A Universal World Model Learned from Large Scale and Diverse Videos

Hanchen Cui
Shanghai Qi Zhi Institute
hanchen.cui@sjtu.edu.cn

Yang Gao
Tsinghua University
Shanghai Qi Zhi Institute
gaoyangiiis@tsinghua.edu.cn

Abstract

World models play a crucial role in model-based reinforcement learning (RL) by providing predictive representations of an agent in an environment and enabling the agent to reason about the future and make more informed decisions. However, there are still two main problems limiting the applications of world models. First, current methods typically train the world models using only massive domain-specific data, making it challenging to generalize to unseen scenarios or adapt to changes in the environments. Second, it is difficult to define the actions when world models are trained using in the wild videos. In this work, we tackle these two problems by learning a general purpose world model from a diverse and large scale real world video dataset with extracted latent actions. Specifically, our approach leverages a pre-trained vision encoder to project the images of two adjacent frames into states; then, extracts the latent actions into a low dimensional space based on vector quantization; finally, a dynamic function is learned using latent actions. Results show that the proposed generic world model can successfully extract latent actions of arbitrary neighboring frames when testing on in the wild video dataset. Furthermore, fine-tuning on only a small amount of in-domain data can significantly improve the accuracy of the generic world model when adapting to unseen environments.

1 Introduction

Reinforcement learning (RL) has shown remarkable success in various domains, but its application to real-world problems is often limited by the high sample complexity and lack of stability. Model-based RL seeks to address these limitations by incorporating powerful world modelsHa and Schmidhuber [2018]. World models offer significant advantages by providing predictive representations of the environment's dynamics. These models enable agents to simulate possible trajectories and make informed decisions, leading to improved sample efficiency, risk-free exploration, and enhanced planning capabilities. Inspired by the great success of pre-training in Computer Vision(CV)Chen et al. [2020] and Natural Language Processing(NLP)Brown et al. [2020], large models can learn rich and generic representations from diverse and large-scale datasets. In addition, the pre-trained models also enjoy strong abilities of generalization, fast adaptation, and even Zero-shot learningRadford et al. [2021]. Therefore, we could naturally ask the question: can we obtain a robust and generalizable world model via pre-training on large scale and diverse datasets?

However, training world models using large scale daily videos is quite challenging, due to the absence of actions. Currently, world models are usually trained by a large amount of domain-specific observation and action pairsHafner et al. [2020]. Actions describe the transition between two adjacent observations and serve as essential information in training dynamic functions. While, this action-required training paradigm generally restricts the available data scale, because most of the video data are observation-only and it is also extremely hard to define abstract and accurate actions in daily

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

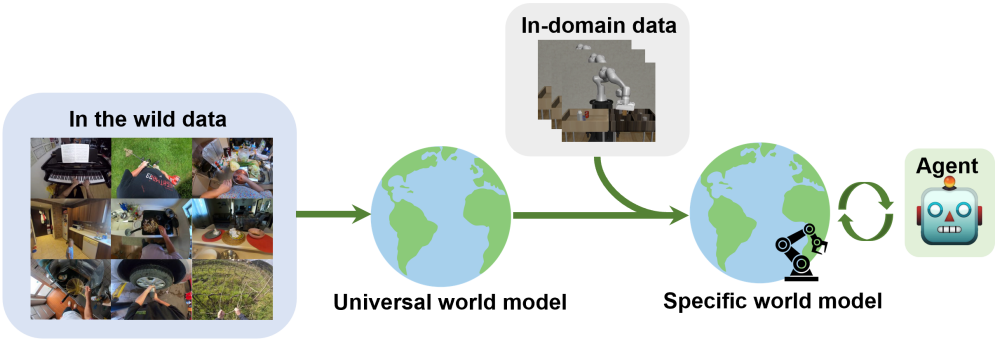

Figure 1: We pre-train the world model using large scale in the wild data and then fine-tune with a small amount of action conditioned in-domain data.

videos. So, how to extract actions from in the wild videos and leverage the abundant internet-scale data has become a promising and valuable topic. Furthermore, data-hungry and poor generalization are other shortcomings in current world models. To learn an accurate and powerful world model, a mass of in-domain data is required, which is usually generated by a specific simulation environment for a long time. Additionally, a well-trained world model commonly surfers from environmental changes, which will lead a significant performance degradation. Unfortunately, the only solution is to generate new data and retrain the world model, which is really a time-consuming and tedious procedure.

