

VideoWorld 2: Learning Transferable Knowledge from Real-world Videos

Zhongwei Ren^{1,2*}, Yunchao Wei^{1,3†}, Xiao Yu¹, Guixun Luo¹,
 Yao Zhao¹, Bingyi Kang², Jiashi Feng², Xiaojie Jin^{2*†‡}

¹Beijing Jiaotong University ²ByteDance Seed ³Beijing Academy of Artificial Intelligence, BAAI

<https://VideoWorld2.github.io/>

Abstract

Learning transferable knowledge from unlabeled video data and applying it in new environments is a fundamental capability of intelligent agents. This work presents **VideoWorld 2**, which extends VideoWorld [51] and provides the first investigation of learning transferable knowledge for complex, long-horizon tasks directly from raw real-world videos. At its core, VideoWorld 2 introduces a dynamics-enhanced Latent Dynamics Model (dLDM) that decouples action dynamics from visual appearance: a pretrained video diffusion model handles visual appearance modeling, enabling the dLDM to learn latent codes that focus on compact and meaningful task-related dynamics. These latent codes are then modeled autoregressively to learn task policies and support long-horizon reasoning. We evaluate VideoWorld 2 on challenging real-world handcraft making tasks, where prior video generation and latent-dynamics models struggle to operate reliably. Remarkably, VideoWorld 2 achieves up to **70% improvement in task success rate** and produces coherent long execution videos. In robotics, we show that VideoWorld 2 can acquire transferable manipulation knowledge from the Open-X dataset, which substantially improves task performance on CALVIN, demonstrating strong cross-domain generalization. This study reveals the potential of learning transferable world knowledge directly from raw videos, with all code, data, and models open-sourced for further research.

1. Introduction

Current AI models primarily learn knowledge[§] from large-scale text data [2, 5, 10, 16, 18, 33, 57, 60, 68]. However, text alone cannot fully capture the rich information of the real visual world, including world dynamics, spatial relationships, and underlying physical laws. In contrast, animals in nature can acquire knowledge directly from visual

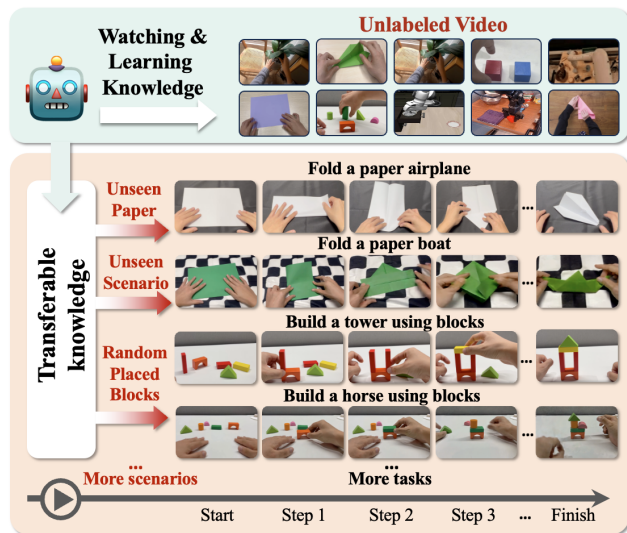


Figure 1. VideoWorld 2 explores how to learn transferable knowledge from unlabeled real-world videos. We construct a handcraft benchmark to evaluate the learned knowledge.

signals, and generalize it to solve tasks across diverse scenarios. For instance, a child can reproduce paper-folding skills demonstrated in a video using different paper materials, without any language instruction. Given the vast abundance of video content available on the internet, enabling AI to learn *generalizable knowledge* [31] from raw video data holds significant promise for scaling their knowledge acquisition and is fundamental to their ability to execute tasks effectively in both real-world and digital environments.

VideoWorld [51] is among the first works to explore learning knowledge from synthetic videos. It investigates the acquisition of rules, as well as reasoning and planning capabilities, from Go game records and simulated robotics environments. The study demonstrates that models can learn such knowledge solely from visual signals using an autoregressive video generation paradigm. However, extending this to real-world domains remains an open challenge. Real-world videos exhibit substantial visual diver-

Work done when Zhongwei interned at ByteDance Seed. † Correspondence to Xiaojie Jin (jinxiaojie@bytedance.com) and Yunchao Wei (yunchao.wei@bjtu.edu.cn). * Equal contribution, ‡ Project lead.

[§]Following prior work [51], we use “knowledge” to refer broadly to the rules, reasoning, and planning abilities required for task completion.

sity, complex action dynamics, and often involve long-horizon, multi-step interactions. These characteristics prevent the training approach and model design of VideoWorld from being directly applied to realistic settings. When presented with minute-long, multi-step real-world task videos, VideoWorld fails to extract the core task-solving knowledge or generalize it to novel scenarios through observation alone—even for tasks such as paper folding that are easily mastered by children (See Sec. 5.3 for details). These limitations naturally lead to the following question: *Can AI models learn transferable knowledge for complex, long-horizon tasks directly from unlabeled real-world videos?*

In this work, we extend VideoWorld to real-world settings. We first establish a Video-Craft benchmark to systematically evaluate the capabilities of existing models across two representative real-world environments. As illustrated in Fig. 1, the first environment focuses on **handicraft making**, which serves as a challenging testbed for learning task knowledge directly from raw videos. These videos involve fine-grained manipulation under diverse environmental appearances, including deformable materials (e.g., paper), viewpoint variation, and frequent occlusions between hands and objects. Moreover, the videos are minute-long and consist of multiple interaction steps, presenting substantially greater complexity and longer temporal horizons than typical entertainment-oriented videos. Second, we study **robotic manipulation** by learning from the Open-X dataset [48], which contains real-world demonstration videos, and evaluating on the CALVIN environment [45]. Since CALVIN features tasks and visual settings that differ from those in Open-X, this setup enables a rigorous assessment of the generalization capability of the learned knowledge. Crucially, learning from these videos requires models to extract task-critical actions from agent dynamics and to derive robust policies that can reason and plan correct actions across diverse environments.

