PhiZero: A World Model Built Around Physical Language

Shuyao Shang **Yuqi Wang**[†§] **Ruopeng Gao**[§] **Xu Chen**[§]
Tieniu Tan **Lue Fan**[✉] **Zhaoxiang Zhang**[✉]

NLPR, Institute of Automation, Chinese Academy of Sciences (CASIA)

https://Phi-Zero.github.io/

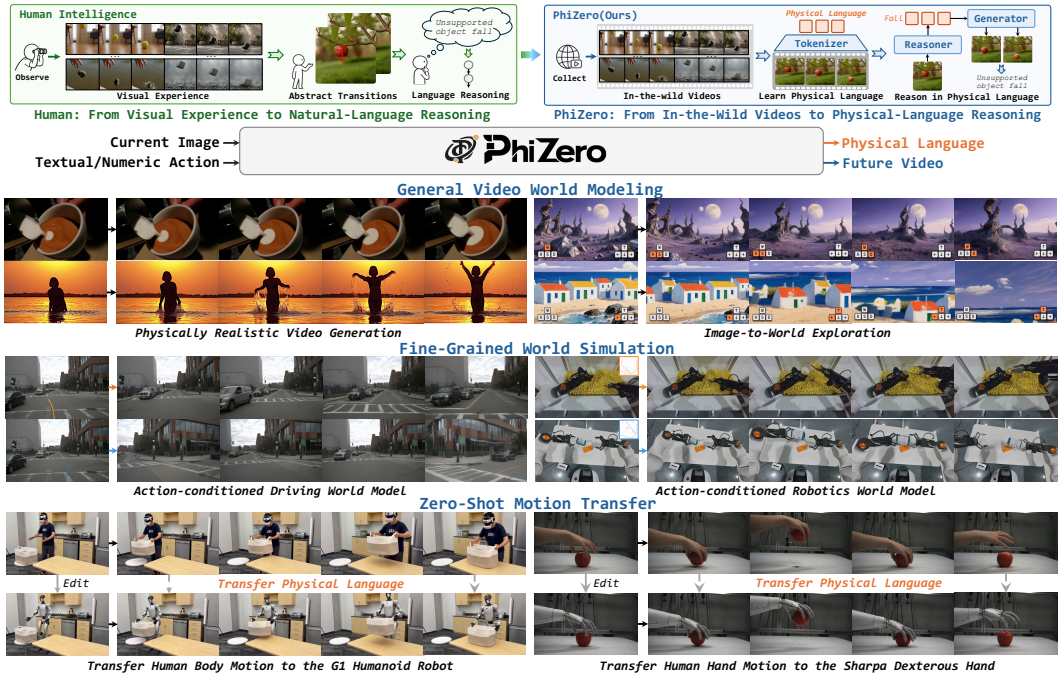


Figure 1: We introduce **PhiZero**, a physical world model that learns a compact discrete *physical language* from in-the-wild videos. It reasons world evolution in this space and renders the inferred transitions into videos, enabling physically realistic and interactive world modeling, fine-grained action-conditioned simulation, and zero-shot motion transfer.

Abstract

We introduce PhiZero, a physical world model built around *physical language*, a compact discrete representation of world-state transitions. Existing physical world models typically predict future videos directly in pixel space, leaving the underlying world dynamics implicit within high-dimensional visual predictors. Motivated by humans' ability to abstract predictive structure from visual experience and organize it in natural language for explicit reasoning, we learn physical language from in-the-wild videos through self-supervision and use it to explicitly reason about how the physical world evolves. Accordingly, PhiZero adopts a reason-then-render paradigm: it first infers future world evolution as a physical-language sequence and then renders the inferred transitions into videos. Extensive experiments across generation and understanding benchmarks validate the ability of PhiZero to model physically coherent world evolution. We further show its potential for realistic and interactive world modeling, fine-grained action-conditioned simulation, and zero-shot motion transfer.

[†]Project Leader. [✉]Corresponding Authors. [§]Independent Researcher.

1 INTRODUCTION

"You see, but you do not observe."
— Sherlock Holmes, *A Scandal in Bohemia*

Video world models can generate future videos with high visual fidelity and are therefore increasingly regarded as promising simulators of the physical world (OpenAI, 2025; Google DeepMind, 2026). For such models to support Physical AI, their central objective must go beyond visual fidelity toward learning how the physical world evolves. However, mainstream approaches are trained primarily through direct prediction in pixel space, leaving the underlying dynamics implicit within high-dimensional visual representations and often resulting in physically inconsistent outcomes.

We draw inspiration from human intelligence. Rather than memorizing individual visual outcomes, humans abstract patterns from visual experience into generalizable knowledge about how the world evolves. Natural language serves as a primary medium for organizing and expressing such knowledge, providing a symbolic space in which humans can explicitly reason. The success of language models across digital domains further demonstrates the scalability of language as a substrate for learning and reasoning. However, when the modeling target shifts from the digital world to the physical world, natural language is often too coarse to faithfully represent the complex state transitions observed in visual experience. We therefore ask: Can we move beyond natural language and learn a finer-grained representation of physical-world transitions that enables explicit reasoning?

To this end, we propose **PHIZERO**, a physical world model built around *physical language*[*]. This language is learned from unlabeled in-the-wild videos through self-supervised learning and captures state-transition patterns across diverse visual experiences (Fig. 1). Once learned, physical language provides an intermediate representation for inferring world evolution before rendering it into video. This reason-then-render paradigm separates dynamics inference from pixel-level synthesis, making world evolution an explicit reasoning target rather than a direct pixel-space prediction.

Learning a physical language that captures how the world evolves requires disentangling world dynamics from visual appearance. We therefore develop a Physical Language Tokenizer that learns this representation through self-supervised video reconstruction. Given a video, the tokenizer encodes its state transitions into a discrete physical-language sequence, which, together with the first frame, conditions a pretrained diffusion decoder to reconstruct the video. The decoder's pretrained generative prior recovers fine-grained visual details, while the first frame anchors static scene appearance, allowing the discrete bottleneck to focus on state changes rather than redundant appearance information. We further introduce a Physical Language Reasoner initialized from a pretrained VLM. Leveraging the VLM's visual-semantic knowledge and commonsense priors, the reasoner autoregressively predicts a physical-language sequence from the current visual state and action intent. The predicted sequence is then rendered into video by the diffusion decoder.

Extensive experiments validate PHIZERO across video-generation benchmarks assessing physical fidelity and video-understanding benchmarks assessing physical plausibility. We further show that physical language provides a general-purpose interface for representing, controlling, and transferring state transitions in the physical world, supporting interactive rollouts under sequential control inputs, fine-grained action-conditioned world modeling, and zero-shot motion transfer (Fig. 1). Together, these results position PHIZERO as a promising framework for controllable and transferable physical-world modeling, with broad potential for Physical AI.

Our main contributions are summarized as follows:

- We introduce *physical language*, a compact discrete representation of state-transition patterns that can be learned at scale from in-the-wild videos through self-supervised learning.
- We develop PHIZERO, a physical world model built around physical language. PHIZERO adopts a reason-then-render paradigm, first inferring future world evolution in physical-language space and then rendering the inferred transitions into videos.
- Extensive experiments validate PHIZERO across both physical generation and understanding. We further demonstrate its potential for interactive rollouts, action-conditioned world simulation, and zero-shot motion transfer.

[*]Here, *physical* refers broadly to physical world, rather than to a symbolic system of physical laws.

2 Related Work

2.1 Physical World Models

Recent work has increasingly developed physical world models by scaling up video generation and prediction to simulate future visual states under language or action conditions. Yang et al. (2023) learns interactive simulators from heterogeneous real-world data, while Wang et al. (2024); Xiang et al. (2024) model world evolution through discrete visual prediction and language-conditioned video generation. Agarwal et al. (2025); Ali et al. (2025); Agarwal et al. (2026) further advance this paradigm through pretrained world foundation models, unified multimodal conditioning, and omnimodal modeling for Physical AI. Other systems improve long-horizon simulation through autoregressive denoising or latent-state prediction followed by video rendering (Zhu et al., 2025; Xiang et al., 2025). These approaches primarily represent world evolution in pixel spaces. In contrast, PhiZero explicitly represents state transitions as a compact discrete physical language, reasons about world evolution in this transition space, and then renders the inferred evolution into video.

2.2 Learning World Dynamics from Unlabeled Videos

A broad line of work has investigated how predictive and actionable representations of world dynamics can be learned from unlabeled videos. Bruce et al. (2024); Gao et al. (2025) discover controllable latent action spaces from observed frame-to-frame changes in Internet videos. In robotics and autonomous driving, Schmidt & Jiang (2024); Ye et al. (2025); Chen et al. (2025); Bu et al. (2025b); Shang et al. (2026) learn latent actions and transfer the resulting motion priors to downstream control tasks. Wu et al. (2023; 2024) learn latent dynamics that capture temporal changes transferable across scenes. Bardes et al. (2024); Assran et al. (2025) predict masked spatiotemporal regions in representation space, demonstrating that large-scale video pretraining can yield representations useful for understanding and planning in the physical world. Ren et al. (2025; 2026) further investigate whether planning capabilities can be acquired directly from task-oriented videos. Unlike these approaches, we learn a compact physical language that captures state-transition patterns in the physical world, rather than latent actions tied to specific domains or tasks.

3 Method

As shown in Fig. 2, PhiZero comprises two complementary components. A Physical Language Tokenizer compresses the state transitions in a video into a discrete physical-language sequence, while a diffusion decoder learns to render the future video conditioned on the first frame and the extracted physical language (Sec. 3.2). A Physical Language Reasoner then predicts the corresponding physical-language sequence from the first frame and a textual action intent (Sec. 3.3).