In this paper, we present a universal world model, summarized in Figure 1, which is learned from large scale out-of-domain action-free videos and can quickly adapt to unseen environments in a sample efficient manner. The key idea of the proposed universal world model is learning to extract the latent actions from two nearby frames in any daily videos instead of utilizing a pre-defined action space. Then, the extracted latent actions enable learning the dynamic function and predicting the future. Moreover, once the universal world model is well trained, it can adapt to any specific environment effectively when fine-tuned on a small amount of in-domain data. Specifically, an action adapter is designed to align the learned latent actions and pre-defined actions in the environment, enabling accurate planning and efficient decision making in world models. Our contributions are summarized as follows:

- We demonstrate an effective learning based approach that can extract latent actions from two nearby frames in any video. Then, a dynamic function is followed to reconstruct the future frame condition on latent actions and past observations. This action extraction method offers a novel path to allow world models to be pre-trained on large scale in the wild videos.
- We show that the pre-trained world models can efficiently adapt to unseen environments when fine-tuned on a handful of domain specific-data. We design an action adapter to neatly align latent actions and pre-defined actions in a specific environment. Results show that this adapter is able to efficiently eliminate the gap between learned actions and pre-defined actions, making it available for planning and self-exploration in the environment.

2 Related works

Self-supervised pre-training in CV and NLP Self-supervised learning(SSL) has emerged as a powerful paradigm in both Computer Vision (CV) and Natural Language Processing (NLP). It leverages unsupervised learning techniques to learn rich representations without the need for manually annotated labelsMisra and Maaten [2020], Liu et al. [2021]. In CV, self-supervised learning has gained momentum for a variety of tasks, including image classificationZhou et al. [2021], object detectionXie et al. [2021], semantic segmentationWang et al. [2021], and depth estimationPillai et al. [2019]. Approaches like contrastive learningChen et al. [2020], where the models learn to maximize the similarity between augmented views of the same image while minimizing similarity with negative samples, have shown promising results in learning robust visual representations. Recently, masked image modeling (MIM)He et al. [2022], Xie et al. [2022], Bao et al. [2021] has shown great potential in self-supervised training. MIM learns strong representations efficiently without complex data augmentations. In particular, He et al. proposes a masked autoencoderHe et al. [2022] trained by reconstructing the masked pixels. In NLP, self-supervised learning has revolutionized language

understanding and generation tasks. Methods like masked language modeling, where words are masked in a sentence, and the models learn to predict the masked words from the context, have been highly successful. Pre-trained language models, such as BERTDevlin et al. [2018] and GPTBrown et al. [2020], serve as powerful tools for transfer learning and have been fine-tuned on various downstream tasks, achieving state-of-the-art performance.

Foundation models for visual motor control Foundation models(FM) are usually large models trained with diverse and large scale data. Thanks to the strong generalization capacity, foundation models are used for visual motor control tasksNair et al. [2022], Xiao et al. [2022], Jiang et al. [2022]. Foundation models can offer a general purpose representation and enable sample efficient policy learning. R3mNair et al. [2022] learns transferable features from a large vision language dataset through contrastive learning methods. In addition, value or reward functions can also be extracted from videos and enable model-based RL. VIPMa et al. [2022] can not only serve as a representation module but also as a generic value function. Furthermore, based on the strong generative capacity of vision models and the powerful planning ability of large language models, foundation models also serve as a robust data augmentation method. ROSIEYu et al. [2023] generates a mass of realistic and task-relevant vision observations to enhance generalization ability. Recent breakthroughs in large language models enable planning for long horizons and complex tasks. Text2motionLin et al. [2023] proposes a language-based planning framework to solve sequential manipulation tasks that require long-horizon reasoning. VoxposerHuang et al. [2023] utilizes large language models collaborating with vision-language models to manipulate in the real world without a training process.

