SHADOWDANCER: TEACHING VIDEO WORLD MODELS ANY ACTION BY LEARNING UNIFIED DYNAMICS REPRESENTATIONS FROM A VIDEO AND ITS SHADOW

Anonymous authors
Paper under double-blind review

ABSTRACT

We present ShadowDancer, an approach to any-action, frame-level control of interactive video world models. The obstacle is representational: existing interfaces either encode an action loosely, leaving how it unfolds for the model to improvise, or encode it exactly through structured signals that serve one family and are hard to acquire, so precise control across diverse dynamics remains impractical. Demonstration videos are the natural remedy, specifying any dynamics frame by frame; yet a video shows its dynamics only through one particular appearance, a single *shadow* of the underlying dynamics, so actions learned from demonstrations transfer poorly to new scenes. ShadowDancer addresses this with two key innovations: (1) *shadow pairs*, video pairs that replay the same dynamics under independently resampled appearance, constructed at scale by our Shadow Library, so that a dynamics family becomes controllable exactly when such pairs can be constructed for it; and (2) *cross-shadow prediction*, which learns actions by predicting one shadow from the other, so that whatever the pairing resamples is discarded by construction and whatever it preserves becomes the action, yielding a unified dynamics representation that drives a block-causal world model. Any demonstrated clip thus becomes a reusable action asset, replayed in new environments without action labels, motion estimators, or fine-tuning. Experiments show improved action transfer and long action rollout over strong latent-action and interactive world model baselines across diverse dynamics families, with an average blinded rollout win rate of 86%. We strongly encourage readers to watch **supplementary videos**.

1 INTRODUCTION

Interactive world models have emerged as a transformative capability in video generation, synthesizing persistent virtual worlds that users can explore in real time (Parker-Holder et al.; He et al., 2025; Sun et al., 2025; Mao et al., 2025; Decart et al., 2024; Team et al., 2026; Gao et al., 2026). However, despite the boundless variety of dynamics a world can host, existing approaches lack a unified means of specifying how those dynamics should unfold. The result is a stubborn trade-off between precision and generality: a behavior can be specified loosely for arbitrary dynamics, or exactly for one narrow family, but rarely both at once. This leaves worlds that can be explored but not yet directed, restricting their applicability for interactive entertainment and world simulation. We therefore target *any-action, frame-level control*: precision, so that a command specifies how an action unfolds frame by frame, and generality, so that the same interface extends to any dynamics family.

The bottleneck is not generative capacity, as video diffusion backbones already synthesize complex articulated motion well (Brooks et al., 2024; Wan et al., 2025; Chen et al., 2025a; Kong et al., 2024); it lies in the *representation* of actions: how the user communicates which dynamics should unfold. Any-action, frame-level control is at its core a representation learning problem, and each existing interface can be read as one candidate representation. Symbolic commands (He et al., 2025; Yu et al., 2025; Decart et al., 2024), designed around logged interactions such as keystrokes, name an event from a finite vocabulary and leave its trajectory for the generator to improvise, so holding a key longer repeats an action yet hardly reshapes it. Free-form text (Zhu et al., 2026; Mao et al., 2025) removes the vocabulary but not the ambiguity, since language describes dynamics rather than supplying it.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

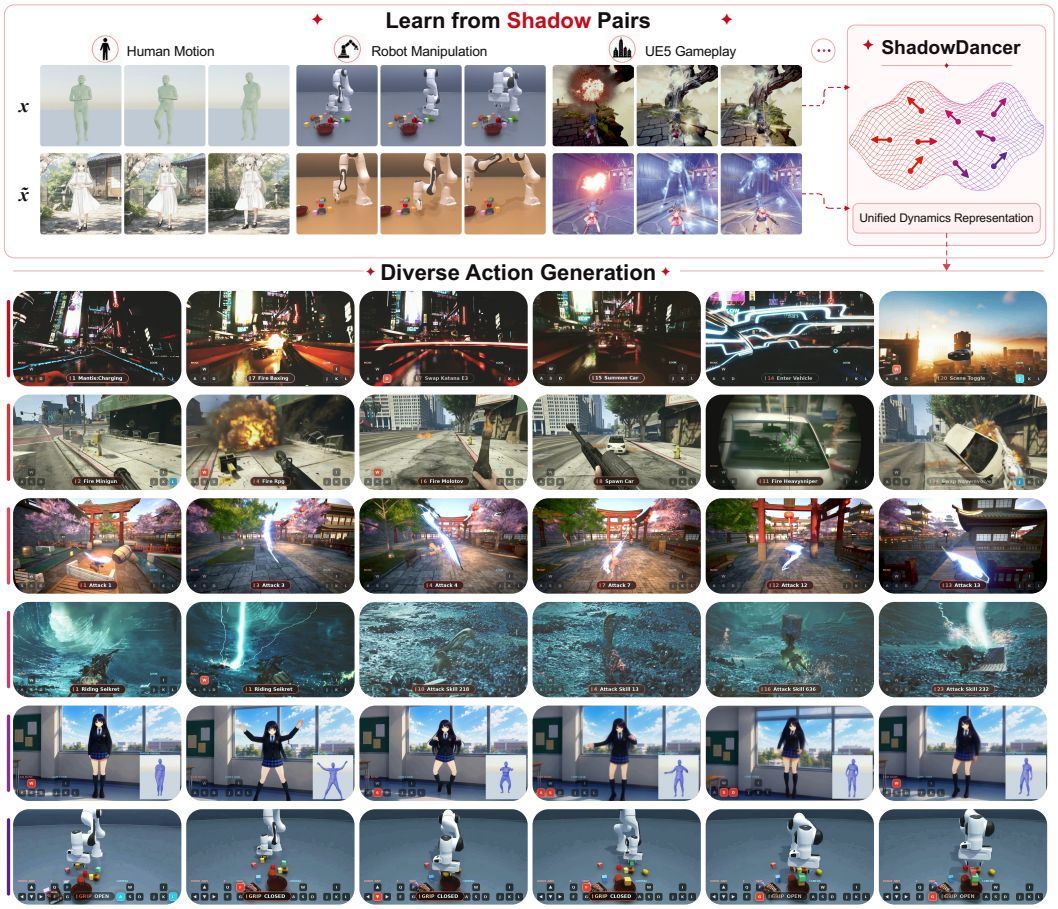

Figure 1: **ShadowDancer learns any action from a video and its shadow.** Top: shadow pairs $(x, \tilde{x})$ replay one dynamics under independently resampled appearance, across heterogeneous sources such as human motion, robot manipulation, and open-world gameplay, and distill it into a unified dynamics representation $z_{1:T-1}$. Bottom: the same latent interface drives diverse commanded actions, spanning first- and third-person combat, driving, locomotion, and robot manipulation.

Structured motion inputs such as 3D human motion (Cao et al., 2025; Fang et al., 2026), point tracks (Lee et al., 2026; Shin et al., 2026), and camera pose trajectories do reach frame-level precision, but each format serves a single family, demands its own conditioning machinery, and depends on specialized yet often brittle estimators, so the precise signal it promises is hard to obtain in practice. What remains is the demonstration itself, a natural specification that is temporally dense and format-free, and the way humans are taught new actions. Latent action models already exploit it, distilling the transitions of a reference clip into continuous latent actions z that condition generation (Edwards et al., 2019; Rybkin et al., 2019; Bruce et al., 2024; Chen et al., 2024; Ye et al., 2025). But critically, such models are trained to reconstruct the very clip they read, and self-reconstruction fundamentally conflicts with *control*: reconstruction asks z merely to explain the observed appearance, whereas control asks it to transfer the underlying dynamics to new appearances (Yang et al., 2025; Garrido et al., 2026; Locatello et al., 2019; Khemakhem et al., 2020; Jiang et al., 2026). Trained without ever observing the same dynamics under two different appearances, the latent cannot tell which traits belong to the action and which merely co-occur with it: it entangles what moves with how it looks indiscriminately, and shrinking it (Gao et al., 2025) or regularizing it toward semantic features (Jiang et al., 2026) operates within the same supervision: invariance can only be encouraged as a penalty, and the unchosen residue can resurface as spurious motion in the generation. What the task calls for is to learn the representation from data in which dynamics is separated from appearance by construction.

To address this challenge, we propose **ShadowDancer**, built on the idea of observing the same dynamics *twice*. Our approach introduces two key innovations. (1) The *Shadow Library*: we realize observing-twice as *Shadow pairs*, where a Shadow of a video is a second render of the same dynamics

under a different appearance, frame-synchronized in motion while subject, scene, materials, and lighting are independently resampled. Constructing shadows is a protocol rather than an engine: replay the dynamics and resample everything else. The Shadow Library implements this protocol across animation suites, open-world games, and robotic simulators. (2) *Cross-shadow prediction*: an encoder extracts latent actions from one video, and a decoder predicts the other shadow from those latents under the shadow's own appearance context. Since the target appearance is already supplied, the prediction requires transferring dynamics rather than reconstructing appearance, making invariance a property of the supervision rather than a penalty. The pairing defines the action itself: whatever the pair shares becomes the controllable factor. What the representation couples to is thus no longer an accident of the data but a choice of the protocol. Pairing on a body motion while resampling the camera yields camera-free control, whereas pairing on a camera trajectory across scenes yields pure camera control. More generally, any dynamics family becomes controllable once a shadow pair can be constructed for it, with no vocabulary to design and no labels to collect. The resulting latent stream is a unified dynamics representation, driving every action family through one interface. It conditions a pretrained video diffusion backbone (Chen et al., 2025a; Wan et al., 2025), which we convert into a block-causal autoregressive generator (Gu et al., 2025; Huang et al., 2025) for interactive rollout. Each demonstrated clip thereby becomes a *variable-length action asset*, the reusable unit of control in ShadowDancer.