3.1 Problem Formulation

Formally, let $\mathbf{V}$ denote the future video, I^0 the first frame representing the current world state, c the textual action intent, and $\mathbf{z}$ the discrete physical language describing the transition from I^0 to $\mathbf{V}$. Physical language serves as a compact intermediate representation of state transitions that captures shared patterns in how the world evolves. We formulate future prediction as the joint modeling of the physical-language sequence and the future video, which factorizes into physical-language reasoning and video rendering:

$$\underbrace{p_{\theta,\psi}\left(\mathbf{V},\mathbf{z} \mid I^0, c\right)}_{\text{PhiZero}} = \underbrace{p_\theta\left(\mathbf{z} \mid I^0, c\right)}_{\substack{\text{Physical-language}\\ \text{Reasoning}}} \underbrace{p_\psi\left(\mathbf{V} \mid I^0, \mathbf{z}\right)}_{\substack{\text{Future-video}\\ \text{Rendering}}}. \tag{1}$$

The Physical Language Reasoner infers how the current state should evolve in the physical-language space, after which the diffusion decoder renders the inferred transition into a future video. This reason-then-render decomposition separates state-transition reasoning from pixel-level synthesis.

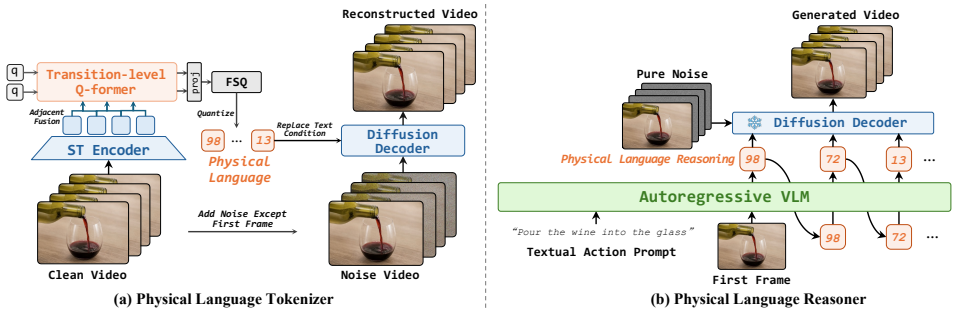


Figure 2: **The overall pipeline of PhiZero.** (a) The Physical Language Tokenizer uses a transition-level Q-Former to compress video state transitions into a compact discrete physical language. A diffusion decoder reconstructs the video conditioned on the physical language and the first frame. (b) The Physical Language Reasoner uses an autoregressive VLM to predict a physical-language sequence from the first frame and a textual action intent. The trained diffusion decoder then renders the inferred state transition into a future video.

3.2 Physical Language Tokenizer

Encoder with Transition-level Q-Former Given a video $\mathbf{V} \in \mathbb{R}^{B\times 3\times T\times H\times W}$, we first employ a spatiotemporal encoder to obtain latent features $\mathbf{x} \in \mathbb{R}^{B\times C\times t\times h\times w}$. Rather than extracting a global representation from the entire video, we explicitly model the transition between each pair of adjacent latent states. This introduces a local temporal inductive bias that reduces the complexity of each compressed transition while preserving the temporal ordering of the full sequence. Let $\mathbf{x}^i \in \mathbb{R}^{B\times C\times h\times w}$ denote the latent state at time step i. For each adjacent pair $(\mathbf{x}^i, \mathbf{x}^{i+1})$, we use a shared Q-Former (Li et al., 2023) to extract a transition representation:

$$\mathbf{q}_i = \text{QFormer}\left(\mathbf{Q}; \mathbf{x}^i, \mathbf{x}^{i+1}\right), \tag{2}$$

where $\mathbf{Q} \in \mathbb{R}^{M\times D_q}$ contains M shared learnable transition queries. By jointly attending to the two adjacent latent states, the Q-Former produces M transition features $\mathbf{q}_i \in \mathbb{R}^{B\times M\times D_q}$ for each temporal interval. We concatenate the features from all adjacent intervals in temporal order to obtain $\mathbf{q} \in \mathbb{R}^{B\times N\times D_q}$, where $N = (t-1)M$. We then discretize the transition features using finite scalar quantization (FSQ) (Mentzer et al., 2024). Specifically, the features are projected into a low-dimensional quantization space and quantized as

$$\mathbf{z} = \text{FSQ}\left(\text{Proj}_{\text{down}}(\mathbf{q})\right), \tag{3}$$

where $\mathbf{z}$ denotes the resulting discrete physical-language sequence. FSQ constructs its vocabulary as the Cartesian product of scalar quantization levels and therefore requires no separately learned codebook. Finally, the quantized representation $\mathbf{z}$ is projected to the hidden dimension d of the diffusion transformer, yielding the physical-language context $\mathbf{P}_c \in \mathbb{R}^{B\times N\times d}$.

Diffusion-prior Decoder A limited-capacity physical-language bottleneck should focus on representing state transitions rather than encoding all fine-grained appearance details required for video reconstruction. We therefore employ a pretrained video diffusion model (Wan et al., 2025) as the decoder, whose strong generative prior can recover realistic appearance and visual details from a compact representation of state transitions. To preserve this pretrained generative prior, we retain the original model architecture and replace only its text condition with the physical-language context $\mathbf{P}_c$. We further provide the first frame I^0 as a clean visual condition and treat it as the source of static appearance. By supplying appearance information directly through the first frame, the discrete bottleneck is encouraged to encode future state changes rather than redundant static visual details. Conditioned on the physical-language context $\mathbf{P}_c$, the first frame I^0, the clean latent $\mathbf{x}_0$, and the diffusion timestep $\tau \sim \mathcal{U}(0,1)$, the decoder reconstructs the target video using the standard flow-matching (Liu et al., 2023b) objective:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x},\,\boldsymbol{\epsilon},\,\tau}\left[\left\|v_\psi(\mathbf{x}_\tau, \tau;\, I^0, \mathbf{P}_c) - (\boldsymbol{\epsilon} - \mathbf{x}_0)\right\|_2^2\right], \quad \mathbf{x}_\tau = (1-\tau)\mathbf{x}_0 + \tau\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

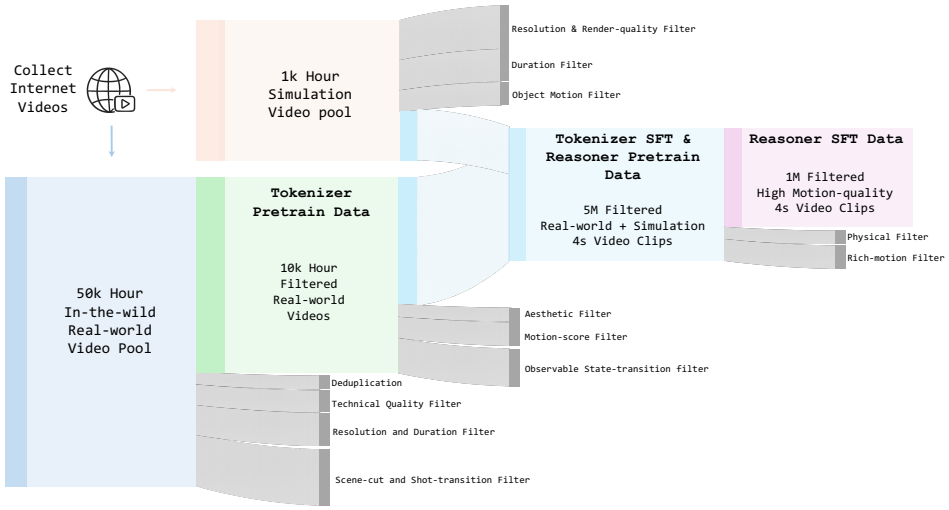


Figure 3: **Hierarchical data curation pipeline.** We collect Internet videos to construct a 50K-hour pool of in-the-wild real-world videos and a 1K-hour pool of simulation videos. Progressive filtering yields 10K hours for tokenizer pretraining, 5M four-second clips for tokenizer SFT and reasoner pretraining, and 1M motion-rich four-second clips for reasoner SFT.

Pure-noise Warm-up The pretrained diffusion decoder may initially ignore the newly introduced physical-language context by relying on partially corrupted target information and its existing denoising prior. To prevent this shortcut, we introduce a pure-noise warm-up stage in which all future-frame latents are initialized from pure noise. The decoder must therefore rely on the physical-language context and the first frame to reconstruct the future video. After warm-up, we restore the standard flow-matching noise schedule. This strategy mitigates the denoising shortcut and provides stronger supervision for learning informative physical-language representations.

Data Curation Pipeline We train the Physical Language Tokenizer in two stages: large-scale pretraining followed by targeted SFT. As shown in Fig. 3, for pretraining, we begin with a real-world video pool containing approximately 50K hours of footage. We remove duplicates and filter out clips with technical defects, including compression artifacts, corrupted frames, and watermarks, as well as those that fail the resolution or duration requirements or contain abrupt shot transitions. This process yields approximately 10K hours of unlabeled videos for tokenizer pretraining. For the subsequent SFT stage, we apply a stricter second-pass filtering procedure based on aesthetic quality, motion magnitude, and the observability of state transitions as assessed by a VLM judge. In parallel, we curate approximately 1K hours of simulated videos by removing samples with low rendering quality, invalid duration, or insufficient object motion. Combining the resulting real-world and simulation data produces approximately 5M four-second video clips for tokenizer SFT. Additional data details are provided in the Appendix A.1.

Curriculum Training Recipe We couple this two-stage training scheme with a joint temporal and spatial curriculum. During pretraining, we process videos at a resolution of 256×448 and progressively increase the clip duration from 1 second to 2 seconds and finally to 4 seconds. This allows the tokenizer to first learn local state changes before modeling longer-range world evolution. We then increase the reconstruction resolution to 512×896 and perform SFT on the curated 5M-clip corpus. Finally, we freeze the tokenizer and refine only the diffusion decoder to further improve reconstruction quality. Additional training details are provided in Appendix A.1.