World models World models in model-based reinforcement learning (RL) refer to the learned models that simulate the dynamics of the environment in which an RL agent operates. Inspired by the great potential of pre-training, the world models can also be pre-trained. While, previous worksHafner et al. [2020], Seo et al. [2023, 2022a] usually learn a domain-specific world model and the training data is generated by a certain environment. This training paradigm can only work well in the specific domain and lead to poor generalization ability. FICCYe et al. [2022] builds a world model through a discrete autoencoder to play Atari games. Recently, some works have tried to explore training the world models using cross-domain or in the wild data. ContextWMWu et al. [2023] trains an action-free video prediction module from in the wild data by decoupling context and dynamics. Next, ContextWM is fine-tuned with action conditioned data in a certain domain similar to APVSeo et al. [2022b]. SWIMMendonca et al. [2023] learns a structured world model from human-centric videos. However, the action space of SWIM is complexly designed with human hand motion extraction. In this work, we propose an end-to-end approach to extract low dimension latent actions from two nearby frames. The latent actions can efficiently adapt to a pre-defined action space of a certain environment by fine-tuning with a handful of in-domain data.

3 Preliminaries

Problem formulation World models describe the dynamics of different states and preserve temporal information of environments. The compact states are extracted from vision observations, which are relatively low dimension and efficient for predicting and planning. Dynamic function models the transition between states in latent space. A decoder is attached to reconstruct the vision observation of the corresponding state. These components are represented as follows:

$$encoder: s_t \sim \mathrm{enc}_\phi\left(s_t \mid o_t\right), \quad dynamics: s_{t+1} \sim p\left(s_{t+1} \mid s_t, a_t\right), \quad decoder: o_t \sim p\left(o_t \mid s_t\right), \tag{1}$$

where o_t means the vision observation and s_t is a compact state learned from image o_t though the encoder, parameterized by ϕ. The dynamic function takes the previous state and action as inputs, represented as s_t and a_t respectively, and predicts future state s_{t+1}.

Masked autoencoder Masked autoencoderHe et al. [2022] is a self-supervised representation method which learns strong representations from raw images without complex data augmentation approaches. The images $o_t \in \mathbb{R}^{H \times W \times C}$ are broken down to a sequence of small patches $p_t \in \mathbb{R}^{N \times (P^2 C)}$. N is the total number of patches in an image defined by $N = \frac{HW}{P^2}$, where P means patch size. Moreover, the sequence of patches is randomly masked with a large portion m. The unmasked patches are represented as $p_t^m \in \mathbb{R}^{M \times (P^2 C)}$, where M means the number of unmasked patches,

$$patchify : p_t \sim f^{patch}(p_t \mid o_t), \quad masking : p_t^m \sim f^{mask}(p_t^m \mid p_t, m). \tag{2}$$

The unmasked patches are projected to D-dimensional vectors through a linear layer and fed into a ViT encoder with additional positional embeddings. Then, the encoded unmasked embeddings combined with learnable tokens are fed into a lightweight ViT decoder to reconstruct pixel values of the masked patches,

$$ViT\ encoder : h_t^m \sim p_\phi\left(h_t^m \mid p_t^m\right), \quad ViT\ decoder : o_t \sim p_\phi(o_t \mid h_t^m). \tag{3}$$

The encoder and decoder are jointly trained via Mean Square Error(MSE) calculated on masked patches.

Vector Quantised Variational AutoEncoder(VQ-VAE) In VQ-VAEVan Den Oord et al. [2017], the encoder network outputs discrete latent codes, which differs from VAE. The codes are determined by the nearest neighbor calculated between the encoder output and a codebook. Besides, the codebook is updated. The codebook is defined as $e \in \mathbb{R}^{K \times D}$, where K is the number of codes in a codebook and D means the dimension of every code. First, the image o_t goes through an encoder to obtain a representation $z_e(o_t)$. Then, the discrete latent variable z is determined by the nearest neighbor look-up using the shared embedding space e as shown below,

$$q(z = k \mid o_t) = \begin{cases} 1 & \text{for k} = \operatorname{argmin}_j \left\| z_e(o_t) - e_j \right\|_2 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Therefore, the input of decoder becomes e_k. Next, a decoder is attached to reconstruct the input image o_t. The complete set of parameters for the model are from the encoder, decoder, and the embedding space e. The parameters of the model are jointly optimized by the following equation,

$$\mathcal{L}_{vq} = \log p\left(o_t \mid z_q(o_t)\right) + \left\| \operatorname{sg}\left[z_e(o_t)\right] - e \right\|_2^2 + \beta \left\| z_e(o_t) - \operatorname{sg}[e] \right\|_2^2, \tag{5}$$

where $sg[\,]$ means stopping gradient and β is a hyperparameter to balance these loss terms.