In use, directing an action takes one demonstration: the dynamics of the clip is extracted in a single frozen-encoder pass, stored as an asset, composed with other assets along the condition stream, and replayed in a new environment. Unlike symbolic interfaces, which specify what should happen and leave how it happens to the generator, ShadowDancer supplies a dense control signal read from the trajectory itself, so the timing, amplitude, and style of an action are inherited from the demonstration. We evaluate this interface in two settings: action transfer, which tests how precisely a demonstration from one environment is re-enacted in another, and long action rollout, which tests how well stored assets compose over time. In summary, our contributions are threefold:

- We introduce the Shadow learning paradigm, which learns, from a video and its Shadow, dynamics representations invariant to everything the pairing resamples, resolves latent-action non-identifiability by construction, and provides a formal analysis of when the shared dynamics is identifiable.
- We build the Shadow Library, a source-agnostic pairing protocol implemented through paired-rendering scripts across animation suites, open-world games, and robotic simulators, supplying shadow supervision for families including human, camera, and object dynamics without a single action label.
- We present ShadowDancer, an any-action interactive world model driven by dense, variable-length action assets, and show improved action transfer and long action rollout over strong latent-action and interactive world model baselines across the evaluated dynamics families.

2 Related Work

Interactive Video World Models. World models predict future observations and support planning or interactive simulation in games, robotics, and driving (Ha & Schmidhuber, 2018; Hafner et al., 2023; Agarwal et al., 2025; Gao et al., 2024), and recent systems generate explorable worlds at impressive fidelity and horizon (Parker-Holder et al.; Decart et al., 2024; He et al., 2025; Sun et al., 2025; Mao et al., 2025; Team et al., 2026; Gao et al., 2026). Their control interfaces, however, are symbolic or textual: frame-level keyboard and mouse states logged from game engines (Decart et al., 2024; Alonso et al., 2024; Yu et al., 2025; Xiao et al., 2025), or free-form text prompts (Zhu et al., 2026; Mao et al., 2025). These commands specify what happens but not how (Sec. 1), and telemetry-logged interfaces further bind each model to the action schema of its data-collection pipeline (He et al., 2025; Tang et al., 2025; Hong et al., 2025; Ye et al., 2026); structured motion inputs (Cao et al., 2025; Fang et al., 2026; Lee et al., 2026; Shin et al., 2026) are precise instead, but per-family and hard to acquire accurately. ShadowDancer instead conditions on a dense per-frame latent extracted from a reference clip, so that the speed, amplitude, and style of an action are supplied by the reference, and any dynamics family with constructible shadow pairs is controlled through one interface.

Latent Actions as Control Interfaces. Latent action models (LAMs) infer controls directly from unlabeled video, serving as interfaces for interactive generation (Bruce et al., 2024; Gao et al., 2025;

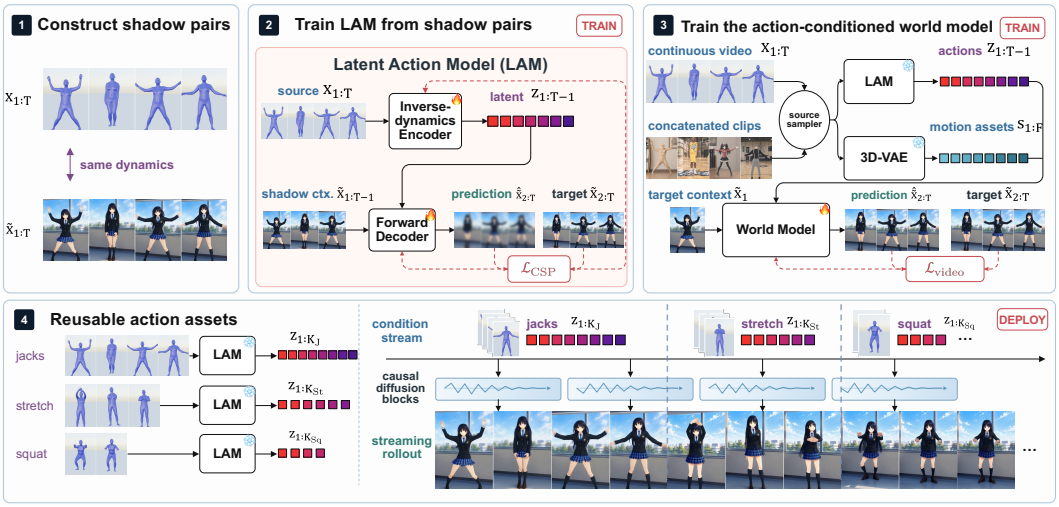

Figure 2: **Overview of ShadowDancer**, taking human motion as the running example; the pipeline is identical for camera, object, and every other dynamics family. (1) A shadow pair renders one dynamics d twice, $x{=}R(d,c)$ and $\tilde{x}{=}R(d,\tilde{c})$, with the remaining factors independently resampled. (2) Cross-shadow prediction trains the LAM: the encoder reads each z_t from the source transition, and the decoder predicts the next shadow frame from $(\tilde{x}_t, z_t)$, so only what the pair shares survives in z. (3) With the LAM and 3D-VAE frozen, a video diffusion backbone is fine-tuned into a block-causal world model, conditioned on the action trajectory z and on source motion assets s, and trained on mixed continuous and concatenated sources. (4) At deployment, variable-length action assets (z, s) are stored, composed along the condition stream, and streamed through causal blocks; a new action costs one pass through the frozen encoders.

Jang et al., 2025), cross-embodiment policy learning (Chen et al., 2024; Ye et al., 2025; Bu et al., 2025; Kim et al., 2025; Chen et al., 2025b;c), and world-model pretraining (Edwards et al., 2019; Rybkin et al., 2019; Wang et al., 2025b; Garrido et al., 2026). Unlike imitation learning, which aims to execute a demonstrated behavior on a target embodiment, our focus is the representation itself: a reusable dynamics signal, extracted from a demonstration, that transfers across appearances and environments. The core difficulty is that inverse-dynamics latents couple control to whatever appearance happens to co-occur with it: clip-local reconstruction admits shortcut solutions through context cues (Yang et al., 2025; Garrido et al., 2026) and is non-identifiable across contexts (Locatello et al., 2019; Khemakhem et al., 2020), so the same dynamics need not map to a consistent latent (Jiang et al., 2026). Existing remedies constrain the latent (information bottlenecks, quantization) (Bruce et al., 2024; Gao et al., 2025) or regularize it toward motion-centric references (Jiang et al., 2026); in practice, regularization reduces but does not eliminate the entanglement, and all of these operate on a single observation of each transition, so the desired invariance is imposed as a penalty rather than exhibited by the data. Shadow pairing changes the data instead: each dynamics is observed twice under independently resampled appearance, so invariance holds by construction and the objective merely has to read it off.

Invariance by Construction. Our supervision transplants a proven recipe from self-supervised learning: a representation is defined by the pairs it is trained on, since what a pair varies is discarded and what it preserves is kept. Contrastive learning drops crops and color while retaining visual identity (Chen et al., 2020), CLIP retains the semantics language can express (Radford et al., 2021), and masked video objectives extend this to spatiotemporal structure (Tong et al., 2022; Assran et al., 2025). In each case the objective does not create the invariance; it reads off the invariance that the pairing already wrote into the data. Dynamics, however, resists pixel-space augmentation: no post-hoc transform of one video can resample its appearance while preserving its motion frame by frame. The shadow pair moves the augmentation into the renderer, which grounds cross-shadow prediction (Sec. 3.2) as a sound way to recover the preserved dynamics.

3 Method

Overview. Fig. 2 summarizes ShadowDancer. The method has three parts, in pipeline order: shadow pairs make dynamics identifiable in the data (Sec. 3.1); *cross-shadow prediction* distills it into the

unified representation z (Sec. 3.2); and a video diffusion backbone, conditioned on z for control and on source assets for motion detail, becomes a block-causal interactive world model whose deployment we describe last (Sec. 3.3). The Shadow Library (Sec. 3.4) implements the pairing protocol at scale. Three objects recur throughout: the LAM extracts a latent trajectory z invariant to everything its pairing resamples (*appearance*, Sec. 3.1); a frozen 3D-VAE encodes the aligned source-detail stream s; and together $a = (z, s)$ is the reusable *action asset* the world model consumes, while the *unified dynamics representation* always denotes z alone, unified in the sense of one encoder, one latent space, and one interface for every dynamics family.

3.1 The Shadow Formulation

We model a video $x = (x_1, \ldots, x_T)$ as the output of a rendering process

$$x = R(d, c), \tag{1}$$

where $d = (d_1, \ldots, d_T)$ is a time-varying *dynamics* trajectory, the factor that drives change between frames, and c collects the remaining generative factors, such as subject identity, scene, materials, and lighting, which we call *appearance*. In this notation our title's metaphor is literal: a video is a *shadow* of its dynamics, one rendering of d under one particular c. Observing x alone, d and c are entangled: distinct (d, c) pairs explain the same pixels, the root cause of the shortcut and non-identifiability failures of latent action learning (Yang et al., 2025; Jiang et al., 2026; Locatello et al., 2019).

A **Shadow pair** comprises two independent renders of the same dynamics,

$$\big(\, x = R(d, c),\ \tilde{x} = R(d, \tilde{c}) \,\big), \quad \tilde{c} \sim p(c), \tag{2}$$

where the two videos are frame-synchronized: frame t of x and frame t of $\tilde{x}$ share the identical d_t. Both videos are shadows of the same d, and each is therefore the Shadow of the other: the relation is symmetric, and "the Shadow of a video" refers to a fellow projection of its dynamics rather than a projection of the video itself. Because $\tilde{c}$ is resampled independently of c, the only factor preserved across the two renders is d: source appearance cannot systematically predict target appearance.