3.3 Physical Language Reasoner

Physical Language Reasoning with a Pretrained VLM Predicting future state transitions requires interpreting a textual action intent in the context of the current visual state and inferring how the world should evolve. Pretrained VLMs encode rich visual semantics and commonsense world knowledge from large-scale image-text data, providing a strong prior for this task. We thus initial-

ize the Physical Language Reasoner from a pretrained VLM (Bai et al., 2025a). To adapt it to the physical-language space learned by the Physical Language Tokenizer, we extend its vocabulary with a distinct atomic symbol for each FSQ index. Given a first frame I^0 and a text prompt c, the Physical Language Reasoner autoregressively predicts a length-N physical-language sequence over the FSQ-induced vocabulary of size K. Its conditional distribution is factorized in temporal order as

$$p_\theta(\mathbf{z} \mid I^0, c) = \prod_{j=1}^{N} p_\theta(z_j \mid I^0, c, z_{<j}). \tag{5}$$

We generate supervision offline by encoding each training video with the frozen Physical Language Tokenizer, $\mathbf{z} = \mathcal{T}_\phi(\mathbf{V})$. Under teacher forcing, the Physical Language Reasoner is optimized with the autoregressive cross-entropy objective

$$\mathcal{L}_{\text{VLM}} = -\sum_{j=1}^{N} \log p_\theta(z_j \mid I^0, c, z_{<j}). \tag{6}$$

Data Curation Pipeline We construct the training data for the Physical Language Reasoner in two stages of increasing specialization. As shown in Fig. 3, for continued pretraining, we reuse the approximately 5M curated four-second real-world and simulation clips used for Physical Language Tokenizer SFT. For each clip, a VLM-generated caption and the first frame serve as inputs, while the frozen Physical Language Tokenizer provides the target physical-language sequence. To prevent the textual condition from revealing the outcome, as shown in Table 14, we prompt the VLM to summarize only the high-level initiating action or interaction in each clip, rather than narrating fine-grained transition details. To construct the subsequent SFT corpus, we further filter this general dataset using a VLM-based rich-motion filter and a physical filter, retaining clips that exhibit salient state changes and physically informative interactions. Together with curated simulator-generated samples, this process yields approximately 1M motion-rich and physically informative four-second clips for reasoner SFT. Additional data details are provided in the Appendix A.2.

Two-stage Training Recipe We train the Physical Language Reasoner in two stages with progressively specialized data distributions. In Stage 1, we initialize the model from a pretrained VLM (Bai et al., 2025a) and perform continued pretraining on the 5M-clip general corpus. This stage adapts the VLM to autoregressive physical-language prediction and establishes the correspondence among textual intent, the current visual state, and the resulting state transition. In Stage 2, we perform SFT on a corpus of approximately 1M motion-rich and physically informative clips. By concentrating training on interactions with salient motion and physically meaningful consequences, this stage improves the physical plausibility and precision of the inferred state transitions while preserving the broad capability for modeling world transitions acquired during continued pretraining. Additional training details are provided in the Appendix A.2.

4 Experiments

4.1 Implementation Details

For the Physical Language Tokenizer, the spatiotemporal encoder follows the Wan2.2 VAE encoder architecture (Wan et al., 2025) and is initialized from its pretrained weights, while the diffusion decoder is initialized from Wan2.2-5B (Wan et al., 2025). We optimize all parameters of the spatiotemporal encoder and transition-level Q-Former, while fine-tuning the diffusion decoder using LoRA (Hu et al., 2022) with rank 32. We configure FSQ with scalar quantization levels $(8, 5, 5, 5, 5, 5)$, yielding a vocabulary of 25K discrete symbols, and use 32 learnable queries in the transition-level Q-Former. A 33-frame video is encoded into nine temporal latent states; extracting 32 transition symbols from each adjacent pair therefore produces a physical-language sequence of length $(9 - 1) \times 32 = 256$. For the Physical Language Reasoner, we initialize the model from the pretrained Qwen3-VL-4B (Bai et al., 2025a) and extend its vocabulary with one atomic symbol for each FSQ index. Additional training configurations and data details are provided in Appendix A.

Table 1: **Physical outcome fidelity on Physics-IQ Verified.**

Model	S-IoU ↑	ST-IoU ↑	WS-IoU ↑	IQ-Score ↑
Wan2.2-5B (Wan et al., 2025)	24.7	22.6	13.3	21.2
Sora 2 (OpenAI, 2025)	37.3	27.0	26.9	26.5
Cosmos3-Nano (Agarwal et al., 2026)	40.4	22.0	24.6	29.1
Wan2.2-14B (Wan et al., 2025)	51.1	20.5	28.5	32.2
Hunyuan-Video (Kong et al., 2024)	47.1	26.9	29.7	33.4
Grok-Video (xAI, 2026)	52.7	21.4	**35.7**	34.8
Cosmos3-Super (Agarwal et al., 2026)	–	–	–	39.5
PhiZero (Ours)	**58.2**	**36.8**	27.6	**41.2**

Table 2: **Physical-law adherence on PhyGround.**

Model	General Quality ↑	Physics Score ↑	Overall ↑
LTX-Video-19B (HaCohen et al., 2026)	2.56	2.54	2.55
LTX-Video-22B (HaCohen et al., 2026)	2.59	2.56	2.57
Wan2.2-5B (Wan et al., 2025)	2.67	2.65	2.66
Cosmos2.5-2B (Gu, 2025)	2.79	2.70	2.74
OmniWeaving (Pan et al., 2026)	2.89	2.78	2.84
Veo3.1 (Google DeepMind, 2026)	**3.01**	2.85	2.93
Wan2.2-14B (Wan et al., 2025)	3.00	2.90	2.95
PhiZero (Ours)	2.93	**3.01**	**2.97**

Table 3: **General world modeling on WorldModelBench.**

Model	Physics Adherence ↑	Common Sense ↑	Total ↑
Pandora (Xiang et al., 2024)	4.05	0.95	6.57
OpenSora-Plan (Lin et al., 2024)	3.85	1.01	6.62
CogVideoX (Yang et al., 2025c)	3.88	0.99	6.75
Mochi (Genmo AI, 2025)	4.14	1.26	7.62
Luma (Luma AI, 2024)	4.13	1.57	7.72
Wan2.2-5B (Wan et al., 2025)	4.51	1.41	7.98
Runway (Runway ML, 2024)	4.27	1.65	8.08
PhiZero (Ours)	**4.88**	**1.71**	**8.19**

Table 4: **Intuitive physics understanding on IntPhys2.**

Model	Easy ↑	Medium ↑	Hard ↑	Overall ↑
Cosmos-4B (Agarwal et al., 2025)	46.00	52.00	48.05	49.41
Qwen2.5-VL (Bai et al., 2025b)	50.96	53.25	51.49	52.27
Gemini-1.5 Pro (Gemini Team, 2023)	58.65	53.00	52.67	52.27
GPT-4o (Achiam et al., 2023)	57.69	54.75	54.17	53.75
VideoMAEv2 (Wang et al., 2023b)	46.00	58.50	52.73	53.75
V-JEPA (Bardes et al., 2024)	52.00	53.00	**57.42**	53.75
Gemini-2.5 Flash (Gemini Team, 2023)	**64.42**	56.75	54.46	55.63
PhiZero (Ours)	60.98	**60.50**	52.38	**56.34**

Table 5: **Physical-plausibility discrimination on LikePhys.**

Model	Rigid ↓	Fluid ↓	Optical ↓	Avg. Error ↓
AnimateDiff-SDXL (Guo et al., 2023)	61.44	53.33	37.65	56.0
ZeroScope (Wang et al., 2023a)	57.32	49.23	41.00	53.3
ModelScope (Wang et al., 2023a)	59.46	47.23	35.65	52.9
Mochi (Genmo AI, 2025)	54.22	47.33	49.15	51.9
CogVideoX-5B (Yang et al., 2025c)	59.00	43.10	26.65	49.8
CogVideoX-2B (Yang et al., 2025c)	56.98	42.00	**25.15**	48.2
Wan2.1-1.3B (Wan et al., 2025)	44.66	57.10	41.65	48.0
LTX-Video-2B (HaCohen et al., 2026)	43.44	**33.43**	61.50	44.7
CogVideoX1.5-5B (Yang et al., 2025c)	44.14	47.90	42.65	43.8
Wan2.1-14B (Wan et al., 2025)	43.34	45.00	38.35	43.8
Hunyuan-Video (Kong et al., 2024)	36.34	48.67	41.15	43.6
PhiZero (Ours)	**29.14**	53.15	37.50	**41.7**

Table 6: **Real-world causal understanding on YoCausal.**

Model	RSI (%) ↑	CCI (%) ↑	Agg. Rank ↓
CogVideoX-2B (Yang et al., 2025c)	41.50	0.93	10.0
Wan2.2-5B (Wan et al., 2025)	51.91	-2.12	9.0
Hunyuan-Video (Kong et al., 2024)	52.05	-0.29	8.0
Wan2.1-1.3B (Wan et al., 2025)	45.51	5.36	7.5
Mochi (Genmo AI, 2025)	49.12	3.85	8.0
CogVideoX1.5-5B (Yang et al., 2025c)	46.83	4.85	8.0
CogVideoX-5B (Yang et al., 2025c)	49.92	5.09	6.5
LTX-Video-13B (HaCohen et al., 2026)	56.48	-4.32	7.0
LTX-Video-2B (HaCohen et al., 2026)	**58.86**	-0.20	5.0
Wan2.1-14B (Wan et al., 2025)	53.24	5.91	3.5
Wan2.2-14B (Wan et al., 2025)	54.19	5.51	3.5
PhiZero (Ours)	55.54	**6.20**	**2.0**

4.2 Main Results

Video Generation Results We evaluate PhiZero on three video-generation benchmarks for world models. Physics-IQ Verified (Rädsch et al., 2026) measures physical outcome fidelity by comparing generated videos with real reference videos using rule-based metrics. PhyGround (Lin et al., 2026) evaluates physical-law adherence with physics-specialized VLM judges, while WorldModelBench (Li et al., 2026a) assesses general world-modeling capability in terms of physics adherence and overall video quality. Additional details on the generation benchmarks, evaluation protocols, and metrics are provided in Appendix B.1. As shown in Tables 1, 2, and 3, PhiZero achieves strong performance across all three generation benchmarks. It obtains the best IQ-Score on Physics-IQ Verified, the highest Physics Score and Overall score on PhyGround, and the best Physics Adherence and Total scores on WorldModelBench. These results demonstrate that PhiZero generates state transitions that closely match real-world physical outcomes across diverse open-domain phenomena.