Forward-Inverse Cycle Consistency (FICC) FICCYe et al. [2022] proposes to leverage action-free videos from Atari games to train a specific world model via cycle consistency loss. Besides cycle consistency, the loss function also includes reconstruction trems of visual observations o_t and difference of observations $o_t - o_{t+1}$. The visual observations are projected to latent space via a CNN network $\mathcal{R}$. Then a VQVAE module is followed to generate one discrete code z as a latent action. Next, $\hat{s}_{t+1}$ is predicted by dynamic function $\mathcal{D}$ based on s_t and z. The cycle consistency loss function is shown below,

$$\mathcal{L}_{\text{cc}} = \overbrace{\underbrace{-\cos(\hat{s}_{t+1}, s_{t+1})}_{\text{similarity}} \underbrace{-\ln p(o_{t+1} - o_t | s_t, z_q)}_{\text{difference reconstruction}}}^{\text{cycle consistency}} \underbrace{-\ln p(o_t | s_t)}_{\text{reconstruction}}, \tag{6}$$

where $s_t = \mathcal{R}(o_t), s_{t+1} = \mathcal{R}(o_{t+1}), z_q = inverse(s_t, s_{t+1}), \hat{s}_{t+1} = \mathcal{D}(s_t, z_q)$. The total loss function is a combination of cycle consistency loss and VQVAE loss, represented as $\mathcal{L} = \mathcal{L}_{\text{cc}} + \alpha\mathcal{L}_{\text{vq}}, \alpha = 1$.

4 Method

In this section, we introduce a generic world model, a framework for extracting latent actions from two adjacent visual observations and adapting the learned latent actions to a pre-defined action space of an environment. The flow chart of the proposed method is shown in Figure 2. Our method consists of: (1) learning a generalized dynamic function from in the wild videos (see Section 4.1), (2) adapting pre-trained world models to the specific domain using only a handful of in-domain data and aligning latent actions with pre-defined environment-specific actions by fine-tuning on a small amount of action-conditioned data (see Section 4.2).

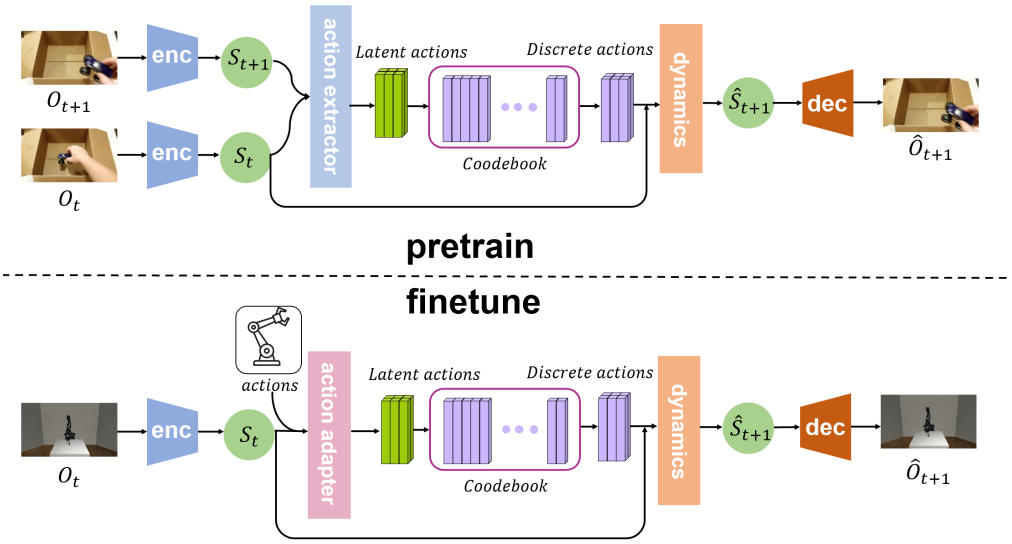

Figure 2: The general framework of training the proposed world models. Up: pre-training by large scale video dataset. Down: fine-tuning by actions conditioned in-domian data.

