ENFORCING CAUSAL ROUTING IN EMBODIED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied world models predict how a robot's environment responds to commanded actions, enabling policy evaluation and planning with varied actions. We show that current models produce such predictions partly by peeking at the future. Frames generated at early timesteps can be corrupted by subsequent actions and the goal, a phenomenon we term *causal leakage*. Existing benchmarks typically replay recorded actions, which match the ground-truth video, so leakage does not lower scores and goes undetected. We introduce Causeway, a world model that routes information along causal paths while preserving visual quality. United Causal Attention aligns the attention structure with this causal order while retaining semantic guidance from language; Biomimetic Sampling Memory sustains prediction fidelity over episode-length rollouts by anchoring the initial observation alongside recent context. To make causal leakage measurable, we introduce the Causal Leakage Probe, to our knowledge the first black-box test of whether subsequent actions affect earlier frames or the goal affects the video directly. Causeway's future-action leakage is 63% below the best baseline's, with no measurable goal leakage under the tested interventions. It leads on six of twelve WorldArena 2.0 metrics and reduces FVD on full-episode DROID rollouts from 113.6 to 8.9 over the same backbone. Code and weights will be released.

1 INTRODUCTION

Embodied world models predict how a robot's environment changes under a sequence of commanded actions, starting from an initial observation (Zhu et al., 2025b; Guo et al., 2026; BWM Team, 2026; Liao et al., 2026). Unlike general video generators assessed mainly by visual realism and prompt adherence (Wang et al., 2025; Seedance, 2026), embodied world models should be judged by whether the predicted video shows the correct consequences of the commanded actions. Recent studies apply them to policy evaluation (Li et al., 2025b; Shang et al., 2026), synthetic data generation (GigaAI, 2025), and planning (Zhang et al., 2026). In these applications, the model must faithfully comply with the commanded actions, including those that deviate from recorded trajectories. Current models, however, do not always meet this requirement. The frames they generate at a given timestep can be corrupted by information from subsequent actions and the task goal. Specifically, even when the commanded actions inevitably lead to task failure, these models may still generate a video with a successful ending (Yang et al., 2026; Peng et al., 2026); and in every baseline we evaluate, changing the subsequent actions alters the frames that precede them (Section 3.4).

We refer to this phenomenon as **causal leakage**. It stems largely from video generators, which denoise all frames of a clip jointly and let every frame attend to the text prompt and to every other frame (Wang et al., 2025; Aditi et al., 2026). Embodied world models inherit this design when they add actions as another condition (BWM Team, 2026; Liao et al., 2026; Zhu et al., 2025a; Li et al., 2025a; Aditi et al., 2026; Bi et al., 2026), so subsequent actions can influence earlier frames through attention, even when each action is aligned with its own frame (Zhu et al., 2025b; Guo et al., 2026). Also, the predicted video attends to the instruction directly (Aditi et al., 2026; Bi et al., 2026), so the task goal opens a second path, steering predictions toward the intended outcome regardless of commanded actions (Section 3.4). These paths have gone largely unnoticed, because existing benchmarks either replay recorded actions or score visual quality, temporal consistency, and instruction following (Hu et al., 2025; Qin et al., 2025; Shang et al., 2026). Under replay, leaked information agrees with the actual outcome, so a model that relies on it still scores well. Therefore,

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.

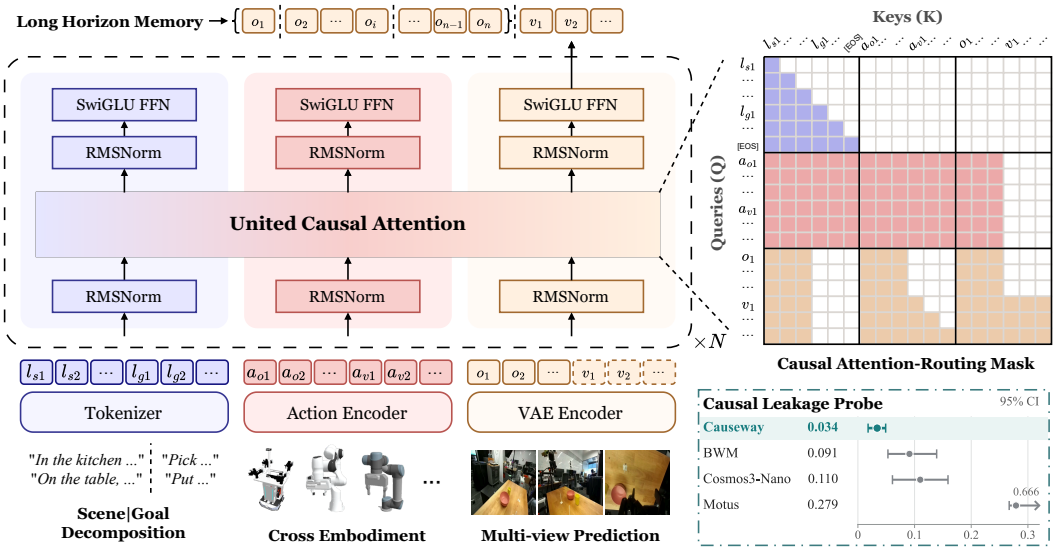


Figure 1: **Overview of Causeway.** Language, action, and video towers interact through United Causal Attention, whose mask (top right) lets frames attend to the scene description and the actions up to their own time. Biomimetic Sampling Memory (top) supports full-episode rollouts. The Causal Leakage Probe (bottom right) shows Causeway has the lowest leakage among compared models.

leakage becomes visible only when the commanded actions depart from the recording, which is precisely the setting required by the applications discussed above.

In this work, we treat the causal order of physical interaction as a design principle for embodied world models. Recent world models generate causally across frames (Huang et al., 2025), which removes access to subsequent actions but constrains the joint generation of frames within chunks. Others drop the instruction (BWM Team, 2026; Li et al., 2026), which discards the semantic information that language provides. We introduce **Causeway**, an embodied world model that applies this principle to cross-modal attention, long-horizon memory, and evaluation. Its United Causal Attention (UCA) lets each generated frame read only the actions up to its own time, routes the goal to the video through the action stream, and keeps generated frames from updating action representations. UCA thus removes the direct leakage paths within each chunk while keeping frame attention bidirectional and the scene description visible. Biomimetic Sampling Memory (BioSaM), inspired by episodic future thinking (Schacter et al., 2017), sustains episode-length rollouts by anchoring the initial scene and current imagined state. The Causal Leakage Probe (CLP) checks what the attention mask alone cannot guarantee. It changes only the future actions, or only the goal, and measures whether the generated video responds.

In experiments, Causeway exhibits the lowest leakage among the compared world models. When only the future actions change, its CLP ratio is 63% below that of the best baseline, and it shows no measurable goal response under the tested interventions. This causal design also preserves video quality. On the WorldArena 2.0 evaluation (Shang et al., 2026), Causeway leads on six of the twelve metrics and raises trajectory accuracy from 47.7 to 60.1 and perspectivity from 82.8 to 97.1 over the best reported scores. On full-episode rollouts from DROID (Khazatsky et al., 2024), it lowers FVD from 113.6 to 8.9 over the same backbone, and retains 87% of its early-rollout exterior SSIM by the end of the episode, versus 55% for the backbone.

Our main contributions are as follows:

- We identify causal leakage in embodied world models, relate it to the attention design inherited from video generators, and show that it occurs in every baseline we evaluate.
- We propose Causeway, whose United Causal Attention removes the direct leakage paths and Biomimetic Sampling Memory sustains episode-length rollouts.

- We introduce CLP, to our knowledge the first test of whether the generated video responds to future actions, or to the task goal when the actions are fixed.

2 METHOD

We first describe the model and training objective (Section 2.1), then United Causal Attention (Section 2.2), Biomimetic Sampling Memory (Section 2.3), and the Causal Leakage Probe (Section 2.4).