Video Understanding Results We further evaluate PhiZero on three video-understanding benchmarks. IntPhys2 (Bordes et al., 2025) evaluates intuitive physics understanding, LikePhys (Yuan et al., 2025) evaluates physical-plausibility discrimination in simulated scenes, and YoCausal (Xie et al., 2026) examines temporal and causal understanding in real-world videos. All three benchmarks adopt a pairwise comparison protocol, requiring the model to distinguish a physically or causally valid video from a matched invalid counterpart. For each pair, we encode both videos into physical-language sequences, compute their log-likelihoods under the Physical Language Reasoner, and select the video with the higher likelihood as valid. Additional details on the understanding benchmarks and our likelihood-based evaluation protocol are provided in Appendix B.2. As shown in Tables 4, 5, and 6, PhiZero achieves competitive performance across all three understanding benchmarks, demonstrating its ability to identify physical inconsistencies and causal violations.

Qualitative Results As shown in Fig. 4, we compare PhiZero with the Wan2.2-5B baseline (Wan et al., 2025) on four representative cases from Physics-IQ Verified (Rädsch et al., 2026). Although Wan2.2-5B often produces visually plausible interactions, it fails to generate their expected physical consequences. For example, the tennis ball strikes the rubber duck, but the duck remains unaffected

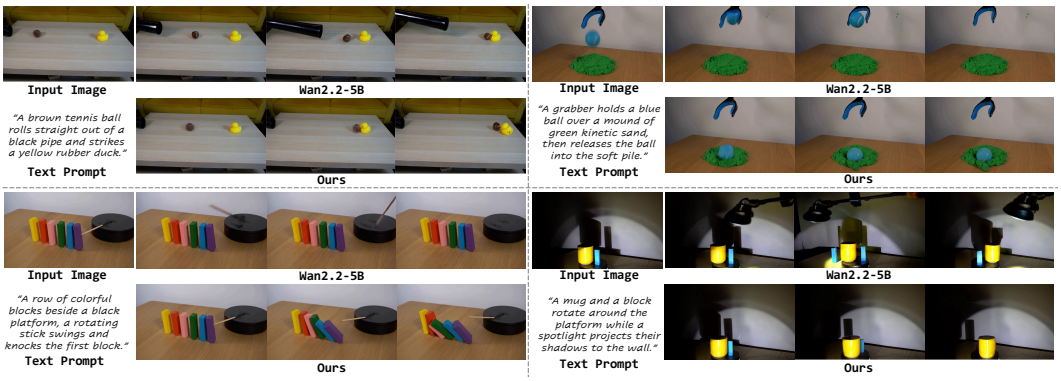


Figure 4: **Qualitative comparison on Physics-IQ Verified.** Compared with the Wan2.2-5B baseline, PHIZERO better captures the physical consequences of interactions, including collision-induced object displacement, gravity-driven deformation, chain reactions, and dynamically changing shadows.

Table 7: **Reconstruction performance of the Physical Language Tokenizer.** Results are reported on 500 four-second real-world videos sampled at 8 FPS and evaluated at a resolution of 512×896.

Model	Tokens ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Wan2.2-5B VAE(Wan et al., 2025)	44800	37.7	0.957	0.042
Video-LaVIT (Jin et al., 2024)	270	23.6	0.835	0.175
VideoFlexTok (Atanov et al., 2026) ($k = 32$)	288	25.2	0.858	0.177
VideoFlexTok (Atanov et al., 2026) ($k = 64$)	576	26.5	0.870	0.167
Ours	**256**	**28.9**	**0.903**	**0.087**

by the collision. In contrast, PHIZERO generates coherent downstream effects, including collision-induced object displacement, deformation under gravity, chain reactions, and changing shadows.

Physical Language Tokenizer Reconstruction Performance We further evaluate the reconstruction performance of the Physical Language Tokenizer on 500 four-second real-world videos sampled at 8 FPS and evaluated at a resolution of 512×896. We report the number of representation tokens excluding the first-frame condition. As shown in Table 7, the Wan2.2 VAE achieves high reconstruction quality but requires 44,800 continuous visual tokens. In contrast, our Physical Language Tokenizer represents the future state transitions in each video with only 256 discrete physical-language symbols and achieves the best reconstruction quality among the highly compressed tokenizers. These results demonstrate that physical language can compactly encode state transitions while retaining sufficient information for high-quality video reconstruction.

4.3 ABLATION STUDIES

Table 8: **Physical Language Tokenizer Ablations**

Model	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Diffusion Dec	26.6	0.875	0.119
w/o Trans. Q-Former	28.2	0.896	0.089
w/o Pure-noise Warm-up	27.9	0.894	0.091
Full	**28.9**	**0.903**	**0.087**

Physical Language Tokenizer Design Ablations We ablate the key design components of the Physical Language Tokenizer in Table 8. Replacing the pretrained diffusion decoder with a conventional deterministic decoder causes the largest performance degradation, showing that the diffusion prior recovers fine-grained appearance details and allows the compact bottleneck to focus on state transitions. Replacing the transition-level Q-Former with a global Q-Former also degrades reconstruction quality, validating the local temporal inductive bias introduced by explicitly modeling transitions between adjacent latent states. Finally, removing the pure-noise warm-up consistently reduces reconstruction performance, indicating that the warm-up encourages the decoder to rely on the physical-language condition rather than relying primarily on its pretrained denoising prior.

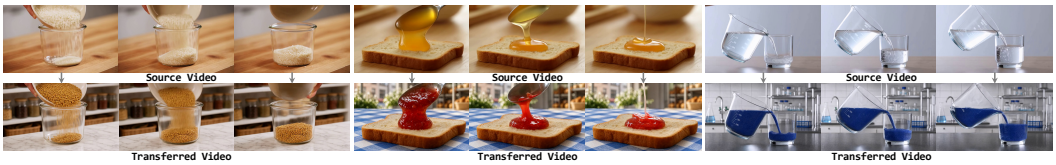
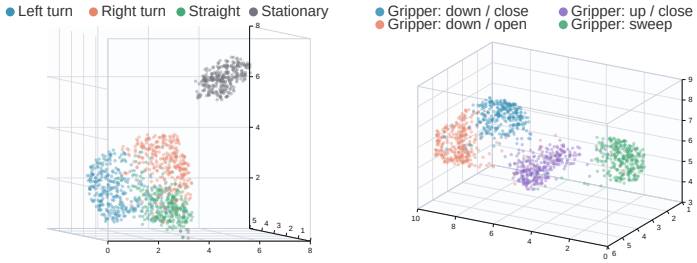


Figure 5: **Transferability of Physical Language.** We encode each source video's state transition into a physical-language sequence and decode the same sequence conditioned on an edited first frame with a different visual appearance. The transferred videos preserve the source evolution across substantial changes in object and scene appearance.



(a) Autonomous driving (b) Robotic manipulation

Figure 6: **Semantic structure of physical language.** We aggregate transition features into clip-level representations, reduce them to 20 dimensions using PCA, and project them into 3D using UMAP. The resulting embeddings naturally organize according to the underlying transition patterns, forming structured manifolds or clusters across the two domains.

Table 9: **Reasoner Ablations**

Method	IQ-Score ↑
Wan2.2-5B (Baseline)	21.2
+ Prompt Enhancement	26.6
Ours w/o Simulation Data	37.7
Ours w/o Two-stage Training	39.2
Ours (Full)	**41.2**

Physical Language Reasoner Design Ablations We evaluate the Physical Language Reasoner on Physics-IQ Verified. As shown in Table 9, prompt enhancement improves the Wan2.2-5B baseline, but still remains substantially below PhiZero, indicating that natural-language reasoning alone is insufficient for modeling fine-grained state transitions. Removing simulation data also reduces performance, confirming that simulator-generated interactions improve the prediction of physically plausible transitions and benefit real-world video generation. This result further highlights the potential to scale physical-language reasoning with simulator-generated data. Finally, replacing two-stage training with joint training degrades performance, showing that dedicated SFT on a curated corpus of motion-rich and physically informative videos helps the model infer more precise and physically plausible state transitions.

4.4 Analysis of Physical Language

Transferability of Physical Language We investigate whether physical language disentangles state transitions from visual appearance. Given a source video, we encode its state transition with the Physical Language Tokenizer, edit the first frame to specify a different target appearance, and decode the unchanged physical-language sequence conditioned on the edited frame. As shown in Fig. 5, the transferred videos preserve transition patterns across changes in appearance and background, including the pouring, viscous spreading, and liquid flow. These results indicate that physical language captures reusable transition information that can be transferred across visual appearances.