To address this benchmark, we first apply VideoWorld directly to real-world videos. Compared to its original training data, these videos exhibit substantially more complex dynamics and appearances. We observe that VideoWorld struggles to extract task-critical actions and to learn policies generalizable to novel environments. As shown in Fig. 2 (bottom), the model’s predictions contain severe errors, including distorted hand poses, incorrect object shapes, and inconsistent environmental appearances, ultimately failing to produce accurate and visually coherent action sequences. Moreover, although state-of-the-art video generation models [1, 36, 59] can generate high-fidelity visual details, they similarly fail to faithfully represent task execution. We conjecture that these models are unable to disentangle task-core actions embedded in visual changes and motion dynamics, and instead overfit to irrelevant visual details. This overfitting leads to degraded task performance and unstable long-

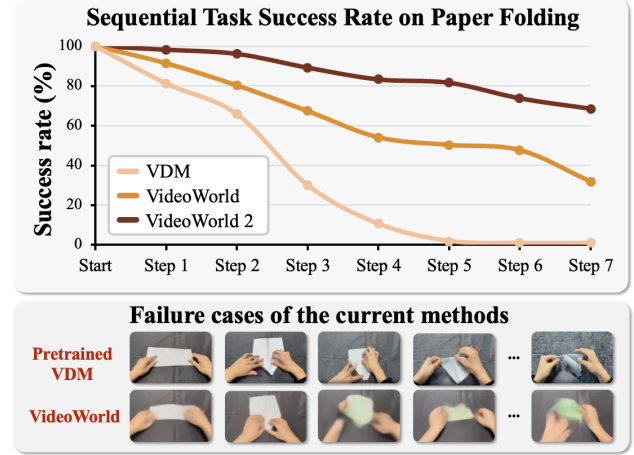


Figure 2. Comparison of different frameworks in sequential success rate for 7-step long-horizon paper folding tasks in Video-CraftBench (detailed in Sec. 4). VDM (e.g., Wan2.2 14B [59]) produces high visual fidelity but fails to learn long-horizon policies. VideoWorld [51] improves policy learning but suffers from poor visual quality. VideoWorld 2 learns more robust dynamics while also achieving better visual quality, enabling generalizable long-horizon knowledge learning from videos.

horizon generation. In contrast, humans naturally prioritize essential actions while filtering out extraneous variations.

Inspired by this, we propose **VideoWorld 2**, which aims to robustly acquire knowledge by explicitly decoupling appearance modeling from action learning. The key innovation over VideoWorld is the incorporation of prior appearance knowledge during action extraction, which prevents irrelevant appearance information from entangling with the learned action representations. We implement this idea through a novel **dynamics-enhanced Latent Dynamics Model (dLDM)**. The dLDM comprises a causal VQ-VAE and a pretrained Video Diffusion Model (VDM). The VQ-VAE compresses future visual changes into discrete latent codes that capture task-relevant actions, while the VDM is responsible for modeling visual appearance, generating high-fidelity reconstructions of task sequences conditioned on these codes. By offloading appearance modeling to the VDM, VideoWorld 2 encourages the latent codes to focus on concise, semantically meaningful, and transferable actions rather than superficial appearance details. As a result, the learned policies are significantly more robust and generalizable than those produced by VideoWorld. As illustrated in Fig. A.2, VideoWorld 2 successfully executes long-horizon handicraft making tasks in unseen environments, producing coherent task execution videos.

Beyond these qualitative demonstrations, we further compare VideoWorld 2 with prior methods on Video-CraftBench. As shown in Fig. 2, VideoWorld 2 excels at long-horizon tasks, outperforming competing methods by learning more generalizable action representations and achieving significantly higher visual quality. In robotics,

pre-training on Open-X further enables VideoWorld 2 to acquire effective manipulation knowledge. This leads to substantial improvements in action prediction on the CALVIN environment, demonstrating that the learned actions and policies transfer robustly across diverse robotic tasks.

Our contributions are summarized as follows:

- We are the first to investigate learning transferable knowledge for complex, long-horizon tasks directly from raw real-world videos. Our findings show that disentangling action dynamics from visual appearance modeling is critical for effective knowledge acquisition.
- We propose the dynamics-enhanced Latent Dynamics Model (dLDM), which decouples task-relevant dynamics from visual appearance, substantially improving the quality and transferability of learned knowledge for both handicraft and robotic manipulation tasks.
- We introduce Video-CraftBench, a new benchmark for fine-grained, long-horizon visual reasoning in real-world handicraft tasks, facilitating future research on knowledge learning from raw videos.

2. Related Work

2.1. Video Generation

Video generation research is dominated by two main paradigms: diffusion-based models for high-fidelity synthesis [20, 36, 46, 59, 69] and autoregressive (AR) models [21, 35, 75] for LLM-inspired [79] causal sequence modeling. Leading platforms like Cosmos [1] explore both branches to support downstream planning [23] and real-time interactivity [69, 72]. VideoWorld 2 combines the strengths of both paradigms. It leverages the appearance priors of diffusion models to *disentangle* core actions from visual changes, thereby enabling autoregressive modeling of long-horizon policies based on these dynamics. Further comparison on the exploration of “disentanglement” in video generation models is provided in the Supp. B.

2.2. World Models

World models, which aim to learn physical dynamics for controllable simulation [22, 82], are approached from several perspectives. The video generation community [19, 30, 32, 49, 54, 55, 73] treats them as controllable synthesizers, emphasizing fidelity and consistency. The reinforcement learning/robotics communities [4, 12, 27–29, 37, 43, 50, 64] view them as learned dynamics simulators for sample-efficient planning. JEPA-style approaches [3, 7, 8, 25, 34, 41, 80, 81] avoid pixel-level reconstruction, instead forecasting in an abstract space to benefit downstream tasks. While these perspectives primarily focus on learning short-term world dynamics for synthesis or planning, VideoWorld 2 addresses the distinct challenge of learning transferable knowledge for complex, long-horizon tasks directly

from unlabeled real-world videos.

2.3. Learning from unlabeled videos

Learning transferable representations of visual dynamics from unlabeled video remains challenging. Some methods [67] rely on scarce paired data, and others explore unsupervised latent action learning via cycle consistency [52, 71], VQ-VAE quantization [15, 70], or future frame prediction [11, 13, 14, 24, 56]. VideoWorld 2 advances these works in two ways. First, unlike existing models focusing on short-horizon transitions and standard reconstruction objectives, our dLDM captures long-horizon dynamics guided by a pre-trained video generator. Second, while prior LAMs target simpler tasks like grasping or 2D games [17, 40, 42], we tackle complex, multi-stage, minute-long tasks (e.g., handicrafts) involving significant appearance variations.

