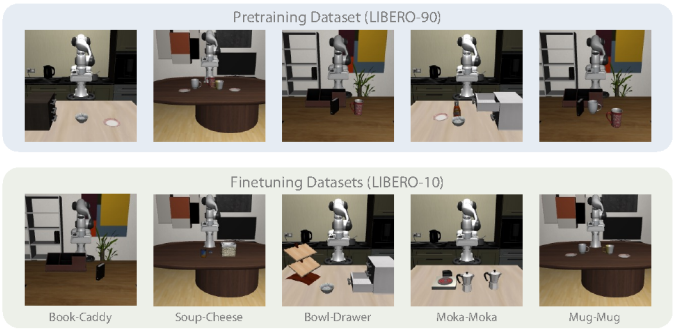

Fig. 7. Visualization of the LIBERO datasets. The pretraining dataset (LIBERO-90) consists of 90 tasks sampled across the kitchen, living room, and study scenes. The finetuning datasets (LIBERO-10) consist of 10 tasks used for evaluation. Tasks from LIBERO-10 are fine-tuned and evaluated under distribution shifts, with unseen initializations and modified configurations.

C. Simulated Experiments

To validate these findings in standard community benchmark settings, we evaluate the methods on the LIBERO [29] simulation benchmark. The LIBERO-100 benchmark consists of 90 training environments across multiple scenes and 10 evaluation environments, each with accompanying expert demonstrations. We combine the demonstrations from the 90 training environments to construct a multitask training dataset, and finetune on a random subset of the evaluation environments, shown in Fig 7. To evaluate the methods' generalization capabilities, we introduce distribution shifts to evaluation environments by enlarging the range of initialization for all objects and removing objects from the scene. The details for this setup is described in Appendix C4.

We pretrain each method on the multitask dataset for 100K gradient steps, and finetune on the downstream tasks for 10K gradient steps. We finetune 3 random seeds for each method on each environment, and evaluate on 50 different initializations. Table II reports the average success rates across initializations with confidence intervals across random seeds. UWM achieves the highest success rates across the evauation tasks in the out-of-distribution setting. DP achieves the second highest performance, followed by GR1 and PAD. These results imply that UWM effectively learns from large robotic datasets, due to its use of pixel reconstruction as an auxiliary signal and the independent diffusion timesteps instilling the model with a causal understanding of actions and observations.

Although our method showed an improvement over baselines, we note that the improvement in OOD scenarios is much

TABLE II
EVALUATION ON LIBERO BENCHMARK.

	Book-Caddy	Soup-Cheese	Bowl-Drawer	Moka-Moka	Mug-Mug	Average
UWM (Ours)	**0.91** ± 0.07	**0.93** ± 0.01	**0.80** ± 0.02	**0.68** ± 0.02	**0.65** ± 0.01	**0.79** ± 0.11
DP	0.73 ± 0.10	0.88 ± 0.02	0.77 ± 0.02	0.65 ± 0.03	0.53 ± 0.05	0.71 ± 0.12
PAD	0.78 ± 0.04	0.47 ± 0.04	0.74 ± 0.05	0.59 ± 0.08	0.25 ± 0.04	0.57 ± 0.19
GR1	0.77 ± 0.03	0.65 ± 0.05	0.62 ± 0.03	0.46 ± 0.04	0.38 ± 0.05	0.58 ± 0.14

less than their real world counterparts. We hypothesize this to be an artifact of current simulations having simpler dynamics than what we see in the real world.

D. Analysis and Ablation Experiments

In this section, we conduct analysis and ablation experiments to help understand the various components and design choices in UWM. We provide additional experiments in Appendix. D.

1) Forward Dynamics: To examine the world modeling component of UWM, we visualize the forward dynamics prediction of UWM on simulated and real-world domains. To generate samples from the forward dynamics model, we perform image diffusion while fixing the action diffusion timestep to 0 and setting the action tokens to be the ground truth actions. As shown in Fig. 8, UWM accurately predicts the image observations conditioned on actions, closely resembling the ground truth image observations. This implies that UWM can effectively model the conditional distribution.