2.1 DESIGN OVERVIEW

Causeway generates video in successive chunks from an initial observation, the action stream, and an optional instruction. Following Cosmos 3 (Aditi et al., 2026), we represent video using a pretrained video autoencoder (Wang et al., 2025) and process language, action, and video tokens with a Mixture-of-Transformers architecture (Liang et al., 2025). When provided, the instruction $\mathbf{c}$ consists of scene and goal spans constructed during data preparation (Section A); otherwise, $\mathbf{c}$ is empty. Each modality has separate attention and feed-forward parameters: the frozen language tower encodes the instruction, the action tower represents the commanded trajectory $\mathbf{a}$, and the video tower predicts visual latents, with cross-stream communication passing through attention. The routing in Section 2.2 determines which actions and language spans each video token may attend to.

For chunk n, let $\mathbf{z}^n$ denote the training latent window and $\mathbf{M}^n$ a binary mask that equals one at conditioning positions and zero at the L target positions. Let $\mathbf{z}^n_{\text{ctx}}$ contain the supplied context latents at conditioning positions. At noise level σ, the input is

$$\tilde{\mathbf{z}}^n_\sigma = \mathbf{M}^n \odot \mathbf{z}^n_{\text{ctx}} + (1 - \mathbf{M}^n) \odot \big((1-\sigma)\mathbf{z}^n + \sigma\boldsymbol{\epsilon}\big), \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \quad \sigma \sim \mathcal{U}[0,1]. \tag{1}$$

During training, context is taken from recorded video and augmented as described in Section 2.3; during inference, it consists of the initial observation and earlier predictions.

The video tower predicts the rectified-flow velocity (Liu et al., 2023), with video supervision restricted to the L target positions. Here, $w(\sigma)$ is the flow-loss weighting function and λ weights the action reconstruction loss:

$$\mathcal{L} = \mathbb{E}\Big[w(\sigma)\Big\|(1 - \mathbf{M}^n) \odot \Big(\mathbf{v}_\theta(\tilde{\mathbf{z}}^n_\sigma, \sigma, \mathbf{M}^n, \mathbf{c}, \mathbf{a}) - (\boldsymbol{\epsilon} - \mathbf{z}^n)\Big)\Big\|_2^2\Big] + \lambda\mathcal{L}_{\text{action}}, \tag{2}$$

where $\mathcal{L}_{\text{action}}$ supervises reconstruction of the clean action inputs. At inference, target positions are initialized with Gaussian noise at $\sigma = 1$ and iteratively denoised to $\sigma = 0$ while the supplied context and actions remain fixed. The decoded predictions then serve as context for subsequent chunks.

2.2 UNITED CAUSAL ATTENTION

United Causal Attention (UCA) specifies which keys each query may attend to (Fig. 1 and Table 1). Let $\mathcal{C}_{\text{scene}}$, $\mathcal{C}_{\text{goal}}$, $\mathcal{A}$, $\mathcal{V}_{\text{cond}}$, and $\mathcal{V}_{\text{gen}}$ denote the sets of scene, goal, action, conditioning-video, and target-video tokens, respectively. For an action query q, the allowed keys are

$$\mathcal{K}_{\text{act}}(q) = \mathcal{C}_{\text{scene}} \cup \mathcal{C}_{\text{goal}} \cup \mathcal{A} \cup \mathcal{V}_{\text{cond}}. \tag{3}$$

For a target-video query q, they are

$$\mathcal{K}_{\text{gen}}(q) = \mathcal{C}_{\text{scene}} \cup \mathcal{V}_{\text{cond}} \cup \mathcal{V}_{\text{gen}} \cup \{k \in \mathcal{A} : \tau(k) \le \tau(q)\}. \tag{4}$$

Here τ is measured in latent-frame coordinates: actions associated with latent frame $p \ge 1$ occupy $(p-1, p]$, so the mask includes that frame's own actions. We retain bidirectional attention among target-video tokens to support joint generation within each chunk. Conditioning-video queries use

$$\mathcal{K}_{\text{cond}}(q) = \mathcal{C}_{\text{scene}} \cup \mathcal{A}_{\text{cond}} \cup \mathcal{V}_{\text{cond}}, \tag{5}$$

where $\mathcal{A}_{\text{cond}}$ contains action tokens designated as conditioning context. Together with Eqs. (3) and (5), this blocks feedback from target-video tokens to action representations, including through the visual context across layers. Action representations within a chunk therefore depend on the supplied inputs rather than the video being generated.

Intervention representation. We treat the commanded action as an input to the dynamics and the target video as its predicted consequence. Blocking target-video feedback encodes this physical

Table 1: **United Causal Attention.** Allowed keys for each query group. Blank entries denote blocked connections. Language attention follows token order, with scene preceding goal.

Query tokens	$\mathcal{C}_{\text{scene}}$	$\mathcal{C}_{\text{goal}}$	$\mathcal{A}$	$\mathcal{V}_{\text{cond}}$	$\mathcal{V}_{\text{gen}}$
language	causal	causal			
action	✓	✓	✓	✓	
conditioning video	✓		$\mathcal{A}_{\text{cond}}$	✓	
target video	✓		$\tau(k) \leq \tau(q)$	✓	✓

direction as a design principle: for fixed commands, language, and visual context, the action representation is independent of the sampled target-video outcome. Actions may still read the supplied visual context, which is fixed for the current chunk. This separation concerns the representation of an intervention; temporal non-anticipation additionally concerns which action times may influence each predicted frame.

Goal–scene decomposition. Scene descriptions specify the observed environment, whereas goals specify intended outcomes. UCA exposes scene tokens to action and video queries, and goal tokens only to action queries outside the language tower. The scene span precedes the goal span, and language attention is causal, so scene-token representations do not incorporate subsequent goal tokens. Goal information therefore reaches the video tower only through action representations.

UCA constrains direct attention paths while retaining bidirectional interactions within the action and target-video streams. Because these interactions can transmit information across layers, we assess temporal leakage and residual goal dependence at the output level using CLP (Section 2.4).

2.3 Biomimetic Sampling Memory

Because open-loop generation must preserve the initial scene while tracking action-induced changes, Biomimetic Sampling Memory (BioSaM) allocates a fixed context budget to both the observed starting state and the most recent predictions (Fig. 2). The design draws an analogy with episodic future thinking, in which remembered experience supports the construction of imagined events (Schacter et al., 2017).

Memory selection. Let b be the starting latent-frame index of the next chunk and O the overlap length. The conditioning window contains the O immediately preceding frames and up to H memory frames from the earlier history. The memory budget includes the initial frame, with the remaining $H-1$ slots divided between k_a anchor-side and k_r recent-side samples. Let $G_k(\ell, u)$ denote the k rounded interior points of a geometric sequence with endpoints ℓ and u. The two sets are

$$\begin{aligned} \mathcal{P}_{\text{anchor}} &= G_{k_a}(1, b-O-1), \\ \mathcal{P}_{\text{recent}} &= \{b-d : d \in G_{k_r}(O+1, b-O-1)\}, \end{aligned} \tag{6}$$

where $k_a + k_r = H-1$, with the extra slot assigned to the anchor side when $H-1$ is odd. The initial frame is retained separately, and both sampled sets exclude the overlap. Duplicates from rounding are replaced by available history frames farthest from the selected samples (Section C). This provides dense coverage near the initial observation and the current chunk, with sparse coverage between them. Selected latents retain their temporal positions rather than being reindexed as consecutive frames.

Training augmentations. Training uses recorded context, whereas inference conditions on earlier predictions and may reach temporal positions beyond those encountered during training. We address these differences with temporal position shifts and conditioning noise. For a training chunk with $b > O+1$, we sample $\Delta \sim \mathcal{U}[0, \alpha(b-O)]$ and round it to an integer. We first set the augmented chunk start to $b' = b + \Delta$ and generate its memory position labels using the same sampling rule as inference. Overlap and target labels are shifted by Δ, preserving their spacing.