3. VideoWorld 2

3.1. Learning Knowledge from Unlabeled Videos

Generative knowledge learning. A video can be viewed as a demonstration trajectory that captures world state transitions and the underlying action policy, which constitute the knowledge to be learned. Accordingly, we propose to use a generative model to capture these latent policies and action dynamics directly from video, without reliance on language supervision. Formally, we define this setting as a tuple $\mathcal{G} = \langle \mathcal{X}, \mathcal{A}, \rho \rangle$, where $\mathcal{X}$ denotes the observation space, $\mathcal{A}$ is the action space, and ρ is a video generator. Given a sequence of video frames $x \in \mathcal{X}$, our goal is to train ρ to model the conditional distribution of the next frame x_{t+1} given the observation history $x_{0:t}$. This formulation enables the acquisition of task knowledge through video generation, eliminating the need for explicit labels.

To transfer the learned knowledge into executable actions for downstream tasks, ρ maps the visual state transitions into action space and thus serves as a policy model $\pi(\cdot|x_{0:t}) : \mathcal{X} \rightarrow \mathcal{A}$, which predicts actions based on historical observations and executes the tasks accordingly.

Basic generative framework. We adopt mainstream video generation models [1, 36, 59] as the basic framework. They typically utilize a VQ-VAE to encode videos into a compressed representation. The generative process operates within this space to predict future states, which are subsequently decoded back to RGB domain. However, these representations require thousands of discrete tokens or continuous embeddings to capture the full spectrum of visual information, inevitably leading to spatiotemporal redundancy and a sparse distribution of knowledge. As demonstrated in VideoWorld, such representations result in an inefficient encoding of the visual changes and dynamics tied to critical decisions and actions, thereby hindering the framework from acquiring essential task knowledge from raw videos.

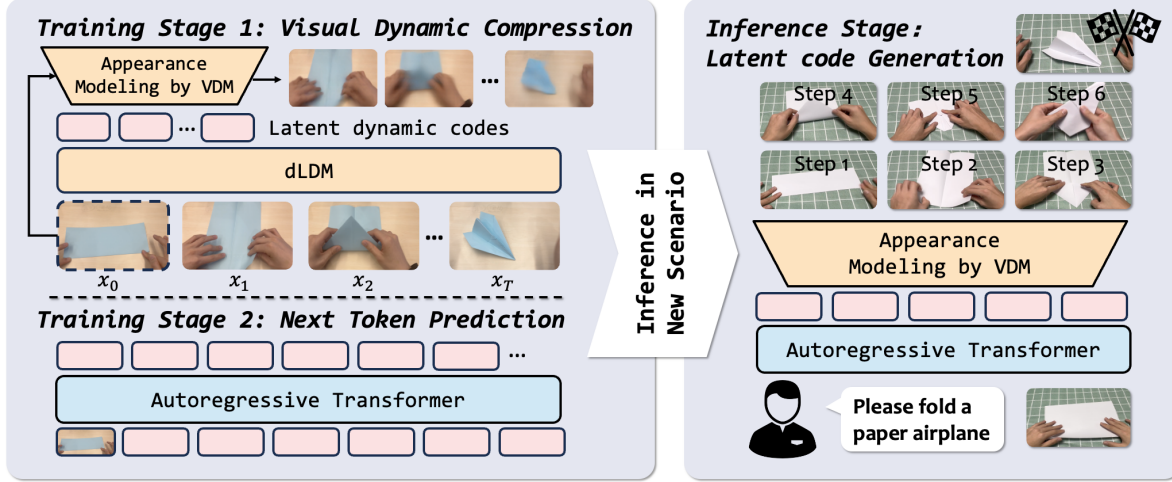


Figure 3. **Overview of the VideoWorld 2 model architecture.** (Left) First, the dLDM compresses future visual changes into compact and generalizable latent codes. These codes are then modeled by an autoregressive transformer. (Right) In inference, the transformer predicts latent codes for a new, unseen environment from the input image, which are subsequently decoded into task execution videos.

Latent dynamic model. To mitigate the above limitation and facilitate the learning of knowledge, VideoWorld introduces a Latent Dynamic Model (LDM) that compresses future visual changes into a set of compact latent codes, effectively capturing the motion dynamics of multi-step actions.

Specifically, as shown in the left panel of Fig. 4, the LDM utilizes a MAGViT-style [74] causal codec. The encoder first maps an input clip x of length T to a feature sequence $f_{0:K}$, where $K = 1 + \lfloor \frac{T-1}{s} \rfloor$ and s is the temporal downsampling stride. Next, it defines N learnable query embeddings $q = \{q^n\}_{n=1}^N$. As shown in Fig. 4 (left), these queries use cross-attention to capture change information in $\{f_{0:k}\}_{k=1}^K$, yielding a continuous representation $z = \{z_k^n\}_{k=1, n=1}^{K, N}$. This representation is then quantized to prevent the LDM from learning shortcuts (e.g., trivially copying f_k to z_k). Finally, the decoder uses f_0 and the quantized z to reconstruct the subsequent frames in a causal manner. The training objective is to minimize the ℓ_2 distance between the original and reconstructed frames. We refer to these embeddings as latent dynamics codes.

3.2. Dynamics-enhanced Latent Dynamic Model

While VideoWorld effectively handles long-term reasoning in synthetic environments by encapsulating temporal dynamics into compact latent codes, it still struggles with real-world application. As shown in the bottom of Fig. B.7, the model fails to generate coherent long-horizon sequences or generalize to new environments. Specifically, applying the latent dynamics learned from handcrafting training videos to novel environments with different desktops, paper materials and arms, frequently results in significant scene drift and erroneous actions. We also evaluate several state-of-the-art video generation models [1, 36, 59]. As shown in Sec. 5.3 and Fig. B.7, although these models can pro-

duce high-fidelity frames, they also fail to reproduce fine-grained, long-horizon skills in new contexts, even when detailed language instructions are provided.

Entanglement of dynamics and appearance. We conjecture that these limitations stem from the insufficient disentanglement of action dynamics and visual appearance. In the basic framework, the joint modeling of these components impedes the effective extraction of task-core action knowledge. Regarding VideoWorld, we find its learned latent codes capture irrelevant visual details—such as background motion, lighting changes, textures, and camera displacement—making the model sensitive to environmental changes, as further discussed in Sec. 5.5. Consequently, this incomplete separation ultimately leads to poor consistency in long-horizon tasks and limits the model’s ability to generalize to new environments.