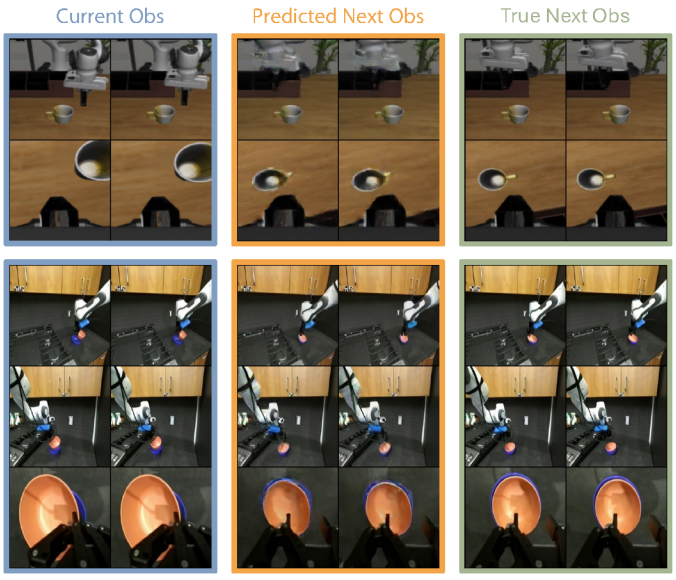

Fig. 8. Visualization of the forward dynamics model predictions. The model accurately predicts the robot and object poses conditioned on the initial observation and actions.

2) Inverse Dynamics: We evaluate the inverse dynamics mode of UWM on trajectory tracking, where we provide a reference expert trajectory and query the inverse dynamics model to track it. Specifically, for each reference trajectory, we reset the simulation environment to match the exact initial state of this trajectory. At each step, we take the ground truth future observations from the trajectory and use the inverse dynamics mode of a finetuned UWM to generate corresponding actions. Table III shows the results of tracking 50 trajectories from the LIBERO training datasets. We find that given the same time limit as the trajectory length, the inverse dynamics model achieves a higher success rate than the policy. This result indicates that actions generated by the inverse dynamics adhere more closely to the reference trajectory. We note that while the policies deviate from the reference trajectories, they eventually recover and solve the tasks given enough time.

TABLE III
TRAJECTORY TRACKING EXPERIMENTS

	Book-Caddy	Soup-Cheese
Policy (1000 steps)	1.00 ± 0.00	0.97 ± 0.01
Policy (trajectory length)	0.47 ± 0.02	0.26 ± 0.02
Inverse dynamics (trajectory length)	0.65 ± 0.01	0.55 ± 0.02

3) Categorized OOD Experiments: We evaluate UWM and DP in several more out-of-distribution (OOD) settings to study their generalization patterns. As shown in Fig. 9, we construct scenes with varied lighting conditions (including static and Disco lights), backgrounds, and clutter. For each scene, we randomly select 5 initializations to evaluate. Results in Table. IV show that across the board, UWM cotrained on videos (co) is significantly more robust than both UWM (pre) and DP pretrained on robot data.

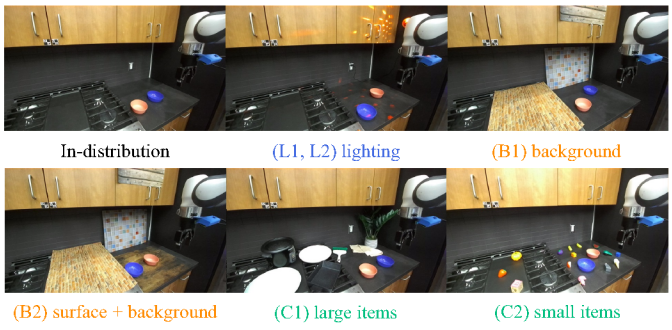

Fig. 9. Visualization of categorized out-of-distribution (OOD) settings. We construct scenes with varied lighting conditions, backgrounds, and clutter to analyze the models' generalization patterns.

4) Real-World Learning from Scratch: To study UWM's ability to scale with pretraining, we train UWM and DP on the task-specific expert demonstrations *from scratch* for the same number of steps as the finetuning stage of the experiments in

TABLE IV
OOD PERFORMANCE ON STACK-BOWLS AND BLOCK-CABINET

	Stack-Bowls			Block-Cabinet		
	UWM (Co)	UWM (Pre)	DP	UWM (Co)	UWM (Pre)	DP
L1	4/5	4/5	2/5	5/5	5/5	3/5
L2	3/5	2/5	2/5	4/5	0/5	0/5
B1	4/5	3/5	3/5	4/5	3/5	2/5
B2	3/5	1/5	2/5	1/5	0/5	0/5
C1	3/5	2/5	2/5	0/5	0/5	0/5
C2	4/5	3/5	1/5	1/5	0/5	0/5
All	**21/30**	**15/30**	**12/30**	**15/30**	**8/30**	**6/30**

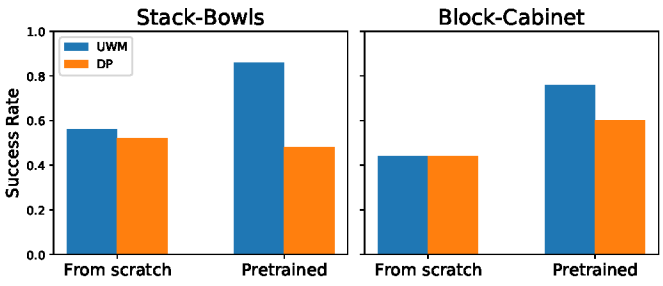

Fig. 10. Training models from scratch vs finetuning pretrained models. UWM scales more effectively from pretraining than DP.