Semantic Structure of Physical Language We analyze the semantic structure of physical language in two domains with clearly defined kinematic patterns: autonomous driving and robotic manipulation. We use Physical Language Tokenizers adapted to the corresponding domains and select four categories with 300 samples per category. For autonomous driving, we select left turn, right turn, straight driving, and stationary clips from nuScenes (Caesar et al., 2020). For robotic manipulation, we select four gripper transition patterns from AGI-Bot RealRobot (Bu et al., 2025a): moving downward while closing, moving upward while closing, moving downward while opening, and sweeping. We aggregate the transition features into clip-level representations, reduce them to 20 dimensions us-

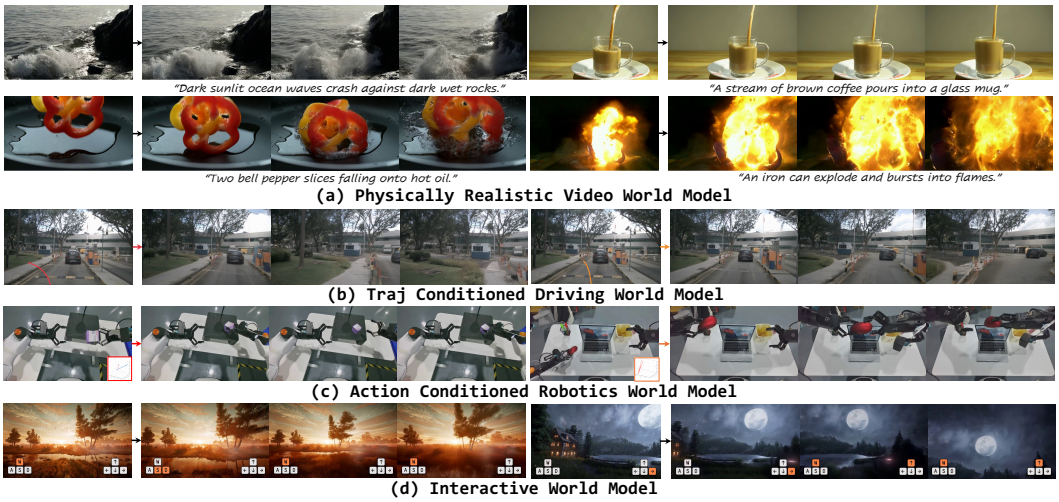


Figure 7: **PhiZero as a physically realistic, controllable, and interactive video world model.** PhiZero generates physically coherent world evolution, follows fine-grained action conditions, and supports interactive rollouts under sequential controls.

ing PCA, and further project them into 3D using UMAP (McInnes et al., 2018). As shown in Fig. 6, the learned representations exhibit clear kinematic organization. In autonomous driving, stationary clips form a distinct cluster, while different steering patterns are arranged along a continuous manifold. In robotic manipulation, the four gripper transition patterns form compact and largely separated clusters. These results indicate that physical language captures meaningful state-transition semantics and organizes videos by how the world changes rather than by their visual content.

4.5 Broader Applications

We further demonstrate the broader application potential of PhiZero. By explicitly representing how the physical world evolves, physical language provides an appearance-disentangled interface for representing, controlling, and transferring state transitions across visual appearances and domains. This interface supports physically realistic video generation, fine-grained action-conditioned world modeling, interactive rollouts, and zero-shot transfer across embodiments and visual domains.

Physically Realistic Video World Model PhiZero can serve as a general video world model that predicts physically coherent future evolution from the current visual state. As shown in Fig. 7(a), it models diverse real-world dynamics, including ocean waves crashing against rocks, liquid pouring into a container, objects falling into hot oil and producing splashes, and a metal can exploding into flames and fragments. These examples capture both the complete temporal evolution of complex events and their fine-grained consequences, demonstrating the ability of PhiZero to model open-domain world dynamics with strong physical coherence.

Controllable and Interactive World Model Existing action-conditioned world models typically map control signals directly to future videos, making precise control over state evolution difficult. In contrast, PhiZero separates state-transition reasoning from pixel synthesis: given a trajectory or action signal, the Physical Language Reasoner first predicts the corresponding physical language, which is then decoded into a future video. We validate this capability on nuScenes (Caesar et al., 2020) and AGI-Bot RealRobot (Bu et al., 2025a). As shown in Fig. 7(b) and (c), PhiZero captures fine-grained variations in driving trajectories, such as different steering magnitudes, and accurately follows action signals to generate precise gripper movements. It also supports interactive world modeling under sequential control inputs. As shown in Fig. 7(d), the model updates the camera viewpoint and position according to successive controls while maintaining temporal consistency. These results demonstrate the potential of PhiZero for fine-grained, controllable, and interactive world modeling. Additional training and inference details for the controllable and interactive world models are provided in Appendix C.1.

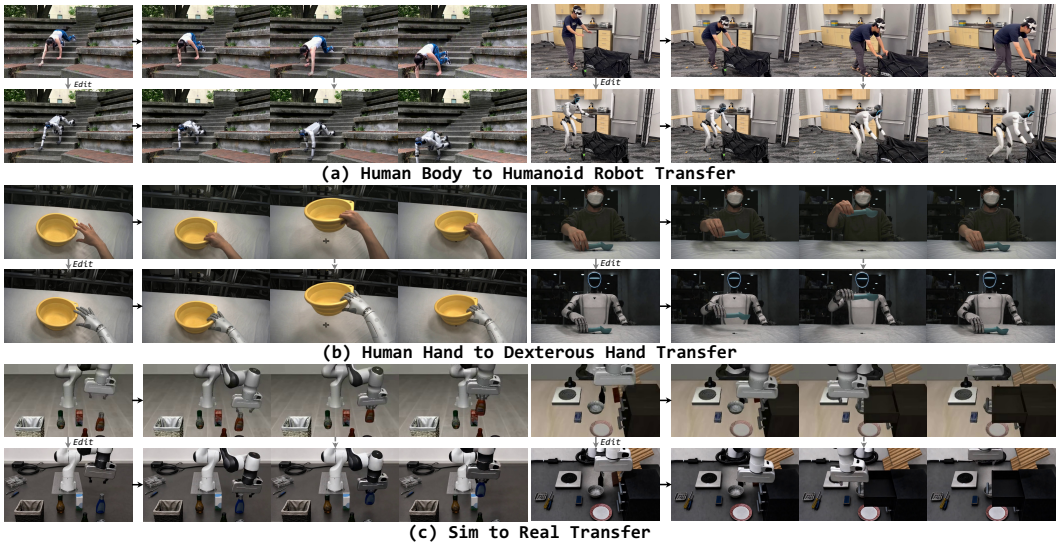


Figure 8: **Zero-shot cross-embodiment and sim-to-real transfer with physical language.** We encode source state transitions into physical-language sequences and render them conditioned on edited first frames specifying new embodiments or visual domains, enabling zero-shot cross-embodiment and sim-to-real transfer.

Zero-Shot Cross-embodiment and Sim-to-real Transfer Because physical language disentangles state transitions from visual appearance, PhiZero supports zero-shot transfer across embodiments and visual domains. As shown in Fig. 8(a) and (b), we encode the state transition of a human-motion video into physical language, edit its first frame by replacing the human body with a Unitree G1 humanoid or the human hand with a Sharpa dexterous hand, and decode the unchanged physical-language sequence conditioned on the edited first frame. This transfers full-body and hand motion patterns to substantially different embodiments without target-specific training. The same formulation also enables sim-to-real transfer. As shown in Fig. 8(c), we transform the first frames of LIBERO (Liu et al., 2023a) videos into a realistic visual domain and render the original simulated state transitions under the new appearance. These results demonstrate that physical language provides a transferable interface across embodiments and visual domains, with potential for demonstration retargeting, cross-morphology transfer, and low-cost generation of realistic interaction data from simulation. Additional adaptation and inference details for zero-shot cross-embodiment and sim-to-real transfer are provided in Appendix C.2.

5 Conclusions

We introduced PhiZero, a physical world model built around physical language. Unlike conventional video models that predict future videos directly in pixel space, PhiZero learns physical language: a compact discrete representation of state-transition patterns from unlabeled in-the-wild videos through self-supervised learning. Building on this representation, PhiZero follows a reason-then-render paradigm: it first predicts future world evolution in the physical-language space and then renders the inferred transitions into videos. Extensive experiments across generation and understanding benchmarks demonstrate that PhiZero generates physically coherent outcomes and identifies physically implausible events. Physical language further provides an interface for fine-grained action-conditioned simulation, interactive rollouts, and zero-shot motion transfer across embodiments and visual domains. These results highlight the potential of physical language for scalable, controllable, and transferable physical-world modeling. We further discuss limitations and future works in the Appendix E.

A Additional Training Details

A.1 Physical Language Tokenizer

We train the Physical Language Tokenizer in two phases: large-scale pretraining followed by targeted SFT, using a joint temporal and spatial curriculum. For pretraining, we aggregate the datasets listed in Table 10. After deduplication, we filter out clips with technical defects, including compression artifacts, corrupted frames, and watermarks, as well as those that fail the resolution or duration requirements or contain abrupt shot transitions. This process yields approximately 10K hours of unlabeled real-world videos for tokenizer pretraining.

Table 10: **Physical Language Tokenizer pretraining data.** We report the duration of each source corpus after data filtering. The mixture contains approximately 10K hours of unlabeled video.

Dataset	Hours
In-house Data	3329
HOIGen-1M (Liu et al., 2025b)	2200
OpenVid-1M (Nan et al., 2025)	2051
Moments in Time (Monfort et al., 2019)	844
Kinetics-710 (Li et al., 2022)	840
Sekai-walking (Li et al., 2026b)	288
SSV2 (Goyal et al., 2017)	245
UltraVideo (Xue et al., 2025)	142
WISA-80K (Wang et al., 2026b)	141
Overall	**10K**

During the first three pretraining stages, we keep both the input and reconstruction resolutions at 256×448 and progressively increase the clip duration from 1 second (9 frames) to 2 seconds (17 frames) and finally to 4 seconds (33 frames). In the fourth pretraining stage, the tokenizer continues to receive inputs at 256×448, while the diffusion decoder reconstructs targets at 512×896. This cross-resolution reconstruction encourages the compact physical-language bottleneck to focus on state transitions while allowing the diffusion decoder to rely on its generative prior to recover high-frequency appearance details.

For tokenizer SFT, we apply a stricter second-pass filtering procedure to the real-world corpus based on aesthetic quality, motion magnitude, and VLM-assessed state-transition observability. We further incorporate simulation videos after filtering out samples with low rendering quality, invalid duration, or insufficient object motion. The resulting corpus contains approximately 5M four-second clips, as summarized in Table 11. We first fine-tune the full model on this curated corpus. In the final refinement stage, we freeze the spatiotemporal encoder, transition-level Q-Former, and FSQ quantizer, and optimize only the diffusion decoder. We additionally apply an entropy regularization loss (Yu et al., 2023) to encourage balanced utilization of the FSQ vocabulary and a REPA loss (Yu et al., 2024a) to improve the semantic quality of the diffusion representations. All training stages use 128 NVIDIA A100 GPUs. Table 12 summarizes the complete training schedule.