Dynamics-enhanced latent dynamic model. To address this issue, we introduce a dynamics-enhanced Latent Dynamic Model (dLDM). The key mechanism is to replace the original LDM decoder with a pretrained Video Diffusion Model (VDM). By leveraging the capacity of VDM for high-fidelity reconstruction to handle the appearance modeling, we allow the LDM encoder and learnable query embeddings to focus exclusively on capturing task-relevant visual changes. Although the VDM does not contain any knowledge of the target task’s dynamics, it is highly effective at producing realistic visual content once given appropriate dynamics guidance, making it well-suited for this decomposition (As shown in Fig. A.2 and Fig. A.6).

Concretely, as shown in the right panel of Fig. 4, the dLDM consists of a causal VQ-VAE that encodes future visual changes into discrete latent codes and a pretrained VDM that reconstructs high-fidelity frames. The latent codes are provided to the VDM through a projection layer

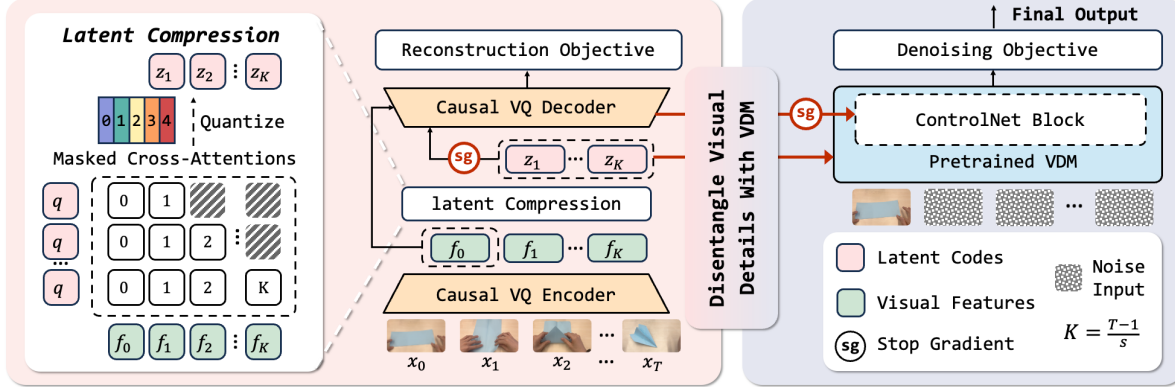


Figure 4. **The proposed dynamics-enhanced latent dynamics model (dLDM).** (Left) Latent dynamic model in VideoWorld [51]. Visual changes between the first and subsequent frames are compressed into a set of latent codes. (right) The dLDM proposed in VideoWorld 2. It employs a pre-trained VDM as an appearance prior, yielding better latent codes and facilitating high-fidelity video output.

and causal cross-attention. Since the VDM handles appearance modeling, the latent codes are relieved from encoding fine-grained visual details and can instead focus on capturing task-relevant dynamics. To maintain temporal correctness, we enforce causal attention in the VDM so that features at time t attend only to information up to time t .

Directly training the VDM to generate future frames from noise would be extremely slow and prone to incorrect motion, as it has never been trained on the target tasks like long-horizon handicraft making. Therefore, we reuse the VQ-VAE decoder to reconstruct latent codes into low-fidelity, motion-rich outputs, providing coarse temporal cues such as hand movements and object displacements. This signal is fed into the VDM via a gradient-stopped, ControlNet-like [76] branch. This stabilizes training and allows the VDM to focus on refining appearance rather than inferring motion from scratch. Additionally, we stop the gradient flow of the decoder to latent codes to prevent the introduction of irrelevant noise. We ablate this in Sec. 5.5.

By leveraging the VDM, the learned latent codes are much less affected by changes in appearance and stay consistent across environments. As shown in Fig. A.6, latent codes corresponding to the similar action exhibit tighter intra-class alignment and reduced cross-environment variance, indicating robust and transferable dynamics.

Auto-regressive transformer. After extracting the latent codes, we use an autoregressive transformer to model the latent dynamics sequence. For each video $x_{0:T}$, the dLDM extracts latent codes $\{z_k^n\}_{k=1, n=1}^{K, N}$. We flatten these into a sequence and train the transformer to predict them, conditioned on initial frame x_0 and task instruction. This allows the model to learn long-term patterns in complex tasks.

During inference (Fig. 3, right), given a single input frame from a new, unseen environment, the transformer predicts future latent dynamics based on its learned task representation, and the dLDM decodes them into coherent long-horizon videos. This allows VideoWorld 2 to transfer its

learned dynamics to new environments and execute action sequences beyond those observed during training.

4. Video-CraftBench

Our objective is to learn transferable knowledge for complex and long-horizon tasks from real-world videos. We prioritize tasks that demand multi-step planning and feature delicate manipulations. To this end, we introduce the Video-CraftBench, a dataset of first-person video tutorials covering five long-horizon handicraft tasks: folding a paper airplane, folding a paper boat, and building a tower/horse/person using blocks.

4.1. Dataset Generation

These handicraft tasks involve fine-grained, visually diverse manipulations that are difficult to describe linguistically. We collected ~ 7 hours (split into ~ 9.5 k clips) of tutorials from manual and web sources, enabling long-horizon knowledge evaluation. Paper folding tasks typically last 40-80s, while block building takes 20-30s. To assess transferability, the test set (~ 150 separately collected videos) features novel backgrounds, paper textures, and block arrangements not seen in training. Supp. C provides more details.