Table. I. As shown in Fig. 10, we find that UWM and DP perform similarly when trained from scratch. However, UWM scales from pretraining more effectively than DP.

5) Ablation of Learning Objectives: To evaluate whether the performance gain of UWM is a result of dynamics prediction or pure reconstruction, we pretrain a UWM to reconstruct the *current* observations instead of the *future* observations. This incentivizes the model to learn about image features, but not about temporal dynamics. Table. V shows that while reconstructing the current observations improves upon the base DP architecture with no image reconstruction, we find it advantageous to reconstruct future observations. This indicates that our model benefits from predicting dynamics rather than purely just image features.

TABLE V
ABLATION OF LEARNING OBJECTIVES

	Stack-Bowls	Block-Cabinet
UWM Reconstruct Future Obs	0.86	0.76
UWM Reconstruct Current Obs	0.70	0.66
DP (No Reconstruction)	0.48	0.60

V. RELATED WORK

*a) **Imitation Learning**:* Imitation learning (IL) for robotics is a paradigm in which robots learn to perform tasks by learning behaviors from experts, typically via teleoperation. A common approach within the imitation learning family is behavior cloning, where supervised learning techniques are applied to replicate expert actions from the provided demonstrations. In particular, these methods are useful for tasks with well defined inputs such as manipulation.

One common challenge for problems cast in the BC framework is the inability to fit multi-modal action distributions [37]. Previous methods have attempted to solve this by attempting to fit multiple pre-defined distributions [31], using architectures amenable to modeling high-dimensional distributions [38, 27, 49, 50], and more recently, generative models such as diffusion [11]. Diffusion models, in particular, have shown to scale favorably to both a large number of demonstrations [40, 6], and dexterous behaviors [15]. Although the diffusion framework has shown the ability to scale, at their core, these formulations rely on access to **high quality** action data. Despite recent efforts from the community to open source large amounts of data [25, 12], the magnitudes of readily available data pales in comparison to the internet scale data that is used to train state of the art foundation models such as LLMs (Large Language Models) and VLMs (Vision Language Models). Alternative formulations to scaling robotic policies focus on leveraging pre-trained foundation models in order to leverage their common-sense reasoning [26, 43] using autoregressive techniques. Although these efforts are promising, they are still heavily reliant on access to high quality action data and focus on increasing generalization.

*b) **Learning from Videos**:* In order to scale large robot foundation models, an appealing approach lies in leveraging video as a source of abundant data. Video data, however, does not contain explicit actions and may contain a significant cross-embodiment gap. In order to address these issues, hand-engineered solutions are often used in order to extract semantic information and map this information to the physical robot. For example, [42, 47] both use keypoints to map actions from video models to the robots themselves. Alternative methods use predicted future points and maps these to rigid body transforms explicitly in order to transfer from internet trained videos to robots [5]. Other work often explicitly track human hand trajectories and and contact patches in order to leverage data from human videos [41, 2].

An alternative approach to leveraging video data relies on large scale pre-training on robot video datasets. For example, [44, 48] use an autoregressive style prediction to pre-train a video and language model which is then fine-tuned on robot actions in a second stage. Other approaches use diffusion models in order to predict and supervise on dense future frames combined with an action diffusion transformer [24]. These use a two-stage process that relies on fine-tuning pre-trained vision models that may not contain robot information. Most importantly, by using a decoupled architecture, they limit the feature sharing capabilities between the video and action data. Closest in spirit to our approach is PAD [20] which trains a joint video-action model using diffusion as its core mechanism. Their approach, however, uses a **shared** diffusion time step between all the modalities which we hypothesize leads to a sub-optimal shared representation that lacks a causal understanding between the underlying video and action models. We show that by having independent diffusion timesteps, our policy performs better in both in distribution and out of distribution scenarios.

*c) **Unified Inference:*** Unified multi-modal models for both decision making and general inference have recently become an emerging topic due to the potential of feature sharing between modalities. [9] explores this topic from the perspective of decision making and shows that masking tokens is an effective way to share information across the decision making process itself. [4] studies this problem from the diffusion perspective on image and text generation. Their results show that by having flexible control of each modality, and thus controlling the marginal, conditional, and joint probabilities, the model is able to do share features and show an increased performance for each individual modality. Our framework builds upon the core insight from this work and studies this from the perspective of joint video and action modeling.