The split between d and c in Eq. (1) is not fixed a priori; it is chosen by the pairing protocol. Replaying a body motion while resampling the camera assigns the camera to c; replaying a camera trajectory while resampling the scene assigns it to d. This operational definition is what makes the paradigm any-action: designating a factor as controllable requires only constructing pairs that preserve it. Shadow pairing therefore differs from data augmentation: augmentation encourages invariance to nuisances chosen within a fixed task, whereas the pairing chooses the task variable itself, by deciding what is preserved. **Accordingly, *appearance* in this paper denotes the factors a pairing resamples, and *dynamics* what it preserves.** Dynamics need not exclude visual traits: a fire spell without its flames is no longer the same action, so the protocol keeps such traits in the representation by simply preserving them across the pair, and they transfer together with the motion.

3.2 Cross-Shadow Prediction

Objective. Cross-shadow prediction turns the pairing constraint into a learning objective. We instantiate it with a latent action model (LAM): an inverse-dynamics encoder models each transition (x_t, x_{t+1}) with a variational posterior $q_\phi(z_t \mid x_t, x_{t+1})$ over latent actions $z_t \in \mathbb{R}^{d_z}$, $t = 1, \ldots, T-1$, and a forward decoder p_θ performs next-frame prediction. Standard LAMs (Gao et al., 2025; Jiang et al., 2026) read and predict the same clip. Our decisive difference is the cross-shadow arrangement: the encoder reads the source transition, while the decoder predicts the other shadow, reconstructing $\tilde{x}_{t+1}$ from the target's own previous frame $\tilde{x}_t$ and the source-derived action z_t. Training minimizes the β-VAE objective (Higgins et al., 2017; Alemi et al., 2017):

$$\mathcal{L}^{\mathrm{CSP}}_{\theta,\phi} = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big(- \mathbb{E}_{q_\phi(z_t \mid x_t, x_{t+1})} \big[\log p_\theta(\tilde{x}_{t+1} \mid \tilde{x}_t, z_t) \big] + \beta \, \mathrm{KL}(q_\phi(z_t \mid x_t, x_{t+1}) \,\|\, p(z_t)) \Big), \tag{3}$$

where $p(z_t)$ is a fixed prior $\mathcal{N}(0, I)$.

Why this works. Source-context information in z_t beyond what is needed to infer d cannot improve population prediction of $\tilde{x}$, whose appearance is supplied by its own context $\tilde{x}_t$ and resampled independently of c: appearance ceases to help, and the bottleneck, pressured by the KL term, has room only for the shared dynamics. Unlike feature-alignment regularizers (Jiang et al., 2026), the predictive supervision itself supplies this cross-context constraint. With probability p a video also pairs with itself ($\tilde{x} = x$), recovering the standard single-video LAM objective as a degenerate case; these self-pairs admit unpaired real video into the same training stream (Sec. 3.4) and fall outside the guarantee below.

Formal guarantee. This preference for dynamics is not only an intuition; we prove it. Whenever the pairing protocol makes the shared dynamics *observable* from the source video, resamples the two sides *independently*, so that source and target are conditionally independent given it, and *separates* distinct dynamics in their effect on the target, the minimal representation sufficient for predicting one shadow from the other is that dynamics itself, up to an invertible reparameterization (Theorem C.2). The three conditions and the proofs are stated formally in Sec. C, which also delimits its scope.

Factor-selective control. Dynamics itself is composite: the observer's own motion (ego, *i.e.* he camera) and the motion of everything observed (scene) are superimposed in every video. Pair construction separates them within one latent space and specifies the readout: pairs preserving only the camera trajectory supervise a `cam` head, pairs preserving only the scene dynamics supervise a `dyn` head, and pairs preserving both supervise a `full` head; what the `dyn` factor contains is itself defined by each family's pairing protocol, the arm trajectory alone in our robot manipulation pairs, the body motion in our human pairs. At inference the same reference video can therefore be read as a pure camera action, a pure scene action, or their joint; the mechanism is detailed in Sec. B.1.

3.3 From Representation to World Model

The conditioning design follows two principles: (i) the action latent z carries the unified, transferable control, and (ii) the source stream s supplies the high-frequency motion detail that an invariance-constrained bottleneck is deliberately too small to carry.

Action-conditioned generation. We fine-tune a pretrained video diffusion transformer (DiT) (Peebles & Xie, 2023; Chen et al., 2025a; Wan et al., 2025) on Shadow Library clips with flow matching (Liu et al., 2023): the frozen LAM's actions are fused into the timestep embedding for AdaLN modulation and attended through a dedicated cross-attention arm, while the source video is encoded by the 3D-VAE and injected as *motion assets*, through channel concatenation at the input convolution and as additional cross-attention context (Sec. B.2). Because source and target form a shadow pair, copying source appearance is not rewarded, so the generator is trained to read the source for motion detail rather than appearance.

Block-causal conversion for interactive rollout. An interactive world model must generate forward-only over long horizons, while diffusion backbones are bidirectional and fixed-length. Following recent causal conversions of video diffusion (Gu et al., 2025; Huang et al., 2025), we fine-tune the bidirectional model into a block-causal generator: latent frames are grouped into blocks, attention is causal across blocks, and denoising is conditioned on the clean history of previous blocks. At inference the model rolls out block by block with a key–value cache, consuming a stream of latent actions and emitting a stream of frames.

Actions as reusable assets. At inference, the model receives a first frame $\mathbf{I}_0$ and a demonstration, given as reference clips or stored action assets: the frozen encoder reads the demonstration into per-frame latent actions z, and the block-causal generator rolls out a video in which the depicted world re-enacts the demonstrated dynamics. Because control is a latent trajectory rather than a per-frame key state, an "action" in ShadowDancer is a *variable-length dynamics segment*: extract $a = (z_{1:K}, s)$ from a clip of any duration, store it, replay it in any world, and compose segments by concatenation, so teaching the model a new action takes one pass through the frozen encoders, with no vocabulary design, labels, or fine-tuning. Blocks group latent frames only for causal denoising, while the action stream stays per-frame, so a single action may span many blocks and its temporal detail is not tied to the block rate. At deployment, discrete commands retrieve and concatenate short stored assets rather than supplying one continuous demonstration; we therefore expose the generator to the same discontinuous conditioning during training (Sec. B.2), so a command at inference is a library lookup matching the input statistics the generator was trained on.

Table 1: **Quantitative comparison of action transfer across dynamics families.** Best in **bold**.

Method	Human motion		First-person combat		Third-person action		Camera control		Robot manipulation	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	ATE↓	RPE↓	PSNR↑	LPIPS↓
Olaf-World (Jiang et al., 2026)	18.2	0.288	13.0	0.532	12.6	0.506	0.072	0.021	14.0	0.478
ShadowDancer (ours)	**22.4**	**0.184**	**17.0**	**0.354**	**17.4**	**0.309**	**0.005**	**0.003**	**22.6**	**0.116**

3.4 The Shadow Library

The shadow protocol, which replays the dynamics and resamples everything else, admits many implementations, and we deliberately build a library rather than a single engine. Animation suites replay articulated human motion across characters, scenes, lighting, and cameras; open-world game environments replay trajectories across districts, weather, and time of day; robotic simulators replay manipulation across arms, objects, and tabletops; camera-trajectory sources replay one path through different scenes. All sources emit the same artifact, frame-synchronized clip sets that share one dynamics, so a new action family integrates by contributing one more shadow script.

Real-world video, which lacks exact shadows, enters the same stream as degenerate self-pairs ($\tilde{x} = x$): it contributes visual diversity and realism priors, while the invariance pressure is supplied by the synthetic pairs. Genuine pairs supply the identifying signal; self-pairs broaden visual support.

4 Experiments

4.1 Implementation Details

Following Gao et al. (2025); Jiang et al. (2026), the LAM has latent dimension d_z=32 and is trained on the Shadow Library at half the world model's spatial resolution ($H/2\times W/2$) with β=0.01; real videos enter as degenerate self-pairs. The world model is built on the SkyReels-V2-1.3B I2V DiT backbone (Chen et al., 2025a), fine-tuned with flow matching and converted to a block-causal generator. All experiments run on NVIDIA H200 GPUs; further implementation details are in Sec. D.

Datasets The Shadow Library spans articulated human motion (SMPL-X sequences re-rendered in Blender across characters, scenes, lighting, and cameras, together with paired renders driven by Cao et al. (2025) that replay one motion under a resampled character and environment), robotic manipulation (ManiSkill (Tao et al., 2024)), first-person open-world games (GTA- and Cyberpunk-style urban scenes with first-person weapon fire and driving), third-person games (Unreal Engine scenes; Monster Hunter-style action), and camera trajectories, both around articulated subjects and through static scenes (DL3DV (Ling et al., 2023)); unpaired real video (OpenX robot corpora (O'Neill et al., 2024), Internet clips (Ju et al., 2024)) joins as self-pairs. Held-out shadow pairs from every source form the evaluation sets; the full composition of the Shadow Library is tabulated in Sec. D.