4.1 World models learned from in the wild videos

In the wild video data enjoys rich sequential dynamic frames collected from diverse scenes, making it possible to optimize a generic world model which can learn the dynamics of the complex real world. However, because daily videos are visually complex, it is difficult to optimize visual representations and dynamics simultaneously. To address this issue, unlike FICC, which optimizes representation network and dynamics together, we leverage a powerful visual encoder, pre-trained on a large dataset, to capture visual information and only the dynamic function is optimized. We propose a universal world model, trained from in the wild video data, to learn dynamic functions for all environments. To begin with, two nearby frames O_t and O_{t+1} are sampled from a randomly selected video and fed to a pre-trained visual encoder to obtain visual features, represented as S_t and S_{t+1}, respectively,

$$Visual\ encoder: S_t \sim p_\phi(O_t), \quad S_{t+1} \sim p_\phi(O_{t+1}). \tag{7}$$

Next, these two representations are concatenated and put into an action extractor module, several transformer layers, to extract compact latent actions A_L represented the dynamic transition between S_t and S_{t+1},

$$Action\ extractor: A_L \sim p_\phi(O_t, O_{t+1}). \tag{8}$$

Then, the extracted latent actions A_L are compressed to several discrete latent codes A_D via a codebook, denoted as $e \in \mathbb{R}^{K\times D}$, to reduce the information of the latent actions. Due to the naturally complexity of in the wild videos, VQVAE output multiple discrete latent codes to represent the actions in the videos. The discrete process is designed as an information bottleneck to avoid serious model collapse. The reason that latent actions must be compact is to prevent shortcut learning which means the latent actions A_L directly copy S_{t+1} and makes the action extractor effectless. Therefore, the information capacity of A_L should be relatively low to encode only the transition between S_t and S_{t+1} instead of the whole state S_{t+1}. The codebook helps to further restrict the information flow and result in a discrete and abstract action A_D,

$$A_D = e_k, \quad \text{where} \quad k = \text{argmin}_j \left\| a_L - e_j \right\|_2. \tag{9}$$

In addition, the discrete actions A_D are combined with S_t to predict the next state, denoted as $\hat{S}_{t+1}$, through a dynamics network,

$$Dynamics: \hat{S}_{t+1} \sim p_\phi(S_t, A_D). \tag{10}$$

The predicted state $\hat{S}_{t+1}$ is fed to a pre-trained decoder, which is jointly optimized with the pre-trained encoder, to reconstruct the input observation O_{t+1}, represented as $\hat{O}_{t+1}$,

$$Visual\ decoder : \hat{O}_{t+1} \sim p_\phi(\hat{S}_{t+1}). \tag{11}$$

In the end, the training objective is much simpler than FICC which designs a relatively complex loss function including cycle consistency and difference reconstruction. Our loss function can be divided into two parts. The first term is to minimize MSE loss between ground true state S_{t+1} and predicted state $\hat{S}_{t+1}$. The second term is to minimize VQVAE loss. The loss function can be expressed as follows:

$$\mathcal{L} = \log p\left(O_t \mid z_q(O_t)\right) + \left\|\mathrm{sg}\left[z_e(O_t)\right] - e\right\|_2^2 + \beta \left\|z_e(O_t) - \mathrm{sg}[e]\right\|_2^2 + \left\|S_{t+1} - \hat{S}_{t+1}\right\|_2^2, \tag{12}$$

where the hyperparameter β is set as 0.25 in our experiment.

4.2 Finetune with in-domain data

The accuracy of transition dynamics is essential for model-based reinforcement learning. The generic world model is trained by plenty of in the wild and diverse video data. Although they enjoy excellent generalization ability, they are still not optimal for a specific environment. Hence, a natural method is proposed to fine-tune the generic world model with a small bunch of domain-specific data.

The main purpose of world models is to enable planning in a specific environment with real actions. In the previous step, the universal world model has been pre-trained by large scale and diverse in the wild video data and fine-tuned with a small amount of in-domain data. However, the world model can only use learned latent actions, which are inconsistent with ground truth actions of the environment, to predict the next frame. Therefore, the goal of fine-tuning with action conditioned data is to establish an efficient map between latent actions and ground truth actions. First, sample a pair of visual observations accompanied by its corresponding actions, noted as O_t, O_{t+1}, and A_{GT}. Next, two observations are fed into a pre-trained encoder, which is the one used in the pre-training phase, to obtain latent representations S_t and S_{t+1}. Then, an action adapter is designed to build a map between ground true actions and learned actions. In detail, the action adapter takes S_t combined with ground true actions A_{GT} as inputs to fit the latent actions produced by the action extractor in the pre-training phase. The simple objective can be formulated as:

$$\mathcal{L}_{adapter} = \left\|A_L - p_\phi(S_t, A_{GT})\right\|_2^2 \tag{13}$$

The following processes are identical to the pre-training stage, including discretizing by VQVAE, predicting the next state, and decoding. In addition, the codebook does not need to be fine-tuned and the only trainable module is the action adapter. After finetuning with action conditioned data, the world model can efficiently obtain the dynamic function of this environment and accurately predict the next frame.