Finally, recent efforts exist in order to combine advantages from both autoregressive and more continuous approaches. [51] combines the ability to do both autoregressive language generation and diffusion based image generation in one framework. Their framework shows efficient scalability and feature sharing. [10] also provides an alternative way to bridge the gap between autoregressive and diffusion techniques. Although the framework provides a mechanism for doing flexible inference that combines the capabilities of continuous and discrete approaches, the multi-modal feature sharing capabilities have not been shown yet.

VI. Discussion

In this work, we present Unified World Models, a diffusion based framework that naturally unifies policy learning and world modeling into a single flexible framework. We instantiate UWM with a coupled conditional diffusion process using separate timesteps for actions and future observations. During training, the model is exposed to all combinations of timesteps covering various conditional and marginal distributions, instilling the model with a understanding of the causal relationship between actions and future observations. This distinguishes UWM from traditional imitation learning approaches, which often lack a nuanced understanding of causal dependencies. Moreover, the independent diffusion timesteps allow for a natural connection between noising and partial masking, enabling the use of action-free videos for co-training, as well as for marginalization and conditioning of the variables by appropriately setting timesteps. The resulting model is then able to flexibly perform inference as a policy, a video prediction model, a forward dynamics model, and an inverse dynamics model. We show through a thorough experimental evaluation that UWM provide significant gains over imitation learning across the board by enhancing large scale pretraining from robotic datasets.

VII. Limitations

While UWM shows promising results, there are several avenues for future investigation. Firstly, the proposed model does not yet learn from large scale human videos, bridging the embodiment gap. Additionally, while UWM shows an improvement on action prediction, the forward dynamics reconstruction may often contain artifacts which may reduce efficacy when planning with this model. We believe this can be addressed by incorporating the latest progress in the generative model literature. Finally, we expect to see further improvement by leveraging denser video prediction.

Acknowledgements

We thank members of the WEIRD lab at UW and the Toyota Research Institute for thoughtful discussions during the course of this work. CZ was supported by funding from the Toyota Research Institute and NSF Grant No. 2212310 during the course of this work.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[2] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics, 2023. URL https://arxiv.org/abs/2304.08488.

[3] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022.

[4] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale, 2023. URL https://arxiv.org/abs/2303.06555.

[5] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation, 2024. URL https://arxiv.org/abs/2405.01527.

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam

Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Micah Carroll, Orr Paradise, Jessy Lin, Raluca Georgescu, Mingfei Sun, David Bignell, Stephanie Milani, Katja Hofmann, Matthew Hausknecht, Anca Dragan, and Sam Devlin. Unimask: Unified inference in sequential decision problems, 2022. URL https://arxiv.org/abs/2211.10869.

[10] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. URL https://arxiv.org/abs/2407.01392.

[11] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[12] Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and rt-x models, 2024. URL https://arxiv.org/abs/2310.08864.

[13] Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *International Conference on Learning Representations*, 2023.

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

[15] Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*, 2024.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[18] 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.

[19] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[20] Yanjiang Guo, Yucheng Hu, Jianke Zhang, Yen-Jen Wang, Xiaoyu Chen, Chaochao Lu, and Jianyu Chen. Prediction with action: Visual policy learning via joint denoising process, 2024. URL https://arxiv.org/abs/2411.18179.

[21] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[23] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

[24] 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, 2024. URL https://arxiv.org/abs/2412.14803.

[25] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Ji-

tendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://arxiv.org/abs/2403.12945.

[26] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[27] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions, 2024. URL https://arxiv.org/abs/2403.03181.

[28] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

[29] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, qiang liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=xzEtNSuDJk.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[31] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.

[32] Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. Ssm meets video diffusion models: Efficient long-term video generation with structured state spaces. *arXiv preprint arXiv:2403.07711*, March 2024.

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022.

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

[35] NVIDIA Research. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, January 2025.

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[37] Stephane Ross and Drew Bagnell. Efficient reductions for imitation learning. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 661–668, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/ross10a.html.

[38] Nur Muhammad Mahi Shafiullah, Zichen Jeff Cui, Ariuntuya Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone, 2022. URL https://arxiv.org/abs/2206.11251.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

[40] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

[41] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play, 2023. URL https://arxiv.org/abs/2302.12422.

[42] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning, 2024. URL https://arxiv.org/abs/2401.00025.

[43] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, and Jian Tang. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation, 2024. URL https://arxiv.org/abs/2409.12514.

[44] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation, 2023. URL https://arxiv.org/abs/2312.13139.

[45] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *The Twelfth International Conference on Learning Representations*, 2024.

[46] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. In *Advances in Neural Information Processing Systems*, 2024.

[47] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learn-

ing by watching: Physical imitation of manipulation skills from human videos, 2021. URL https://arxiv.org/abs/2101.07241.

[48] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2024. URL https://arxiv.org/abs/2410.11758.

[49] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

[50] Tony Z. Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Kamyar Ghasemipour, Chelsea Finn, and Ayzaan Wahid. Aloha unleashed: A simple recipe for robot dexterity, 2024. URL https://arxiv.org/abs/2410.13126.