We then fill these labels with content from the recorded episode. A history label p' receives the latent at $\text{round}(p'/s)$, clipped to the valid history range, where $s = (b'-O)/(b-O)$; the initial frame remains at position zero. An overlap or target label p' receives the latent at $p' - \Delta$, with associated actions mapped consistently. Thus, the augmented window uses the position labels of a longer rollout while preserving the original local dynamics in the overlap and target chunk.

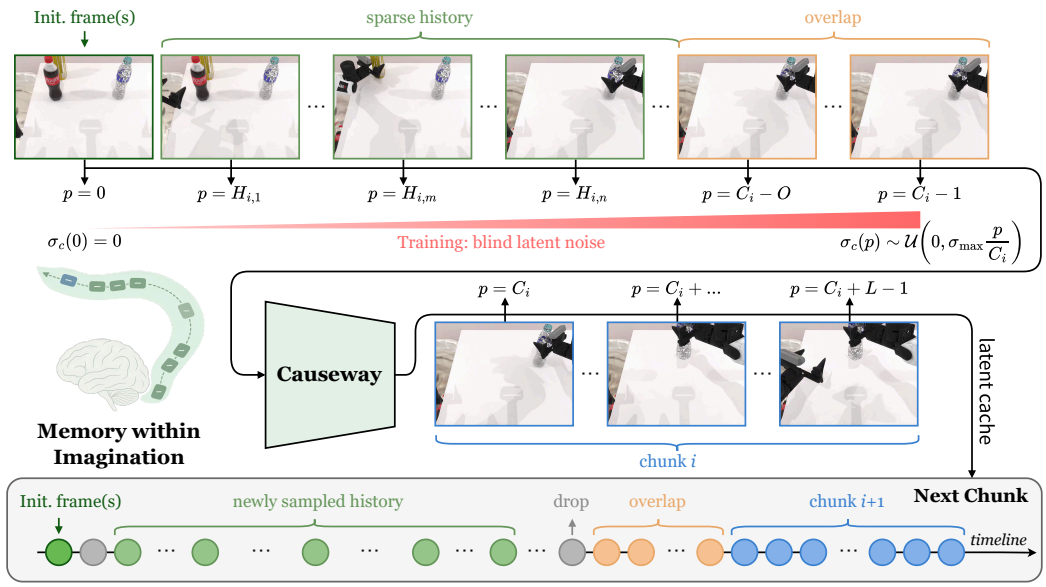


Figure 2: **Biomimetic Sampling Memory.** Each chunk conditions on the initial frame, sparse history with denser sampling near two ends, and an overlap with the preceding chunk. During training, conditioning noise introduces imperfect visual context.

To expose the model to imperfect visual context, we also corrupt each conditioning latent:

$$\begin{aligned} \mathbf{z}_{\mathrm{ctx},p'} &= (1-\sigma_{c,p'})\mathbf{z}_{\mathrm{src},p'} + \sigma_{c,p'}\boldsymbol{\epsilon}_{p'}, \\ \sigma_{c,p'} &\sim \mathcal{U}\left[0, \sigma_o \frac{p'}{b'}\right], \qquad \boldsymbol{\epsilon}_{p'} \sim \mathcal{N}(0, I), \end{aligned} \tag{7}$$

where $\mathbf{z}_{\mathrm{src},p'}$ is the recorded latent assigned to label p' and $0 \le p' < b'$. Without a temporal shift, $b' = b$. The maximum corruption level σ_o controls the conditioning-noise range, which increases toward the current chunk while keeping the initial frame clean. The sampled corruption level is not supplied to the network. This trains the model to condition on imperfect context without requiring a noise-level estimate at inference.

2.4 Causal Leakage Probe

Standard video metrics do not directly test whether changing future actions alters earlier predictions. If these earlier frames change in a way that depends on which intervention was made, they contain information about actions not yet issued. The Causal Leakage Probe (CLP) measures this intervention-specific response through controlled changes to model inputs. It requires reproducible sampling and timestamped output videos, without access to model internals.

Action interventions. For each episode and conditioning time t_c, we generate a baseline video from the recorded actions and construct K intervened streams. Each intervention holds selected action channels at their last commanded values before a future window, leaving all other entries unchanged. The classes hold either all selected channels or individual channel groups. Classes that suppress negligible recorded motion are dropped, and points with too few classes are skipped using a model-independent gate based on recorded actions. Cross-model comparisons use their common probe points. Let $t_\star$ denote the timestamp of the first intervened action row. We measure responses over matched intervals $[t_\star - h, t_\star)$ and $[t_\star, t_\star + h)$, termed the *forbidden* and *licensed* regions. Frames are selected on a common time grid, so the measurement is defined in physical time rather than model-specific chunks or latent frames.

Response energy. At a fixed probe point i, let $\mathbf{x}_{k,s}$ and $\mathbf{x}_{0,s}$ denote the intervened and baseline videos generated with the same observation, text, and seed s. We compute

$$\boldsymbol{\delta}_{k,s} = \phi(\mathbf{x}_{k,s}) - \phi(\mathbf{x}_{0,s}), \qquad \mathbf{d}_{k,s} = \boldsymbol{\delta}_{k,s} - \frac{1}{K}\sum_{j=1}^{K} \boldsymbol{\delta}_{j,s}, \tag{8}$$

where ϕ normalizes pixel values, resizes frames to a common resolution, and flattens them. Centering within each seed removes the response shared by all classes, retaining differences associated with which intervention was made. With S sampling seeds and $\bar{\mathbf{d}}_k = S^{-1}\sum_s \mathbf{d}_{k,s}$, the energy in frame set R is estimated as

$$\widehat{E}_{i,R} = \frac{1}{K}\sum_{k=1}^{K} \|\bar{\mathbf{d}}_k\|_R^2 - \frac{1}{KS(S-1)}\sum_{k=1}^{K}\sum_{s=1}^{S} \|\mathbf{d}_{k,s} - \bar{\mathbf{d}}_k\|_R^2. \tag{9}$$

Here $\|\cdot\|_R^2$ sums squared feature values over the selected frames. For $S \geq 2$ independent sampling seeds, the second term removes the finite-sample contribution to the squared class means, yielding an unbiased estimate of their population energy. For episode e, we average over probe points $\mathcal{P}_e$: $\bar{E}_{e,R} = |\mathcal{P}_e|^{-1}\sum_{i\in\mathcal{P}_e} \widehat{E}_{i,R}$. The action leakage score is

$$\mathrm{CLP}_A = \frac{\sum_e \bar{E}_{e,\mathrm{forbidden}}}{\sum_e \bar{E}_{e,\mathrm{licensed}}}. \tag{10}$$

A score near zero indicates little measured leakage; one indicates equal response in both regions.

Goal interventions. To test residual goal dependence, we hold the observation and actions fixed and vary the goal prompt over K alternatives. We normalize the resulting energy by the action response on the same frames, obtained by holding selected action channels from the first eligible row after t_c through the evaluation window, with the prompt fixed:

$$\mathrm{CLP}_L = \frac{\sum_e \bar{E}_{e,\mathrm{goal}}}{\sum_e \bar{E}_{e,\mathrm{action}}}. \tag{11}$$

Both terms use the same probe points, sampling seeds, evaluation frames, and energy estimator.

Controls. We verify reproducibility by repeating a baseline generation with identical inputs and seed. Each ratio is reported only when its aggregate action-response denominator is positive and exceeds a permutation-based threshold; otherwise it is *not measurable*. Class labels are permuted independently within each seed, and confidence intervals are obtained by bootstrapping episodes.