4.2 Action Transfer Across Dynamics Families

Our first experiment tests the central claim in one protocol: a single model with one latent interface re-enacts every family of dynamics in a new environment. For every family in the Shadow Library (human motion, first-person combat, third-person action, robot manipulation, and camera control), we hold out shadow pairs, extract the action from video x, and generate from the first frame of its Shadow $\tilde{x}$, which realizes the identical, frame-synchronized dynamics. Four families are scored by frame-aligned reconstruction against $\tilde{x}$ (PSNR, LPIPS (Zhang et al., 2018)); camera control is scored by trajectory error (ATE/RPE (Sturm et al., 2012)) between the target trajectory and poses recovered from the generations with VGGT (Wang et al., 2025a). The baseline is Olaf-World (Jiang et al., 2026), the state-of-the-art self-reconstruction latent-action model: we match the training data and world-model backbone, inject its latent through the same routes, and retain its original single-head, z-only design, which has no source-asset stream. Table 1 reports the results: ShadowDancer leads on every family by a wide margin. The gap is visible in Fig. 3: Olaf-World routinely produces warped, twisting subjects and spurious camera motion, because its entangled latent cannot cleanly extract the intended dynamics from the source, so what it transfers is the action mixed with whatever appearance happened to accompany it; with the shadow-trained latent, the same interface re-enacts the demonstrated dynamics faithfully in the new environment. Two checks in Sec. D corroborate this. Before any generation is involved, linear probes on z show that cross-shadow training preserves action content while reducing

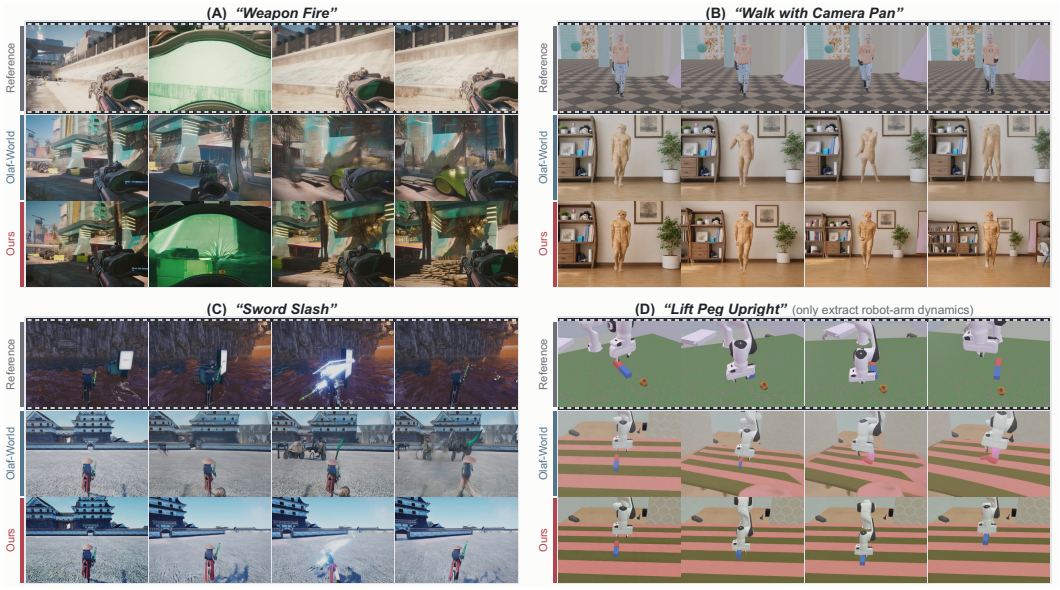

Figure 3: **Qualitative action transfer against Olaf-World.** Top: reference videos providing the action across four families; below: frame-synchronized generations in the new environment. Olaf-World warps subjects and leaks spurious motion; ShadowDancer re-enacts the reference dynamics faithfully.

Table 2: **Long action rollout.** $>50\%$ favors ours.

ShadowDancer (ours) over	Action Control	Action Fidelity	Long-horizon Consistency
Olaf-World (Jiang et al., 2026)	94%	95%	94%
Yume-1.5 (Mao et al., 2025)	88%	86%	91%
LingBot-World 2.0 (Gao et al., 2026)	78%	83%	64%

Table 3: **Impact of key components.**

Setting	z	Paired z	Assets (s)	PSNR↑	LPIPS↓
(a) Olaf-World (z only)	✓	×	×	12.44	0.555
(b) + pairing	✓	✓	×	14.75	0.448
(c) assets, no action	×	–	✓	15.07	0.420
(a′) assets + unpaired z	✓	×	✓	14.92	0.428
(d) ShadowDancer (ours)	✓	✓	✓	**16.35**	**0.376**

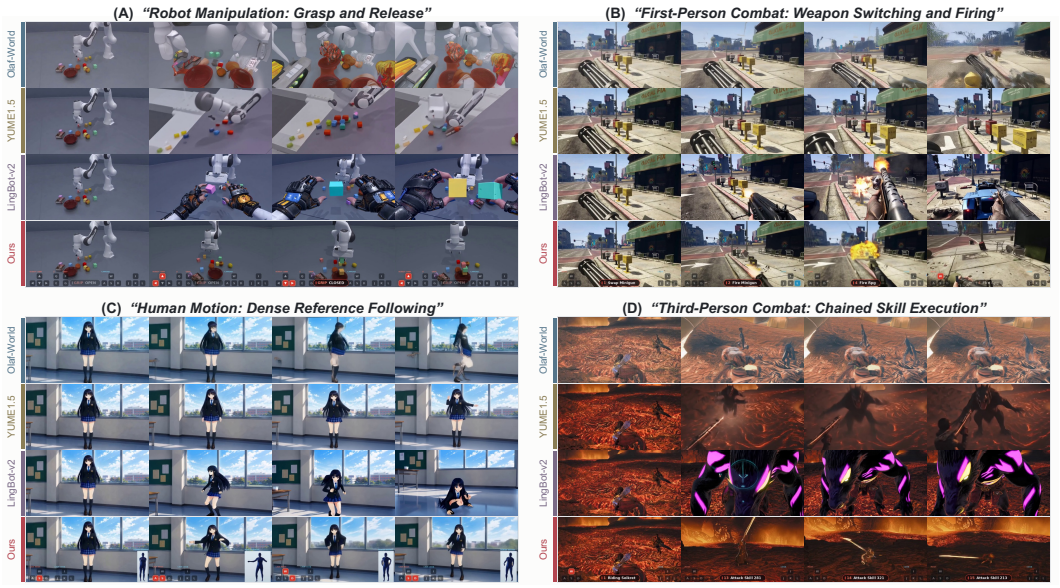

Figure 4: **Qualitative long action rollouts.** Each comparison starts from a shared first frame and follows the same command stream; rows are methods, columns are time steps along the rollout. Where the baselines drift or degrade, ShadowDancer stays aligned.

leakage of the resampled scene; and the gain is not bought by blur, since ShadowDancer roughly halves the Fréchet Video Distance of Olaf-World on three reconstruction families.

4.3 Long Action Rollout

Our second experiment evaluates whether learned action assets remain composable over long horizons. Discrete commands are mapped to canonical action assets by lookup and streamed into long rollouts that chain several commands per video, spanning navigation, aiming and firing, and skill casts across first- and third-person worlds, together with human-motion and robot-manipulation streams. Unlike the transfer setting, a chained command stream has no ground-truth realization, so control must be *judged* rather than measured: following the preference protocol of Zhu et al. (2026), each ShadowDancer rollout is compared head-to-head against a baseline's in a blinded two-alternative forced choice (2AFC), judged by a VLM (Fable 5) over all pairwise comparisons, with a random 20% of the judgments human-audited (Sec. D). The comparison runs along three action-centric axes: *action control* (is the commanded action performed at all), *action fidelity* (is the specific weapon/object/motion preserved), and *long-horizon consistency* (no drift or degradation as the rollout extends); visual fidelity is excluded as orthogonal to control. We compare against the state-of-the-art open interactive world models LingBot-World 2.0 (Gao et al., 2026), driven by a camera-pose trajectory with text-triggered events, and Yume-1.5 (Mao et al., 2025), driven by text and keystrokes, alongside Olaf-World's latent interface on our backbone; each receives the same command stream through its native interface. The comparison is deliberately system-level: the interfaces receive different information by design, and how faithfully a user's commands survive each interface is precisely what is being measured. Table 2 reports the favor rates, and Fig. 4 shows representative rollouts (in full in the supplementary videos); where the baselines drift, hallucinate, or degrade over the horizon, ShadowDancer stays aligned to the command stream.

4.4 Ablation Studies

Effectiveness of shadow pairing and source assets As detailed in Table 3, we study the contribution of each component on the transfer split of Sec. 4.2: *(a)* the prior z-only latent interface, a separately trained model following the Olaf-World recipe (Jiang et al., 2026) on our data; *(b)* the same interface with the LAM trained on shadow pairs instead of self-reconstruction; *(c)* source assets s without the action z; *(a′)* the assets with an *unpaired* z from the Olaf-recipe LAM; and *(d)* the full model; (b), (c), and (d) are inference-time ablations of one checkpoint, while (a) and (a′) are separately trained under the matched recipe. Pairing is the foundation: once the latent is constrained to carry dynamics alone, the same interface becomes transferable ((a) vs. (b)). Assets are a strong carrier on their own, since the source video already contains the action and its spatial detail, so (c) matches or surpasses the bottlenecked z. The decisive comparison is (a′) vs. (d): they share both a latent and the assets and differ only in whether the latent was trained with shadow pairing, and the paired latent improves PSNR by 1.4; an unpaired latent, by contrast, adds nothing over assets alone ((a′) vs. (c)). The top-line gain is thus attributable to cross-shadow pairing, not to the assets alone or to conditioning on any latent.

Transferring unseen actions Actions introduced *after* training through game mods, an unseen character and a greatsword whose attack is a swing rather than a shot, replay faithfully on held-out maps from one recorded demonstration each (Sec. D.7): the encoder reads dynamics rather than memorized categories, so coverage grows with new demonstrations rather than new training runs.

