YoCausal: How Far is Video Generation from World Model? A Causality Perspective

Anonymous authors
Paper under double-blind review

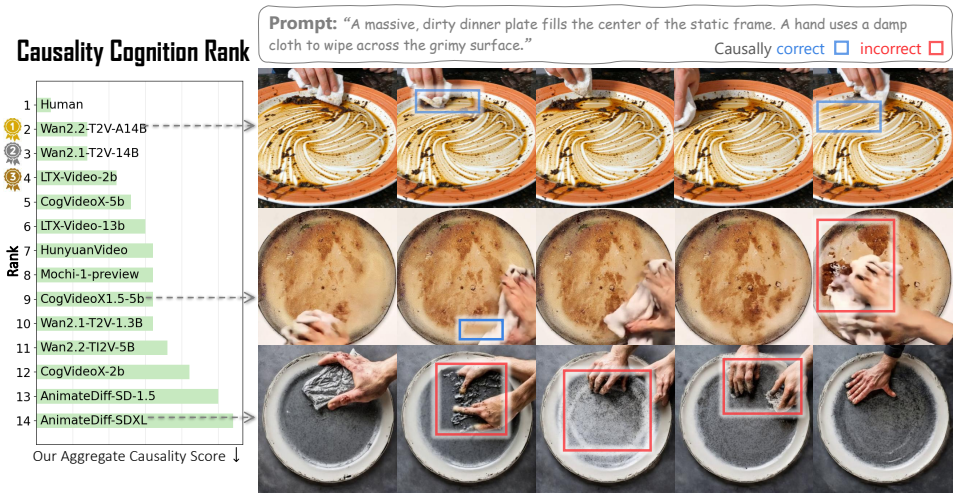


Figure 1: **YoCausal benchmark. (Left)** We evaluate and rank 13 SOTA video diffusion models alongside a human baseline based on our Causality Score (lower is better). **(Right)** Visualizations of a causal event generation strongly align with our quantitative rankings. The top-ranked VDM, Wan2.2-A14B(Wan et al., 2025), successfully captures the cause-and-effect relationship by progressively removing the dirt (**blue boxes**). In contrast, mid- to low-ranked models like CogVideoX1.5-5B(Hong et al., 2022; Yang et al., 2024) and AnimateDiff-SDXL(Guo et al., 2023) struggle to generate the underlying causal logic, with causal errors such as dirt reappearing in an illogically (**red boxes**). This shows our benchmark probes causal cognition of video generation model.

Abstract

As video diffusion models (VDMs) advance toward world models, a key question arises: do they truly understand causality? Existing benchmarks mostly rely on synthetic data, limiting real-world generalization due to the sim-to-real gap. We present YoCausal, a two-level benchmark inspired by the Violation of Expectation (VoE) paradigm from cognitive science. By temporally reversing real-world videos at zero cost as natural counterfactual samples, YoCausal establishes an arbitrarily extensible evaluation protocol. Level 1 introduces the Reverse Surprise Index (RSI), quantifying arrow-of-time perception via denoising loss. Level 2 introduces the Causality Cognition Index (CCI), which leverages a VLM to stratify datasets into causal and non-causal subsets, disentangling genuine causal reasoning from temporal bias. Evaluation of 13 state-of-the-art VDMs reveals that perceiving the arrow of time does not imply understanding causality, and a significant gap persists relative to human-level causal cognition.

1 Introduction

A long-standing aspiration of AI is to build machines that truly model the world (LeCun et al., 2022; Ha & Schmidhuber, 2018; Matsuo et al., 2022; Hu et al., 2023; Bruce et al., 2024; Lin et al., 2023). One important ability of such *world models* is to capture *causality*[1].

Video generation models have been regarded as a direct path to world modeling (Liu et al., 2024b; Esser et al., 2023; Qin et al., 2024; Li et al., 2025b; Zhu et al., 2024). Trained on vast real-world data,

[1]In this paper, we focus on intuitively observable cause-and-effect (i.e., event A leads to event B), rather than structural causal models or interventions (Neuberg, 2003; Didelez & Pigeot, 2001).

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.

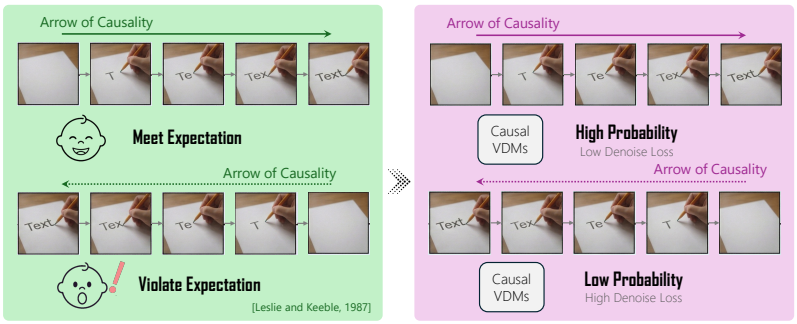


Figure 2: **Conceptual overview of YoCausal benchmark.** We draw inspiration from the **Violation of Expectation (VoE)** paradigm in cognitive science. **(Left)** Infants show surprise when seeing videos played in reverse (bottom), violating their causal cognition (Leslie & Keeble, 1987) (⇢). **(Right)** We transfer this paradigm to generative models: treating the learned data distribution as the model's cognition, a causally-aware VDM should assign lower probability (higher denoising loss) to counterfactual reversed videos than to forward ones.

they can and produce highly realistic video. However, a fundamental question remains: **do current video generation models actually understand causality?**

Previous research on "world knowledge" has focused on adherence to physical laws (Riochet et al., 2018; Bordes et al., 2025; Yuan et al., 2025; Motamed et al., 2026; Wu et al., 2016; Bear et al., 2021; Kang et al., 2024; Bansal et al., 2024; 2025; Meng et al., 2024). However, a genuine world model must go beyond physics to comprehend broader causality. Existing physics benchmarks face a limitation: to isolate physical variables, they rely on synthetic data or small collections of controlled laboratory recordings, creating a *sim-to-real gap* that limits assessment of real-world generalization.

To bridge this gap, we draw on the *Violation of Expectation (VoE)* paradigm from cognitive science (Margoni et al., 2024). In a seminal study (Fig. 2), Leslie and Keeble (Leslie & Keeble, 1987) assessed whether infants perceive causality by showing temporally reversed videos

Building on this, we propose **YoCausal**, a two-level benchmark for causal cognition in VDMs. At Level 1, the *Reverse Surprise Index (RSI)* measures the proportion a model assigns lower likelihood to the reversed than the forward video. However, RSI alone cannot isolate causal cognition, as RSI may come from both the *arrow of time* and causality. Level 2 therefore introduces the *Causality Cognition Index (CCI)*, which contrasts RSI on causal versus non-causal videos: a causally aware model should be more surprised by reversed causal events.

A key advantage is our *arbitrarily extensible dataset* design. Any real-world video can be temporally reversed at zero cost to produce a counterfactual sample, freeing the benchmark from the fixed synthetic scenes or controlled recordings of prior work. This bridges the sim-to-real gap. We further provide performance of human annotators as a reference for model performance.

Comprehensive evaluation across 13 state-of-the-art VDMs reveals four key insights: (1) while some advanced models exhibit preliminary causality cognition, a significant human-model gap remains; (2) perceiving the arrow of time is not equivalent to understanding causality; (3) causal cognition correlates partially with intuitive physics but not with aesthetic quality and (4) scaling parameters and advancing architectures improve causal cognition, indicating that scaling laws(Yin et al., 2025; Kaplan et al., 2020) extend to this higher-order reasoning.

In summary, our contributions are as follows:

- The first likelihood-based, VoE-grounded causality benchmark for VDMs, built on a scalable real-world dataset free from sim-to-real gaps.
- A cognitive-science-grounded two-level framework that disentangles arrow-of-time perception from causal cognition.
- Evidence that current open-source VDMs lack causal understanding, revealing a critical gap toward world models and providing guidance.

2 Related Work

Video Diffusion Models. Video synthesis has scaled from UNet (Ho et al., 2022; Wang et al., 2023; Blattmann et al., 2023; Singer et al., 2022) to DiT (Peebles & Xie, 2023) backbones (Zheng et al.,

2024; Peng et al., 2025) and commercial systems (Brooks et al., 2024; Polyak et al., 2024) with debated zero-shot reasoning (Wiedemer et al., 2025; Guo et al., 2026), motivating VDMs as *world models* (LeCun et al., 2022; Ha & Schmidhuber, 2018) for interactive simulation and planning (Bruce et al., 2024; Yang et al., 2023; Valevski et al., 2024; Agarwal et al., 2025; Hafner et al., 2020; 2019; Du et al., 2023). A world model differs by its causal structure (Pearl et al., 2003; Schölkopf et al., 2021), which robust generalization under distribution shift requires (Richens & Everitt, 2024). Whether VDMs acquire it remains open (Kang et al., 2024; Motamed et al., 2026; Bai et al., 2025; Zhang et al., 2026), and memorization of content and motion (Carlini et al., 2023; Somepalli et al., 2023; Rahman et al., 2025; Chen et al., 2024) makes realistic rendering weak evidence of understanding.

Video Generation Evaluation. Evaluation has evolved from distribution metrics (Unterthiner et al., 2018; Ge et al., 2024) to multi-dimensional suites (Huang et al., 2024; Zheng et al., 2025; Liu et al., 2024a; Sun et al., 2025) and temporal, physical, and world-model benchmarks (Liu et al., 2024c; Cai et al., 2024; Yuan et al., 2024; Bansal et al., 2024; Meng et al., 2024; Guo et al., 2025; Chen et al., 2025a; Upadhyay et al., 2026; Thozhiyoor et al., 2026; Han et al., 2026b; Duan et al., 2025); concurrent work frames evaluation causally via prompt, appearance, viewpoint, or action interventions (What-If World (Cai et al., 2026), CRONOS (Begiristain et al., 2026), Omni-WorldBench (Wu et al., 2026)). These score *generated* videos with VLM judges (He et al., 2024) that inherit known biases (Zheng et al., 2023), or with pixel comparisons (Zhang et al., 2025).