A separate temporal-shift control advances the supplied actions while retaining the original scoring windows, moving the intervention into the forbidden window, where an increased response is expected. Section B provides the full protocol, prompt alternatives, threshold construction, and validation on synthetic sources and a simulator renderer. CLP assesses dependence under the selected interventions and windows rather than establishing invariance for arbitrary inputs.

3 Experiments

We evaluate Causeway on embodied video quality, long-horizon prediction, and intervention-specific leakage. Ablations assess attention routing, memory selection, and training augmentations.

3.1 Setup

Model and inference. We initialize the language tower from the Cosmos3-Nano Reasoner (Aditi et al., 2026) and the other two towers from its Generator. Across embodiments, we keep the language tower frozen while optimizing the action and video towers and their interfaces on the corresponding data (Section A). For evaluation, Causeway generates full-episode rollouts from the initial observation, recorded actions, and an instruction, using its own predictions as visual context. We use 35 UniPC steps (Zhao et al., 2023) per chunk without classifier-free guidance.

Benchmarks and baselines. WorldArena 2.0 (Shang et al., 2026) evaluates embodied video quality on RoboTwin 2.0 (Chen et al., 2025), providing initial observations, instructions, and actions

Table 2: **WorldArena 2.0 video quality evaluation on RoboTwin 2.0.** Higher is better. Baseline scores are reported in WorldArena 2.0 (Shang et al., 2026); Cosmos-Pred 2.5 (T/A) denotes Cosmos Predict 2.5 contitioned on text/action. Causeway scores are from the official evaluation of our submission. Best in bold; second best underlined.

Model	Visual quality			Content consistency			Physics adherence		3D accuracy		Controllability	
	Image quality	Aesthetic quality	JEPA similarity	Subject consist.	Backgr. consist.	Photom. consist.	Interaction quality	Trajectory accuracy	Depth accuracy	Perspec-tivity	Instruction following	Semantic alignment
Text-driven video generators												
CogVideoX (Yang et al., 2025)	35.8	37.8	93.8	80.8	87.7	35.8	59.4	35.3	91.0	78.3	72.7	**89.8**
Wan 2.2 (Wang et al., 2025)	38.8	39.6	75.8	83.9	90.4	**47.8**	51.8	16.3	77.7	76.6	53.8	88.8
Wan 2.6 (Wang et al., 2025)	**68.2**	44.3	72.3	75.2	86.9	19.0	72.8	11.8	71.4	80.3	85.4	87.3
Veo 3.1 (Google DeepMind, 2025)	66.0	**46.3**	56.9	78.8	87.1	32.5	78.7	12.3	74.2	82.8	**93.3**	86.1
Vidar (Feng et al., 2025)	41.4	40.7	56.1	76.3	83.0	23.5	53.5	19.3	78.7	75.9	59.1	88.3
Cosmos-Pred 2.5 (T) (NVIDIA, 2025)	66.7	45.0	31.3	74.9	85.1	13.8	38.7	8.2	70.5	79.6	26.6	77.3
Action-conditioned embodied world models												
Cosmos-Pred 2.5 (A) (NVIDIA, 2025)	44.9	35.8	93.0	82.0	88.9	35.3	55.0	29.4	88.6	76.4	58.4	88.8
GigaWorld-0 (GigaAI, 2025)	50.4	39.9	44.1	73.0	85.6	17.6	53.7	15.5	63.2	76.0	61.6	85.9
Genie Envisioner (Liao et al., 2026)	23.1	32.9	33.4	77.6	90.2	20.1	20.5	6.8	86.6	52.8	20.3	85.4
TesserAct (Zhen et al., 2025)	33.2	45.9	45.8	82.5	**92.4**	24.9	58.0	14.0	71.6	79.2	61.5	87.8
WoW (Chi et al., 2025)	45.9	38.7	74.4	81.6	90.2	21.7	55.6	20.6	72.8	76.7	56.9	88.4
IRASim (Zhu et al., 2025b)	34.9	36.2	93.3	83.1	90.7	35.2	56.6	36.4	93.1	77.9	66.0	88.5
Ctrl-World (Guo et al., 2026)	35.2	38.9	91.8	84.1	90.6	17.3	62.1	47.7	93.0	79.6	72.7	89.1
Ours	67.5	45.3	**96.5**	**88.0**	92.3	32.4	**81.5**	**60.1**	**98.9**	**97.1**	88.6	89.1

without reference videos. DROID (Khazatsky et al., 2024) provides recorded videos for evaluating prediction fidelity and accumulated rollout error. We reserve 2,048 DROID episodes of up to 30 seconds for evaluation. Baselines include Cosmos3-Nano (Aditi et al., 2026) and Ctrl-World (Guo et al., 2026). The official Cosmos3-Nano checkpoint was also trained on DROID. We evaluate Genie Envisioner (Liao et al., 2026) zero-shot.

3.2 Embodied Video Quality

Table 2 reports twelve metrics from the WorldArena 2.0 video quality evaluation. Causeway ranks first on six metrics and second on four, including a tie. Its largest gains over the best reported baselines are in trajectory accuracy, from 47.7 to 60.1, and perspectivity, from 82.8 to 97.1. It also leads in interaction quality and depth accuracy. These gains span motion, interaction, and geometric consistency, while image and aesthetic quality remain within one point of the best reported scores. Image quality is 0.7 points below Wan 2.6, and aesthetic quality is 1.0 point below Veo 3.1. CLP complements this evaluation by testing whether changes in subsequent actions affect earlier frames (Section 3.4).

3.3 Open-Loop Fidelity Over Whole Episodes

We evaluate full-episode DROID rollouts against the recorded videos, using the same episodes, resolution (320×180), and frame rate (5 fps) for all models. Table 3 reports PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), and FVD (Unterthiner et al., 2019), separately for exterior and wrist views.

Using the same backbone, Causeway improves on the official Cosmos3-Nano checkpoint across all reported metrics. For exterior views, PSNR increases from 18.16 to 24.29 dB, LPIPS decreases from 0.220 to 0.073, and FVD falls from 113.6 to 8.9. Against Ctrl-World, Causeway improves nine of ten view-specific metrics; wrist FVD decreases from 52.8 to 17.2.

The gains extend to late rollout stages (Fig. 3). Exterior SSIM at 95% of episode progress retains 0.87 of its value at 5%, compared with 0.55 for the same Cosmos3-Nano backbone. Retention is computed from curves averaged over episodes and the two exterior views after alignment by normalized progress. Adding language improves PSNR, SSIM, and LPIPS in both views, while FID and FVD change little, so most of the improvement over the official checkpoint is already present without it.

3.4 Temporal Leakage and Goal Dependence

Protocol and controls. We run CLP on 32 common probe points from 11 RoboTwin 2.0 episodes, using four shared sampling seeds and the physical-time protocol of Section 2.4. Table 4 reports the

Table 3: **Full-episode video prediction on DROID.** Exterior scores average the two third-person cameras; wrist scores are reported separately. Best in bold.