5 Experiment

We evaluate the proposed generic world model on reconstructing various video datasets and simulation environments, including Ego4DGrauman et al. [2022], something-something V2Goyal et al. [2017], and Robosuite benchmarksZhu et al. [2020]. Specifically, we aim to investigate the following questions:

- Can the generic world model reconstruct the next frame of any in the wild video with high quality?
- Can fine-tuning with domain-specific data improve the reconstruction performance of this environment?
- Can the generic world model accurately predict the next visual observations with ground truth actions by fine-tuning with action conditioned data?

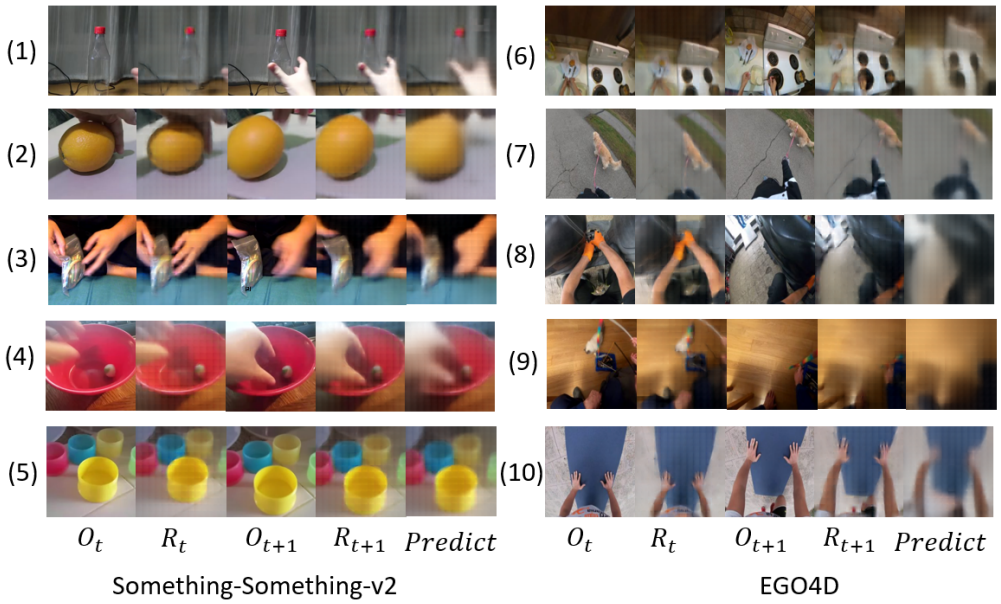

Figure 3: The visualization results of the pre-trained world model from Something-Something-v2 and EGO4D datasets. O_t, R_t, O_{t+1}, and R_{t+1} mean two observations and their corresponding reconstructions. $Predict$ represents the predicted future frames through the world model.

Table 1: The LPIPS scores between the predicted frame and two observations in SSv2 and EGO4D datasets.

SSv2	(1)	(2)	(3)	(4)	(5)
O_t	0.280	0.437	0.350	0.323	0.416
O_{t+1}	**0.193**	**0.240**	**0.224**	**0.201**	**0.153**
EGO4D	(6)	(7)	(8)	(9)	(10)
O_t	0.470	0.378	0.540	0.427	0.397
O_{t+1}	**0.370**	**0.247**	**0.385**	**0.232**	**0.284**

5.1 Experimental setup

Pre-training datasets We select Something-Something-v2 (SSv2) and EGO4D as our pre-training datatsets. The Something-Something-v2 dataset is a collection of 220,847 labeled video clips of humans performing pre-defined, basic actions with everyday objects, such as putting something on a surface, Moving something up, and Pushing something from left to right. These videos are defined to finish some relatively specific tasks and include most of our daily motions. Furthermore, EGO4D is a massive-scale and egocentric dataset collected across 74 worldwide locations and 9 countries, with over 3,670 hours of daily-life activity video. The EGO4D dataset is pretty challenging for the reason that it is collected purely in the wild and does not represent a specific motion, even with frequent view shifting.