[51] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

APPENDIX

A. Additional Implementation Details

1) Model Architecture: We base our implementation of UWM on the diffusion transformer architecture with AdaLN conditioning [33]. The inputs to the model are $(o, a_{t_a}, o'_{t_{o'}}, t_a, t_{o'})$, where $o := \{o^i_{0:h_o}\}_{i=1}^{n_c}$ is a sequence of observations from n_c camera views, $a_{t_a} := a_{h_o:h_o+h_a}$ is a sequence of noisy actions, $o'_{t_{o'}} := \{o^i_{h_o+h_a:2h_o+h_a}\}_{i=1}^{n_c}$ is a sequence of noisy observations from each camera view *after* the actions, and t_a, t'_o are diffusion timesteps. The current observations are encoded into features $\{\phi^i_{0:h_o}\}_{i=1}^{n_c}$ using a ResNet-18 [21] encoder, which is initialized the using ImageNet [16] pretrained weights and updated throughout training. The timesteps t_a and $t_{o'}$ are encoded into features ψ_a, $\psi_{o'}$ via a sinusoidal embedding network [22]. The image features are flattened and concatenated with the timestep embeddings and used to condition each transformer block via AdaLN layers [33].

The input sequence to the transformer consists of encoded tokens from a_{t_a}, $o'_{t_{o'}}$ and additional register tokens $r_{1:N_r}$. The actions are encoded to tokens using a shallow MLP encoder share across timesteps. For images, we follow the latent diffusion paradigm [36] and downsample the raw image observations into latent space using a frozen VAE from Stable Diffusion XL [34]. The image latents are patchified into patch embeddings using a 3D convolution layer. We concatenate the action embeddings, the image patch embeddings, and the learnable register tokens along the sequence dimension and pass them as input to the transformer model. We add a learnable positional embedding to the inputs to encode positional information. To decode action and image noise predictions from the model outputs, we take the respective tokens (discarding registers) and decode then using shallow MLP networks. Note that the image noise predictions are in the latent space, and we only decode the final image prediction at the end of the sampling procedure.

2) Training and Inference Details: Given a transition tuple (o, a, o') from sampled from the dataset, we first apply random cropping and augmentations to the image observations. The cropping and augmentation parameters are kept temporally consistent across o and o' but differ from camera view to camera view. We then sample action and observation diffusion timesteps t_a, t'_o *independently* from the uniform distribution $\mathcal{U}[0, T)$. These are used to sample noisy actions a_{t_a} and observations $o'_{t'_o}$ according to the forward diffusion process. The tuple $(o, a_{t_a}, o'_{t_{o'}}, t_a, t_{o'})$ is passed as input to the model, which outputs the action and observation noise predictions $\epsilon_a, \epsilon_{o'}$. We train the model by optimizing the diffusion loss outlined in Eq. 1.

To co-train the model on video data, we combine a robot dataset and a video dataset to get a mixture dataset and sample batches of transition tuples from the mixture dataset uniformly at random. Each batch contains a mixture of video data and action data. For the action-free video samples in each batch, we manually set the corresponding action diffusion timesteps to $t_a = T$, and impute the missing actions with random actions drawn from the unit Gaussian distribution. The action diffusion loss is computed across all samples in a batch (both robot samples and video samples).

TABLE VI
HYPERPARAMETERS

Parameter	Value
Model	
Observation Length h_o	2
Observation Encoder	ResNet-18
Image VAE	SDXL
Image Shape	[224, 224, 3]
Latent Image Shape	[28, 28, 4]
Patch Shape	[4, 4, 2]
Action Length h_a	16
Rollout Length h'_a	8
Embed Dim	768
Timestep Embed Dim	512
Depth	12
Num Heads	12
MLP Ratio	4
QKV Bias	True
Num Registers N_r	8
Diffusion	
Beta Schedule	squaredcos_cap_v2
Num Training Diffusion Steps	100
Num Inference Diffusion Steps	10
Sampler	DDIM
Clip Sample	True
Training	
Batch Size (Distributed)	36 × 4 (pretraining) 36 × 2 (finetuning)
Optimizer	AdamW
Learning Rate	$1e^{-4}$
Weight Decay	$1e^{-6}$
Betas	[0.9, 0.999]
Epsilon	$1e^{-8}$
LR Schedule	constant (pretraining) cosine w/ warmup (finetuning)
LR Warmup Steps	1000
Action Loss Weight w_a	1.0
Image Loss Weight $w_{o'}$	1.0

We optimize the model using the AdamW [30] optimizer. For pretraining experiments, we use a constant learning rate. For finetuning experiments, we use a cosine annealed learning rate with warmup. We sample from the reverse diffusion processes using the DDIM [39] sampler to speed up inference. At deployment, we execute the the first h'_a action predictions and replan. We provide all model and training hyperparameters in Table VI.