4.2. Evaluation

Our evaluation focuses on two main aspects: the accuracy of key actions and the visual quality of the generated videos. Accordingly, we use the following metrics for evaluation:

- **Sequential task success rate:** We decompose paper folding into 7 key steps (Fig. A.5) and train a DINOv2-based classifier [47] to detect completion. The classifier prioritizes action correctness while disregarding appearance consistency (e.g., appearance drift from the initial frame). Training utilizes annotated frames from the training set, test set, and manually verified generated videos, labeled as specific step or “failed”. For evaluation, we generate 3

Method	Fine-tuning	Sequential Paper folding Success Rate (%)							Block Success Rate (%)			Visual Quality	
		1	2	3	4	5	6	7	Human	Tower	Horse	SSIM↑	LPIPS↓
<i>Pre-trained Video Generation Model</i>													
Cosmos AR 4B [1]	Craft-text	68.4	56.7	11.5	3.3	0.0	0.0	0.0	10.1	18.0	12.0	0.643	0.312
Cosmos DiT 2B [1]	Craft-text	73.4	63.3	20.0	6.7	0.0	0.0	0.0	24.2	21.3	19.7	0.680	0.264
Hunyuan-13B [36]	Craft-text	76.9	68.1	27.5	5.8	0.0	0.0	0.0	30.9	38.4	31.5	0.703	0.255
Wan 2.2 14B [59]	Craft-text	81.2	75.0	30.4	10.6	0.0	0.0	0.0	39.7	42.6	34.1	0.719	0.237
<i>Learning with Latent Action/Dynamic Model</i>													
LAPA [70]	Craft				N.A.					N.A.		N.A.	
Moto [15]	Craft	19.1	11.7	3.3	0.0	0.0	0.0	0.0	11.5	10.1	9.8	0.585	0.394
AdaWorld [70]	Craft	43.6	39.8	27.4	10.8	0.0	0.0	0.0	20.7	13.1	15.0	0.611	0.378
VideoWorld [51]	Craft	<u>70.3</u>	<u>66.7</u>	<u>42.5</u>	<u>21.3</u>	<u>6.7</u>	0.0	0.0	<u>23.8</u>	<u>33.9</u>	<u>27.8</u>	<u>0.680</u>	<u>0.351</u>
VideoWorld 2	Craft	97.2	95.3	90.0	83.3	81.4	74.6	68.8	70.0	81.5	80.9	0.770	0.205
iVideoGPT [65]	OpenX & Craft	23.1	18.7	13.3	3.7	0.0	0.0	0.0	15.3	11.0	12.6	0.588	0.390
Moto [15]	OpenX & Craft	43.1	35.3	30.7	25.5	18.3	9.7	0.0	17.4	15.3	16.0	0.596	0.387
AdaWorld [70]	OpenX & Craft	49.5	41.6	34.8	30.7	22.3	19.8	13.0	37.4	29.8	29.1	0.624	0.365
CoLA [62]	OpenX & Craft	83.5	74.4	<u>69.1</u>	<u>64.8</u>	<u>52.3</u>	<u>49.8</u>	<u>40.2</u>	<u>54.1</u>	<u>52.4</u>	<u>49.9</u>	<u>0.668</u>	<u>0.289</u>
VideoWorld [51]	OpenX & Craft	<u>91.7</u>	<u>75.0</u>	68.2	63.1	51.7	48.2	31.9	47.3	<u>52.7</u>	49.8	0.601	<u>0.389</u>
VideoWorld 2	OpenX & Craft	98.2	96.4	90.1	86.7	83.3	81.7	72.3	74.0	83.0	85.8	0.774	0.193

Table 1. **Comparison on Video-CraftBench.** Craft-text denotes using step-by-step textual description.

video rollouts per test case conditioned on the first frame. A step is successful only if all preceding steps are complete. For the shorter block-building tasks, a similar classifier is trained, but only to verify the final state.

- **Visual quality:** We also assess visual quality using standard LPIPS [77] and SSIM [63] metrics to quantify visual fidelity and content consistency with the input.

5. Experiment

5.1. Implementation Details

We employ the NVIDIA Cosmos AR 4B model [1] as our auto-regressive transformer, repurposing its next-token prediction capability to predict our latent codes. For the appearance prior, we utilize the Cosmos DiT 2B model [1], leveraging its high-fidelity I2V capabilities to generate 93-frame (~5s at 16 fps), 480px videos. The dLDM also processes a 93-frame clip at a time by default, using a vocabulary size of 1000 (FSQ levels [8, 5, 5, 5]) and an embedding length $N = 4$. See Supp. A for more training details.

5.2. Benchmarks and Baselines

Benchmarks. We evaluate VideoWorld 2 on two benchmarks: Video-CraftBench assesses learning from long-horizon, real-world tasks with fine-grained actions. For robotics, we train our latent representation on the large-scale OpenX dataset and test its knowledge transfer to CALVIN. CALVIN features 34 tasks and uses a long-horizon, sequential evaluation protocol similar to our paper folding setup: models must complete a 5-task sequence, where success is contingent on all preceding tasks.

Baselines. We compare against three strong baselines:

- Pre-trained video generation model: We select four state-of-the-art models: NVIDIA Cosmos AR 4B [1], NVIDIA

Cosmos DiT 2B [1], Wan2.2 14B [59], and Hunyuan-Video 13B [36]. We fine-tune them on Video-CraftBench. To align with their original training, we utilize Qwen2.5-VL 72B [6] to generate detailed step-wise language annotations as auxiliary guidance. During inference, they generate execution videos conditioned on an initial image and text, producing subsequent clips autoregressively.

- Latent action models: We select five concurrent works [15, 24, 62, 65, 70] that utilize latent action models. These methods compress inter-frame visual changes into latent codes, serving as a pre-training objective for both manipulation and video reconstruction. We adapt their latent extraction methodologies to our benchmark to assess their ability to extract knowledge from real-world videos.
- VideoWorld [51]: It is the most relevant baseline to our work, using the original latent dynamic model.

5.3. Results on Video-CraftBench

Training with Video-CraftBench only. In Tab. 1, we first evaluate pre-trained video generation models fine-tuned on Video-CraftBench (row 1-4). While these models achieve high success rates (>68%) in the first step of the paper folding task and up to 38% in the block stacking task, performance deteriorates rapidly for complete, minute-long sequences. By step 4 of paper folding (Column 7), success rates drop to $\leq 10.6\%$, with a total failure in subsequent steps. Crucially, this failure persists even though we have provided detailed step-wise text descriptions during training and inference to leverage their text-conditioning capabilities. This demonstrates that current models struggle to learn complex, long-horizon knowledge from real-world videos, underscoring the need for a more effective paradigm.