Visual dynamic environment We examine the visual dynamic function of the world model in a robot control environment. Robosuite is a simulation framework powered by the MuJoCo physics engineTodorov et al. [2012] for robot learning. It also offers a suite of benchmark environments for reproducible research. It contains seven robot models, eight gripper models, six controller modes, and nine standardized tasks. We randomly sample a video clip with actions from Robosuite and predict the next visual observation with past observations and actions.

Implementation details In pre-training, an image sequence is sampled from a video clip with a time interval of 0.2 seconds. The size of input image is $224 \times 224 \times 3$ by randomly cropping from raw observations. The pre-trained encoder and lightweight decoder are taken from MAE, a self-supervised framework trained by reconstructing masked patches. The structure of the action extractor and dynamics network are a 4-layer ViT and a 2-layer ViT, respectively. The discretization module is a VQVAE, with a codebook of 1024 individual codes with 64 dimensions. We use the

random restart methodDhariwal et al. [2020] to train and update the codebook, which dramatically smooths the training process and improves the sample efficiency. The parameters of the pre-trained encoder and decoder are frozen during the training process. The other parameters including action extractor, VQVAE, and dynamics are jointly optimized by the AdamW optimizer with a learning rate of 1×10^{-4}. Additionally, the learning rate is warmed up and cosine scheduling decayed to ensure better convergence. In addition, other parts of the network consisting of the encoder, decoder, and VQVAE, share the parameters with the ones in the pre-training process, which means the action adapter is the only module to be trained. The parameters of the action adapter is optimized by Adam with a learning rate of 3×10^{-4}.

Reconstruction evaluation metrics We select the LPIPS scoreZhang et al. [2018] to measure the similarity between ground truth visual observations and predicted images by the world model. Unlike traditional metrics such as Mean Squared Error (MSE) and Structure Similarity Index (SSIM), which tend to focus on pixel level difference, LPIPS takes into account perceptual features extracted from deep neural networks, making it more aligned with human visual perception. It is based on the idea that features learned by deep convolutional neural networks (CNNs) such as AlexNetKrizhevsky et al. [2012] or VGGSimonyan and Zisserman [2014], particularly those pre-trained on large image datasets, capture meaningful information about image content and structure. The key insight behind LPIPS is that image patches with similar high-level features are likely to appear visually similar to humans. The LPIPS score is calculated as the Euclidean distance between the feature representations of two images. A lower LPIPS score indicates higher perceptual similarity between the images.

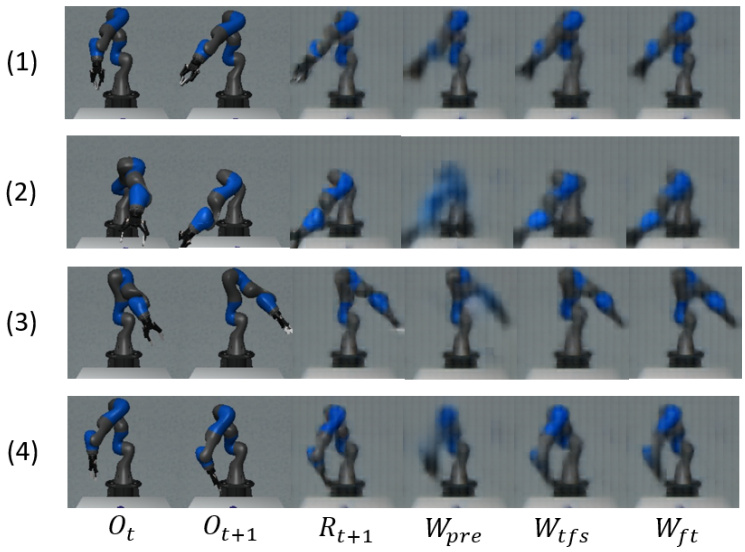

Figure 4: The visualization comparison of the predicted frames between the world model pre-trained by diverse data, in-domain data from scratch, and fine-tuned by in-domain data after pre-training, denoted by W_{pre}, W_{tfs}, and W_{ft} respectively.