Intuitive Physics and VoE. Rooted in perceptual causality (Michotte, 2017) and infant core knowledge, including causal launching (Baillargeon et al., 1985; Spelke et al., 1992; Stahl & Feigenson, 2015; Spelke & Kinzler, 2007; Téglás et al., 2011; Baillargeon, 2004; Ullman et al., 2017; Oakes & Cohen, 1990; Saxe & Carey, 2006), the *violation-of-expectation* (VoE) paradigm (Margoni et al., 2024) measures cognition via surprise at counterfactuals, also in adults viewing naturalistic video (Plikat et al., 2025); we define surprise as an observation's improbability under an internal model (Friston, 2010) rather than the belief update it induces (Itti & Baldi, 2009). Since human-like intelligence requires intuitive causal theories (Lake et al., 2017), VoE has been applied to discriminative models on synthetic benchmarks (Riochet et al., 2021; Bordes et al., 2025; Dasgupta et al., 2021; Smith et al., 2019; Baradel et al., 2019; Gandhi et al., 2021; Shu et al., 2021; Garrido et al., 2025; Piloto et al., 2022; Battaglia et al., 2013), to latent world-model prediction error (Maes et al., 2026), and to generative likelihoods (Buschoff et al., 2023). LikePhys (Yuan et al., 2025) extends it to VDMs via the denoising loss as an ELBO-grounded likelihood proxy (Chao et al., 2022; Li et al., 2023; Clark & Jaini, 2023; Song et al., 2020; 2021; Kingma et al., 2021; Nichol & Dhariwal, 2021; Kingma & Gao, 2023), best used in relative form since such likelihoods can be dominated by low-level statistics (Nalisnick et al., 2018; Ren et al., 2019; Graham et al., 2023); related probes use a generator's flow residual (Hassan et al., 2026) or internal states (Esmati et al., 2026).

Arrow of Time and Causality in Video Understanding. Temporal directionality (Layzer, 1975) serves as a self-supervised signal (Misra et al., 2016; Lee et al., 2017; Wei et al., 2018; Xu et al., 2019; Pickup et al., 2014) yet remains hard for multimodal models (Bagad et al., 2023; Shangguan et al., 2025; Zhang et al., 2024; Xue et al., 2025; Cores et al., 2024; Li et al., 2024b; Han et al., 2026a; Yu et al., 2026), with VLMs near chance on forward/backward stimuli from human psychophysics (Matta et al., 2026; Hanyu et al., 2023). Video understanding (Fu et al., 2025; Li et al., 2024a; Patraucean et al., 2023; Chandrasegaran et al., 2024; Li et al., 2024c) and causal reasoning (Liu et al., 2025) benchmarks span synthetic (Yi et al., 2019; Ates et al., 2022) and real-world (Li et al., 2022; Xiao et al., 2021; Foss et al., 2025; Chen et al., 2025b; Chi et al., 2024; Wang et al., 2025) video QA and language counterfactuals (Zečević et al., 2023; Kiciman et al., 2023; Jin et al., 2023), with Bayesian surprise recently brought to video LLMs (Ravi et al., 2026). All target *discriminative* tasks; on the generative side, prior work steers VDMs backward in time (Feng et al., 2024; Liu et al., 2026) or diagnoses under-developed irreversible processes in generated videos (Xu et al., 2026).

Positioning. Existing evaluations, including the concurrent What-If World, CRONOS, and Omni-WorldBench, either judge generated videos, entangling causality with appearance, or target discriminative models and synthetic physics. YoCausal fills this gap by measuring whether a *generative* VDM's learned prior has internalized causal structure, read directly from its own likelihood on real-world videos without generation or an external judge. To our knowledge, it is the first likelihood-based, VoE-grounded causality benchmark for VDMs on real-world videos.

3 METHOD

Fig. 3 overviews our framework. We describe the *extensible dataset* construction (Sec. 3.1), formalize the link between a diffusion model's *"surprise"* and its *denoising loss* (Sec. 3.2), then introduce the

Table 1: **Comparison with existing physics-law evaluation benchmarks.** Prior benchmarks rely on synthetic data or controlled recordings, limiting scene diversity despite large video counts (e.g., 3M videos from only 70 scenes in PhyWorld). YoCausal incorporates *any* real-world video at zero cost and is extensible, achieving the highest scene diversity among all listed benchmarks. ↑ indicates that video count and scene coverage grow continuously as new subsets are added. (2D): simulation in two-dimensional. (Laboratory): videos recorded under controlled laboratory setups.

	Video type	# Video	# Video scene
PhyWorld (Kang et al., 2024)	Synthetic (2D)	**3M**	70
LikePhys (Yuan et al., 2025)	Synthetic	120	12
Physion (Bear et al., 2021)	Synthetic	10400	260
IntPhys2 (Bordes et al., 2025)	Synthetic	1416	344
Phys101 (Wu et al., 2016)	Real-World (Laboratory)	2500	101
Physics IQ (Motamed et al., 2026)	Real-World (Laboratory)	396	132
Ours	Real-World	1232↑	**1232**↑

Reverse Surprise Index (RSI) for arrow-of-time perception (Sec. 3.3) and the *Causality Cognition Index* (CCI) for disentangling genuine causal understanding (Sec. 3.4).

3.1 Dataset Construction

As mentioned in Section 1, we use reversed video to validate models' causal cognition ability. Our benchmark can utilize *any* real-world video at zero cost. This enables building a benchmark of arbitrary scale and scene diversity without synthetic rendering or controlled setups.

As shown in Fig. 3(a), we construct a dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_i, \dots\}$ of thematic subsets. Unlike closed-form benchmarks, our design is arbitrarily extensible: new subsets can be seamlessly added. In this paper, we use four representative subsets of everyday scenes: **General** $\mathcal{D}_{General}$ (unconstrained daily-life events), **Physics** $\mathcal{D}_{Physics}$ (mechanics, optics, thermodynamics, etc.), **Human Action** $\mathcal{D}_{Human}$ (diverse human activities), and **Animal Action** $\mathcal{D}_{Animal}$ (various animal behaviors), sourced from existing datasets: Moment in Time (Monfort et al., 2019), Physics IQ(Motamed et al., 2026), Kinetics(Kay et al., 2017) and Animal Kingdom (Ng et al., 2022) (details are provided in Appendix A). Performance breakdowns across subsets reveal each model's domain-specific strengths and weaknesses. Notably, future researchers can integrate additional domains (e.g.,, tool use) to keep the benchmark evolving alongside model capabilities.

As summarized in Table 1, prior physics benchmarks rely on synthetic data or small-scale controlled recordings, whereas our method incorporates any real-world videos at zero cost. Consequently, YoCausal achieves significant breakthroughs in scale and real-world scene coverage.

3.2 Formulating Surprise via Denoising Loss

We quantify a VDM's surprise at a video by its denoising loss. Following *violation of expectation* (VoE) paradigm from cognitive science (Leslie & Keeble, 1987), ,where the degree of surprise at expectation-violating events reveals whether an internal model has formed. We treat the learned distribution as the model's internal expectation: lower assigned probability means greater surprise. This framing allows us to probe a VDM's cognitive priors including causal understanding.

Concretely, a VDM learns the data distribution $p(x)$ of videos by training a neural network ϵ_θ to denoise noisy inputs x_t. Its denoising loss is formulated as Eq. (1):

$$\mathcal{L}_{\text{denoise}}(\theta; x_t) = \mathbb{E}_{t\sim\mathcal{U}(1,T),\, \epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \gtrsim \mathbb{E}_{x_0}[-\log p_\theta(x_0)]. \tag{1}$$

By variational inference theory (Ho et al., 2020; Sohl-Dickstein et al., 2015), this loss upper-bounds the negative log-likelihood (NLL). This means that the denoising loss can serve as an empirical proxy for the NLL of the video sequence x_0 within the model's learned distribution. Specifically, a higher denoising loss thus indicates lower model-assigned probability, letting us translate "degree of surprise" into magnitude of denoising loss as a quantifiable metric. The validity of denoising loss as a likelihood proxy has been established from prior work (Li et al., 2023; Clark & Jaini, 2023; Yuan et al., 2025; Song et al., 2020; 2021; Kingma et al., 2021).

3.3 Level 1: Measuring Arrow-of-Time Perception via RSI

Our first-level metric asks a simple question: is a model more surprised by a video played backward? Following infant studies (Leslie & Keeble, 1987), a model that has internalized the arrow of time and causality, like an infant, should assign lower likelihood to the reversed video x^r than to the forward

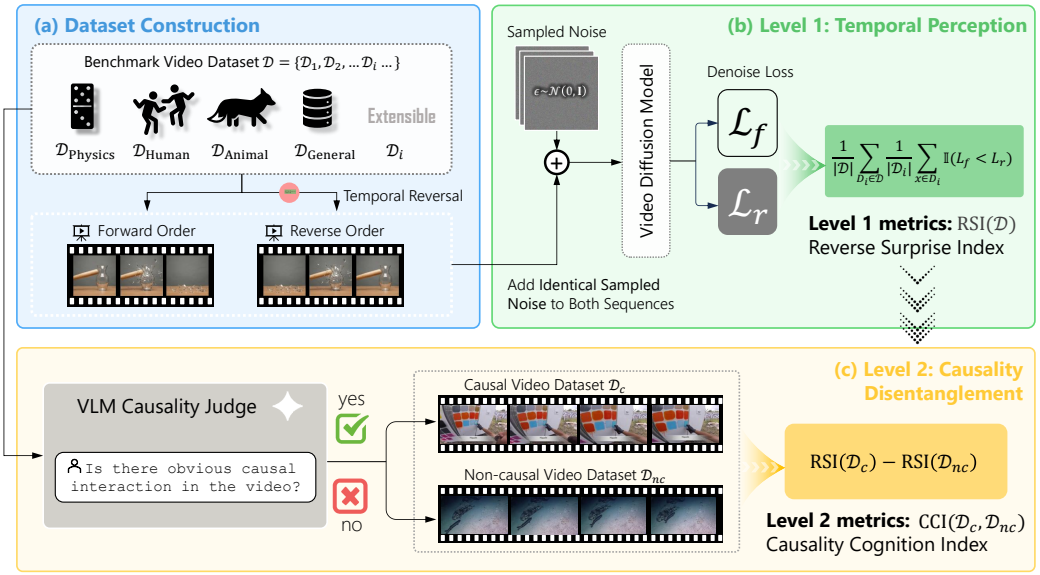


Figure 3: **Overview of the YoCausal evaluation framework. (a) Dataset Construction:** We construct an infinitely extensible benchmark by using real-world videos from different domains. By applying zero-cost temporal reversal, we generate natural counterfactual pairs (forward x^f and reverse x^r). **(b) Level 1 Temporal Perception:** Identical sampled noise ϵ is added to both sequences and compute their denoising losses. Reverse Surprise Index (RSI), quantifies the model's perception of the arrow of time by measuring the proportion of instances where the reversed video has a higher loss ($\mathcal{L}_r > \mathcal{L}_f$). **(c) Level 2 Causality Disentanglement:** To disentangle genuine causal cognition from statistical temporal biases, a Vision-Language Model (VLM) divides the dataset into causal ($\mathcal{D}_c$) and non-causal ($\mathcal{D}_{nc}$) subsets. The Level-2 metric, Causality Cognition Index (CCI), is computed as the difference in RSI between these subsets, isolating the model's genuine causal cognition ability.

one x^f. Building on the surprise signal in Section 3.2, We can express$p_\theta(x^f) > p_\theta(x^r)$ equivalently as $\mathcal{L}_{\text{denoise}}(\theta; x^r) > \mathcal{L}_{\text{denoise}}(\theta; x^f)$.