Subsequently, we evaluate existing methods that leverage latent action or dynamic models. Following their pro-

Idx	Method	Pretraining Type	Pretraining	Fine-tuning	Sequential Task Success Rate (%)					
					1	2	3	4	5	Avg. Len.
<i>In-Domain Latent Pretraining</i>										
1	Transformer (Oracle)	-	-	ABCD→D	80.9	55.6	44.5	31.3	24.6	2.36
2	Transformer (Oracle)	-	-	10% data	50.5	35.4	20.1	5.20	0	1.11
3	LAPA [70]	Latent	ABCD→D	10% data	74.4	45.8	25.2	15.3	2.30	1.49
4	VideoWorld 2	Latent	ABCD→D	10% data	75.8	47.9	31.8	20.4	9.70	1.87
<i>Cross-Domain Latent Pretraining</i>										
5	Transformer (Oracle)	Video	OpenX	ABCD→D	85.9	60.4	46.0	30.7	23.0	2.46
6	LAPA [70]	Latent	OpenX	ABCD→D	84.0	58.8	46.2	35.4	27.0	2.51
7	VideoWorld 2	Latent	OpenX	ABCD→D	88.5	64.6	55.8	47.5	30.9	2.88

Table 2. **Comparison on CALVIN benchmark.** “10% data” means 2k trajectories randomly sampled from the ABCD→D training set. “Oracle” means directly training the transformer using ground-truth action label as supervision.

tocols, we train these baselines on Video-CraftBench and train our AR transformer to predict their latent codes. Note that due to LAPA [70]’s structural constraints, decoding its codes into long-horizon sequences causes severe degradation. Thus, we mark its results as “N.A.”. Among the other baselines, Moto [15] uses a pre-trained vision encoder to extract dynamics, AdaWorld [24] utilizes an auxiliary diffusion head for video synthesis, and VideoWorld [51] features capabilities for long-horizon dynamics compression and generation. Although they generate coherent videos, they fail to generalize to novel environments, with none successfully completing the full paper folding sequence under unseen settings (e.g., different desktops or paper styles). As shown in Fig. B.7, their predictions exhibit substantial action errors and scene distortions, indicating that their latent codes overfit to irrelevant appearance information.

In contrast, VideoWorld 2 (row 9) generates complete and coherent task sequences in test environments. Despite lacking large-scale pre-training, training solely on Video-CraftBench attains success rates of 68.8% for paper folding and up to 81.5% for block stacking. This demonstrates that our dLDM efficiently extracts core task-relevant action information while filtering out extraneous details like background clutter, object variations, and camera noise. Furthermore, by leveraging VDM appearance priors, VideoWorld 2 also produces videos with better SSIM and LPIPS metrics. Fig. A.2 and Supp. D visualize our generated sequences.

Data scaling with Open-X. We next evaluate whether VideoWorld 2 scales with larger data by incorporating the OpenX dataset [48]. Compared to Video-CraftBench, OpenX features a vast array of robotic agents and environments, providing an ideal testbed to assess VideoWorld 2’s ability to filter out agent-specific and environmental factors while scaling its action extraction capabilities.

Specifically, we train dLDM and baseline models on a combination of OpenX and Video-CraftBench, then train the AR transformer exclusively on Video-CraftBench using the resulting latent codes. As shown in Tab. 1 (row 10–12), baselines show improved quality and higher success rates. However, Moto and iVideoGPT still fail to complete the full

folding sequence. While VideoWorld generates complete sequences, its final-step success rate remains low at 31.9%. We also examine CoLA [62], a concurrent work that also uses a VDM to optimize latent action codes. CoLA is limited to short 2-frame transitions and ignores the structured temporal cues from coarse VAE outputs. Adopting its training scheme yields a limited 40.2% final-step success rate. In contrast, VideoWorld 2 further improves performance, achieving 72.3% (paper folding) and 85.8% (block stacking) final-step success. As shown in Fig. A.1, similar latent codes in OpenX and Video-CraftBench correspond to similar motion patterns across differing agents and environments, demonstrating robust transferability.

5.4. Results on CALVIN

In-domain latent pre-training on CALVIN. To demonstrate that our latent codes aid in learning manipulation skills, we conduct a latent pre-training experiment, following LAPA. We first pre-train the AR transformer on latent codes from 22k CALVIN trajectories, then fine-tune it on only 2k ground-truth action labels. As shown in Tab. 2, this pre-training strategy significantly boosts VideoWorld 2’s performance (Idx 4) on long-horizon tasks compared to the baseline (Idx 2). Notably, its performance approaches the model trained on the 22k full action label dataset (Idx 1), demonstrating high data efficiency. While LAPA (Idx 3) also benefits from this protocol, it still lags behind VideoWorld 2, highlighting the effectiveness of our latent codes.

Cross-domain latent pre-training on CALVIN. We further evaluate transferability by pre-training on the 1.3M OpenX dataset and fine-tuning on 22k CALVIN trajectories. As shown in Tab. 2, this pre-training significantly boosts success rates for LAPA (Idx 6) and VideoWorld 2 (Idx 7) compared to training only on labeled CALVIN data (Idx 1). Consistent with previous findings, VideoWorld 2 benefits more from this pre-training on long-horizon tasks. We also include a video next-token prediction baseline pre-trained on OpenX (Idx 5). The results show that our latent pre-training is more effective, validating it as a more efficient knowledge transfer paradigm than pre-training directly on

Pre-trained VDM	Decoder Stop-Grad	Ctrl- Net	Video-CraftBench		
			Paper	Block	LPIPS↓
			0.0	28.5	0.312
✓			30.3	45.2	0.297
✓	✓		47.3	54.7	0.275
✓		✓	51.1	52.0	0.213
✓	✓	✓	68.8	77.5	0.205

(a) dLDM architecture.

dLDM Codebook size	Video-CraftBench			CALVIN Avg.Len.
	Paper	Block	LPIPS↓	
baseline	0.0	28.5	0.312	1.11
8	20.1	29.9	0.258	1.65
1000	68.8	77.5	0.205	1.87
4096	50.4	59.6	0.208	1.90
64,000	29.4	36.0	0.230	1.89

(c) Different codebook size of dLDM.

Query length N	Video-CraftBench			CALVIN Avg.Len.
	Paper	Block	LPIPS↓	
baseline	0.0	28.5	0.312	1.11
1	41.9	65.0	0.221	1.53
2	55.1	69.5	0.210	1.64
4	68.8	77.5	0.205	1.87
8	65.0	77.5	0.195	1.88

(b) Different query embeddings lengths on Video-CraftBench.