Model	Exterior views					Wrist view				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓
Cosmos3-Nano (Aditi et al., 2026)	18.16	0.652	0.220	17.88	113.6	15.58	0.508	0.437	15.10	90.6
Genie Envisioner (zero-shot) (Liao et al., 2026)	19.21	0.759	0.184	15.61	199.5	13.88	0.440	0.555	74.23	255.7
Ctrl-World (Guo et al., 2026)	23.85	0.843	0.082	**2.48**	12.9	18.61	0.627	0.305	11.73	52.8
Ours (w/o language)	24.09	0.848	0.075	3.21	**8.9**	18.66	0.635	0.275	5.60	**17.0**
Ours	**24.29**	**0.855**	**0.073**	3.20	**8.9**	**19.05**	**0.639**	**0.269**	**5.56**	17.2

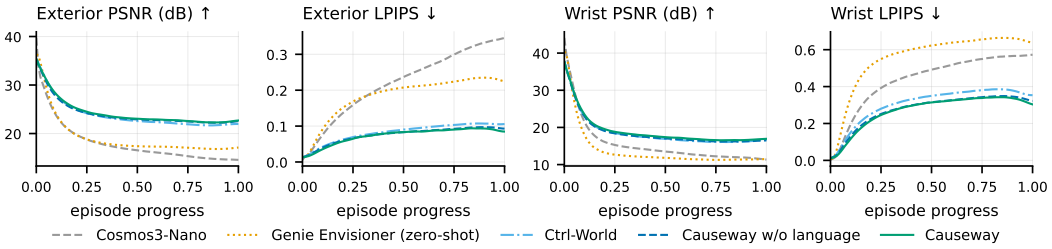


Figure 3: **Prediction error over episode progress.** PSNR and LPIPS on held-out DROID episodes. Curves are aligned by normalized episode progress and averaged over episodes.

action and language probes. The RoboTwin renderer is a deterministic reference: its response before the hold is exactly zero. All four learned models pass the aggregate licensed-response gate; Motus passes the pointwise gate on 14 of 32 points.

Action leakage. Causeway obtains $\mathrm{CLP}_A = 0.034$, compared with 0.091 for BWM, 0.110 for Cosmos3-Nano, and 0.279 for Motus, 63% below the best learned baseline. Motus's wide, right-skewed interval is consistent with its weak licensed response. Removing the temporal action mask raises the point estimate to 0.044; this is a descriptive comparison on the common sample, not a claim that the two intervals are separated. Allowing generated-video tokens to update the action representation, starting from the w/o-mask configuration, raises CLP_A to 0.101 $[0.053, 0.148]$. This comparison keeps the temporal mask disabled and associates added video feedback with a higher leakage point estimate.

Absolute response energies. The probe-point means in Table 6 show that, relative to w/o mask, Causeway has 9% lower centered forbidden energy and 8% higher licensed energy. Its uncentered baseline-relative forbidden difference is also 32% lower, with nearly unchanged licensed difference. These diagnostics support reduced pre-intervention response beyond a larger denominator or class centering alone. Their per-point averaging differs from the episode aggregation used for the main CLP ratio.

Goal dependence. With actions fixed, Causeway and its w/o-mask variant both obtain $\mathrm{CLP}_L = 0.00$ at the reported precision. This is a routing validation rather than a mask ablation: both variants route the goal through the action representation. Cosmos3-Nano and Motus, whose video pathways read the prompt directly, obtain $\mathrm{CLP}_L = 0.76$ and 1.99, respectively. Motus's wide interval $[0.49, 44.9]$ indicates substantial uncertainty in the magnitude of its normalized goal response.

Shift control. Advancing the actions supplied to the full Causeway model by 0.133 s raises CLP_A from 0.034 to 0.222 ($[0.173, 0.270]$), with all 32 points passing the positive control. The 6.5-fold increase shows that the probe detects an earlier action-dependent response in the same model.

3.5 Ablation

Attention routing and training augmentations. Each trained variant in Table 5 uses the same seeds as the full model. Removing UCA changes frame-wise metrics modestly but nearly doubles FVD (8.9 to 16.9) and lowers retention by 0.11 (0.87 to 0.76), indicating a larger effect on video-level

Table 4: **Causal Leakage Probe on RoboTwin 2.0.** Brackets show 95% episode-cluster bootstrap intervals. Positive controls are counted on 32 common action-probe points. Best in bold.

Model	Action probe		Language probe
	CLP_A [95% CI] ↓	Positive control (passed/common)	CLP_L [95% CI] ↓
RoboTwin 2.0 (ref) (Chen et al., 2025)	0.000	32/32	n/a
Motus (Bi et al., 2026)	0.279 [0.267, 0.666]	14/32	1.99 [0.49, 44.9]
BWM (BWM Team, 2026)	0.091 [0.053, 0.139]	32/32	n/a
Cosmos3-Nano (Aditi et al., 2026)	0.110 [0.061, 0.159]	32/32	0.76 [0.06, 3.81]
Ours	**0.034** [0.018, 0.049]	32/32	**0.00** [0.00, 0.00]
Ours (w/o mask)	0.044 [0.019, 0.080]	32/32	**0.00** [0.00, 0.00]
Ours (w/o mask + video feedback)	0.101 [0.053, 0.148]	32/32	**0.00** [0.00, 0.00]
Ours (shift)	0.222 [0.173, 0.270]	32/32	–

Table 5: **Ablation on DROID** (exterior views), evaluated under the same protocol as Table 3. Block A retrains without one component; block B changes only the memory schedule; block C restricts attention in one tower. ctx_m is the maximum number of latent frames in any generation window; s/chunk is the mean sampling time over chunks, using 35 UniPC steps on one H20 GPU.

Variant	PSNR↑	SSIM↑	LPIPS↓	FVD↓	Retention↑	ctx_m	s/chunk
Ours	24.29	0.855	0.073	8.9	0.87	24	33.5
A. Trained without one component							
Ours (w/o UCA)	23.92	0.844	0.077	16.9	0.76	24	39.0
Ours (w/o conditioning noise; $\sigma_o = 0$)	24.10	0.851	0.073	10.0	0.81	24	33.5
Ours (w/o temporal shift; $\alpha = 0$)	24.26	0.855	0.073	9.1	0.85	24	33.5
B. Memory schedule at inference							
Ours (overlap & first frame only)	22.38	0.815	0.087	18.6	0.73	11	16.7
Ours (uniform sampling; $H = 14$)	24.11	0.850	0.071	9.4	0.82	24	33.4
Ours (full history)	24.40	0.858	0.074	9.9	0.89	113	117.1
C. Trained with stronger causal restrictions							
Ours (causal action self-attention)	24.06	0.854	0.073	10.1	0.85	24	33.2
Ours (causal target-video self-attention)	22.97	0.813	0.080	17.8	0.79	24	31.4

quality and sustained fidelity. This row reverts the full attention modification; the w/o-mask row in Table 4 removes only the temporal action mask. Removing conditioning noise lowers retention to 0.81, while removing temporal shifts has a smaller effect (0.85). The stronger within-tower restrictions expose a quality–causality trade-off. Causal action self-attention changes PSNR/SSIM from $24.29/0.855$ to $24.06/0.854$ and increases FVD from 8.9 to 10.1; causal target-video self-attention has a larger effect, reducing PSNR/SSIM to $22.97/0.813$ and increasing FVD to 17.8. These variants change temporal communication within a tower, so the results quantify their generation-quality cost rather than proving output-level non-anticipation.

Memory selection and efficiency. Using only the first frame and overlap reduces sampling time but lowers PSNR from 24.29 to 22.38 dB and retention from 0.87 to 0.73. At the same budget, uniform sampling slightly improves LPIPS (0.071 versus 0.073), but yields lower retention (0.82 versus 0.87) and higher FVD (9.4 versus 8.9). Full history improves retention to 0.89, but has slightly higher LPIPS (0.074) and FVD (9.9). BioSaM approaches this retention with 21% of its maximum context (24 versus 113 latent frames) and 29% of its mean sampling time (33.5 versus 117.1,s).

4 Conclusion and Future Work

We introduced Causeway, which treats causal order as both a design constraint and an evaluation criterion. United Causal Attention restricts which actions and language each frame may read, Biomimetic Sampling Memory sustains open-loop rollouts, and the Causal Leakage Probe tests future-action dependence from generated video. Causeway reduces leakage while preserving frame-wise quality in the proposed routing, whereas stronger within-tower causal restrictions expose a quality–causality trade-off. Future work will extend the model to jointly predict actions and consequences and will test additional physical modalities in closed loop.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, formatting checks, research ideation, literature review, code writing and review, statistical analysis, and preparing input prompts. We did not use generative AI tools to run experiments or independently produce the reported results. We reviewed all AI-assisted text, code, analyses, figures, and claims, and take responsibility for the final content.