3.3.1 Reverse Surprise Index.

We propose the *Reverse Surprise Index* (RSI Eq. (2)) as our Level-1 metric. For each video $x_{i,j} \in \mathcal{D}_i$, let $x^f_{i,j}$ and $x^r_{i,j}$ denote its forward and reversed sequences, respectively. We uniformly sample $K{=}10$ timesteps from the diffusion process and apply identical Gaussian noise $\boldsymbol{\epsilon}$ to both sequences at each timestep (Fig. 3(b)), then average the resulting losses. Fixing the timesteps and noise ensures identical denoising difficulty. More details are provided in Appendix D. Concretely, RSI measures the proportion of videos for which the model correctly assigns a lower denoising loss to the forward sequence. For a dataset $\mathcal{D}$ composed of sub-datasets $\{\mathcal{D}_i\}$, we compute the average across subset:

$$\text{RSI}(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{D}_i \in \mathcal{D}} \frac{1}{|\mathcal{D}_i|}\sum_{x_{i,j} \in \mathcal{D}_i} \mathbb{1}\big[\mathcal{L}_{\text{denoise}}(\theta;\, x^r_{i,j}) > \mathcal{L}_{\text{denoise}}(\theta;\, x^f_{i,j})\big], \tag{2}$$

where $\mathbb{1}[\cdot]$ is the indicator function and $\text{RSI} \in [0, 1]$; higher values indicate stronger perception of the arrow of time and causality.

3.4 Level 2: Disentangling Causality via CCI

We introduce the Causality Cognition Index (CCI) to disentangle a model's understanding of causality from its sensitivity to temporal direction.

Blind spot of RSI. RSI cannot reveal why a model is surprised: reversal flips both the arrow of time and causality. Leslie & Keeble (1987) resolved this ambiguity in infants using hand-crafted causal and non-causal stimuli (Fig. 4), showing that infants were more surprised by reversed causal videos, which implies that they perceive causality.

3.4.1 Causality Cognition Index.

Following Leslie & Keeble (1987), we need causal and non-causal videos. Real-world videos naturally vary in *causal salience* (Fig. 4): breaking glass exhibits a clear causal chain, while a car

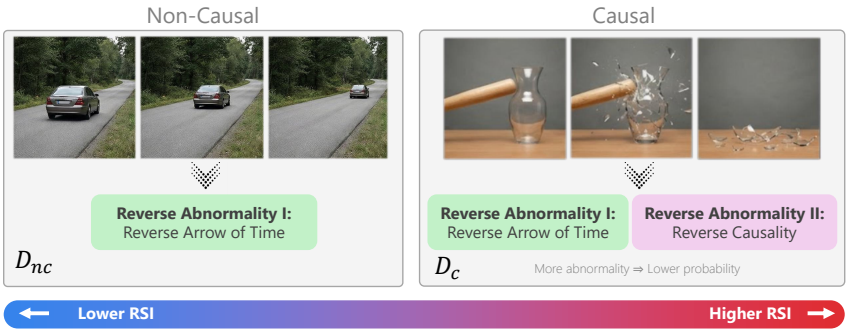


Figure 4: **Key idea behind CCI.** Reversing a non-causal video (e.g., a car cruising) introduces only one source of abnormality: reversed arrow of time. While, a causal video (e.g., a hammer shattering a vase) introduces an additional source: reversed causality. A causally aware model should perceive more abnormality in the reversed causal video, exhibiting greater surprise (higher RSI). The gap $\mathrm{CCI} = \mathrm{RSI}(\mathcal{D}_c) - \mathrm{RSI}(\mathcal{D}_{nc})$ thus disentangles the model's sensitivity to causality from its sensitivity to statistical temporal patterns.

cruising on a highway does not. Therefore, there is no need to craft synthetic videos, and we directly partition $\mathcal{D}$ into a causal subset $\mathcal{D}_c$ and a non-causal subset $\mathcal{D}_{nc}$ based on whether cause-effect interactions are present.

A causally aware model should therefore show higher RSI on $\mathcal{D}_c$ than on $\mathcal{D}_{nc}$, and we define CCI as Eq. (3). A higher CCI indicates a model captures reversed causality cues.

$$\mathrm{CCI}(\mathcal{D}) = \mathrm{RSI}(\mathcal{D}_c) - \mathrm{RSI}(\mathcal{D}_{nc}). \tag{3}$$

As shown in Fig. 3(c), since constructing CCI only requires detecting whether causality exists, which is easier than judging its correctness, we automate dataset splitting with an advanced Vision-Language Model (VLM) using a carefully designed prompt (see Appendix E) to ensure scalability. We validate the reliability of this split in Section 5.

Finally, we must emphasize that CCI is a relative index and must be jointly interpreted with RSI: high RSI but low CCI suggests the model only perceives the statistical arrow of time; high CCI but low RSI renders the CCI unreliable due to poor temporal grounding. We introduce an aggregate ranking combining both metrics in Sec. 4.3.

4 Experiment

We employ YoCausal to evaluate the causal cognition of current open-source video diffusion models, presenting Level 1 RSI results (Section 4.1), Level 2 CCI results (Section 4.2), an aggregate ranking (Section 4.3) that jointly considers both metrics and cross-metric analysis (Section 4.4).

Settings. We evaluate 13 state-of-the-art open-source text-to-video diffusion models spanning diverse architectures and scales (Guo et al., 2023; Hong et al., 2022; Yang et al., 2024; Team, 2024; Kong et al., 2024; Wan et al., 2025; HaCohen et al., 2024) (details in Appendix B). To ensure each model operates at its best, all inference configurations strictly follow official recommendations, including FPS, resolutions and so on. Videos are preprocessed through reasonable resizing and temporal adjustments to match each model's specifications. Detailed per-model settings and preprocessing procedures are provided in Appendices B and C.

4.1 RSI Results

In Fig. 5, we evaluate all 13 models on Level 1 RSI, with human annotators judging 1,232 videos' temporal directions as a refrence (Hanyu et al., 2023) (details in Appendix G). Humans achieve the highest RSI across all subsets except Human Action. Several models surpass the 50% random-guess baseline with 90% confidence (bootstrap test), yet a significant gap remains relative to human performance. Higher-fidelity models (e.g., LTX-Video-13B, Wan2.1/2.2-14B) tend to score higher, whereas lower-fidelity models (e.g., AnimateDiff-SDv1.5/SDXL) exhibit weaker temporal perception.

Per-subset results reveal cross-domain variation due to (1) differing cue strength across subsets: for example, videos in $D_{Physics}$ contain unambiguous anomalies when reversed, and (2) domain-specific biases from training data: for instance, most models perform well on D_{Human} given the abundance of human activity videos online. These highlight the value of evaluating across diverse subsets.

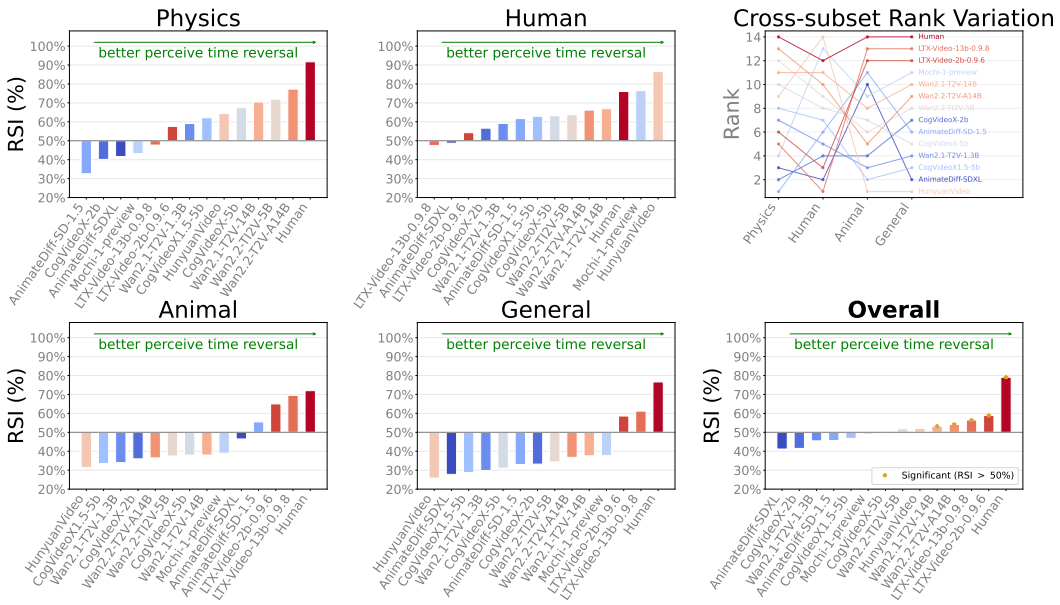


Figure 5: **Level-1 RSI results.** RSI scores of 13 VDMs and human annotators across four subsets and the full dataset. Several models surpass the 50% random-guess baseline with 90% confidence (bootstrap test) but still lag considerably behind human performance. Cross-subset variations highlight the necessity of evaluating across diverse domains.

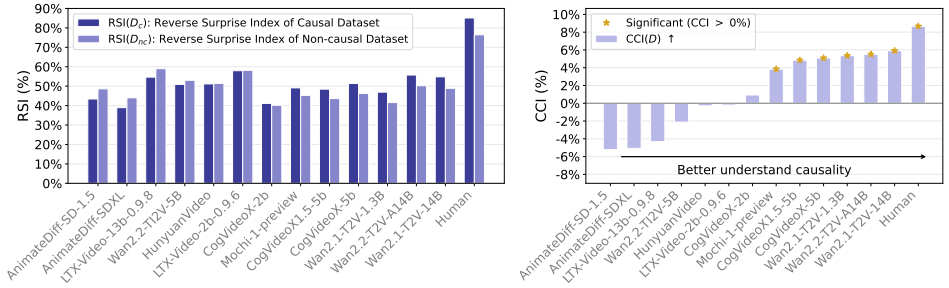


Figure 6: **Level-2 CCI results.** Left: RSI scores of 13 VDMs and human annotators on the causal subset $\mathcal{D}_c$ (dark blue) and non-causal subset $\mathcal{D}_{nc}$ (light blue). Right: the resulting CCI.

Notably, some models score below the 50%, suggesting their learned distributions capture local visual smoothness without internalizing the arrow of time and yield no preference, or even a slight inverse preference for forward sequences. A similar observation is reported in LikePhys (Yuan et al., 2025).

4.2 CCI Results

We further analyze Level 2 CCI in Fig. 6. Humans achieve the highest CCI, though the margin over models is modest since humans already saturate both subsets. Several models attain positive CCI with 90% confidence (bootstrap), demonstrating preliminary causal perception; the top performers concentrate in the Wan and CogVideo families, hinting that shared training data and architectural choices may give rise to emergent causal understanding. Crucially, models ranking high on RSI—LTX-Video-2B/13B and HunyuanVideo—score poorly on CCI, confirming that our two-level framework disentangles causal cognition from mere arrow-of-time perception.

Negative CCI in some models reflects the same distributional deficiency noted in Section 4.1: lacking an internalized causality, they treat reversed causal and non-causal sequences as equally anomalous.