5 Conclusion

ShadowDancer enables any-action, frame-level control of interactive video world models by demonstration. Through shadow pairs and cross-shadow prediction, it learns a unified dynamics representation that discards, by construction, whatever the pairing resamples and keeps whatever it preserves, so users can extract the dynamics of any clip as a reusable action asset and replay it in new environments without labels, estimators, or fine-tuning. Our experiments demonstrate clear improvements over latent-action and interactive world model baselines in both action transfer and long action rollout, across dynamics families ranging from human motion and camera control to gameplay and robot manipulation. ShadowDancer opens new possibilities for interactive world models, which can now be driven by showing rather than telling, from entertainment to simulation-based training of embodied agents. We discuss limitations and future directions in Sec. A.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing (grammar, phrasing, copy-editing, and condensing of author-written text, including adapting the manuscript to the conference template) and to polish code (refactoring and cleanup of author-written implementation and evaluation scripts). We did not use generative AI tools to generate research ideas, to design the method or the experiments, to produce or analyze experimental results, or to write the substantive content of the paper or the core implementation. All AI-assisted edits to text were reviewed by the authors, and all AI-assisted edits to code were reviewed and tested for correctness by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

ShadowDancer lowers the effort required to direct video generation: any recorded clip becomes a reusable action asset that can be replayed in a new environment. The same property that makes the interface general also makes misuse easier, since a behavior demonstrated once can be transferred onto a different subject or scene without labels or fine-tuning. We note two mitigations relevant to our setting. First, the Shadow Library is built almost entirely from synthetic sources (animation suites, game engines, and robotic simulators) whose assets we control, and the real video we use enters only as unpaired self-pairs from existing public corpora; we introduce no new human-subject data and collect no personally identifying information. Second, the model transfers dynamics rather than identity: the representation is trained to discard exactly the appearance factors that a pairing resamples, so it carries motion, not the likeness of the person who performed it. We nonetheless recommend that any deployment of action-transfer world models pair generation with provenance signalling such as watermarking, and we will release artifacts under terms that prohibit non-consensual depiction of real individuals.

REPRODUCIBILITY STATEMENT

We have made the following efforts to support reproduction. The formal claim of Sec. 3.2 is stated with its full set of assumptions and proved in Sec. C, which also delimits the scope of the "any dynamics" claim. Sec. B.2 specifies, route by route, the conditioning architecture summarized in Sec. 3.3, and Sec. B.1 the factor-selective readout. Backbone, optimizer, schedule, resolution, and block-causal conversion settings are given in Sec. D, together with the full composition of the Shadow Library (per-source clip counts and the pairing protocol used for each), the held-out evaluation splits, and the exact prompts and judging protocol of the blinded 2AFC rollout study. The pairing protocols are scripts over publicly available engines and simulators rather than a new proprietary dataset, so a source can be regenerated from its description. We will release the code, including the Shadow Library generation scripts and the evaluation harness.

REFERENCES

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.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *NeurIPS*, 2024.

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.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video

generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

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 *ICML*, 2024.

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*, 2025.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Unifying precisely 3D-enhanced camera and human motion controls for video generation. In *SIGGRAPH Asia*, 2025.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. IGOR: Image-GOal Representations are the Atomic Control Units for Foundation Models in Embodied AI. *arXiv preprint arXiv:2411.00785*, 2024.

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. villa-X: enhancing latent action modeling in vision-language-action models. *arXiv preprint arXiv:2507.23682*, 2025b.

Yi Chen, Yuying Ge, Weiliang Tang, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *ICCV*, 2025c.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. URL https://oasis-model.github.io.

Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *ICML*, 2019.

Zhixue Fang, Xu He, Songlin Tang, Haoxian Zhang, Qingfeng Li, Xiaoqiang Liu, Pengfei Wan, and Kun Gai. 3D-aware implicit motion control for view-adaptive human video generation. *arXiv preprint arXiv:2602.03796*, 2026.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *NeurIPS*, 2024.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *ICML*, 2025.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

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.

Google DeepMind. Veo. Model page. URL https://deepmind.google/models/veo/.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, pp. 217–227, 2020.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

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

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*, 2026.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *NeurIPS*, 2024.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTATS*, 2020.

Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. UniSkill: Imitating human videos via cross-embodiment skill representations. *arXiv preprint arXiv:2505.08787*, 2025.

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.

Yao-Chih Lee, Zhoutong Zhang, Jiahui Huang, Jui-Hsien Wang, Joon-Young Lee, Jia-Bin Huang, Eli Shechtman, and Zhengqi Li. Generative video motion editing with 3D point tracks. In *CVPR*, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *arXiv preprint arXiv:2312.16256*, 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. https://deepmind.google/models/genie/. Blog post.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Oleh Rybkin, Karl Pertsch, Andrew Jaegle, Konstantinos G. Derpanis, and Kostas Daniilidis. Learning what you can do before doing anything. In *ICLR*, 2019.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. MotionStream: Real-time video generation with interactive motion controls. In *ICLR*, 2026.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 573–580, 2012.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, and Qinglin Lu. Hunyuan-GameCraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467, 2021.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

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*, 2025b.

Halbert White. Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, 3(5):535–549, 1990. doi: 10.1016/0893-6080(90)90004-5.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Jiange Yang, Yansong Shi, Haoyi Zhu, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. CoMo: Learning continuous latent motion from internet videos for scalable robot learning. *arXiv preprint arXiv:2505.17006*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent Action Pretraining from Videos. In *ICLR*, 2025.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. GameFactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Shangwen Zhu, Qianyu Peng, Zhao Pu, Zhilei Shu, Xiangrui Ke, Zhaohu Xing, Zizhao Tong, Zeqing Wang, Xinyu Cui, Zian Zheng, Huangji Wang, Jian Zhao, Yeying Jin, Fan Cheng, and Ruili Feng. Incantation: Natural language as the action interface for multi-entity video world models. *arXiv preprint arXiv:2605.18601*, 2026.

APPENDIX

A LIMITATIONS AND FUTURE DIRECTIONS

Limitation ShadowDancer depends on two preconditions. First, action assets must be prepared in advance: at deployment the model consumes reference clips or stored assets, so a dynamics for which no demonstration exists must first be captured or authored. Second, shadow pairs are easy to construct in games and simulators, which replay the same dynamics under resampled appearance on demand, but are much harder to collect in the real world, where such re-runs are rarely possible; real video therefore enters training only as self-pairs, contributing visual realism rather than the identifying signal.

Future directions Both preconditions point to generative remedies. First, asset generation: because the encoder is appearance-invariant by construction, a demonstration does not have to be filmed. Off-the-shelf video generation (Google DeepMind; Wan et al., 2025) is not sufficient by itself, however: a usable asset set consists of many clips of the same subject that stay strictly consistent in appearance and starting state and differ only in the performed action, whereas current models generate each video independently. Designing an *asset-generation* video model with this cross-clip consistency would let the action library scale with the generative ecosystem rather than with engine coverage. Second, shadows for real video: since the pairing is a protocol rather than an engine, video editing models that change a clip's appearance while keeping its motion could manufacture shadows of real footage, extending the identifying supervision from synthetic worlds to the real domain.

B MORE MODEL DETAILS

B.1 FACTOR-SELECTIVE READOUT OF THE UNIFIED REPRESENTATION

The main text (Sec. 3.2) states that ego and scene dynamics are separated within one latent space; this section gives the mechanism. The inverse-dynamics encoder follows the standard LAM architecture (Gao et al., 2025; Jiang et al., 2026): a spatiotemporal transformer with causal temporal attention whose readout at frame $t{+}1$ infers z_t, so in implementation the posterior conditions on the full causal prefix, $q_\phi(z_t \mid x_{1:t+1})$, rather than the transition alone; clips of any length are encoded in a single pass. We use *three prompt slots*: three learnable tokens (`cam`, `dyn`, and `full`) are prepended to every frame's patch sequence, attend jointly through the shared encoder, and each owns a variational readout head. The pairing protocol routes supervision among them: pairs preserving only the camera trajectory (scene and subject resampled or frozen) supervise the `cam` head; pairs preserving only scene motion (camera resampled) supervise the `dyn` head; pairs preserving both supervise the `full` head. A per-sample mask $(m_{\text{cam}}, m_{\text{dyn}}) \in \{0,1\}^2$ selects one head, so each head is supervised only by the factor its pairs preserve. What constitutes the `dyn` factor is not fixed a priori but defined by each family's pairing protocol: our robot manipulation pairs preserve the arm trajectory alone, so for this family `dyn` reads out arm dynamics and nothing else in the scene; our human pairs preserve the body motion. This is the protocol nature of the Shadow Library: designating a factor as `dyn` requires only constructing pairs that preserve it. The per-source control channel $(m_{\text{cam}}, m_{\text{dyn}})$ is listed in the composition table of Sec. D. All heads share one encoder and one latent space $\mathcal{Z}$, so downstream consumption is identical; yet at inference they are independent dials: the same reference video yields a pure camera action, a pure scene action, or their joint, and a body motion, a camera move, and an object interaction are simply different trajectories z.