REFERENCES

Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson G. Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, and Cyrus Hogg. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35101–35113, 2026.

BWM Team. BWM: A low-cost high-fidelity world simulator for robot learning. *arXiv preprint arXiv:2607.29302*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, Zezhong Qian, Anthony Chen, Qiang Zhou, Yueru Jia, Jiaming Liu, Yong Dai, Qingpo Wuwu, Chengyu Bai, Yu-Kai Wang, Ying Li, Lizhang Chen, Yong Bao, Zhiyuan Jiang, Jiacheng Zhu, Kai Tang, Ruichuan An, Yulin Luo, Qiuxuan Feng, Siyuan Zhou, Chi-Min Chan, Chengkai Hou, Wei Xue, Sirui Han, Yike Guo, Shanghang Zhang, and Jian Tang. WoW: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

GigaAI. GigaWorld-0: World models as data engine to empower embodied AI. *arXiv preprint arXiv:2511.19861*, 2025.

Google DeepMind. Veo 3.1, 2025. URL https://deepmind.google/models/veo/. Model card.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Yue Hu, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R. Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J. Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems (RSS)*, 2024.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025a.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. WorldEval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*, 2025b.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng YAN, Maoqing Yao, and Guanghui Ren. Genie Envisioner: A unified world foundation platform for robotic manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

NVIDIA. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025.

Quanquan Peng, Yutong Liang, Rui Yan, Nicklas Hansen, and Xiaolong Wang. FACT: Failure-aware causal training for world-action models. In *Conference on Robot Learning (CoRL)*, 2026.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, and Ruimao Zhang. WorldSimBench: Towards video generation models as world simulators. In *Forty-second International Conference on Machine Learning*, 2025.

Daniel L Schacter, Roland G Benoit, and Karl K Szpunar. Episodic future thinking: Mechanisms and functions. *Current Opinion in Behavioral Sciences*, 2017.

Team Seedance. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. RoboScape: Physics-informed embodied world model. *arXiv preprint arXiv:2506.23135*, 2025.

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, Haisheng Su, Weizhen He, Wei Wu, Haoyi Duan, Gordon Wetzstein, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, Chen Gao, and Yong Li. WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform. *arXiv preprint arXiv:2605.17912*, 2026.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation. In *International Conference on Learning Representations Workshop*, 2019.

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, Xiaofeng Meng, 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.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. MiraBench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiahan Zhang, Muqing Jiang, Nanru Dai, TaiMing Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-In-World: World models in a closed-loop world. In *The Fourteenth International Conference on Learning Representations*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. UniPC: A unified predictor-corrector framework for fast sampling of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Robotics: Science and Systems (RSS)*, 2025a.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844, 2025b.

A Experimental Setup

Initialization and training. Causeway's language tower is initialized from the Cosmos3-Nano Reasoner (Aditi et al., 2026), and its action and video towers from the Generator. The language tower remains frozen while the action and video towers and their interfaces are optimized. This initialization preserves the pretrained language and visual generation capabilities while allowing the action–video interface to adapt to the causal routing used by Causeway. We optimize the trainable towers and interfaces at learning rate 3×10^{-5} using a cosine learning-rate schedule. The action reconstruction term in Eq. (2) uses $\lambda = 0.2$.

Cross-embodiment instantiation. We instantiate Causeway for Aloha–AgileX in RoboTwin 2.0 and Franka in DROID, using the same three-tower architecture, United Causal Attention, and Biomimetic Sampling Memory. The routing rules, memory construction, and training augmentations are shared across these instantiations; only the data-specific action interface is adapted to each embodiment. Both use temporal position shifts and conditioning noise during training.

Scene and goal preparation. We use Qwen 3.7-Plus to construct scene and goal text from the initial frame and the task description. The prompt requires the scene to describe subjects, surroundings, lighting, and visual composition consistently with the initial frame. The goal expands the task description into stepwise actions, including the locations of manipulated objects. Resolution, aspect ratio, and conditioning frame rate are appended to the scene text. The scene precedes the goal in the language stream, and their token boundary determines which language tokens video queries may attend to under UCA. At inference, users may instead supply optional scene and goal text.

Scene–Goal Prompt

```
<instructions>
Your function is to operate as a prompt upsampler for an world action model.
You will be provided with several inputs: (a) an attached starting frame image,
which serves as the definitive visual ground truth for subjects, setting,
lighting, and color palette; (b) this instruction block; (c) a
<video_description> detailing the scene's temporal and action-based intent;
and (d) an <output_json_template> schema. Your sole output is one fenced JSON
object. This object must populate every required field from the template.
Fields pertaining to visual information (subjects, background_setting,
lighting) must be entirely consistent with the attached image and must not
contradict it.
</instructions>

<video_description>
{description}
</video_description>

<output_json_template>
{
  "scene": "full visual description of all key subjects; subjects description,
  the background environment in which the scene takes place; lighting;
  cinematography.",
  "goal": "a group of step by step actions extended from the video description.
  contain the location information of the manipulated objects. For robotic
  arms, use 'one robotic arm / another robotic arm'. (e.g. 'Robotic arm enters
  the frame, extends toward bottle at the top left, grasps it, lifts it, moves
  to black dustbin at the far left, and drops it in')"
}
</output_json_template>
```

The generated instruction is represented by two fields: a scene field describing the visual state and a goal field describing the intended action sequence. This separation is applied before tokenization and is used only to determine the language spans exposed by UCA; it does not alter the commanded action sequence.

Scene–goal prompt. The following prompt is used with Qwen 3.7-Plus to construct the scene and goal fields from the initial frame and task description. The placeholder {`description`} is replaced by the task description. The augmentor appends resolution, aspect ratio, and conditioning frame rate to the scene field after generation.

At inference, users may replace either field with optional custom scene or goal text; the resulting spans are serialized in the same scene–goal order.

Inference. Causeway generates each chunk with 35 UniPC steps (Zhao et al., 2023) without classifier-free guidance. The initial observation is the observed visual anchor, and subsequent chunks condition on earlier predictions. The benchmark-specific evaluation protocols are described in the main experiments section.

B Causal Leakage Probe: Protocol and Validation

This appendix gives the full protocol behind Section 2.4. The probe uses paired generations, controlled interventions, and a common physical-time frame grid.

Time windows and features. For each episode and conditioning time t_c, the action intervention window is $[t_c + 0.5, t_c + 0.9)$ seconds. Let $t_\star$ denote the timestamp of the first intervened action row. We evaluate the response on two matched regions:

$$\text{forbidden} = [t_c + 0.1, t_c + 0.5), \qquad \text{licensed} = [t_c + 0.5, t_c + 0.9).$$

One frame is selected from each 0.2-second bin, nearest its center. Thus, both regions contain frames on the same physical-time grid. Points without a frame in any required bin are excluded.

Pixels are mapped to $[-1, 1]$, resized to 480×640 with antialiasing, and flattened before computing responses. All generations within a probe point use the same observation, text, sampling seeds, and evaluation grid.

Action interventions and eligibility. For the action probe, we generate one baseline video and $K_A = 4$ intervened videos. Each intervention holds selected action channels at their last commanded values before the intervention window and leaves all other entries unchanged. The four classes hold all selected moving channels or one of three channel groups.

The three groups are formed by sorting channels according to their recorded motion and assigning them in round-robin order. A model-specific adapter maps this common intervention protocol to the model's native action interface. A class is retained only when the recorded motion suppressed by the hold exceeds 0.01 radians, and a probe point requires at least three retained classes. An end-effector adapter additionally checks that its selected channels move. Cross-model comparisons use the intersection of eligible probe points. The baseline generation is separate from the K_A intervention classes.