4.3 Aggregating Arrow of Time and Causality Cognition

As discussed in Section 3.4, RSI and CCI should not be interpreted in isolation, and robust causal understanding requires strong performance on both metrics. To provide a single, intuitive measure of overall causal cognition, we combine the two indices with direct arithmetic combination. We propose a heuristic aggregate rank by summing each model's ranks on RSI and CCI as aggregate causality

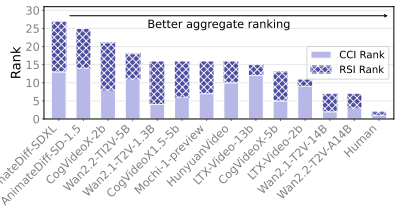


Figure 7: **Aggregate ranking of causal cognition.** We sum each model's RSI and CCI ranks to obtain an aggregate causality score (lower is better). Ties are broken by RSI rank.

Table 2: **Cross-metric analysis.** Kendall's τ between our aggregate causal cognition ranking and external benchmarks/model features. High correlation with human preference validates our benchmark; positive correlation with (Yuan et al., 2025) indicates causal cognition relates to physical understanding; low correlation with aesthetic quality rules out visual-appeal confounding; strong correlation with release date and parameters supports scaling laws and architectural evolution.

			VBench (Huang et al., 2024)						
	Human Preference	LikePhys (Yuan et al., 2025)	Aesthetic Quality	Subject Consistency	Background Consistency	Motion Smoothness	Temporal Flickering	Release Date	# of Parameters
Kendall's τ	0.6667	0.5111	0.1026	0.3846	0.1282	0.2308	0.1538	0.4576	0.6584
p-value	0.0009	0.0466	0.6754	0.0763	0.5900	0.3062	0.5098	0.0318	0.0024

score and sort the totals. Since RSI serves as the prerequisite foundation for causal cognition, ties are broken in favor of the model with the higher RSI rank.

As shown in Fig. 7, this aggregate ranking provides a holistic view of each model's causal cognition capability and will serve as the quantitative basis for the cross-metric analyses (Section 4.4).

4.4 Cross-Metric Analysis

To validate our benchmark and explore the interplay between causality and other model capabilities, we compute Kendall's rank correlation coefficient τ(Abdi et al., 2007) between our aggregate rank and model rankings on several external metrics and model properties, as summarized in Table 2.

Human Causality Preference. To verify that our benchmark captures causal understanding, we conduct a user study in which models generate videos from causally rich prompts and human evaluators compare them by causality plausibility (details in Appendix H). As shown in Table 2, our benchmark exhibits a high correlation (τ = 0.6667) with human preference.

Intuitive Physics Correlation. Clarifying the relationship between causal cognition and intuitive physics will provide important guidance for future model improvement. We compute the rank correlation with LikePhys (Yuan et al., 2025), a benchmark for intuitive physics in VDMs. Table 2 reveals positive correlation ($\tau = 0.5111$), implying that physical laws constrain object interactions and thereby influence causality. However, the moderate magnitude shows causal cognition is not reducible to physical intuition alone, meaning that separately assessing causal capabilities remains essential, highlighting the unique contribution of our benchmark.

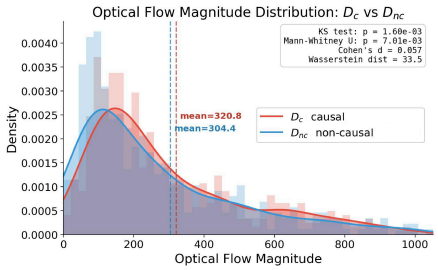
VBench Correlation. In addition, we examine the relationship between causality and visual generation quality by computing correlations with five VBench metrics (Huang et al., 2024). As shown in Table 2, causality shows low correlation with aesthetic quality ($\tau = 0.1026$), suggesting that our benchmark is not confounded by visual appeal. The subject consistency index show mild correlations: Inconsistency violates causal expectations.

Scaling Law and Generation Evolution. As shown in Table 2, release date (τ=0.4576) and model parameters (τ=0.6584) correlate significantly with causality cognition, implying scaling laws (Yin et al., 2025; Kaplan et al., 2020) and architectural evolution improve causal cognition (Appendix J).

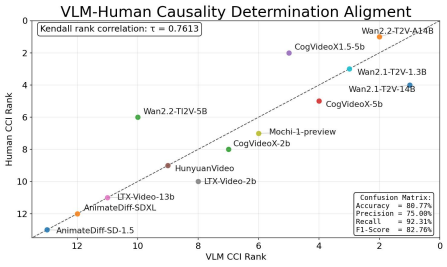
5 Benchmark Reliability and Validity

5.1 VLM Reliability

CCI relies on a VLM, Gemini 3.0 Pro (Team et al., 2023) by default, to partition the dataset (Sec. 3.4), so its validity depends on the reliability of VLMs. We validate it from three aspects: agreement with human annotation, independence from low-level motion cues, and robustness to the choice of VLMs.
Human Alignment. VLM labels agree well with human annotation (F1 = 82.76%), and CCI-based model rankings from VLM and human labels are highly consistent (Kendall's $\tau = 0.761$, Fig. 8b). Moreover, $\mathcal{D}_c$ and $\mathcal{D}_{nc}$ show negligible difference in optical-flow magnitude (Cohen's $d = 0.057$, Fig. 8a), indicating the VLM partitions videos by causality rather than low-level motion cues.



(a) RAFT (Teed & Deng, 2020)) Optical flow magnitude distributions of $\mathcal{D}_c$ and $\mathcal{D}_{nc}$.



(b) VLM-human Alignment.

Figure 8: **Validating VLM-based causal/non-causal partitioning.** (a) The effect size between two optical flow distribution is negligible (Cohen's d=0.057<0.2), confirming the VLM infers causality from high-level semantics rather than low-level motion cues. (b) Comparing CCI rankings and the confusion matrix from a VLM partition over the full dataset against human annotators (30 clearly causal, 30 clearly non-causal videos), the high Kendall correlation (τ=0.7613) and F1-score (82.76%) confirm VLM-based labeling as a reliable proxy for human judgment.

Table 3: **VLM sensitivity** Kendall's τ between aggregate rankings induced by three different VLMs. The consistently high correlations confirm our benchmark is robust to the choice of VLM judge.

Kendall's τ / p-value	Gemini 3.0 Pro (Team et al., 2023)	GPT-4o (Hurst et al., 2024)	Qwen 3.5 9B (Team, 2026; Yang et al., 2025)
Gemini 3.0 Pro	1.000 / 0.0000	0.6923 / 0.0005	0.6666 / 0.0009
GPT-4o	0.6923 / 0.0005	1.000 / 0.0000	0.6666 / 0.0009
Qwen 3.5 9B	0.6666 / 0.0009	0.6666 / 0.0009	1.000 / 0.0000

VLM Sensitivity. Re-partitioning with GPT-4o (Hurst et al., 2024) and the open-weight Qwen3.5-9B (Team, 2026; Yang et al., 2025) yields aggregate rankings consistent with Gemini 3.0 Pro's ($\tau \in [0.667, 0.692]$, $p < 0.001$; Table 3). YoCausal is thus robust to the choice of VLM judge, and its partition should improve as VLMs advance.

5.2 Confound Exclusion

Model-side. For each video, the forward and reversed clips share the same model, preprocessing, timesteps, and noise, so these factors cancel in the pairwise comparison. Since RSI is a win rate rather than a raw loss, it is also insensitive to model-specific loss scales and thus comparable across models; CCI inherits this property.

Video-side. Static factors (appearance, lighting) are identical across the pair and cancel out. However, temporal factors (motion distribution and video smoothness), change under reversal. CCI addresses this by differencing RSI across $\mathcal{D}_c$ and $\mathcal{D}_{nc}$: when the two subsets share similar temporal statistics, these factors cancel out. As shown in Sec. 5.1, the two subsets have nearly identical motion-magnitude distributions, suggesting motion and smoothness does not account for the CCI.

6 Conclusion and Limitation

This paper presents YoCausal, the first benchmark for evaluating causal cognition in VDMs. Inspired by the VoE paradigm from cognitive science, we leverage temporally reversed real-world videos as natural counterfactual samples, establishing an arbitrarily extensible evaluation protocol free from synthetic data or controlled settings. Experiments across 13 open-source VDMs reveal that: (1) perceiving the arrow of time is not equivalent to understanding causality; (2) even the best models exhibit a substantial gap relative to the human upper bound; (3) causal cognition is related to yet not reducible to intuitive physics and shows low correlation with aesthetic quality, confirming YoCausal captures a unique evaluation dimension, and (4) both parameter scaling and architectural evolution improve causal understanding. While physical understanding and scaling up models may partially improve causal perception (Li et al., 2025a), we hope YoCausal motivates the community (Croitoru et al., 2023) to treat causal cognition as a distinct objective and advance this capability in the future.

Limitation. Our method struggles with temporally symmetric events (e.g., Newton's cradle), where forward and reversed sequences are visually near-identical, rendering RSI ineffective. Moreover, computing denoising losses requires access to model weights, limiting external evaluation of closed-source models, though model providers can still apply YoCausal internally to diagnose and improve causal cognition. Addressing these limitations remains future work.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on the presentation of our experimental methodology. We have not used generative AI tools for generating synthetic datasets, developing conceptual frameworks, proposing hypotheses, implementing methods, assisting with translation, cleaning datasets, analyzing data, or interpreting results, and formulating or proving mathematical claims is not applicable to this work. Additionally, we used generative AI tools for editing the paper to improve readability and conciseness, formatting tables and references, and debugging LaTeX compilation. We have reviewed all AI-assisted work: all edited text was checked by the authors against the original content and experimental results, and no numerical result was produced or altered by AI tools. Note that the VLM judges (Gemini 3.0 Pro Team et al. (2023), GPT-4o Hurst et al. (2024), and Qwen3.5-9B Team (2026); Yang et al. (2025)) used to partition the dataset are a component of the proposed benchmark (Section 3.4 and Appendix E) rather than research assistance. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves human participants in two studies: annotating the temporal direction of all 1,232 benchmark videos to establish a human upper bound (Appendix G), and a pairwise causality preference study with 40 participants over 780 generated videos (Appendix H). All participants were informed of the purpose of the study and participated voluntarily, and no personally identifiable information was collected. All benchmark videos are drawn from publicly available research datasets, namely Moments in Time Monfort et al. (2019), Physics IQ Motamed et al. (2026), Kinetics-400 Kay et al. (2017), and Animal Kingdom Ng et al. (2022) (Appendix A), and are used in accordance with their respective licenses for non-commercial research. Our benchmark evaluates, rather than improves, the generative capability of video models, and we do not foresee direct harmful applications.