Compression Length T	Video-CraftBench			CALVIN Avg.Len.
	Paper	Block	LPIPS↓	
2	19.1	38.7	0.267	1.55
9	55.4	68.7	0.231	1.61
49	65.3	76.2	0.208	1.80
93	68.8	77.5	0.205	1.87
177	69.0	76.8	0.190	1.79

(d) Different compression length on Video-CraftBench.

Table 3. **Ablation studies.** We train with Video-CraftBench only and evaluate CALVIN in its in-domain setting.

videos. Supp. A provides more training details.

5.5. Ablation Study

Code transferability is enhanced by VDM. Tab. 3a (rows 1–2) first ablates the VDM prior. Incorporating VDM enhances code transferability and visual quality, yielding a $\sim 30\%$ increase in success rate and improved LPIPS scores. This validates our decoupling design. To illustrate this, Fig. A.6 visualizes latent codes from 4,000 CALVIN and Bridge (included in OpenX) trajectories [58] via UMAP, labeled by robot action (up, down, left, right). With the VDM (Fig. A.6, bottom), latent codes for the same action align much more tightly across environments, indicating more consistent and transferable dynamics. Conversely, without the VDM (Fig. A.6, top), codes separate strongly by environment, revealing a loss of cross-environment consistency.

Effect of the original decoder. In Sec.3 and Supp. A, we first warm up dLDM with the original decoder during training. Subsequently, we discard the VQ-VAE reconstruction loss to prevent noise injection. However, we observe that after warm-up, the decoder can reconstruct latent codes into videos that preserve coherent object motion (e.g., hand movements and displacements) despite blurry details. We therefore utilize these reconstructions as VDM conditions through a ControlNet-like structure to enhance output quality and stabilize training. Tab.3a further investigates this.

First, we stop the gradient from the decoder without using the reconstructed video (row 3) to verify if this decoder introduces noise. Compared to row 2, this yields a $\sim 20\%$ success rate gain, suggesting that the original decoder indeed introduces noise that degrades latent representation. Second, incorporating the reconstructed video as conditions (rows 4–5) further stabilizes outputs, improving LPIPS by ~ 0.9 and success rates by $\sim 20\%$. This benefit is more pronounced in paper folding than block stacking, highlighting our method’s efficacy for long-horizon video generation.

Query embedding length of dLDM. Tab. 3b ablates the number of dLDM query embeddings N . A larger N cap-

tures more information but also risks encoding noise and increases the sequence length for the AR transformer. We find that $N = 1$ already achieves a respectable task success rate, while $N = 4$ provides the best performance balance. Increasing N to 8 offers a slight LPIPS benefit but introduces more noise, which reduces the overall success rate.

dLDM codebook size. Tab. 3c examines the impact of the dLDM codebook size (adjusted via FSQ levels) on policy learning. We find that CALVIN, with its simpler action space, benefits from a relatively small codebook. In contrast, the more complex Video-CraftBench requires a larger codebook to achieve significant improvements. However, an excessively large codebook risks encoding extraneous noise and can hinder dLDM training convergence, a finding consistent with VideoWorld [51].

Horizon length in dLDM. Tab. 3d ablates the dLDM context length T . At $T = 2$ (a LAPA-like setting), codes remain transferable, but long-range performance on paper folding is poor due to insufficient temporal perception. This minimal context also provides low-quality motion guidance for the VDM, leading to poor LPIPS scores. Performance improves as T increases, plateauing at $T = 93$, matching the maximum context length of the Cosmos VDM.

6. Conclusion

This work explores knowledge learning for complex, long-horizon tasks from raw videos through experiments on Video-CraftBench and robotic manipulation environments. We find that decoupling visual appearance from core actions is crucial for real-world knowledge learning. Based on this finding, we propose VideoWorld 2, a model featuring a dynamics-enhanced latent dynamics model (dLDM) that leverages a pre-trained VDM to learn generalizable, transferable policies directly from video. While our method shows significant potential, we leave its continued scaling to future work, moving toward the goal of enabling AI to learn the vast knowledge encapsulated in the real world.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.92470203, U23A20314), and Beijing Natural Science Foundation (No. L242022). Xiaojie Jin’s work was supported by the Talent Fund of Beijing Jiaotong University under Grant No. 2025XKRC015.

References

- [1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *CoRR*, 2025. 2, 3, 4, 6
- [2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 1
- [3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zhohus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 3
- [4] Junyeob Baek, Yi-Fu Wu, Gautam Singh, and Sungjin Ahn. Dreamweaver: Learning compositional world models from pixels. *arXiv preprint arXiv:2501.14174*, 2025. 3
- [5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1
- [6] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6
- [7] Federico Baldassarre, Marc Szafraniec, Basile Terver, Vasil Khalidov, Francisco Massa, Yann LeCun, Patrick Labatut, Maximilian Seitzer, and Piotr Bojanowski. Back to the features: Dino as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025. 3
- [8] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. 3
- [9] Juan Luis Gonzalez Bello, Xu Yao, Alex Whelan, Kyle Olaszewski, Hyeonwoo Kim, and Pablo Garrido. Videospats: Video spatiotemporal splines for disentangled occlusion, appearance and motion modeling and editing. *arXiv preprint arXiv:2504.07146*, 2025. 2
- [10] 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. 1
- [11] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3
- [12] Maxime Burchi and Radu Timofte. Accurate and efficient world modeling with masked latent transformers. *arXiv preprint arXiv:2507.04075*, 2025. 3
- [13] Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024. 3
- [14] Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. Villa-x: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025. 3
- [15] Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *ICCV*, pages 19752–19763, 2025. 3, 6, 7
- [16] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. 1
- [17] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020. 3
- [18] Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blisstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1
- [19] Yufeng Cui, Honghao Chen, Haoge Deng, Xu Huang, Xinghang Li, Jirong Liu, Yang Liu, Zhuoyan Luo, Jinsheng Wang, Wenxuan Wang, et al. Emu3. 5: Native multimodal models are world learners. *arXiv preprint arXiv:2510.26583*, 2025. 3
- [20] DeepMind. Veo – deepmind’s video evolution model. <https://deepmind.google/models/veo/>, 2025. Accessed: November 11, 2025. 3
- [21] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 3
- [22] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025. 3
- [23] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel.