Tips for Tuning UWM While UWM is generally stable with respect to hyperparameters, we find that for pretraining on highly multimodal datasets, increasing the number of registers helps improve performance. For new datasets, we recommend trying the default hyperparameters first and then tuning the number of registers for potential performance gains.

3) Training Compute: Training a UWM on the DROID dataset for 100K gradient steps with the hyperparameters shown in Table VI takes 24 hours on 4 NVIDIA A100 GPUs using Pytorch DDP.

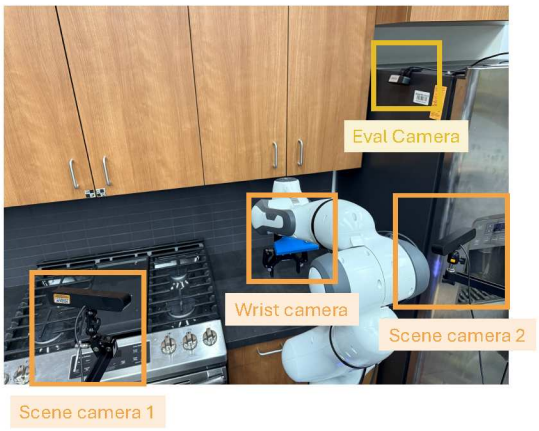

Fig. 11. Setup of the robot experiments. We adopt the DROID [25] setup which consists of two scene cameras and one wrist camera. We use an additional evaluation camera to track the initialization of evaluation seeds.

B. Baseline Details

1) Diffusion Policies: We base our implementation of diffusion policies on the UWM model. We remove the image tokens, image diffusion timestep, and registers and keep everything else identical. This is equivalent to the Transformer version of the original diffusion policy [11] and similar to the architecture in [15].

2) PAD: We base our implementation of PAD on the UWM model, replacing coupled action-image diffusion with joint diffusion, and condition the model by concatenating the clean current observations to the noisy future observation predictions along the channel dimension. The diffusion timestep is still passed into the transformer via AdaLN. While the original PAD [20] method predicts consecutive actions and future frames, we adapt it to predict sequences of actions and the following observations (same as UWM) to isolate the effect of key design differences such as joint video-action diffusion and conditioning method.

3) GR1: We use a custom implementation of the GR1 model adapted to have the same input-output format as UWM. Instead of regressing consecutive actions and observations, we predict a sequence of actions and the following image observations. GR1 conditions on the current observations by passing the ViT encoded observation tokens through a Perceiver resampler from Flamingo [1], and then concatenating the resulting tokens to the input sequence of the transformer model. The rest of input sequence for the transformer consists of learnable action and observation tokens. The output tokens are passed into respective decoders (MLP for actions, DiT decoder for image patches) to regress the modalities.

C. Additional Details on Real-World Experiments

1) Robot Setup: We conduct real-world experiments using a Franka Panda robot in the DROID [25] setup. As shown in Fig. 11 the robot's observation space consists of two scene cameras and a wrist camera (visualized in Fig. 13. We additionally mount an overhead camera to track the initializations during evaluation. The robot operates at a control frequency of 10 Hz, allowing us to have responsive and smooth task execution. The action space is defined by a delta end-effector (EE) pose, which specifies incremental positional and rotational adjustments relative to the current pose. Additionally, the gripper state is represented using a single continuous dimension, where 0 indicates the gripper is open and 1 indicates the gripper is closed.

TABLE VII
TASK-SPECIFIC PARAMETERS

	# demos	# finetuning steps	# eval conditions
Stack-Bowls	50	10K	50
Block-Cabinet	50	10K	50
Paper-Towel	100	20K	50
Hang-Towel	50	10K	50
Rice-Cooker	150	50K	20

2) Tasks: We provide a detailed description of each real-world task shown in Fig. 5 and the task-specific settings in Table VII.