A.2 Physical Language Reasoner

We train the Physical Language Reasoner in two stages. In Stage 1, we perform continued pretraining on the same 5M four-second clips used for Physical Language Tokenizer SFT. For each clip, the first frame and a VLM-generated high-level action caption are provided as inputs, while the frozen Physical Language Tokenizer encodes the full video into the target physical-language sequence. In Stage 2, we perform motion-focused SFT on a corpus of approximately 1M clips, comprising 800K samples selected from the 5M-clip corpus using VLM-based motion-quality and physical-interaction filters and 200K additional simulator-generated samples. This stage concentrates training on salient and physically informative state transitions. The first frame is provided at a resolution of 512×896, and the action captions are generated using the prompt in Table 14. Both training stages use 128 NVIDIA A100 GPUs, with their hyperparameters summarized in Table 13.

Table 11: **Physical Language Tokenizer fine-tuning data.** We distinguish between real-world and simulation sources and report the number of four-second clips contributed by each dataset.

Dataset	Type	Clips
In-house Data	Real	2545K
HOIGen-1M (Liu et al., 2025b)	Real	574K
Moments in Time (Monfort et al., 2019)	Real	528K
OpenVid-1M (Nan et al., 2025)	Real	516K
SSV2 (Goyal et al., 2017)	Real	180K
Sekai-walking (Li et al., 2026b)	Real	173K
Kinetics-710 (Li et al., 2022)	Real	128K
UltraVideo (Xue et al., 2025)	Real	82K
WISA-80K (Wang et al., 2026b)	Real	44K
Physics101 (Wu et al., 2016)	Simulation	1K
Phyco (Narayanan et al., 2026)	Simulation	126K
ComPhy (Chen et al., 2022)	Simulation	60K
Cosmos3 (Agarwal et al., 2026)	Simulation	51K
CLEVRER (Yi et al., 2019)	Simulation	20K
Physion++ (Tung et al., 2023)	Simulation	10K
Physion (Bear et al., 2021)	Simulation	9K
Overall	–	**5M**

Table 12: **Physical Language Tokenizer training schedule.** Video and optimization settings for curriculum pretraining, full-model fine-tuning, and decoder-only refinement.

Configuration	Pretrain Stage 1	Pretrain Stage 2	Pretrain Stage 3	Pretrain Stage 4	Full SFT	Decoder-only SFT
Frames	9	17	33	33	33	33
Duration	1s	2s	4s	4s	4s	4s
Global batch size	256					
Input resolution	256×448					
Output resolution	256×448	256×448	256×448	512×896	512×896	512×896
Training steps	150K	100K	100K	50K	50K	50K
Initial pure-noise steps	50K	20K	20K	0	0	0
Warm-up steps	1K					
LoRA rank	32					
FSQ entropy-loss weight	0.005					
REPA-loss weight	0.1					
Learning rate	2×10^{-5}					
AdamW betas	$\beta_1 = 0.9,\ \beta_2 = 0.95$					
Maximum gradient norm	1.0					

Table 13: **Physical Language Reasoner training settings.** Continued pretraining adapts the VLM to physical-language prediction, while SFT specializes it for physically informative state transitions.

Configuration	Continued Pretraining	Motion-focused SFT
Global batch size	512	
First-frame resolution	512×896	
Training steps	50K	10K
Warm-up steps	1K	
Learning rate	5×10^{-5}	1×10^{-5}
AdamW betas	$\beta_1 = 0.9,\ \beta_2 = 0.95$	
Maximum gradient norm	1.0	

Table 14: **Captioning prompt.** Prompt used to generate textual conditions for the four-second clips used to train the Physical Language Reasoner.

System prompt

```
You are a professional video captioning expert. Your task is to generate a
concise high-level action condition for the video clip, used to train a
physical world model.

Caption requirements:

    • Write one concise flowing English sentence or paragraph.
    • Keep it 25-60 words.
    • Use present tense and describe only visible content.
    • Cover the main subject, action, setting, camera/framing, lighting,
      colors, and visual style when observable.

Critical rules:

    • Prioritize spatial precision.
    • Focus on the high-level initiating action or interaction intent rather
      than narrating the full temporal evolution. Avoid fine-grained
      intermediate motions, state changes, and secondary physical effects.
```

User prompt

```
Watch this video carefully and generate a caption strictly following your
instructions.
```

B Additional Evaluation Details

B.1 Physical Video Generation Evaluation

We evaluate PhiZero on three physical video-generation benchmarks. For all benchmarks, the model takes the provided first frame and textual condition as input, and we follow the official evaluation protocols and released evaluators.

Physics-IQ Verified Physics-IQ Verified (Rädsch et al., 2026) evaluates whether generated videos reproduce the outcomes of 66 controlled real-world physical experiments. The generated video is compared with the real reference using Spatial IoU (S-IoU), Spatiotemporal IoU (ST-IoU), and Weighted Spatial IoU (WS-IoU), which respectively evaluate the location, timing, and temporal frequency of scene changes. IQ-Score aggregates these metrics after normalization by the variation between repeated real experiments.

PhyGround PhyGround (Lin et al., 2026) contains 250 curated image–text conditions covering 13 physical laws in solid mechanics, fluid dynamics, and optics. We follow the official protocol using the released PhyJudge-9B evaluator and sample generated videos at 4 FPS. Each video is scored from 1 to 5 according to criterion-specific rubrics. General Quality evaluates prompt alignment, temporal validity, and object consistency, while Physics Score measures adherence to the physical laws applicable to each sample. Overall is the average of General Quality and Physics Score.

WorldModelBench WorldModelBench (Li et al., 2026a) evaluates general world-modeling capability using 350 image–text conditions from seven domains and 56 subdomains, including autonomous driving, robotics, human activities, industrial scenes, gaming, animation, and natural environments. We follow the official protocol using the released 2B-VLM judge. Physics Adherence evaluates five physical-consistency criteria and ranges from 0 to 5, while Common Sense evaluates spatial and temporal visual quality and ranges from 0 to 2. The Total score is obtained by summing all metrics evaluated by the benchmark.

B.2 Physical Video Understanding Evaluation

IntPhys2 IntPhys2 (Bordes et al., 2025) evaluates intuitive physics through matched possible and impossible videos involving object permanence, immutability, spatiotemporal continuity, and solidity. Its main set contains 1,012 videos organized into 253 quadruplets, with two possible and two impossible outcomes per scene. We compare the physical-language likelihoods of each matched possible–impossible pair and report the percentage of pairs for which the possible video receives the higher likelihood. Results are reported separately on the Easy, Medium, and Hard subsets, together with the Overall accuracy on the complete main set.

LikePhys LikePhys (Yuan et al., 2025) contains controlled valid–invalid video pairs from 12 scenarios. The videos within each comparison are matched in visual appearance and differ primarily in whether the relevant physical principles are satisfied. Following its likelihood-preference protocol, Plausibility Preference Error (PPE) is the percentage of comparisons for which an invalid video receives an equal or higher likelihood than a valid video; therefore, lower values indicate stronger physical understanding. We report PPE for rigid-body mechanics, fluid mechanics, and optical effects, as well as Avg. Error averaged over all 12 scenarios.

YoCausal YoCausal (Xie et al., 2026) contains 1,232 real-world video pairs. Each natural forward video is paired with its temporally reversed counterpart, and the same caption is used for both directions. The Reversal Surprise Index (RSI) is the mean forward-video preference rate across the four source datasets, where a forward preference occurs when the natural video receives the higher physical-language likelihood; 50% corresponds to chance. The benchmark further partitions videos into causal and non-causal subsets using VLM annotations and defines the Causality Cognition Index as $\mathrm{CCI} = \mathrm{RSI}(\mathcal{D}_c) - \mathrm{RSI}(\mathcal{D}_{nc})$. Aggregate Rank combines the model rankings on RSI and CCI, with a lower rank indicating better overall performance.

Evaluation Protocol All three benchmarks are evaluated through pairwise likelihood comparison under a shared text condition. Following the official evaluation protocol, for LikePhys, we use the official prompt template provided for each scenario, which is shared by all valid and invalid variants of that scenario; for YoCausal, we use the benchmark-provided prompt for the forward video and its temporally reversed counterpart. IntPhys2 does not provide instance-level text descriptions. We therefore use a VLM to generate one scene-level caption for each video quadruplet. The VLM observes only the first frames of the four videos, without access to future frames or validity labels, and is instructed to describe only their shared visible scene and initial configuration. We also apply the resulting caption unchanged to all four videos in the quadruplet. Formally, let $(\mathbf{V}^+, \mathbf{V}^-)$ denote a matched pair and c_{pair} its shared text condition. For each video, we use its own first frame $I^0_\pm$ and the identical text condition c_{pair}, encode it into a physical-language sequence $\mathbf{z}^\pm = \mathcal{T}_\phi(\mathbf{V}^\pm)$, and compute

$$s(\mathbf{V}^\pm, c_{\text{pair}}) = \sum_{j=1}^{N} \log p_\theta \left(z_j^\pm \mid I^0_\pm, c_{\text{pair}}, z^\pm_{<j} \right). \tag{7}$$

Thus, the two likelihoods within each comparison differ only in the evaluated video and its corresponding first frame, while the text condition remains identical. The video with the higher sequence likelihood is selected as the physically valid or temporally natural one.

C Training and Inference Details for Broader Applications

C.1 Controllable and Interactive World Model

Robotics World Model For robotic manipulation, we adapt PhiZero on the AGI-Bot RealRobot dataset (Bu et al., 2025a). We first fine-tune the Physical Language Tokenizer on RealRobot videos so that the learned physical language captures domain-specific robotic-arm and gripper transitions. For each four-second clip, the corresponding trajectories of all robot-arm joints and the gripper pose are sampled at 8 FPS. At every time step, these values are arranged in a fixed order and serialized into numerical text, and is used as the action condition for Physical Language Reasoner fine-tuning. At inference time, a desired joint and gripper trajectory is converted into the same textual format and provided together with the current observation. The reasoner translates this action sequence

into physical language, and the domain-adapted decoder renders the resulting robotic interaction as a future video.