REPRODUCIBILITY STATEMENT

All 13 evaluated video diffusion models are publicly available, and their inference configurations follow official defaults (Appendix B and Table 5). Dataset construction, including source datasets, sampling, and clip trimming, is described in Appendix A, and the unified video preprocessing pipeline in Appendix C. The complete two-level evaluation procedure is formalized in Algorithm 1, with implementation details of RSI (timestep and noise sampling, text conditioning) in Appendix D and of CCI, including the exact VLM prompt used for causal partitioning, in Appendix E; the effect of text conditioning is further analyzed in Appendix F. Since aggregate rankings remain consistent under the open-weight Qwen3.5-9B judge (Section 5.1 and Table 3), the full pipeline can be reproduced without proprietary models. Complete numerical results for all models are reported in Appendix I. Anonymized code and an interactive HTML interface of video results are provided in the supplementary material.

REFERENCES

Hervé Abdi et al. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510, 2007.

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.

Alan Agresti. Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):287–297, 1992.

Tayfun Ates, M Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. Craft: A benchmark for causal reasoning about forces and interactions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2602–2627, 2022.

Piyush Bagad, Makarand Tapaswi, and Cees GM Snoek. Test of time: Instilling video-language models with a sense of time. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2503–2516. IEEE, 2023.

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible videos. *arXiv preprint arXiv:2503.14378*, 2025.

Renée Baillargeon. Infants' physical world. *Current directions in psychological science*, 13(3):89–94, 2004.

Renee Baillargeon, Elizabeth S Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019.

Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the national academy of sciences*, 110(45):18327–18332, 2013.

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

León Begiristain, Olaf Dünkel, and Adam Kortylewski. Cronos: Benchmarking counterfactual physical consistency in video models. *arXiv preprint arXiv:2605.23699*, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

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

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Luca M Schulze Buschoff, Eric Schulz, and Marcel Binz. The acquisition of physical knowledge in generative neural networks. In *International Conference on Machine Learning*, pp. 30321–30341. PMLR, 2023.

Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. What-if world: A causal benchmark for general world models in embodied scenarios. *arXiv preprint arXiv:2605.27589*, 2026.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. *arXiv preprint arXiv:2203.14206*, 2022.

Chen Chen, Enhuai Liu, Daochang Liu, Mubarak Shah, and Chang Xu. Investigating memorization in video diffusion models. *arXiv preprint arXiv:2410.21669*, 2024.

Yongfan Chen, Xiuwen Zhu, and Tianyu Li. A physical coherence benchmark for evaluating video generation models via optical flow-guided frame prediction. *arXiv preprint arXiv:2502.05503*, 2025a.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. Countervqa: Evaluating and improving counterfactual reasoning in vision-language models for video understanding. *arXiv preprint arXiv:2511.19923*, 2025b.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? *Advances in Neural Information Processing Systems*, 37:96640–96670, 2024.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36:58921–58937, 2023.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, and Cheston Tan. Avoe: a synthetic 3d dataset on understanding violation of expectation for artificial cognition. *arXiv preprint arXiv:2110.05836*, 2021.

Vanessa Didelez and Iris Pigeot. Causality: models, reasoning, and inference, 2001.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. *arXiv preprint arXiv:2606.05328*, 2026.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7346–7356, 2023.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J Black, and Xuaner Zhang. Explorative inbetweening of time and space. In *European Conference on Computer Vision*, pp. 378–395. Springer, 2024.

Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025.

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2): 127–138, 2010.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Kanishk Gandhi, Gala Stojnic, Brenden M Lake, and Moira R Dillon. Baby intuitions benchmark (bib): Discerning the goals, preferences, and actions of others. *Advances in neural information processing systems*, 34:9963–9976, 2021.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7277–7288, 2024.

Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2948–2957, 2023.

Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2vphysbench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

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

Ziyu Guo, Xinyan Chen, Renrui Zhang, Ruichuan An, Yu Qi, Dongzhi Jiang, Xiangtai Li, Manyuan Zhang, Hongsheng Li, and Pheng-Ann Heng. Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9175–9184, 2026.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

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

Peitao Han, Fei Cheng, Lis K Pereira, Qianying Liu, and Shigeru Kitazawa. Tracing the arrow of time: Diagnosing temporal information flow in video-llms. *arXiv preprint arXiv:2605.07568*, 2026a.

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. Oscbench: Benchmarking object state change in text-to-video generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30867–30884, 2026b.

Nao Hanyu, Kei Watanabe, and Shigeru Kitazawa. Ready to detect a reversal of time's arrow: a psychophysical study using short video clips in daily scenes. *Royal Society open science*, 10(4), 2023.

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation. *arXiv preprint arXiv:2605.28230*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10): 1295–1306, 2009.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065, 2023.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

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.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

David Layzer. The arrow of time. *Scientific American*, 233(6):56–69, 1975.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 667–676. IEEE, 2017.

Alan M Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3): 265–288, 1987.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Chenyu Li, Oscar Michel, Xichen Pan, Sainan Liu, Mike Roberts, and Saining Xie. Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. *arXiv preprint arXiv:2503.09595*, 2025a.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025b.

Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21273–21282, 2022.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024b.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *arXiv preprint arXiv:2404.12966*, 3, 2024c.

Zhengyang Liang, Hao He, Ceyuan Yang, and Bo Dai. Scaling laws for diffusion transformers. *arXiv preprint arXiv:2410.08184*, 2024.

Jessy Lin, Yuqing Du, Olivia Watkins, Danijar Hafner, Pieter Abbeel, Dan Klein, and Anca Dragan. Learning to model the world with language. *arXiv preprint arXiv:2308.01399*, 2023.

Lingyu Liu, Yaxiong Wang, Li Zhu, and Zhedong Zheng. Can video diffusion models predict past frames? bidirectional cycle consistency for reversible interpolation. *arXiv preprint arXiv:2604.01700*, 2026.

Xinxin Liu, Zhaopan Xu, Ming Li, Kai Wang, Yong Jae Lee, and Yuzhang Shang. Can world simulators reason? gen-vire: A generative visual reasoning benchmark. *arXiv preprint arXiv:2511.13853*, 2025.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024a.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024c.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Francesco Margoni, Luca Surian, and Renée Baillargeon. The violation-of-expectation paradigm: A conceptual overview. *Psychological Review*, 131(3):716, 2024.

Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 152:267–275, 2022.

Shiho Matta, Lis Kanashiro Pereira, Peitao Han, Shigeru Kitazawa, and Fei Cheng. Which way does time flow? a psychophysics-grounded evaluation for vision–language models. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, volume 11, pp. 9449–9459, 2026.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

Albert Michotte. *The perception of causality*. Routledge, 2017.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pp. 527–544. Springer, 2016.

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

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 948–958, 2026.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Leland Gerson Neuberg. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19023–19034, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

Lisa M Oakes and Leslie B Cohen. Infant perception of a causal event. *Cognitive Development*, 5(2): 193–207, 1990.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36: 42748–42761, 2023.

Judea Pearl et al. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video generation model in $200 k. *arXiv preprint arXiv:2503.09642*, 2025.

Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2035–2042, 2014.

Luis S Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature human behaviour*, 6(9): 1257–1267, 2022.

Vincent Plikat, Pablo R Grassi, Julius Frack, and Andreas Bartels. Hierarchical surprise signals in naturalistic violation of expectations. *Imaging Neuroscience*, 3:imag_a_00459, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

Aimon Rahman, Malsha V Perera, and Vishal M Patel. Frame by familiar frame: Understanding replication in video diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2766–2776. IEEE, 2025.

Sahithya Ravi, Aditya Chinchure, Raymond Ng, Leonid Sigal, and Vered Shwartz. Spike-rl: Video-llms meet bayesian surprise. In *International Conference on Learning Representations*, volume 2026, pp. 121915–121935, 2026.

Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *International Conference on Learning Representations*, volume 2024, pp. 15786–15817, 2024.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. Intphys 2019: A benchmark for visual intuitive physics understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5016–5025, 2021.

Rebecca Saxe and Susan Carey. The perception of causality in infancy. *Acta psychologica*, 123(1-2): 144–165, 2006.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International conference on machine learning*, pp. 9614–9625. PMLR, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. *Advances in neural information processing systems*, 32, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058. IEEE, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

Elizabeth S Spelke, Karen Breinlinger, Janet Macomber, and Kristen Jacobson. Origins of knowledge. *Psychological review*, 99(4):605, 1992.

Aimee E Stahl and Lisa Feigenson. Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230):91–94, 2015.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8406–8416, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Genmo Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *science*, 332(6033): 1054–1059, 2011.

Varun Varma Thozhiyoor, Shivam Tripathi, Venkatesh Babu Radhakrishnan, and Anand Bhattad. Principia: Relational physics tests for video models. *arXiv preprint arXiv:2609.04200*, 2026.

Gerhard Tutz. Bradley-terry-luce models with an ordered response. *Journal of mathematical psychology*, 30(3):306–316, 1986.

Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9):649–665, 2017.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Rishi Upadhyay, Howard Zhang, Jim Solomon, Ayush Agrawal, Pranay Boreddy, Shruti Satya Narayana, Yunhao Ba, Alex Wong, Celso M de Melo, and Achuta Kadambi. Worldbench: Disambiguating physics for diagnostic evaluation of world models. *arXiv preprint arXiv:2601.21282*, 2026.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

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.

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

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. Timecausality: Evaluating the causal ability in time dimension for vision language models. *arXiv preprint arXiv:2505.15435*, 2025.

Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8052–8060, 2018.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

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

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, et al. Omni-worldbench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9777–9786, 2021.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10326–10335. IEEE, 2019.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Diagnosing under-development of irreversible processes in video generation. *arXiv preprint arXiv:2608.00617*, 2026.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

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

Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, et al. Towards precise scaling laws for video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18155–18165, 2025.

Fangxu Yu, Ziyao Lu, Liqiang Niu, Fandong Meng, and Jie Zhou. Arrowgev: Grounding events in video via learning the arrow of time. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34657–34671, 2026.

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

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 2023.

Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck WE Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, et al. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, volume 2026, pp. 55660–55699, 2026.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Bohan Li, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, et al. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024.

APPENDIX OVERVIEW

This appendix provides additional details and complete numerical results to complement the main manuscript. First, we describe the dataset construction protocol in Appendix A. Then, we detail the inference configurations and specifications of all 13 evaluated models in Appendix B, followed by the unified video preprocessing pipeline in Appendix C. Next, we formalize the Reverse Surprise Index (RSI) and Causality Cognition Index (CCI) algorithms in Appendix D and Appendix E, respectively. We further discuss the potential prompt bias and how our two-level metric design addresses it in Appendix F. Finally, we present the human annotation procedure and human preference study design in Appendix G and Appendix H, and provide the complete numerical tabular results for all metrics in Appendix I. In addition, we provide an interactive HTML interface to show our video results.