5.2 Experimental results

Dynamics for in the wild data Figure 3 shows some examples of the evaluation of dynamic functions trained from in the wild video data. To verify the generalization ability of the world model, we evaluate our method on Something-Something-v2 and EGO4D datasets, which are known as commonplace and diverse datasets. The transition dynamics of real world is quite complex, including new object appearing or disappearing, view shifting, and object pose variation. The visualization results show that the proposed world model trained on diverse datasets can effectively predict the future frame based on the current frame and latent actions. The slight blur of the reconstructed and predicted frames comes from the pre-trained MAE decoder. To quantitively analyze the results, we calculate the LPIPS score between the predicted frame and two observations. To make a fair evaluation, these

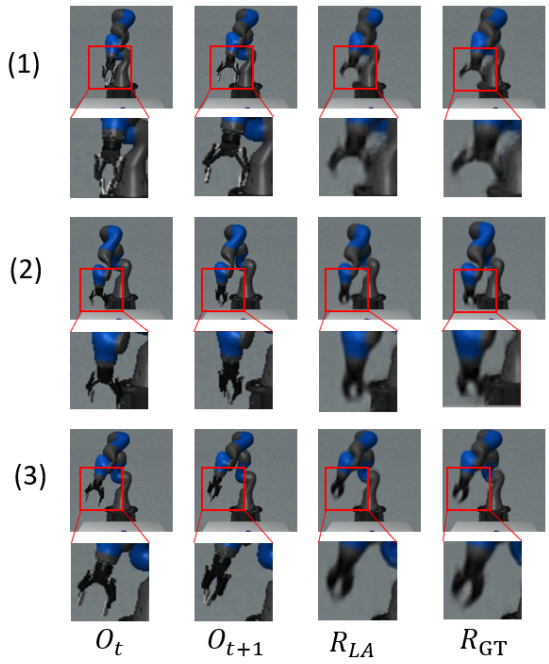

Figure 5: The visualization results of action conditioned adaptation, with enlarged gripper section. R_{LA} and R_{GT} mean reconstruction by latent actions and reconstruction by ground truth actions.

two observations are equally blurred because our purpose is to measure semantic similarity instead of fine-grained details. The quantitive results are reported in Table 1. The LPIPS scores show that the predicted frame are more resemble to the future observation indicating the world model effectively learn the dynamics of in the wild video instead of copying the current observation.

Fine-tuning with in-domain data In the pre-training process, the models are trained by in the wild data. Therefore, it might not be optimal for a specific environment. Fine-tuning serves as a promising way to address this sub-optimal problem. Figure 4 shows the results of future frames predicted by the world model trained by complete diverse data, fine-tuned with a hand of in-domain data, and entire in-domain data. The results indicate that the pre-trained world model obviously suffers from sub-optimal issue and results in a degraded observation. However, after fine-tuning with a small amount of data, only 2k paired frames, the image quality of predicted observations is significantly enhanced, even matching the performance of the model trained totally by in-domain data. consequently, pre-trained by large scale and diverse in the wild data and fine-tuned with a small group of domain-specific data is an efficient configuration to train the world model.

Fine-tuning with action conditioned data To enable planning in the world model, the learned latent actions should align with real actions, which are defined by the simulation environment. So, we design a simple action adapter to learn the map between latent actions and real actions. The results are shown in Figure 5. For elaborately comparing the differences, which are usually tiny between two adjacent frames, we additionally fine-tuned the parameters of the decoder to obtain better reconstruction performance. After fine-tuning with action conditioned data, the learned world model can successfully predict future frames based on real actions.

6 Discussion

In this paper, we propose a general purpose world model trained by a large scale and diverse video dataset, which can learn latent actions from two observations and predict the future frame in a large range of diverse video. Then, when fine-tuned with a handful of in-domain data and action conditioned data, the latent actions can be strictly aligned to real actions and accurately predict the future frame. For limitations, currently, the generic dynamic function of the proposed world model is only examined by visualization and quantitive analysis. It also should be tested by a model-based RL algorithm in simulation environments and we will accomplish this part in future works.

References

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2020.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

Sudeep Pillai, Rareş Ambruş, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9250–9256. IEEE, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *arXiv preprint arXiv:2303.12153*, 2023.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. In *Conference on Robot Learning*, pages 1332–1344. PMLR, 2023.

Younggyo Seo, Junsu Kim, Stephen James, Kimin Lee, Jinwoo Shin, and Pieter Abbeel. Multi-view masked autoencoders for visual control. 2022a.

Weirui Ye, Yunsheng Zhang, Pieter Abbeel, and Yang Gao. Become a proficient player with limited data through watching pure videos. In *The Eleventh International Conference on Learning Representations*, 2022.

Jialong Wu, Haoyu Ma, Chaoyi Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *arXiv preprint arXiv:2305.18499*, 2023.

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pages 19561–19579. PMLR, 2022b.

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

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.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.