- Learning universal policies via text-guided video generation. *NeurIPS*, 36:9156–9172, 2023. 3
- [24] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025. 3, 6, 7
- [25] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024. 3
- [26] Yuying Ge, Yizhuo Li, Yixiao Ge, and Ying Shan. Divot: Diffusion powers video tokenizer for comprehension and generation. *arXiv preprint arXiv:2412.04432*, 2024. 2
- [27] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 3
- [28] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.
- [29] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3
- [30] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. 3
- [31] William Huitt and John Hummel. Piaget’s theory of cognitive development. *Educational psychology interactive*, 3(2): 1–5, 2003. 1
- [32] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025. 3
- [33] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 1
- [34] Efsthios Karypidis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. Dino-foresight: Looking into the future with dino. *arXiv preprint arXiv:2412.11673*, 2024. 3
- [35] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3
- [36] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 3, 4, 6
- [37] Maria Krinner, Elie Aljalbout, Angel Romero, and Davide Scaramuzza. Accelerating model-based reinforcement learning with state-space world models. *arXiv preprint arXiv:2502.20168*, 2025. 3
- [38] Pulkit Kumar, Namitha Padmanabhan, Luke Luo, Sai Saketh Rambhatla, and Abhinav Shrivastava. Trajectory-aligned space-time tokens for few-shot action recognition. In *ECCV*, pages 474–493. Springer, 2024. 2
- [39] Ruineng Li, Daitao Xing, Huiming Sun, Yuanzhou Ha, Jinglin Shen, and Chiuman Ho. Tokenmotion: Decoupled motion control via token disentanglement for human-centric video generation. In *CVPR*, pages 1951–1961, 2025. 2
- [40] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *Conference on Robot Learning*, pages 3705–3728. PMLR, 2025. 3
- [41] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024. 3
- [42] Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023. 3
- [43] Haoyu Ma, Jialong Wu, Ningya Feng, Chenjun Xiao, Dong Li, Jianye Hao, Jianmin Wang, and Mingsheng Long. Harmonydream: Task harmonization inside world models. *arXiv preprint arXiv:2310.00344*, 2023. 3
- [44] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2
- [45] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7327–7334, 2022. 2, 4
- [46] OpenAI. Sora 2: Advancing video generation models. <https://openai.com/index/sora-2/>, 2025. Accessed: November 11, 2025. 3
- [47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, 2024. 5
- [48] Abby O’Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903, 2024. 2, 7
- [49] J Parker-Holder, P Ball, J Bruce, V Dasagi, K Holzheimer, C Kaplanis, A Moufarek, G Scully, J Shar, J Shi, et al. Genie 2: A large-scale foundation world model. URL: <https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model>, 2024. 3
- [50] Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *arXiv preprint arXiv:2505.20171*, 2025. 3

- [51] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *CVPR*, pages 29029–29039, 2025. 1, 2, 5, 6, 7, 8, 4
- [52] Dominik Schmidt and Minqi Jiang. Learning to act without actions. *arXiv preprint arXiv:2312.10812*, 2023. 3
- [53] Yifan Shen, Peiyuan Zhu, Zijian Li, Shaoan Xie, Namrata Deka, Zongfang Liu, Zeyu Tang, Guangyi Chen, and Kun Zhang. Controllable video generation with provable disentanglement. *arXiv preprint arXiv:2502.02690*, 2025. 2
- [54] Qiao Sun, Liujia Yang, Wei Tang, Wei Huang, Kaixin Xu, Yongchao Chen, Mingyu Liu, Jiange Yang, Haoyi Zhu, Yating Wang, et al. Learning primitive embodied world models: Towards scalable robotic learning. *arXiv preprint arXiv:2508.20840*, 2025. 3
- [55] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 3
- [56] Bahey Tharwat, Yara Nasser, Ali Abouzeid, and Ian Reid. Latent action pretraining through world modeling. *arXiv preprint arXiv:2509.18428*, 2025. 3
- [57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 1
- [58] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023. 8, 4
- [59] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwei Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 4, 6
- [60] Jiacong Wang, Bohong Wu, Haiyong Jiang, Zhou Xun, Xin Xiao, Haoyuan Guo, and Jun Xiao. World to code: Multimodal data generation via self-instructed compositional captioning and filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4608–4623, 2024. 1
- [61] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, pages 5264–5273, 2020. 2
- [62] Yucen Wang, Fengming Zhang, De-Chuan Zhan, Li Zhao, Kaixin Wang, and Jiang Bian. Co-evolving latent action world models. *arXiv preprint arXiv:2510.26433*, 2025. 6, 7
- [63] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6
- [64] Zizhao Wang, Kaixin Wang, Li Zhao, Peter Stone, and Jiang Bian. Dyn-o: Building structured world models with object-centric representations. *arXiv preprint arXiv:2507.03298*, 2025. 3
- [65] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *NeurIPS*, 37:68082–68119, 2024. 6
- [66] Jianwen Xie, Ruiqi Gao, Zilong Zheng, Song-Chun Zhu, and Ying Nian Wu. Motion-based generator model: Unsupervised disentanglement of appearance, trackable and intrackable motions in dynamic patterns. In *AAAI*, pages 12442–12451, 2020. 2
- [67] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on robot learning*, pages 3536–3555. PMLR, 2023. 3
- [68] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1
- [69] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- [70] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, SeJune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. 3, 6, 7, 1
- [71] 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. 3
- [72] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video. *arXiv preprint arXiv:2504.21853*, 2025. 3
- [73] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025. 3
- [74] Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *ICLR*, 2024. 4
- [75] Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, et al. Lumos-1: On autoregressive video generation from a unified model perspective. *arXiv preprint arXiv:2507.08801*, 2025. 3
- [76] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, pages 3836–3847, 2023. 5
- [77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep

- features as a perceptual metric. In *CVPR*, pages 586–595, 2018. [6](#)
- [78] Long Zhao, Sanghyun Woo, Ziyu Wan, Yandong Li, Han Zhang, Boqing Gong, Hartwig Adam, Xuhui Jia, and Ting Liu. Epsilon-vae: Denoising as visual decoding. *arXiv preprint arXiv:2410.04081*, 2025. [2](#)
- [79] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023. [3](#)
- [80] Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025. [3](#)
- [81] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. [3](#)
- [82] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. [3](#)