A DETAILS ON DATASET CONSTRUCTION

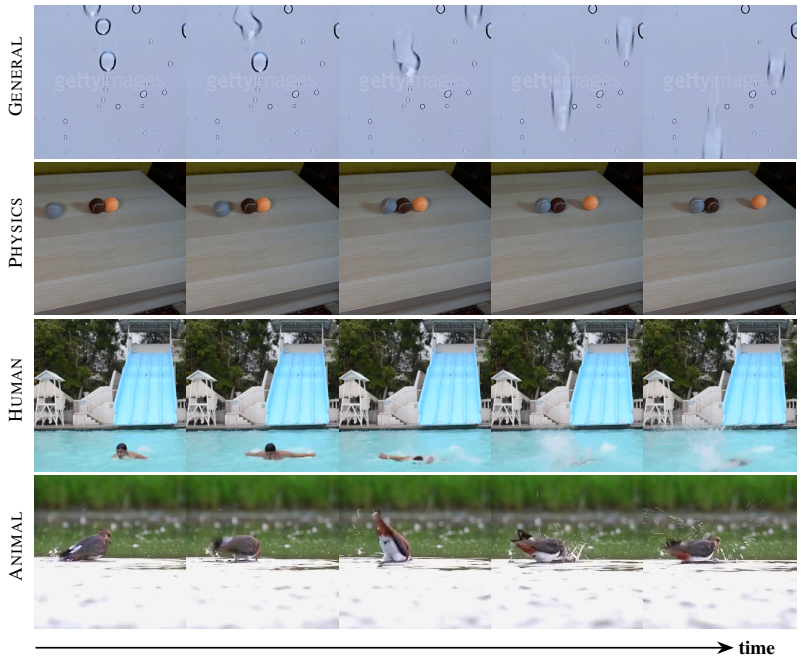


Figure 9: **Qualitative results of our four diverse causal domains.** Each row illustrates the featured video in the subsets.

Our benchmark dataset $\mathcal{D}$ comprises four thematic subsets, each sourced from a well-established real-world video collection. We describe the construction protocol for each subset below, followed by a discussion of the dataset's overall scale and extensibility.

General Subset ($\mathcal{D}_{\textbf{General}}$)**.** We randomly sample 500 videos from the Moments in Time dataset (Monfort et al., 2019), retaining each clip at its original duration of 3 seconds. Moments in Time Monfort et al. (2019) is a large-scale collection of approximately one million 3-second clips curated by MIT, covering 339 action and event categories that span human activities, animal behaviors, natural phenomena, and object state changes. Its broad semantic diversity makes it well-suited for evaluating general everyday causality without domain-specific bias.

Physics Subset ($\mathcal{D}_{\textbf{Physics}}$)**.** We source 132 videos from the Physics IQ dataset (Motamed et al., 2026), selecting one camera viewpoint per scene from the dataset's 132 physical scenarios and trimming each clip to the first 5 seconds following the official evaluation protocol. Physics IQ is specifically designed to assess intuitive physical understanding in video generative models, covering phenomena such as gravity, fluid motion, and collision dynamics. Each clip presents an unambiguous physical event, making this subset particularly rich in temporally asymmetric causal cues.

Human Action Subset ($\mathcal{D}_{\textbf{Human}}$)**.** We randomly sample one video per action class from Kinetics-400 (Kay et al., 2017), yielding 400 clips in total, each trimmed to the first 3 seconds. Kinetics-400 is a canonical human action recognition benchmark curated by DeepMind, comprising approximately 300K YouTube clips across 400 action categories such as swimming, playing guitar, and handshaking. The abundance of diverse human-centric activities provides strong coverage of goal-directed, agent-caused interactions where the causal arrow is tied to purposeful body motion.

Animal Action Subset ($\mathcal{D}_{\textbf{Animal}}$)**.** We randomly sample 200 videos from the Animal Kingdom dataset (Ng et al., 2022), trimming each clip to the first 3 seconds. Animal Kingdom is a large-scale dataset annotated with a rich taxonomy of animal behaviors, including foraging, aggression, courtship, and locomotion, across a highly diverse set of species. This subset extends our benchmark beyond anthropocentric scenarios to assess whether causal cognition generalizes to non-human agents and natural world dynamics.

Scale and Extensibility. As discussed in the main paper, our benchmark is inherently extensible, as any real-world video can be incorporated at zero additional cost via temporal reversal. In the current work, we evaluate a total of 1,232 videos across the four subsets described above. This scale was chosen under practical computational constraints: evaluating 13 video diffusion models (several with billions of parameters) requires substantial GPU overhead, and the computational resource we can afford only supports us to evaluate with thousands of videos. Future researchers with access to greater resources, or those benchmarking a single target model, are encouraged to expand the dataset's scale and diversity by incorporating additional domains (e.g., tool use, cooking, or sports) to keep the benchmark evolving alongside model capabilities.

A summary of the subset composition is provided in Table 4.

Table 4: **Composition of the YoCausal benchmark dataset.** Each subset is sourced from an existing real-world video collection, with clips uniformly trimmed to the specified duration.

Subset	Source Dataset	Clip Duration	# Videos
$\mathcal{D}_{\text{General}}$	Moments in Time (Monfort et al., 2019)	3 s	500
$\mathcal{D}_{\text{Physics}}$	Physics IQ (Motamed et al., 2026)	5 s (first 5 s)	132
$\mathcal{D}_{\text{Human}}$	Kinetics-400 (Kay et al., 2017)	3 s (first 3 s)	400
$\mathcal{D}_{\text{Animal}}$	Animal Kingdom (Ng et al., 2022)	3 s (first 3 s)	200
Total	—	—	**1,232**

B Models Setting

We evaluate 13 state-of-the-art open-source text-to-video diffusion models spanning diverse architectures and parameter scales: AnimateDiff-SDv1.5/SDXL (Guo et al., 2023), CogVideoX-2B/5B (Hong et al., 2022; Yang et al., 2024), CogVideoX1.5-5B (Hong et al., 2022; Yang et al., 2024), Mochi-1-preview (Team, 2024), HunyuanVideo (Kong et al., 2024), Wan2.1-T2V-1.3B/14B (Wan et al., 2025), Wan2.2-TI2V-5B/T2V-A14B (Wan et al., 2025), and LTX-Video-2B/13B (HaCohen et al., 2024).

To ensure each model operates under its optimal conditions, all inference configurations strictly follow the official recommended defaults for each model (Table 5), including output resolution, number of frames, and frames per second (FPS). It is worth noting that classifier-free guidance (CFG) (Ho & Salimans, 2022) is not applied during evaluation: rather than performing full denoising generation, we directly compute the MSE between the model's predicted noise $\hat{\epsilon}$ and the sampled Gaussian noise ϵ at each timestep. In addition, we report model specifications (e.g., spatial compression ratios) in Table 5 to further investigate the relationship between model features and causal cognition performance.

C Video Preprocessing

Since each model operates under its officially recommended settings (Appendix B), input specifications inevitably differ across models in spatial resolution, frame rate, and temporal length. To ensure a fair and consistent evaluation, we apply a unified preprocessing pipeline encompassing three stages: resolution adaptation, FPS resampling, and long video handling.

Table 5: **Specifications of evaluated video generation models, grouped by model series.** For each model, we report the parameter count, prediction target, spatial and temporal compression ratios of the VAE, frame window size, frames per second (FPS), and default output resolution.

Model Name	Params.	Prediction Target	Spatial Compress	Temporal Compress	Frame Window	FPS	Default Resolution
Wan							
Wan 2.2 A14B	14B	Flow	8×	4×	81	16	1280×720
Wan 2.2 5B	5B	Flow	8×	4×	121	24	1280×720
Wan 2.1 14B	14B	Flow	8×	4×	81	16	1280×720
Wan 2.1 1.3B	1.3B	Flow	8×	4×	81	16	832×480
Hunyuan							
HunyuanVideo	13B	Flow	8×	4×	129	24	1280×720
Mochi							
Mochi 1	10B	Flow	8×	6×	163	30	848×480
CogVideoX							
CogVideoX-5B	5B	v	8×	4×	81	16	720×480
CogVideoX-2B	2B	v	8×	4×	81	16	720×480
CogVideoX1.5-5B	5B	v	8×	4×	81	16	1360×768
AnimateDiff							
AnimateDiff SD 1.5	1.4B	ϵ	8×	1×	16	8	512×512
AnimateDiff SDXL	3.5B	ϵ	8×	1×	16	8	1024×1024
LTXV							
LTXV-13B-0.9.8	13B	Flow	32×	8×	121	30	1216×704
LTXV-2B-0.9.6	2B	Flow	32×	8×	121	30	1216×704

Resolution Adaptation. Since different models require different input resolutions and aspect ratios, we employ two adaptation strategies depending on the model's design. For models that only support a fixed resolution, we first rescale the input video such that its shorter side matches the target dimension, and then apply a center crop to obtain the model's officially specified resolution and aspect ratio. For models that support multiple aspect-ratio buckets (e.g., the Wan series and HunyuanVideo), we automatically select the bucket resolution closest to the original video's aspect ratio.

FPS Resampling. We use FFmpeg to resample all source videos to each model's prescribed frame rate, ensuring that the temporal sampling rate is consistent with the model's training configuration.

Long Video Handling. Different models have different maximum frame windows (i.e., the number of frames they can process in a single forward pass). When the total number of frames in a source video exceeds a model's frame window limit, naively truncating the video would compromise evaluation fairness, as the model would only observe a partial segment. To address this, we partition the video along the temporal axis into consecutive frame-window-sized segments. If the last segment is shorter than a full window, we pad it by prepending the necessary number of frames from the preceding segment as temporal context. Importantly, this context portion is used solely to provide temporal information and is excluded from the denoising loss accumulation. Finally, we sum the denoising losses across all windows for the entire video, and use this aggregated value as the basis for comparing forward and reversed sequences.

D Details on RSI Algorithm

The Reverse Surprise Index (RSI) quantifies a model's ability to perceive the arrow of time, serving as our Level-1 metric. The complete procedure is formalized in Algorithm 1 (Stage 1 and Stage 2).

Timestep and Noise Sampling. The denoising loss is defined as an expectation over both diffusion timesteps t and sampled noise ϵ. To approximate this expectation, we uniformly sample $K{=}10$ timesteps from $[1, T]$, excluding the boundaries $t{=}0$ and $t{=}T_{\max}$ (fully clean and fully noised states). For each timestep, we draw $N_\epsilon{=}1$ noise sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and apply it identically to both the forward sequence x^f and the reversed sequence x^r, ensuring that the two versions face

identical denoising difficulty. For latent diffusion models, the noise is sampled in the latent space; for pixel-space models, it is sampled at the full spatial resolution.

Following LikePhys(Yuan et al., 2025), we condition both the forward and reversed sequences on the same text prompt, i.e., the original caption of the forward video in the dataset. Writing a separate caption for the reversed clip would conflate causal understanding with the ability to follow unrealistic instructions. While a null prompt is also viable, conditioning on the meaningful caption provides a sharper denoising signal at low-SNR timesteps, avoding the randomness of this evaluation method and giving more stable evaluation process. We also empirically validate that RSI and CCI remain significant under null prompts (see Appendix F), confirming that our metrics capture genuine temporal-causal structure rather than text-video misalignment artifacts.

We set $K{=}10$ and $N_\epsilon{=}1$ primarily due to computational constraints: our evaluation spans 13 large-scale models over 1,232 videos, making exhaustive sampling prohibitive given our limited computational resources. As noted in Appendix A, future researchers with sufficient computational resources are encouraged to increase both K and N_ϵ, yielding estimates closer to the true expected denoising loss.