B.2 CONDITIONING ARCHITECTURE

This section details the conditioning routes summarized in Sec. 3.3 of the main text. The backbone operates on latents of a causal 3D video VAE with temporal compression $r{=}4$; we therefore group the per-frame actions by latent frame, $\mathcal{Z}_f = (z_{r(f-1)+1}, \ldots, z_{rf})$, following (Yu et al., 2025). The modulation route fuses the group-averaged action $\bar{z}_f$ into the per-latent-frame timestep embedding,

$$e_f = \psi(\tau) + \alpha_z \, W_z \, \bar{z}_f, \tag{B.1}$$

where $\psi(\tau)$ embeds the diffusion timestep and e_f drives the adaptive layer norm of every DiT block, providing global, low-frequency control. The cross-attention route preserves all r action tokens: the

spatial hidden states h_f of latent frame f attend to a per-frame context $\mathcal{C}_f$ through a dedicated arm,

$$h_f \leftarrow h_f + W_o \,\mathrm{Attn}\big(h_f,\, \mathcal{C}_f\big), \tag{B.2}$$

with $\mathcal{C}_f = [\, W_a \mathcal{Z}_f \,]$ initially holding the projected action tokens. W_o is zero-initialized, so the cross-attention arm is an exact no-op at step 0; the modulation gate α_z is instead initialized to a positive constant, since a zero-initialized gate received vanishing gradient and stalled in practice.

At deployment a command is a library lookup rather than one continuous demonstration, so during training we replace the continuous source, with some probability, by a concatenation of *canonical action chunks* drawn from a fixed asset library; a command at inference then matches the input statistics the generator was trained on.

The bottleneck d_z=32 is a feature rather than a limitation: invariance demands that z be too small to smuggle appearance, but it is then also too small to carry the high-frequency content of a motion, such as the precise articulation of limbs and the exact timing of contacts. We therefore feed the source video itself to the generator as assets. Let $s = \mathcal{E}_{\mathrm{vae}}(x^{\mathrm{src}})$ be the source's 3D-VAE latents. On the input side, s is channel-concatenated with the noisy latents u and a binary availability mask m, then embedded by the input 3D convolution:

$$h^{(0)} = \mathrm{Conv3D}\big(\,[\,u;\, s;\, m\,]\,\big); \tag{B.3}$$

on the context side, each latent frame s_f is patchified into tokens that join the cross-attention context of Eq. (B.2),

$$\mathcal{C}_f = \big[\, W_a \mathcal{Z}_f;\;\; W_s \,\mathrm{patchify}(s_f)\,\big], \tag{B.4}$$

alongside a per-frame semantic embedding of the source in a third, parallel arm.

C Formal Guarantee for Shadow Learning

This section makes precise the sense in which a shadow pair makes an arbitrary dynamics family learnable. An unconditional statement is impossible: if two dynamics produce exactly the same observations, no algorithm can distinguish them; if a finite-dimensional code is too small for the intrinsic dynamics dimension, no continuous encoder can store them. We therefore state the necessary observability, separation, and capacity conditions explicitly. The result is agnostic to the renderer and to whether the dynamics describe a human, camera, robot, object, or an entire multi-action trajectory.

Our argument is related in spirit to results showing that paired views can isolate invariant content from independently changing style (von Kügelgen et al., 2021) and can restore identifiability in multi-view nonlinear models (Gresele et al., 2020). Here we do not assume independent latent components or attempt to recover a privileged coordinate system. We prove the task-specific statement needed by cross-shadow prediction: the coarsest statistic sufficient for predicting one shadow from the other is exactly the shared dynamics, up to an invertible reparameterization.

C.1 Setup and Assumptions

Let D denote the dynamics to be represented. It may be one transition d_t, a variable-length segment $d_{1:T}$, or a full rollout. Let X be a source video, let B collect the target-side information already available to the decoder (for example, the previous target frame or clean block history), and let Y be the target next frame or block. All variables take values in standard Borel spaces, so the conditional distributions below exist. Define the target transition kernel

$$K_d(b) := \mathbb{P}(Y \in \cdot \mid B = b, D = d). \tag{C.1}$$

Three explicit conditions capture the shadow construction.

Assumption A1 (Independent context resampling)**.** *The source is conditionally independent of the target rendering once the shared dynamics is fixed:*

$$X \perp\!\!\!\perp (B, Y) \mid D. \tag{C.2}$$

This is the probabilistic statement of "replay the dynamics, resample everything else." It allows both views to depend arbitrarily on D.

Assumption A2 (Source observability)**.** *There is a measurable map h such that*

$$D = h(X) \qquad \text{almost surely.} \tag{C.3}$$

Thus the requested dynamics is actually visible in the source clip. This is necessary for any deterministic encoder of one video to recover it.

Assumption A3 (Target overlap and separation)**.** *There is a probability measure ν on the target-context space such that $\mathbb{P}(B \in \cdot \mid D = d)$ is equivalent to ν for almost every d; that is, every dynamics is rendered over the same target contexts up to null sets. Moreover, distinct dynamics induce distinct transition kernels:*

$$K_d = K_{d'} \quad \Longrightarrow \quad d = d'. \tag{C.4}$$

Here equality means equality ν-almost everywhere. We also assume that $d \mapsto [K_d]_\nu$ is a measurable map into a standard Borel coding of these kernels. Equivalently, every pair $d \neq d'$ differs in its effect in a set of target contexts of nonzero probability. Separation is necessary. The stated mutual absolute continuity is a transparent sufficient coverage condition that prevents the decoder-side context B from perfectly revealing D.

Definition C.1 (Cross-shadow representation)**.** A representation $Z = s(X)$ is *cross-shadow sufficient* when

$$Y \perp\!\!\!\perp X \mid (B, Z), \tag{C.5}$$

and is called *minimal* when every other cross-shadow sufficient statistic determines it. Minimality does not impose a particular coordinate system; two minimal statistics may differ by a bijection.

C.2 Identification Theorem

Theorem C.2 (Shadow identification)**.** *Under Assumptions A1–A3, define the conditional-law statistic*

$$\Gamma(x) := [K_{h(x)}]_\nu. \tag{C.6}$$

Then: (i) $\Gamma(X)$ is cross-shadow sufficient and minimal; (ii) it is invariant to every source-context factor resampled by the shadow construction; and (iii) every cross-shadow sufficient representation $Z = s(X)$ determines D. Consequently, every minimal sufficient Z and D determine one another and are equivalent up to a one-to-one reparameterization.

Proof. Because $D = h(X)$ and $X \perp\!\!\!\perp (B, Y) \mid D$, for almost every x and target context b,

$$\begin{aligned} &\mathbb{P}(Y \in \cdot \mid B = b, X = x) && \text{(C.7)}\\ &\quad = \mathbb{P}(Y \in \cdot \mid B = b, D = h(x)) && \text{(C.8)}\\ &\quad = K_{h(x)}(b). && \text{(C.9)} \end{aligned}$$

The conditional law of Y given (B, X) therefore depends on X only through $\Gamma(X)$, which proves sufficiency. The same identity shows invariance: source videos with the same D have the same Γ regardless of source context.

It remains to prove minimality without allowing the target-side variable B to do the identification. Let $Z = s(X)$ be any cross-shadow sufficient statistic and write $L_z(b) = \mathbb{P}(Y \in \cdot \mid B = b, Z = z)$. Sufficiency and Eq. (C.9) give

$$K_D(B) = L_Z(B) \qquad \text{almost surely.} \tag{C.10}$$

Because Z is a function of X, Assumption A1 implies $B \perp\!\!\!\perp Z \mid D$. Therefore, for almost every pair (d, z) in the support of (D, Z), Eq. (C.10) holds under $\mathbb{P}(B \in \cdot \mid D = d)$. Equivalence of this measure to ν in Assumption A3 upgrades this to the kernel-class identity

$$[K_D]_\nu = [L_Z]_\nu \qquad \text{almost surely.} \tag{C.11}$$

The Borel injection $d \mapsto [K_d]_\nu$ has a measurable inverse on its image, so

$$D = (d \mapsto [K_d]_\nu)^{-1}([L_Z]_\nu) =: r(Z) \qquad \text{almost surely.} \tag{C.12}$$

Consequently, $\Gamma(X) = [K_{r(Z)}]_\nu$ is a measurable function of every sufficient Z, proving minimality. Conversely, the same inverse recovers D from $\Gamma(X)$. Any other minimal sufficient Z is also a function of Γ, so Z and D are bijective up to null sets. □

Corollary C.3 (Cross-shadow likelihood learns a dynamics representation)**.** *Assume the relevant conditional laws admit densities with respect to a common dominating measure and have finite population log loss. Let $Z = s(X)$ and allow the decoder to range over all conditional distributions $q(Y \mid B, Z)$. Under Assumptions A1–A3, Z attains the same population negative log-likelihood as a decoder given the full source X if and only if Z is cross-shadow sufficient. Therefore every Bayes-optimal code contains D, and every minimal Bayes-optimal code is a one-to-one reparameterization of D.*

Proof. For a fixed Z, the optimal decoder is the true conditional law $\mathbb{P}(Y \mid B, Z)$. Since Z is a function of X, the gap from the full-source Bayes risk is

$$\mathcal{R}^*(Z) - \mathcal{R}^*(X) = \mathbb{E}[\mathrm{KL}(\mathbb{P}(Y \mid B, X) \,\|\, \mathbb{P}(Y \mid B, Z))] \tag{C.13}$$

$$\geq 0. \tag{C.14}$$

Equality holds exactly when $Y \perp\!\!\!\perp X \mid (B, Z)$. The conclusion then follows from Theorem C.2. □

The theorem identifies a representation, not a preferred axis system: if φ is any bijection, $\varphi(D)$ is equally valid. This is the strongest identifiability one can request of an unlabeled latent. A nonminimal Bayes-optimal encoder may additionally store source appearance, but that information is provably unnecessary for cross-shadow prediction and is removed by passing to the minimal sufficient statistic. In our model, the finite variational bottleneck is the mechanism that favors this minimal solution; the theorem does not claim that a KL penalty or a particular SGD run uniquely guarantees it.