- **Stack-Bowls:** the robot needs to pick up the red bowl on the counter and place it in the blue bowl. The positions of the bowls are randomized across the counter top. A rollout is successful if the red bowl is placed securely inside the blue bowl. For the OOD setup, we open the top cabinet, the bottom drawer, and place unseen objects on the counter and stovetop.
- **Block-Cabinet:** the robot needs to (1) open the left cabinet door by grasping the handle, and (2) pick up the red block from the counter top and place it on the bottom level of the cabinet. The position of the red block is randomized across the counter. A rollout is successful if the block is placed securely in the cabinet. For the OOD setup, we open the bottom drawer and place unseen objects on the counter and stovetop.
- **Paper-Towel:** the robot is tasked to take out a paper towel placed in the open cabinet and place it vertically on a base plate on the counter. The position of the paper towel is randomized across the cabinet shelf, and position of the base plate is randomized across the counter top. Success is counted if the paper towel is placed securely on the base plate and does not topple. For the OOD setup, we open the bottom drawer and place unseen objects on the counter and stovetop.
- **Hang-Towel:** the robot is tasked to pick up a towel from the counter and hang it on a hook on the cabinet. The position and shape of the towel are randomized during data collection. For evaluation, we fold the towel carefully to ensure standardization (Fig. 12). A rollout is successful if the towel hangs on the hook and does not slip off. For the OOD setup, we open the bottom drawer and place unseen objects on the counter and stovetop.
- **Rice-Cooker:** this is a multistage task that involves (1) picking up a cup of rice, (2) pouring the rice into the bowl, (3) placing the cup back on the counter, (4)

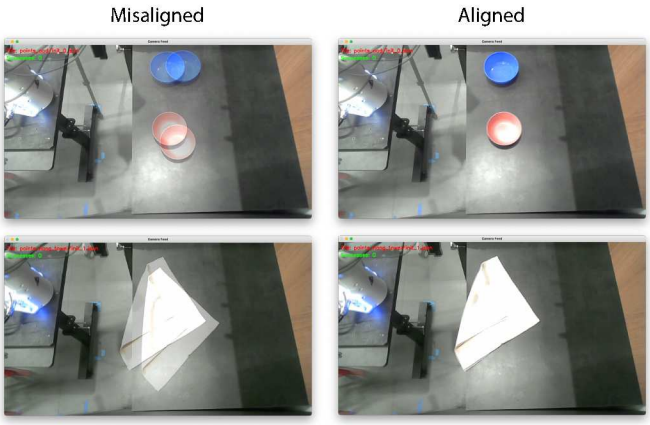

Fig. 12. Screenshots of the evaluation tracker. We use the same interface to track the randomization for all real robot tasks.

picking up the bowl and placing it in the rice cooker. The positions of all objects are randomized. A rollout is successful if there is minimal spill of rice and the bowl is placed securely in the rice cooker. We find this task to be particularly challenging and hence only evaluate on 20 initializations that are close to the dataset distribution. We do not evaluate this task in OOD settings.

3) Evaluation Protocol: To ensure fairness of real-robot evaluations, we use an overhead camera and a Python program to systematically track randomizations. As shown in Fig. 12, the program overlays the reference frame onto the current frame, so the user can adjust the objects to match the reference frame. All tasks except Rice-Cooker are evaluated on 50 randomly sampled configurations. We find Rice-Cooker particularly challengeing and evaluate on 20 configurations close to the data distribution. These initializations are frozen across all evaluations. To mitigate the effects of camera shake due to the mounting mechanism, each method is given three attempts per initialization, making for a more robust evaluation across trials.

4) Failure Modes: We provide a description of some common failure modes in the real-world experiments. Although we utilized three cameras to maximize coverage (Fig. 13), certain angles resulted in objects being visible to only one camera. These limited viewpoints made some initializations more challenging for the robot to complete the tasks successfully. Additionally, variability in object behavior contributed to task failures. For instance, in the Paper-Towel task, the robot often places the paper towel on the wooden platform, but the angle of placement may cause the paper towel to topple over. In the Stack-Bowls task, a source of failure for baseline methods is their inability in distinguishing between the blue bowl and the distractor when attempting to locate the blue bowl after picking up the pink bowl. This issue does not occur with our proposed method.

5) Simulated Environments: LIBERO [29] is a simulated robotic benchmark designed to evaluate lifelong learning algorithms. It involves controlling a 7-DoF Franka Panda

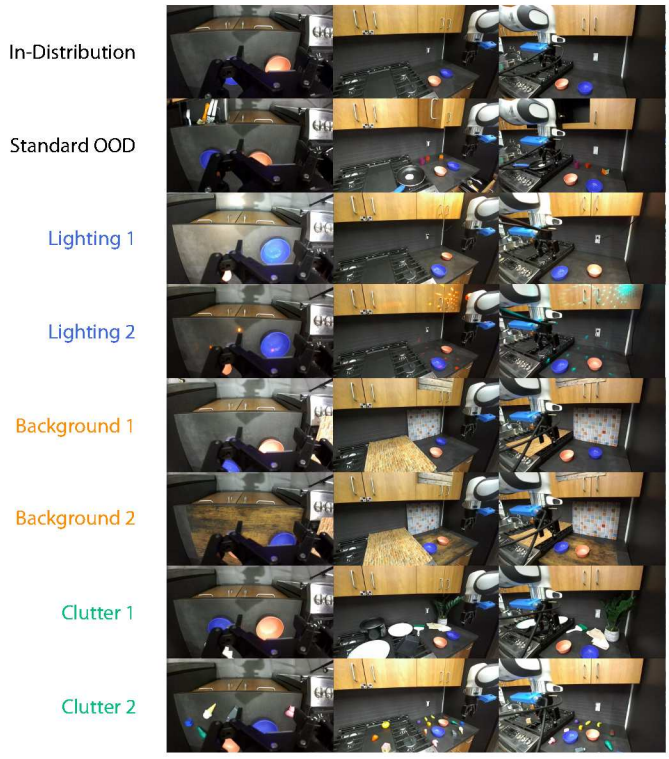

Fig. 13. Visualization of the robot's perspective in in-distribution, standard out-of-distribution (Table. I), and categorized out-of-distribution (Table. IV) scenarios.