Goal interventions. For the language probe, the action sequence and observation remain fixed. We use $K_L = 4$ goal classes: the original goal and three variants formed by prepending one of the following sentences:

- *Keep both arms completely still and do not move.*
- *Do not perform the task; both arms stay where they are.*
- *Hold the current pose without moving either arm.*

The original goal is the reference class, so its response relative to the reference generation is zero by construction. The three prefixed goals are paired with the same observation, actions, and sampling seeds. The four class responses are then centered jointly before estimating the language energy.

Algorithm 1 Causal Leakage Probe procedure.

Require: model $\mathcal{W}$, common eligible probe points, seeds, intervention classes, physical-time frame grid

for each action probe point $i = (e, t_c)$ **do**
  Generate a separate baseline for each seed and verify one repeated generation is identical.
  Generate the K_A action interventions, keeping observation, text, and seeds fixed.
  Select forbidden and licensed frames on the common time grid.
  Compute the two action energies using Eqs. (8) and (9).
end for
for each language probe point **do**
  Generate the $K_L = 4$ goal classes under fixed observation, actions, and matched seeds.
  Generate the four immediate-onset action holds under the original goal and matched seeds.
  Compute language and action energies on the licensed frames.
end for
Permute class labels within seeds to obtain pointwise thresholds.
Average energies and thresholds within episodes, then aggregate across episodes.
Report each ratio only if its denominator is positive and exceeds its aggregate threshold.
Bootstrap episodes for confidence intervals.

The numerator of CLP_L is evaluated on the licensed frame grid $[t_c + 0.5, t_c + 0.9)$. Its action denominator uses four action holds that begin immediately after t_c and continue through the evaluation window, with the original goal fixed. The denominator is evaluated on the same licensed frames, probe points, seeds, and estimator. This immediate-onset denominator makes the language and action energies comparable at the same physical time.

Estimator and aggregation. We use the response and energy estimators defined in Eqs. (8) and (9). For the action probe, the resulting episode-level ratio is Eq. (10); the language ratio is Eq. (11). Energy estimates are retained without clipping. Negative estimates close to zero reflect finite-seed variability rather than negative population energy.

For each probe point and region, class labels are independently permuted within each seed to form a permutation reference. Let $q_{i,R}$ denote its 95th percentile for region R. Using the same episode averaging as for the measured energy, we define

$$Q_R = \sum_e \frac{1}{|\mathcal{P}_e|} \sum_{i \in \mathcal{P}_e} q_{i,R}. \tag{12}$$

A reported ratio must have a positive aggregate action-response denominator that exceeds the corresponding threshold Q_R. This is a conservative gate based on the sum of pointwise permutation quantiles; it is deliberately kept separate from the ratio estimator itself. Points that fail the individual gate are not selectively removed from the aggregate. Confidence intervals use episode-cluster bootstrap resampling, keeping each episode's numerator and denominator paired.

Validation. The synthetic tests target the same content-energy estimator used by CLP, rather than the auxiliary decodability diagnostics. A causal source produces a licensed response but no forbidden response; a leaky source produces responses in both regions; constant class-shared responses and seed-random perturbations remain at the permutation null and are reported as not measurable. These tests verify the intended ordering of the estimator under known response patterns.

The RoboTwin renderer provides a non-learned causal reference. Paired trajectories share the same initial state and action prefix, so frames before the first intervened action remain unchanged. A temporal-shift control advances the supplied actions by four rows while retaining the original scoring windows. The response therefore moves into the forbidden region, providing a known temporal-misalignment control for the same measurement protocol.

These controls validate sensitivity and reproducibility under the tested interventions and time windows. They do not establish invariance for arbitrary inputs.

Table 6: **CLP energy diagnostics.** E_F, E_L are absolute centered energies before ratio normalization; D_F, D_L are uncentered baseline-relative squared differences. Q_L is the licensed-response permutation threshold. All entries are means over the 32 common probe points and are shown in units of 10^6. For the deterministic renderer, $Q_L = 0$ is the deterministic floor, not a seed-permutation quantile. The shifted rows are intervention-drift positive controls rather than additional models.

Model	E_F	E_L	D_F	D_L	Q_L
RoboTwin 2.0 (reference)	0	0.046080	0	0.096844	$0^\dagger$
RoboTwin 2.0 (4-row shift)	0.002104	0.065356	0.003935	0.137770	$0^\dagger$
Motus	0.00000089	0.0000031	0.0000199	0.0000501	0.0000014
BWM	0.002788	0.028346	0.020877	0.095565	0.009462
Cosmos3-Nano	0.003335	0.030317	0.026671	0.178328	0.007437
Ours (w/o temporal mask)	0.002084	0.054608	0.025836	0.166487	0.017421
Ours	0.001892	0.058785	0.017551	0.164967	0.017530
Ours (w/o temporal mask + shift)	0.008426	0.039854	0.046837	0.142310	0.012392

Interpretation and complementary energies. Within-seed centering in Eq. (8) gives

$$\mathbf{d}_{k,s} = \phi(\mathbf{x}_{k,s}) - \frac{1}{K}\sum_{j=1}^{K}\phi(\mathbf{x}_{j,s}).$$

CLP measures differences between intervention classes in their mean response across seeds. A component shared by all classes cancels. The baseline-relative response $\boldsymbol{\delta}_{k,s}$ additionally retains such shared changes. For complementary diagnostics, define

$$\widehat{D}_{i,R} = \frac{1}{KS}\sum_{k,s}\|\boldsymbol{\delta}_{k,s}\|_R^2.$$

Unlike $\widehat{E}_{i,R}$, this quantity retains variation across seeds and is not a debiased estimate of the squared class means. Both use matched seeds, the same features, and the same scoring frames. For the main CLP score, let $E_R = \sum_e \bar{E}_{e,R}$ and $D_R = \sum_e |\mathcal{P}_e|^{-1}\sum_{i\in\mathcal{P}_e}\widehat{D}_{i,R}$. Table 6 instead reports the arithmetic mean over the 32 common probe points so that the deterministic renderer and learned models share one per-point scale. Reporting E_F and E_L separately distinguishes a smaller forbidden response from a larger licensed denominator; D_F and D_L describe uncentered baseline-relative changes. Here F and L abbreviate the forbidden and licensed regions.

Energy diagnostics. In the units of Table 6, adding the temporal action mask reduces mean centered forbidden energy from 0.002084 to 0.001892, while licensed energy increases from 0.054608 to 0.058785. The uncentered forbidden difference falls from 0.025836 to 0.017551, with licensed difference changing from 0.166487 to 0.164967. The reduced forbidden response therefore remains visible without intervention-class centering. These are descriptive comparisons of point means; their ratios need not equal the episode-aggregated scores in Table 4.

Motus has a much smaller licensed response than the other learned models in both diagnostics, consistent with the uncertainty of its normalized CLP score. The renderer has zero forbidden response under the aligned protocol and a nonzero response after the four-row shift. The shifted learned-model row here uses the w/o-temporal-mask variant; it is distinct from the full-model shift control in Table 4.

C BioSaM Sampling and Augmentation Details

This appendix specifies the boundary conventions used by Biomimetic Sampling Memory (BioSaM).