Driving World Model We follow the same procedure to construct the trajectory-conditioned driving world model on nuScenes (Caesar et al., 2020). We first fine-tune the Physical Language Tokenizer on nuScenes videos to adapt the representation and decoder to driving-scene evolution. The future four-second ego-vehicle trajectory is sampled at 8 FPS and serialized into a temporally ordered numerical text sequence. The first frame and textualized ego trajectory are then used to fine-tune the Physical Language Reasoner, while the physical-language sequence extracted from the corresponding future video serves as the prediction target. During inference, a candidate ego trajectory is converted into the same numerical text format. The reasoner predicts how the scene should evolve under this trajectory in physical-language space, after which the decoder renders the corresponding future driving video. This formulation allows the model to follow fine-grained trajectory variations, including different steering directions and magnitudes, without modifying the general reason-then-render architecture.

Interactive World Model For interactive world modeling, control inputs are expressed directly in natural language as instructions describing the desired change in camera viewpoint or camera position. Since the current model generates four-second clips, we extend generation beyond this fixed horizon using a sliding-window autoregressive rollout. Starting from the current frame, the model predicts the next four-second video segment, and the final frame of the generated segment is subsequently used as the first-frame condition for the next window. Repeating this process produces longer videos and allows the camera trajectory to be modified progressively through sequential natural-language controls.

C.2 Motion Transfer

Transfer Protocol Our motion-transfer pipeline does not require paired videos. Instead, we briefly fine-tune the Physical Language Tokenizer on videos from each source domain, enabling it to more accurately encode the corresponding motion patterns into physical language. After this source-domain adaptation, motion transfer can be performed directly without further training. Given a source video, we first encode its state transition into a physical-language sequence. We then use GPT-Image 2.0 to edit the first frame, replacing the original subject or visual appearance with the desired target embodiment or target domain. The unchanged physical-language sequence is subsequently decoded conditioned on the edited first frame. In this way, the first frame specifies what the target scene should look like, while physical language specifies how it should evolve over time.

Human-body Motion Transfer For human-body motion transfer, we construct a subset containing diverse and salient full-body human motions. The source videos are collected from (Chen et al., 2026c; Allshire et al., 2025). We briefly fine-tune the Physical Language Tokenizer on this subset to adapt it to human-body motion. During inference, the source human-motion video is encoded into physical language, while its first frame is edited to replace the human subject with the target humanoid robot. Decoding the original physical-language sequence from the edited first frame transfers the full-body motion to the humanoid embodiment without paired human–robot training data.

Human-hand Motion Transfer For human-hand to dexterous-hand transfer, we construct a subset of human-hand videos with rich hand motions and object interactions from (Lim et al., 2026). Although it provides paired human and robot data, we use only the human-hand videos for tokenizer adaptation and do not use the provided pairings or cross-embodiment correspondences. At inference time, we edit the first frame to replace the human hand with the target dexterous hand and decode the source physical-language sequence under this edited appearance condition. This transfers fine-grained hand motions and interaction patterns without paired supervision.

Sim-to-real Transfer For sim-to-real transfer, we briefly fine-tune the Physical Language Tokenizer on simulated interaction videos from LIBERO (Liu et al., 2023a). No paired real-world videos are used during adaptation. Given a simulated source video, we encode its state transition into physical language and use GPT-Image 2.0 to transform the first frame into a realistic visual appearance.

We then decode the unchanged physical-language sequence conditioned on the edited realistic frame, producing a realistic-looking video that preserves the interaction demonstrated in simulation.

D Additional Related Work

D.1 Latent-action World Models

Latent-action models infer compact transition variables from observation sequences, typically through an inverse-dynamics bottleneck coupled with a forward predictor. Early approaches learn discrete latent actions from adjacent observations and use them to pretrain controllable world models or downstream policies (Cui & Gao, 2023; Schmidt & Jiang, 2024; Bruce et al., 2024; Ye et al., 2025; Gao et al., 2025). Recent work extends this paradigm along several complementary directions. Garrido et al. (2026) scale latent-action world models to heterogeneous in-the-wild videos and study constrained continuous representations. Jiang et al. (2026b) align latent-action semantics across visual contexts through control-effect alignment, while Wang et al. (2026c) factorize scene dynamics into factor-wise latent actions for multi-entity environments. Zhang et al. (2026a) explicitly disentangle dynamics-relevant structure from visual content. Other methods connect latent actions to mixed-data control or unified modeling: Alles et al. (2025) align action-free and action-conditioned trajectories for offline reinforcement learning, Wang et al. (2025a) jointly evolve latent-action inference and a pretrained video world model, and Bi et al. (2025) integrate understanding, video generation, and action prediction within a unified architecture. Chen et al. (2026a) introduce a unified physical language that aligns human and humanoid actions through visually anchored cross-reconstruction for policy learning and world modeling. Despite these advances, latent-action world models are predominantly developed for control, often targeting specific embodiments and tasks or relatively constrained environments. In contrast, PhiZero uses this transition-modeling paradigm to represent the open-domain evolution of the physical world itself, learning a physical language from diverse in-the-wild videos and explicitly reasoning over it for both future-video generation and state-transition understanding.

D.2 Video Tokenizers

Video tokenizers compress videos into compact representations by exploiting spatial and temporal redundancy. A major line of work separates persistent visual content from temporal variation. Sun et al. (2023) decomposes videos into scene, object, and motion tokens, while Jin et al. (2024) interleaves keyframe and motion tokens. Yu et al. (2024b); Tian et al. (2025); Wang et al. (2025b); Liu et al. (2025a) factorize videos through content frames, reference images, or hierarchical latent variables, and Wang et al. (2026a); Chen et al. (2026b) further summarize shared appearance with keyframe or time-invariant tokens while compactly encoding temporal residuals. Recent work increasingly employs diffusion-based decoders, allowing compact representations to omit fine visual details that can be recovered by a generative prior. Ge et al. (2025); Li et al. (2024) condition pretrained video diffusion models on compressed spatiotemporal features, while Yang et al. (2025a) reconstructs videos with a conditional causal diffusion decoder. Atanov et al. (2026) learns variable-length coarse-to-fine tokens with a generative flow decoder, and Teng et al. (2026) combines query-based 1D latents with a pixel-space diffusion decoder for adaptive compression. Closely related, Ressler-Antal et al. (2025) learns appearance-disentangled motion representations through video reconstruction for open-world motion transfer and motion understanding. Our Physical Language Tokenizer similarly adopts a generative decoder, but the learned physical language is not used solely for video reconstruction. It also serves as an explicit prediction target through which the model reasons about future world evolution before rendering the inferred transition into video.

D.3 Physically Realistic Video Generation

Existing approaches to improving the physical consistency of video generation broadly follow three directions. The first incorporates explicit constraints derived from physics simulators. For example, Montanaro et al. (2024) uses simulator-derived optical flow, Liu et al. (2024) performs rigid-body simulation, and Foo et al. (2026) integrates a 3D physics simulator. The second transfers physical priors from external models or simulated data, often through representation alignment (Zhang et al., 2026b; Bhowmik et al., 2025; Kim et al., 2026). Along this line of work, Xiong et al. (2026) incorporates 3D geometric cues obtained from simulation, while Jiang et al. (2026a) learns a motion prior

from VAE latents. The third introduces explicit physical knowledge or intermediate reasoning into video generation. Wang et al. (2026b); Zhang et al. (2025) inject physical knowledge during training, whereas Yang et al. (2025b) uses a VLM to reason about object trajectories before generation. In contrast, PhiZero does not rely on simulator-derived constraints, predefined physical variables, or explicit physical rules as supervisory signals. Instead, it learns physical language from large-scale unlabeled videos through self-supervision, abstracting patterns of state transition.

E Limitations and Future Work

Limitations Despite the promising results, several limitations remain. First, physical language is currently learned as an empirical representation of state transitions rather than an explicit formulation of symbolic physical laws. This data-driven formulation enables scalable learning from diverse unlabeled videos, but the resulting discrete symbols are not yet directly grounded in interpretable physical variables or formal laws. Second, because physical language is learned primarily from observational videos, its coverage is constrained by what can be visually observed in the training data. World transitions that are visually ambiguous or largely inaccessible to vision, such as tactile interactions and microscopic particle dynamics, may therefore be more difficult to model. Finally, due to limitations in data and computational resources, our current implementation uses relatively small-scale models and a training corpus that remains limited relative to the diversity of the physical world. Nevertheless, the strong performance achieved at this scale suggests that larger models and more diverse training data could further improve the capabilities of physical-language-based world modeling.

Future Work Future work can extend physical language along several directions. First, beyond video generation, physical language may serve as an explicit intermediate representation for improving spatiotemporal understanding in VLMs and supporting more effective planning in embodied policies. Second, its transferability across appearances and embodiments opens a promising avenue for alleviating the scarcity of real-robot interaction data. In particular, state-transition patterns learned from large-scale human videos could potentially be transferred to robotic embodiments, enabling embodied systems to benefit from abundant human data with limited robot-specific supervision. Third, we plan to extend physical-language modeling beyond the current fixed-duration clips through hierarchical or recurrent prediction, enabling long-horizon world modeling while maintaining temporal consistency and faithfully tracking the evolving scene state. Finally, we will further scale PhiZero with stronger backbones, increased computational resources, and larger and more diverse training corpora to improve the performance of physical-language-based world modeling.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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. *arXiv preprint arXiv:2501.03575*, 2025.

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Marvin Alles, Xingyuan Zhang, Patrick van der Smagt, and Philip Becker-Ehmck. Latent action world models for control with unlabeled trajectories. *arXiv preprint arXiv:2512.10016*, 2025.

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

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

Andrei Atanov, Jesse Allardice, Roman Bachmann, Oğuzhan Fatih Kar, R Devon Hjelm, David Griffiths, Peter Fu, Afshin Dehghan, and Amir Zamir. Videoflextok: Flexible-length coarse-to-fine video tokenization. In *Forty-third International Conference on Machine Learning*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: latent video prediction for visual representation learning. In *URL https://openreview.net/forum*, 2024.