robot to complete various tasks across different scenes. The LIBERO-100 benchmark consists of 100 tasks distributed across three scenes (kitchen, living room, study), each with 50 accompanying expert demonstrations. The 100 tasks are split into 90 tasks for training (LIBERO-90) and 10 tasks for evaluation (LIBERO-10)

For our experiments, we use the combined LIBERO-90 dataset as the pretraining data, totaling 4500 trajectories. We evaluate on a random subset of 5 tasks from LIBERO-10. For each task, we finetune the pretrained models on 50 expert demonstrations. To evaluate the generalization capabilities of the methods, we modify the simulation configuration to introduce distribution shifts during the evaluation. Specifically, we increased the initialization range of each object by 0.03 to generate unseen initializations and removed background objects to introduce visual distribution shifts. Visualizations of the evaluation environments are shown in Fig 7, and we provide a description of the evaluation tasks below.

1) Book-Caddy: the robot needs to pick up the book from the table top and place it in the back of a caddy.
2) Soup-Cheese: the robot needs to place the alphabet soup and the cheese in the basket in sequence.
3) Bowl-Drawer: the robot needs to pick up the bowl, place it in the bottom drawer, and close the drawer.
4) Moka-Moka: the robot needs to pick up the two Moka cups from the table and place them on the electric stove.
5) Mug-Mug: the robot needs to place the left mug in the left plate and place the right mug in the right plate.

TABLE VIII
ABLATION OF DESIGN CHOICES

	Book-Caddy	Soup-Cheese
UWM w/ 8 registers	**0.88** $\pm$ **0.04**	**0.90** $\pm$ **0.02**
UWM w/ 4 registers	0.83 $\pm$ 0.05	0.86 $\pm$ 0.03
UWM w/o registers	0.81 $\pm$ 0.07	0.85 $\pm$ 0.03
Cross attention UWM	0.78 $\pm$ 0.05	0.86 $\pm$ 0.04

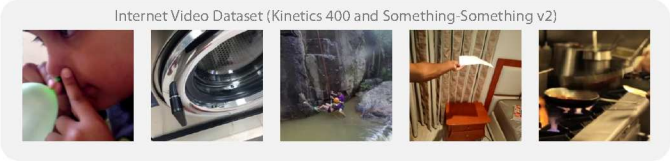

Fig. 14. Visualization of Internet video dataset. We curate the dataset by combining human activity videos from Kinetics-400 [8] and Something-Something-v2 [18].

TABLE IX
COTRAINING ON INTERNET VIDEOS

	Stack-Bowls	Block-Cabinet
UWM Robot Data + Robot Videos	0.92	0.84
UWM Robot Data + Internet Videos	0.88	0.80
UWM Robot Data	0.86	0.76

D. Additional Experiments

1) Ablations of Design Choices: To understand the effect of UWM's design choices, we conduct ablation studies on two simulated tasks from the LIBERO environment. Specifically, we want to (1) understand the effect of registers on task performance, and (2) compare the use of AdaLN for observation conditioning with cross attention [17]. For each model, we train them on the single-task datasets from scratch (without pretraining), and evaluate on 50 initializations across 3 seeds.

Results in Table VIII show that adding registers to the transformer help improve the model performance. We hypothesize that adding registers facilitate the exchange of information between actions and latent image patches, which are distinct modalities. We also found that replacing AdaLN conditioning with cross attention results in worse performance. One possible explanation is that action prediction tasks benefit more from AdaLN's global modulation than from the per-token local modulation provided by cross-attention. We note that this finding may not apply to other modalities such as language.

2) Learning from Internet videos: We evaluate whether UWM can leverage knowledge from Internet videos by including a mixture of Kinetics-400 [8] and Something-Something-v2 [18] dataset in the training, which contain video clips of human activities (Fig. 14). Since the DROID setup has 3 camera views, we use random crops of the same video to impute the missing camera views. Results in Table IX indicate that cotraining on Internet videos shows some improvement on training only on robot data, but cotraining with in-domain robot videos still performs better. We expect these gains to be amplified in more challenging tasks and testing conditions.