Memory selection. Let b be the starting latent-frame index of the next chunk. The $H-1$ non-anchor memory slots are divided between an anchor-side set and a recent-side set:

$$k_a = \left\lceil \frac{H-1}{2} \right\rceil, \qquad k_r = \left\lfloor \frac{H-1}{2} \right\rfloor.$$

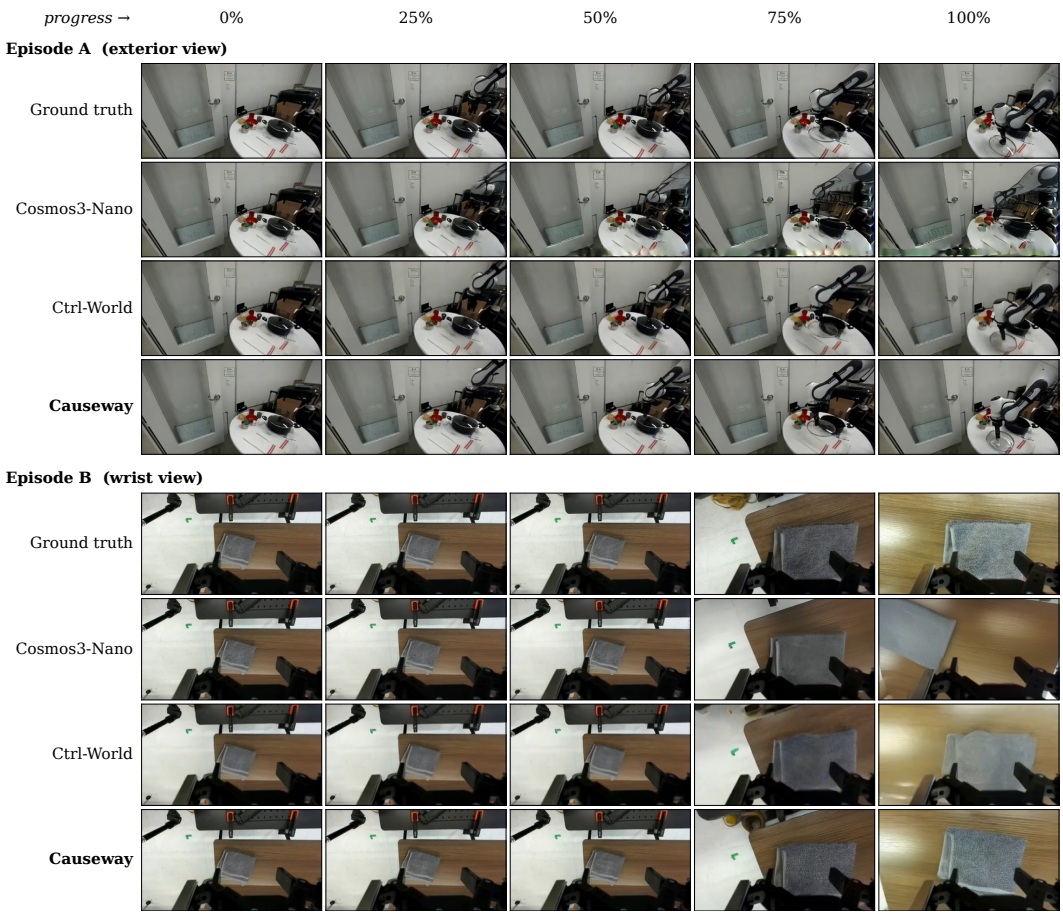


Figure 4: **Qualitative long-horizon rollouts on DROID.** Rows compare the recorded video and open-loop predictions from the same initial observation and recorded actions. Columns show matched points at normalized episode progress; subsequent visual context comes from each model's own predictions. The exterior example illustrates scene and robot-structure preservation; the wrist-view example shows late- rollout differences in object location and appearance.

The main configuration uses $L = 8$, $O = 2$, and $H = 14$. Here H includes the initial frame and excludes the O overlap frames.

For integers $k \geq 0$ and positive endpoints $0 < \ell < u$, let

$$G_k(\ell, u) = \left\{ \text{round}\left(\ell \left(\frac{u}{\ell} \right)^{j/(k+1)} \right) : j = 1, \ldots, k \right\}$$

denote the rounded interior points of a geometrically spaced sequence between ℓ and u. In the two-ended schedule, the anchor-side history samples $G_{k_a}(1, b - O - 1)$. The recent-side samples use recency distances from $O + 1$ through $b - O - 1$ and are mapped back from b. Both sides exclude the initial frame and the overlap; the initial frame is retained separately.

Rounding can produce duplicate positions or fewer valid positions than the requested budget. We first deduplicate the rounded positions. Vacant slots are then filled sequentially with available history positions that maximize the distance to their nearest already selected non-anchor sample. Candidates exclude the initial frame and the overlap, and ties are resolved by ascending latent-frame index. This is the implementation of the duplicate-repair rule stated in Section 2.3.

For the initial chunks, the schedule is limited by the available history. If $b \leq O$, there is no history and no overlap before the first chunk. If $O < b < H + O$, the chunk has overlap but fewer than

$H-1$ available non-anchor history slots; the schedule retains all available history positions and the overlap without duplicating the initial frame.

Temporal-position augmentation. The model default uses $\alpha = 1.0$ for the temporal-position shift. Chunks with $b \leq O+1$ have no history to stretch and are not shifted. For other chunks, we sample and round

$$\Delta \sim \mathcal{U}[0, \alpha(b-O)].$$

The augmented chunk begins at $b' = b + \Delta$. We construct the memory schedule in the augmented coordinates using the same rule as at inference. History position labels p' are mapped to recorded history positions by

$$\text{round}(p'/s), \qquad s = \frac{b'-O}{b-O},$$

and clipped to the valid history range. The initial frame remains at position zero. Overlap and target position labels are mapped to recorded positions $p' - \Delta$, and their associated actions are shifted consistently.

Conditioning-noise augmentation. The model default uses $\sigma_o = 0.15$ for the conditioning-noise cap. For each conditioning latent at augmented position p', we sample

$$\sigma_{c,p'} \sim \mathcal{U}\left[0, \sigma_o \frac{p'}{b'}\right].$$

The latent is mixed with independent Gaussian noise according to Eq. (7). The initial frame remains clean, while the maximum corruption increases toward the current chunk. The sampled corruption level is not supplied to the network.

D Qualitative Long-Horizon Rollouts

The quantitative DROID analysis in the main text measures prediction fidelity over complete episodes. This appendix provides a visual comparison of the same long-horizon behavior. Episode A (exterior view) is "Remove the lid from the pot and put it on the table" and lasts 13.07 s; Episode B (wrist view) is "Move the grey towel to the left" and lasts 10.47 s. Text conditioning follows each model's evaluation setup.

E World Model as an RL Environment

WorldArena 2.0 (Shang et al., 2026) Track 2 evaluates whether a world model is useful as an interactive environment for model-based reinforcement learning. The official evaluator fixes an initial robot policy, reward pipeline, and optimization budget. During training, the current policy proposes action chunks; the world-model service predicts the corresponding future visual observations, which become the next observations for subsequent policy interactions. The optimized policy is then evaluated in an independent held-out RoboTwin 2.0 environment. The primary result is downstream policy success, rather than reconstruction error or perceived video quality.

Table 7: **WorldArena 2.0 Leaderboard Track 2 on Adjust Bottle.** Success rate (%) of the optimized policy in the held-out environment. Baseline entries are taken from the publicly available WorldArena 2.0 Track 2 leaderboard; our entry is from the same leaderboard. All entries use the same task and evaluation protocol; higher is better.

World model	Success rate (%) ↑
Cosmos-Predict 2.5 (action) (NVIDIA, 2025)	55.73
RoboScape (Shang et al., 2025)	60.74
IRASim (Zhu et al., 2025b)	61.33
Ctrl-World (Guo et al., 2026)	61.46
Ours	**62.50**

We report the Adjust Bottle task under the fixed Track 2 protocol. All results, including Causeway's, are taken from the public Track 2 leaderboard and report the optimized policy's success rate (%) in the held-out environment.

The table evaluates whether generated interaction data supports policy improvement and transfer to the held-out simulator. It does not isolate which visual or causal property produces the downstream difference; the CLP and video-quality evaluations provide complementary diagnostics.