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Aritra Bhowmik, Denis Korzhenkov, Cees GM Snoek, Amirhossein Habibian, and Mohsen Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models. *arXiv preprint arXiv:2510.19022*, 2025.

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

Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments. *arXiv preprint arXiv:2506.09849*, 2025.

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.

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

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025b.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Boyu Chen, Yi Chen, Lu Qiu, Jerry Bai, Yuying Ge, and Yixiao Ge. Unit: Toward a unified physical language for human-to-humanoid policy learning and world modeling. *arXiv preprint arXiv:2604.19734*, 2026a.

Weiliang Chen, Yuanhui Huang, Xuebo Wang, and Yueqi Duan. Tivtok: Broadcasting time-invariant tokens for scalable video tokenization. *arXiv preprint arXiv:2606.17590*, 2026b.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19752–19763, 2025.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022.

Zhenyang Chen, Chuizheng Kong, Chuye Zhang, Yuanshao Yang, Lawrence Y Zhu, Shreyas Kousik, and Danfei Xu. Warp: Whole-body retargeting for learning from offline human demonstrations. *arXiv preprint arXiv:2606.29940*, 2026c.

Hanchen Cui and Yang Gao. A universal world model learned from large scale and diverse videos. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4301–4311, 2026.

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.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. *arXiv preprint arXiv:2601.05230*, 2026.

Yuying Ge, Yizhuo Li, Yixiao Ge, and Ying Shan. Divot: diffusion powers video tokenizer for comprehension and generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13606–13617, 2025.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. URL https://arxiv.org/abs/2312.11805.

Genmo AI. Genmo AI Blog. https://www.genmo.ai/blog, 2025. Accessed: April 29, 2026.

Google DeepMind. Veo 3.1. https://deepmind.google/models/veo/, 2026. Accessed: April 29, 2026.

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*, pp. 5842–5850, 2017.

Jinwei Gu. Cosmos world foundation models for physical ai. In *Proceedings of the 3rd International Workshop on Rich Media With Generative AI*, pp. 39–39, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Bo Jiang, Depu Meng, Yihan Hu, Yichen Xie, Tianshuo Xu, and Wei Zhan. Lamo: Self-supervised latent motion priors for physical realism in video generation. *arXiv preprint arXiv:2605.23878*, 2026a.

Yuxin Jiang, Yuchao Gu, Ivor W Tsang, and Mike Zheng Shou. Olaf-world: Orienting latent actions for video world modeling. *arXiv preprint arXiv:2602.10104*, 2026b.

Yang Jin, Zhicheng Sun, Kun Xu, Liwei Chen, Hao Jiang, Quzhe Huang, Chengru Song, Yuliang Liu, Di Zhang, Yang Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024.

Manjin Kim, Suha Kwak, and Minsu Cho. Tempered self-similarity alignment for physically plausible video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5148–5158, 2026.

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.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022.

Yizhuo Li, Yuying Ge, Yixiao Ge, Ying Shan, and Ping Luo. Dicode: Diffusion-compressed deep tokens for autoregressive video generation with language models. *arXiv preprint arXiv:2412.04446*, 2024.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026b.

Jongbin Lim, Taeyun Ha, Mingi Choi, Jisoo Kim, Byungjun Kim, Subin Jeon, and Hanbyul Joo. Hrdexdb: A paired human-robot dataset for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2604.14944*, 2026.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y Lu, Enfu Nan, Hokin Deng, et al. Phyground: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023a.

Huaize Liu, Wenzhang Sun, Qiyuan Zhang, Donglin Di, Biao Gong, Hao Li, Chen Wei, and Changqing Zou. Hi-vae: Efficient video autoencoding with global and detailed motion. *arXiv preprint arXiv:2506.07136*, 2025a.

Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24001–24010, 2025b.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023b.

Luma AI. Luma Dream Machine: AI video generator. https://dream-machine.lumalabs.ai/, 2024. Accessed: April 29, 2026.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

Antonio Montanaro, Luca Savant Aira, Emanuele Aiello, Diego Valsesia, and Enrico Magli. Motioncraft: Physics-based zero-shot video generation. *Advances in Neural Information Processing Systems*, 37:123155–123181, 2024.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, volume 2025, pp. 1045–1064, 2025.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. Phyco: Learning controllable physical priors for generative motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41892–41902, 2026.

OpenAI. Sora 2 system card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf, 2025. Accessed: January 21, 2026.

Kaihang Pan, Qi Tian, Jianwei Zhang, Weijie Kong, Jiangfeng Xiong, Yanxin Long, Shixue Zhang, Haiyi Qiu, Tan Wang, Zheqi Lv, et al. Omniweaving: Towards unified video generation with free-form composition and reasoning. *arXiv preprint arXiv:2603.24458*, 2026.

Tim Rädsch, Yuki M Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T Lüth. Physics-iq verified. *arXiv preprint arXiv:2606.18943*, 2026.

Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29029–29039, 2025.

Zhongwei Ren, Yunchao Wei, Xiao Yu, Guixun Luo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld 2: Learning transferable knowledge from real-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40569–40580, 2026.

Thomas Ressler-Antal, Frank Fundel, Malek Ben Alaya, Stefan Andreas Baumann, Felix Krause, Ming Gui, and Björn Ommer. Dismo: Disentangled motion representations for open-world motion transfer. *arXiv preprint arXiv:2511.23428*, 2025.

Runway ML. Introducing Gen-3 Alpha. https://runwayml.com/research/introducing-gen-3-alpha, 2024. Accessed: April 29, 2026.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, volume 2024, pp. 9379–9395, 2024.

Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, et al. Dynvla: Learning world dynamics for action reasoning in autonomous driving. *arXiv preprint arXiv:2603.11041*, 2026.

Mingzhen Sun, Weining Wang, Xinxin Zhu, and Jing Liu. Moso: Decomposing motion, scene and object for video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18727–18737, 2023.

Yao Teng, Minxuan Lin, Xian Liu, Shuai Wang, Xiao Yang, and Xihui Liu. Adaptive 1d video diffusion autoencoder. *arXiv preprint arXiv:2602.04220*, 2026.

Rui Tian, Qi Dai, Jianmin Bao, Kai Qiu, Yifan Yang, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Reducio! generating 1k video within 16 seconds using extremely compressed motion latents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19237–19247, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Feng Wang, Yichun Shi, Ceyuan Yang, Qiushan Guo, Jingxiang Sun, Alan Yuille, and Peng Wang. Vtok: A unified video tokenizer with decoupled spatial-temporal latents. *arXiv preprint arXiv:2602.04202*, 2026a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Jiasong Feng, Zhanjie Zhang, Wanyuan Pang, and Xiaodan Liang. Wisa: World simulator assistant for physics-aware text-to-video generation. *Advances in Neural Information Processing Systems*, 38:5388–5416, 2026b.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14549–14560, 2023b.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024.

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*, 2025a.

Yuchi Wang, Junliang Guo, Xinyi Xie, Tianyu He, Xu Sun, and Jiang Bian. Vidtwin: Video vae with decoupled structure and dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22922–22932, 2025b.

Zizhao Wang, Chang Shi, Jiaheng Hu, Kevin Rohling, Roberto Martín-Martín, Amy Zhang, and Peter Stone. Factored latent action world models. *arXiv preprint arXiv:2602.16229*, 2026c.

Jiajun Wu, Joseph J Lim, Hongyi Zhang, Joshua B Tenenbaum, and William T Freeman. Physics 101: Learning physical object properties from unlabeled videos. In *BMVC*, volume 2, pp. 7, 2016.

Jialong Wu, Haoyu Ma, Chaoyi Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. *Advances in Neural Information Processing Systems*, 36:39719–39743, 2023.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

xAI. Grok Imagine API: State-of-the-art video generation across quality, cost, and latency. https://x.ai/news/grok-imagine-api, 2026. Accessed: April 29, 2026.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

You-Zhe Xie, Yu-Hsuan Li, Jie-Ying Lee, Kaipeng Zhang, Yu-Lun Liu, and Zhixiang Wang. Yocausal: How far is video generation from world model? a causality perspective. *arXiv preprint arXiv:2605.30346*, 2026.

Zhexiao Xiong, Yizhi Song, Liu He, Wei Xiong, Yu Yuan, Feng Qiao, and Nathan Jacobs. Physalign: Physics-coherent image-to-video generation through feature and 3d representation alignment. *arXiv preprint arXiv:2603.13770*, 2026.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Nianzu Yang, Pandeng Li, Liming Zhao, Yang Li, Chen-Wei Xie, Yehui Tang, Xudong Lu, Zhihang Liu, Yun Zheng, Yu Liu, et al. Rethinking video tokenization: A conditioned diffusion-based approach. *arXiv preprint arXiv:2503.03708*, 2025a.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Xindi Yang, Baolu Li, Yiming Zhang, Zhenfei Yin, Lei Bai, Liqian Ma, Zhiyong Wang, Jianfei Cai, Tien-Tsin Wong, Huchuan Lu, et al. Vlipp: Towards physically plausible video generation with vision and language informed physical prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12360–12370, 2025b.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025c.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 2025.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024a.

Sihyun Yu, Weili Nie, De-An Huang, Boyi Li, Jinwoo Shin, et al. Efficient video diffusion models via content-frame motion-latent decomposition. In *International Conference on Learning Representations*, volume 2024, pp. 9642–9670, 2024b.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. Likephys: Evaluating intuitive physics understanding in video diffusion models via likelihood preference. *arXiv preprint arXiv:2510.11512*, 2025.

Ke Zhang, Cihan Xiao, Yiqun Mei, Jiacong Xu, and Vishal M Patel. Think before you diffuse: Llms-guided physics-aware video generation. *arXiv e-prints*, pp. arXiv–2505, 2025.

Tianqiu Zhang, Muyang Lyu, Yufan Zhang, Fang Fang, and Si Wu. Dila: Disentangled latent action world models. *arXiv preprint arXiv:2605.15725*, 2026a.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026b.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025.