Decision Criterion. For each video, we average the denoising losses over all K timesteps and $N_\epsilon{=}1$ noise samples. If $\mathcal{L}(\theta;\, x^f) < \mathcal{L}(\theta;\, x^r)$, the model is deemed to have correctly identified the temporal direction. RSI is then the proportion of videos satisfying this condition; 50% corresponds to chance level.

Algorithm 1 YoCausal: Two-Level Causality Evaluation for Video Diffusion Models

Require: Video dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$; pretrained VDM ϵ_θ; K uniformly sampled timesteps; N_ϵ noise samples per timestep; VLM causality classifier $\mathcal{C}_{\text{vlm}}$
Ensure: RSI($\mathcal{D}$), CCI($\mathcal{D}$)

// Stage 1: Compute Denoising Losses
for each $\mathcal{D}_i \in \mathcal{D}$, each video $x \in \mathcal{D}_i$ **do**
  Set $x^f \leftarrow x$; $x^r \leftarrow$ TEMPORAL-REVERSE(x)
  for each direction $\tilde{x} \in \{x^f, x^r\}$ **do**
    $\mathcal{L}(\tilde{x}) \leftarrow 0$; sample $\{t_k\}_{k=1}^{K}$ uniformly from $[1, T]$
    for $k = 1$ **to** K, $m = 1$ **to** N_ϵ **do**
      Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
      $\tilde{x}_{t_k} \leftarrow \sqrt{\bar{\alpha}_{t_k}}\,\tilde{x} + \sqrt{1-\bar{\alpha}_{t_k}}\,\epsilon$
      $\hat{\epsilon} \leftarrow \epsilon_\theta(\tilde{x}_{t_k}, t_k)$; $\mathcal{L}(\tilde{x}) \mathrel{+}= \|\epsilon - \hat{\epsilon}\|_2^2$
    end for
    $\mathcal{L}(\tilde{x}) \leftarrow \mathcal{L}(\tilde{x}) \,/\, (K \times N_\epsilon)$
  end for
end for
// Stage 2: Level-1 — Reverse Surprise Index (RSI)
for each $\mathcal{D}_i \in \mathcal{D}$ **do**
  $S_i \leftarrow \sum_{x \in \mathcal{D}_i} \mathbf{1}\big[\mathcal{L}(x^r) > \mathcal{L}(x^f)\big]$; RSI($\mathcal{D}_i$) $\leftarrow S_i \,/\, |\mathcal{D}_i|$
end for
RSI($\mathcal{D}$) $\leftarrow \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}_i \in \mathcal{D}}$ RSI($\mathcal{D}_i$)
// Stage 3: VLM Causality Stratification
$\mathcal{D}_c \leftarrow \emptyset$; $\mathcal{D}_{nc} \leftarrow \emptyset$
for each $\mathcal{D}_i \in \mathcal{D}$, each $x \in \mathcal{D}_i$ **do**
  if $\mathcal{C}_{\text{vlm}}(x) =$ TRUE **then**
    $\mathcal{D}_c \leftarrow \mathcal{D}_c \cup \{x\}$
  else
    $\mathcal{D}_{nc} \leftarrow \mathcal{D}_{nc} \cup \{x\}$
  end if
end for
// Stage 4: Level-2 — Causality Cognition Index (CCI)
Compute RSI($\mathcal{D}_c$) and RSI($\mathcal{D}_{nc}$) as in Stage 2
CCI($\mathcal{D}$) $\leftarrow$ RSI($\mathcal{D}_c$) $-$ RSI($\mathcal{D}_{nc}$)
return RSI($\mathcal{D}$), CCI($\mathcal{D}$)

E Details on CCI Algorithm

As discussed in the main paper, in order to isolate the pure causal cognition signal from the overall arrow-of-time perception captured by RSI, we propose the Causality Cognition Index (CCI). It is defined as the difference in RSI between the causal and non-causal subsets of the dataset: $\text{CCI}(\mathcal{D}) = \text{RSI}(\mathcal{D}_c) - \text{RSI}(\mathcal{D}_{nc})$. where $\mathcal{D}_c$ denotes the subset of videos containing salient cause-and-effect interactions, and $\mathcal{D}_{nc}$ denotes the subset without such interactions. The intuition behind this differential design is as follows: the surprise induced by reversing a non-causal video originates primarily from the reversal of the statistical arrow of time, whereas the surprise from reversing a causal video additionally includes the physical implausibility of inverted causality. By taking the difference, CCI cancels out the shared confounding factor: arrow-of-time sensitivity and isolates the model's independent sensitivity to causal violations. The complete procedure is formalized in Algorithm 1 (Stage 4).

VLM-Based Causal/Non-Causal Partitioning. Computing CCI requires partitioning the dataset $\mathcal{D}$ into $\mathcal{D}_c$ and $\mathcal{D}_{nc}$. To ensure the scalability of our benchmark, we employ an advanced Vision-Language Model (VLM) to automatically determine whether each video contains obvious causal interactions, rather than relying on manual annotation (Algorithm 1 (Stage 3)). The visualization result is shown in Fig. 10. The rationale for this automated strategy is that judging *whether causality exists* in a video is a substantially easier task than judging *whether a causal relation is correct*, making VLMs well-suited for this binary classification. In the main paper, we validate the reliability of this approach from two complementary perspectives, confirming that VLM-based labeling serves as a reliable proxy for human judgment.

VLM Configuration and Prompt. We use Gemini 3.0 Pro (Team et al., 2023) as the VLM judge. Each video is provided as visual input alongside the following prompt:

Prompt

```
You are an expert in spatio-temporal video analysis.  Your task is to analyze the
provided video and determine whether it exhibits "Causality."
```

 # Definition of Causality In this context, "Causality" is defined strictly as observable cause-and-effect mechanisms within spatio-temporal dynamics.

- It means Event A explicitly and visibly causes Event B.
- It goes beyond strict Newtonian physics (e.g., collisions or gravity) to include complex logical event sequences (e.g., a pencil moving on paper causing text to appear, or a person flipping a switch causing a light to turn on).

 # Output Format Provide your final answer in the following JSON format:

```
{
  "reasoning": "Step-by-step reasoning evaluating if this
  specific resulting_event is strictly dependent on the initiating_event.",
  "video_has_general_causality": true/false,
  "confidence": 1-5
}
```

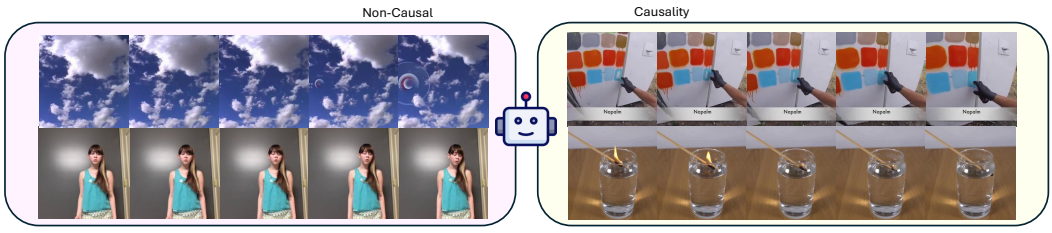


Figure 10: **Examples from the causal and non-causal subsets.** Non-causal videos (left) and Causal videos (right) are split by VLMs.

F Prompt Bias

In our experimental design, both the forward and reversed videos are conditioned on the same text prompt, namely the original caption describing the forward video, avoiding the randomness under

low-SNR timesteps with null prompt. A natural concern is whether the elevated denoising loss on reversed videos merely reflects a semantic mismatch between the visual content and the prompt, rather than the model's sensitivity to causal violations. We address this concern from an empirical ablation under a null-prompt setting.

Empirical validation: null-prompt ablation. We design an ablation to empirically validate that our metrics are not driven by text-video misalignment.

Hypothesis. If the elevated denoising loss on reversed videos primarily reflected text-video misalignment, then removing the prompt conditioning should eliminate the bulk of the RSI/CCI signal. Conversely, if the signal genuinely originates from the model's internalized temporal and causal cognition, then RSI and CCI under null prompts should be close to values under forward video prompt and preserve their discriminative power.

Setup. We re-evaluate three representative models under a *null-prompt* setting, where both forward and reversed sequences are conditioned on an empty caption.

Table 6: **Null-prompt ablation results.** We compare RSI and CCI under the forward prompt setting (used in our main experiments) and the null prompt setting on three representative models from a range of model families and capabilities. Both metrics under null prompts are still close to the values under forward prompt, and the discriminative structure is preserved, supporting our claim that RSI and CCI are not artifacts of text-video misalignment.

	HunyuanVideo	CogVideoX-5b	LTX-Video-2b-0.9.6
RSI w/ forward prompt	52.05%	49.92%	58.86%
RSI w/ null prompt	51.17%	47.55%	56.92%
CCI w/ forward prompt	-0.29%	5.09%	-0.20%
CCI w/ null prompt	-2.95%	6.17%	-0.30%

Results. As reported in Table 6, removing the prompt produces only negligible changes in RSI and CCI. Crucially, the discriminative structure of both metrics is preserved, meaning the sign remain consistent with the forward-prompt setting.

This invariance directly falsifies the misalignment hypothesis. If text-video misalignment were the dominant source of elevated reversed-video loss, removing the prompt would pull all index to randomness line across all models. The observed pattern is incompatible. The dominant signal originates from the model's internalized cognition, not from prompt conditioning. This empirically complements our argument: the forward prompt serves as a stable, easily reproducible conditioning choice for the main benchmark, and is not the source of the cognition signal we measure.

We think using our benchmark with null prompt is also a reasonable choice, it may cancel out the slight text-misalignment but introdeuces randomness at low-SNR timesteps. In constrast, using our benchmark with forward prompt will cancel out the randomness at low-SNR timesteps but introduces little text-misalignment. That is a trade-off and in this paper we choose the more stable one (with forward prompt).

G Details on Human Annotating

To establish a human upper bound for the YoCausal benchmark, we recruit human annotators to label the temporal direction of all 1,232 videos. The labeling protocol is designed to maintain methodological parity with the Reverse Surprise Index (RSI) evaluation pipeline applied to models. Below, we describe the annotation procedure and the design of the "Unknown" option.

Annotation Procedure and Decision Protocol. As shown in Fig. 11, annotators first read the corresponding text prompt, ensuring that human judgments are made under the same semantic constraints as model evaluations. Each annotator then sequentially watches the forward and reverse versions of the same video, presented in randomized order, and after viewing both clips, determines which one is the reversed version.

The rationale behind this sequential design mirrors how VDMs process the two versions: rather than simultaneously comparing both videos side by side, a VDM independently computes the denoising loss for each direction and infers the temporal order by comparing the resulting loss values, which is indirect comparison through its internal learned prior. Our annotation protocol follows the same principle: annotators watch each version independently at different pages and form an internal judgment about the degree of abnormality in each clip. The final decision of which version is reversed is then based on comparing these two independently perceptions of anomaly. The human annotator's internal sense of abnormality functionally corresponds to the model's denoising loss.