Remark C.4 (Deterministic representation versus variational channel)**.** Theorems C.2 and Corollary C.3 concern a deterministic statistic $Z = s(X)$, such as encoder parameters or a posterior mean. The sampled Gaussian channel $q_\phi(z_t \mid x_{1:t+1})$ and the β-weighted KL term in our implementation form a rate–distortion approximation to this ideal. They favor a compact code but do not, by themselves, guarantee exact sufficiency or minimality at finite rate.

C.3 Realization and Learning with Differentiable Networks

The previous theorem is nonparametric. We next show that, under standard regularity conditions, its representation and predictor can be realized by differentiable neural networks.

Theorem C.5 (Differentiable neural representation)**.** *Assume Assumptions A1–A3. In addition, suppose that the supports of X, B, and D are compact subsets of finite-dimensional Euclidean spaces; h is continuous; D admits a continuous embedding $e : \mathcal{D} \to \mathbb{R}^m$ into the chosen latent dimension; and the target rendering is*

$$Y = F(B, D) \tag{C.15}$$

for a continuous F. Then there exist sequences of differentiable neural networks (E_n, G_n) such that

$$\sup_x \|E_n(x) - e(h(x))\|_2 \longrightarrow 0 \tag{C.16}$$

and

$$\sup_{b,x} \|G_n(b, E_n(x)) - F(b, h(x))\|_2 \longrightarrow 0. \tag{C.17}$$

Thus a differentiable neural encoder can represent D up to an invertible coordinate change while discarding source context, and its decoder can attain arbitrarily small cross-shadow reconstruction error.

Proof. Set $E_0 = e \circ h$. Since e is a continuous injection from a compact space into a Hausdorff space, its inverse is continuous on $e(\mathcal{D})$. Hence

$$G_0(b, z) = F\big(b, e^{-1}(z)\big), \qquad z \in e(\mathcal{D}), \tag{C.18}$$

is continuous and satisfies $G_0(B, E_0(X)) = F(B, D) = Y$. Extend G_0 continuously from the compact set $\mathcal{B} \times e(\mathcal{D})$ to a compact Euclidean neighborhood. Universal approximation with a smooth, bounded, nonconstant sigmoidal activation (such as logistic or $\tanh$) then gives sequences of differentiable networks that approximate E_0 and this extension of G_0 uniformly (Hornik, 1991). Uniform continuity of the extension of G_0 turns both approximations into Eqs. (C.16) and (C.17). □

Corollary C.6 (Exact reconstruction identifies the representation)**.** *In the deterministic setting of Theorem C.5, suppose a code $Z = s(X)$ and decoder G attain*

$$\mathbb{E}\|Y - G(B, Z)\|_2^2 = 0. \tag{C.19}$$

Then Z determines D almost surely. If Z is minimal among exact reconstruction codes, it is a one-to-one reparameterization of D.

Proof. Equation (C.19) implies $Y = G(B, Z)$ almost surely, hence $Y \perp\!\!\!\perp X \mid (B, Z)$. The claim follows from Theorem C.2. Conversely, in this deterministic setting any sufficient Z admits the exact decoder $G(B, Z) = \mathbb{E}[Y \mid B, Z] = Y$, so minimal exact-reconstruction codes and minimal sufficient codes coincide. □

Theorem C.5 is a realizability statement. For stochastic renderers, the same construction applies if the rendering randomness is included in B and available to the decoder; alternatively one may approximate the full conditional kernel with a neural density decoder and use Corollary C.3. There is also a standard statistical learning counterpart: for i.i.d. shadow pairs, bounded second moments, globally optimized empirical squared loss, and a neural-network sieve whose capacity grows at a controlled rate, neural nonparametric regression is risk-consistent (White, 1990). This consistency result learns the composite Bayes predictor; it does not by itself imply convergence of the internal encoder. The representation conclusion follows separately, at a minimal population optimum, from Theorem C.2 and Corollaries C.3 and C.6. It does not assert finite-sample recovery, a particular convergence rate, convergence of every encoder parameterization, or that arbitrary nonconvex optimization finds a global solution.

C.4 Scope of the "Any Dynamics" Claim

The identification result places no semantic restriction on D and no component-independence assumption on the renderer. For heterogeneous action families, the guarantee applies family by family: let F index the pairing protocol (the family), and let Assumptions A1–A3 hold conditionally on F with a family-specific context measure ν_F. The overlap condition of Assumption A3 is then required only *within* a family, never across families; a human context need never support a robot dynamics. Since F is itself observable from the source (which family a clip shows is visible in the clip), conditioning on F loses nothing, and the theorem identifies D within every family. What is shared *across* families—one encoder and one latent space $\mathcal{Z}$—is an architectural choice rather than a consequence of the theorem: it is what makes the per-family identified representations unified in practice, as the cross-family transfer results of the main text support empirically. Therefore, any video dynamics for which one can (i) replay the dynamics while independently resampling the other generative factors, (ii) observe the requested dynamics in the source, and (iii) render overlapping contexts, within its family, in which distinct dynamics have distinct effects admits a context-invariant cross-shadow representation. Under the additional finite-dimensionality, compactness, continuity, and latent-capacity conditions of Theorem C.5, that representation and its decoder admit differentiable neural realizations to arbitrary accuracy.

If separation fails, define

$$d \sim d' \quad \Longleftrightarrow \quad K_d = K_{d'}. \tag{C.20}$$

Provided this equivalence class is itself observable from X, the same proof identifies $[D]$—exactly the part of the dynamics that can affect a shadow— and no predictive method can do better. If source observability fails even for $[D]$, a deterministic representation of one source video is impossible; in general only the conditional-law statistic $x \mapsto \mathbb{P}(Y \in \cdot \mid B = \cdot, X = x)$ is identified (it may be representable as a mixture over equivalence classes). These boundaries prevent "any dynamics" from being read as a promise to recover information that no video contains.

Finally, the guarantee applies to exact shadows satisfying Assumption A1. The degenerate self-pairs used to import unpaired real video are an empirical addition and do not, by themselves, provide the identifiability guarantee.

Table 4: **Composition of the Shadow Library.** Shadow pairs are two frame-synchronized renders of one dynamics trajectory with the remaining generative factors resampled; unpaired corpora enter as self-pairs. (cam, dyn) is the control channel each source supervises. Counts are approximate.

Category	Data source	Pairing	(cam, dyn)	#Seqs	#Frames	Mixture
Human motion	SMPL-X re-render, body dynamics only	shadow pair	(0, 1)	∼1k	∼1.1M	13.6%
	SMPL-X re-render, camera only	shadow pair	(1, 0)	∼1.1k	∼1.2M	10.9%
	SMPL-X re-render, camera + body	shadow pair	(1, 1)	∼1k	∼1.2M	8.2%
	Paired character renders (Cao et al., 2025)	shadow pair	(1, 1)	∼2k	∼0.3M	8.8%
Robot manipulation	ManiSkill (Tao et al., 2024), arm only	shadow pair	(0, 1)	∼1.1k	∼1M	2.7%
	ManiSkill (Tao et al., 2024), camera only	shadow pair	(1, 0)	∼1.2k	∼1.1M	6.8%
	ManiSkill (Tao et al., 2024), camera + arm	shadow pair	(1, 1)	∼1.2k	∼1.1M	5.4%
First-person games	GTA-style urban driving and weapon fire	shadow pair	(1, 1)	∼0.9k	∼0.4M	8.8%
	Cyberpunk-style night city	shadow pair	(1, 1)	∼3.4k	∼1.6M	8.2%
Third-person games	Unreal Engine scenes (melee, spellcasting)	shadow pair	(1, 1)	∼30 scenes	∼0.6M	–
	Monster Hunter-style action	shadow pair	(1, 1)	∼1k	∼0.5M	–
Static-scene camera	DL3DV (Ling et al., 2023)	self-pair	(1, 0)	∼1k scenes	∼0.3M	6.8%
Unpaired real video	MiraData Internet clips (Ju et al., 2024)	self-pair	(1, 1)	∼5k	∼6.5M	5.4%
	OpenX BridgeData (O'Neill et al., 2024)	self-pair	(0, 1)	∼29k	∼0.8M	0.1%
	OpenX FurnitureBench (O'Neill et al., 2024)	self-pair	(0, 1)	∼5k	∼2.2M	4.1%
	OpenX Berkeley UR5 (O'Neill et al., 2024)	self-pair	(0, 1)	∼1k	∼0.1M	2.7%
	OpenX Jaco Play (O'Neill et al., 2024)	self-pair	(0, 1)	∼1.1k	∼0.1M	2.7%
	OpenX Stanford HYDRA (O'Neill et al., 2024)	self-pair	(0, 1)	∼0.6k	∼0.3M	2.7%
	OpenX RoboTurk (O'Neill et al., 2024)	self-pair	(0, 1)	∼2k	∼0.1M	2.0%

D Additional Experimental Details and Ablations

D.1 Composition of the Shadow Library

Table 4 summarizes the sources of the Shadow Library: how each source is paired, which control channel it supervises via the per-sample mask $(m_{\text{cam}}, m_{\text{dyn}})$ of Sec. B.1, and its approximate scale and share of the sampling mixture.

D.2 Training and Inference Details

The LAM is trained with β=0.01 and a fixed prior $\mathcal{N}(0, I)$, always at half the world model's spatial resolution; real videos enter the stream as degenerate self-pairs. The self-pair probability is 0.5 for the human body-dynamics source and 0 for the other paired sources; since the unpaired real-video and static-scene camera sources are always self-paired, roughly one third of training samples are self-pairs. The world model is first fine-tuned bidirectionally with flow matching, then converted to a block-causal generator with 3 latent frames per block and rolled out with a key–value cache at inference. The action-transfer comparisons (Table 1) use the model at 480×720 (LAM at 240×360); the long-rollout model (Table 2) generates at 544×960 (LAM at 272×480), warm-started from a lower-resolution checkpoint and fine-tuned at the target resolution for 10k steps on $8\times$H200 GPUs with fully-sharded data parallelism. At inference we sample with 30 flow-matching denoising steps and classifier-free guidance 5.0.

D.3 Evaluation Details

In Table 1, the human-motion, first-person, and third-person families are scored on 81-frame clips from the first frame, averaging 45–50 held-out pairs each; robot manipulation uses a smaller set of held-out ManiSkill pairs, and the camera split uses held-out camera-trajectory pairs. Since monocular pose recovery is scale-ambiguous, the VGGT-recovered trajectory is aligned to the target with a similarity transform (Sim(3)) before computing ATE/RPE. The ablation of Table 3 uses a 12-pair subset of the transfer split under the same protocol, covering the first-person combat and third-person action families; the reported numbers average the two. **Long-rollout judging protocol** The comparison of Table 2 uses 16 long command streams. For each stream, ShadowDancer and the baseline generate from the same first frame under the same commands; both rollouts are cut at the same boundaries into 240-frame segments, and corresponding segments are compared, giving roughly 64 comparison pairs per baseline. Each pair is anonymized as A and B with randomized assignment and judged by a VLM (Fable 5) in a forced choice on each of the three axes, with no tie option; a

Table 5: **LAM representation probes** on held-out shadow pairs (24 pairs) and UE clips (60 clips; character chance $= 0.04$, scene chance $= 0.20$). The Olaf-recipe row follows (Jiang et al., 2026): self-reconstruction with feature alignment and a single head.

LAM training	Cross/self MSE ratio↓		Linear probe on z	
	1st-person	3rd-person	Char. (content)↑	Scene (appear.)↓
Self-recon.+reg., single head	1.265	1.067	0.050	0.650
Cross-shadow, 3-prompt (ours)	**1.017**	1.103	**0.450**	**0.433**

human audit reviews a random 20% of the judgments to validate the VLM's decisions. The judge receives both segments, the command list, and an instruction of the following form:

> You are judging two video rollouts, A and B, generated by two different systems from the same first frame under the same sequence of action commands: `<commands>`. Judge the actions, not the image quality. Answer three independent questions; for each you must answer exactly "A" or "B", even if the difference is small. (1) *Action control*: in which video are the commanded actions actually performed, at the commanded times? (2) *Action fidelity*: which video better preserves the specific weapon, object, and motion of each commanded action? (3) *Long-horizon consistency*: which video stays coherent for longer, without drifting away from the commanded behavior, freezing, or degrading as the rollout proceeds? Ignore differences in visual style, sharpness, or aesthetics, except where degradation makes the action itself unreadable. Output exactly three lines: `control: A|B`, `fidelity: A|B`, `consistency: A|B`.

D.4 Representation-Level Probes

Before touching the world model, we probe the latent z itself on held-out shadow pairs (Table 5). First, the *cross-pair transfer ratio*: with the LAM's own decoder, we reconstruct the target from its visual context using either the target's own action ($z(\tilde{x})$, "self") or the shadow source's ($z(x)$, "cross"); the ratio $\mathrm{MSE}_{\text{cross}}/\mathrm{MSE}_{\text{self}}$ measures how much a cross-appearance z degrades reconstruction. Second, two linear probes on pooled z over held-out UE clips: decoding the *character* and the *scene*. In this evaluation set each character is given its own distinct action set, so character identity is readable from motion alone and the character probe measures action content (higher is better), while the scene probe measures pure appearance (lower is better). Since characters differ in appearance as well as in motion, character accuracy alone could in principle reflect either; the baseline contrast below rules out the appearance route: the self-reconstruction latent leaks *more* appearance by the scene probe yet decodes character only at chance, so character accuracy tracks the motion set rather than looks. The cross-shadow latent transfers almost losslessly on first-person pairs (ratio 1.02 vs. 1.27) and dominates both probes: it decodes character at $11\times$ chance while leaking less scene identity, whereas the self-reconstruction latent is at chance on content yet leaks *more* appearance — regularization does not substitute for pairing. Note that the cross/self ratio is informative only for a content-bearing latent: the baseline decodes content at chance, so its near-unity third-person ratio reflects an uninformative z, for which self and cross reconstructions are equally unguided, rather than successful transfer.

D.5 Distribution-Level Quality

Reconstruction metrics could in principle be won by a conservative model that stays close to the reference at the cost of realism. We rule this out with Fréchet Video Distance (FVD) on the same generations as Table 1, computed over 16-frame windows (stride 8) on the three large reconstruction families (Table 6); the camera family is scored by trajectory error rather than reconstruction, and the robot split is too small for a stable estimate. ShadowDancer roughly halves the FVD of Olaf-World on every family, so its reconstruction advantage comes with better, not worse, distribution-level realism.

D.6 What the Action Latent Carries

The conditioning consumes both the latent z and the source assets s; this probe tests, at inference, which one carries the action. For each family we take two shadow pairs A and B from different sources and, from the base input (z_A, s_A), form four interventions: replacing the latent (z_B, s_A), shuffling the latent along time, replacing the assets (z_A, s_B), and degrading the assets to a quarter

Table 6: **Distribution-level quality (FVD↓)** on the same generations as Table 1.

Family	Olaf-World	ShadowDancer
First-person combat	555.8	**318.3**
Third-person action	842.0	**422.2**
Human motion	511.6	**253.9**

Table 7: **What the action latent carries.** PSNR against target A under each intervention, 3–4 couples per family; Δ relative to *base*.

Family	base	$z \leftarrow B$	shuffle z	$s \leftarrow B$	degrade s
First-person combat	14.77	-2.59	-1.57	-2.48	-0.36
Third-person action	16.55	-3.01	-0.33	-2.15	-0.41
Camera	21.73	-1.92	-2.09	-3.28	-0.13
Robot arm	23.71	-0.98	-1.95	-0.15	$+0.20$
Body motion	22.25	-0.53	$+0.19$	-0.79	-0.37

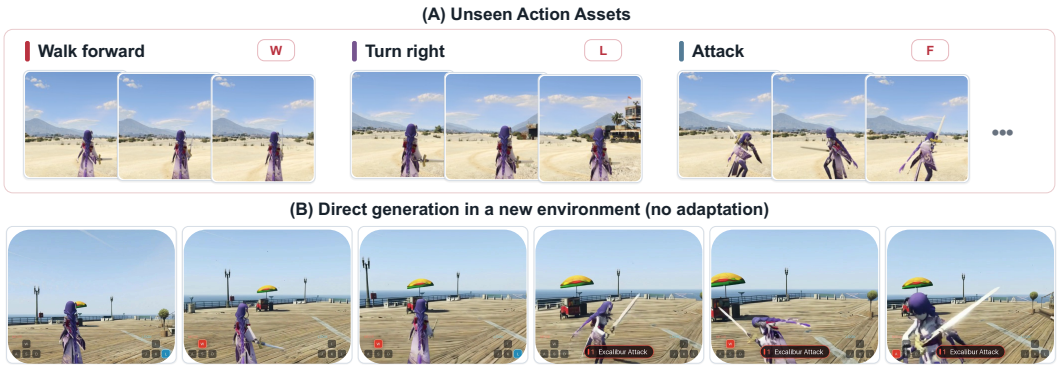

Figure 5: **Transferring unseen actions:** assets recorded from a modded, unseen character (A) drive generation in a new environment (B).

resolution (z_A, $s_A^{\downarrow 4}$). We report PSNR against target A (Table 7). On four of five families, replacing or shuffling z drops PSNR sharply while degrading s barely moves it: the commanded action lives in the latent, and the assets supply appearance and high-frequency detail.[1]

D.7 Transferring Unseen Actions

A latent interface is only as general as its behavior on dynamics it has never seen, so we test transfer on actions introduced through game mods after training: an unseen player character and a two-handed sword whose attack is a swing rather than a shot; neither asset appears in the Shadow Library, nor does the motion either one produces. We record walking, turning, and sword attacks from the modded character as action assets. The source clips are recorded on one map and the extracted actions replayed on held-out maps, so transfer to a new environment is built into the protocol; Fig. 5 shows the results. The modded dynamics survive this replay: this character's gait and this sword's swing, motion the model never observed, are reproduced on a held-out map, indicating that the encoder reads dynamics rather than memorized action categories. This generalization is where the scaling potential of the paradigm lies: a behavior demonstrated after training joins the action library at the cost of one pass through the frozen encoders, so coverage grows with new demonstrations rather than with new training runs.

D.8 Architecture Ablations

Latent size and self-pair ratio The choice d_z=32 is inherited from prior latent-action models (Gao et al., 2025; Jiang et al., 2026), where this size was validated as the best operating point for self-reconstruction; since real video enters our training stream as self-pairs under the same reconstruction objective (Sec. 3.4), their operating point carries over, and it is consistent with the invariance argument of Sec. B.2 that the bottleneck should stay too small to smuggle appearance. A budget-matched sweep confirmed this choice: the representation probes were insensitive to $d_z \in \{16, 64\}$ and to forcing the self-pair ratio, while reconstruction favored $d_z \leq 32$, so we retain d_z=32.

[1] On the body-motion family the source render already depicts the motion in full, so the assets alone can clone it and the probe cannot separate the two channels; the character-render source shares this property and is likewise not probed. This is a limit of the probe on exact-replay families, consistent with the assets being a strong carrier ((c) in Table 3).