Furthermore, to direct annotators' attention toward high-level causal reasoning rather than low-level visual artifacts, we limit viewing to at most three replays for each of the forward and reverse versions.

Design and Modeling of the "Unknown" Option. In the real world, certain events exhibit negligible directional cues (e.g., static scenes, periodic motions), making it inherently difficult to distinguish forward from reversed playback. To deal with such cases, the annotation interface provides an "Unknown" option, preventing annotators from being forced into arbitrary decisions. Across all 1,232 samples, approximately 20% of videos are marked as "Unknown."

To maintain consistency with model behavior under uncertainty, samples labeled as "Unknown" are assigned an expected win rate of 0.5 (equivalent to the random-guess baseline) when computing the overall RSI score.

We provide a screenshot of the annotation web interface in Fig. 11:

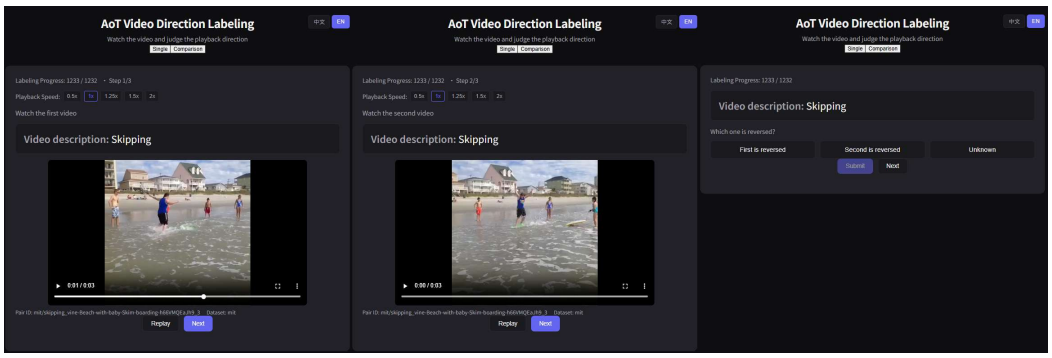


Figure 11: **Website for human annotation.** Annotators are shown a text prompt (top) and sequentially watch two versions of the same video—forward and reversed—presented in randomized order (left and middle panels). After viewing both clips, annotators select which version is reversed, with an "Unknown" option available.

H Details on Human Preference

To verify the consistency between YoCausal's evaluation results and human subjective judgments, we design a *Human Causality Preference* user study. The study encompasses model selection, prompt curation, evaluation procedure, and scoring, each of which is detailed below.

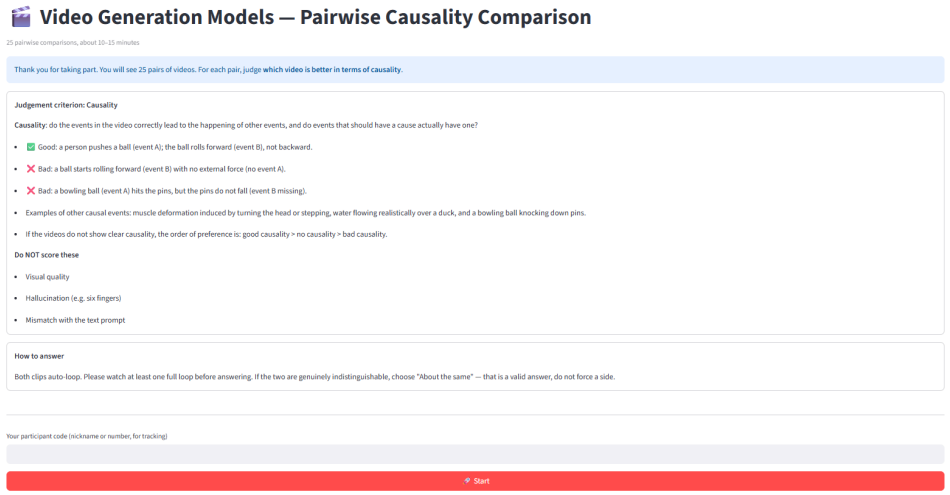
Model Selection. Since the study adopts a pairwise comparison protocol, participants only need to compare two videos at a time. This allows us to include all 13 evaluated models rather than a subset of representative models, so that the human ranking can be directly compared with the full YoCausal ranking.

Prompt Curation. We select 15 prompts with relatively identifiable causal interactions from each of the four subsets ($\mathcal{D}_{General}$, $\mathcal{D}_{Physics}$, $\mathcal{D}_{Human}$, $\mathcal{D}_{Animal}$), yielding a total of 60 prompts. Each of the 13 models then generates a video for every prompt, producing $13 \times 60 = 780$ videos in total.

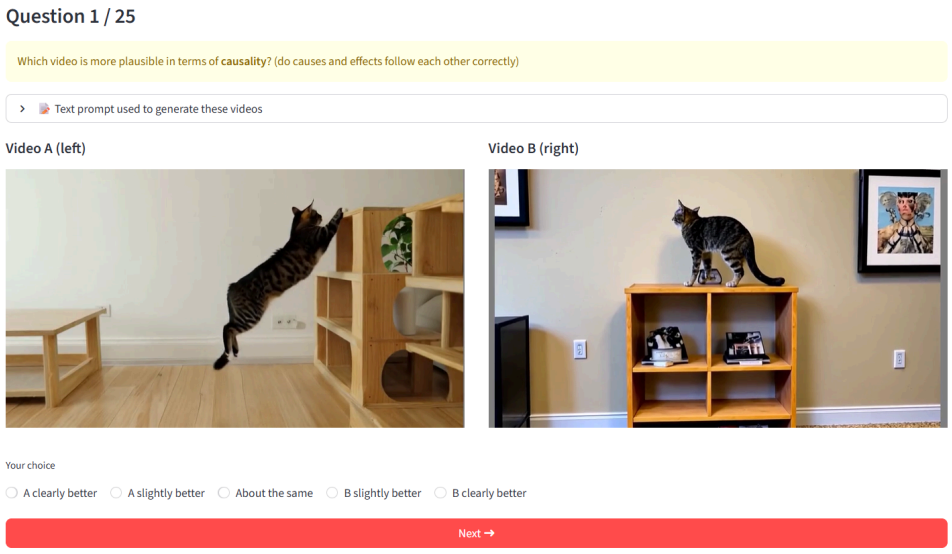
Evaluation Procedure. We deploy an online user-study website for the evaluation. On each page, the website presents two videos generated by two different models from the same prompt, displayed side by side with randomized left/right placement. Participants are asked to judge which video depicts more plausible causal interactions on a 5-point scale: *A clearly better*, *A slightly better*, *About the*

same, *B slightly better*, and *B clearly better*. Since videos generated by different models may exhibit similar levels of causal plausibility, the *About the same* option avoids forcing artificial distinctions. We recruit 40 participants in total, and each participant completes 25 comparisons, yielding 1,000 pairwise judgments. The comparisons are assigned so that all $\binom{13}{2} = 78$ model pairs are covered, each pair is compared on multiple distinct prompts.

We provide a screenshot of the human preference web interface in Fig. 12:



(a)



(b)

Figure 12: **Website for human preference ranking.** (a) Instructions and important notes for participants. (b) The comparing interface, where participants choose the video which perform better causal understanding.

Scoring and Aggregation. We first convert each judgment into a score $y \in \{-2, -1, 0, +1, +2\}$ that indicates how much video B is preferred over video A: *A clearly better* $\rightarrow -2$, *A slightly better* $\rightarrow -1$, *About the same* $\rightarrow 0$, *B slightly better* $\rightarrow +1$, and *B clearly better* $\rightarrow +2$. We then use a *cumulative-logit ordinal paired-comparison model* (Tutz, 1986; Agresti, 1992), an extension of the Bradley–Terry model (Bradley & Terry, 1952) that supports ties and graded preferences through ordered thresholds, to estimate a human preference score s_m for each model m. We further extend it

with participant-specific and prompt-specific scale terms and a position term, described below. These scores are not computed by a closed-form formula. Instead, they are the values that best explain the observed judgments, obtained in two steps: we first model the probability of each possible judgment, and then estimate the scores that make the observed judgments most likely.

Modeling a single judgment. Consider a comparison between model a (shown as video A) and model b (shown as video B), rated by participant r on prompt p. The model first computes a preference strength

$$\eta = (s_b - s_a) \cdot \exp(w_r + \lambda_p) + \beta, \tag{4}$$

which summarizes how strongly B is expected to be preferred over A in this particular comparison. The terms in η are unknown parameters, defined as follows:

- $s_b - s_a$ is the score gap between the two models, which is the quantity of interest.
- w_r is a participant term that captures how strongly participant r reacts to differences between videos. A participant with a small w_r tends to answer 0 even when the difference is noticeable, while a participant with a large w_r tends to answer ± 2. Modeling this prevents unusually cautious or decisive participants from distorting the model scores.
- λ_p is a prompt term that captures how easily prompt p reveals differences between models. For prompts whose causal interactions are easy to judge, λ_p is larger, so the same score gap leads to stronger preferences.
- β is a position term that captures any systematic preference for the left (A) or right (B) video, independent of content.

The preference strength η is then converted into a probability for each of the five judgments using two thresholds $0 < c_1 < c_2$:

$$P(y \le k) = \sigma(\theta_k - \eta), \qquad (\theta_{-2}, \theta_{-1}, \theta_0, \theta_{+1}) = (-c_2, -c_1, c_1, c_2), \tag{5}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function and $P(y \le +2) = 1$. The probability of each individual judgment is the difference between consecutive cumulative probabilities; for example, $P(y = 0) = \sigma(c_1 - \eta) - \sigma(-c_1 - \eta)$. As a result, a large positive η makes $+2$ most likely, a large negative η makes -2 most likely, and $P(y = 0)$ is largest when $\eta = 0$, i.e., when the two models have similar scores. The threshold c_1 determines where *About the same* ends and *slightly better* begins, and c_2 determines where *slightly better* ends and *clearly better* begins.

Estimating the parameters. All unknown parameters, namely the model scores s_m, the participant terms w_r, the prompt terms λ_p, the position term β, and the thresholds c_1, c_2, are estimated jointly from the 1,000 judgments (36 participants $\times$ 25 comparisons). We choose the values that maximize the total log-probability of the judgments actually observed:

$$\max_{\{s_m\},\{w_r\},\{\lambda_p\},\beta,c_1,c_2} \sum_{t=1}^{1000} \log P(y_t \mid \eta_t) \; + \; \log P_{\text{prior}}, \tag{6}$$

where y_t and η_t are the judgment and preference strength of the t-th comparison. Intuitively, if B is frequently preferred over A, the optimization increases s_b relative to s_a so that the model assigns higher probability to these judgments. The prior term places weak Gaussian priors $s_m \sim \mathcal{N}(0, 1)$, $w_r, \lambda_p \sim \mathcal{N}(0, 0.35^2)$, and $\beta \sim \mathcal{N}(0, 0.25^2)$, and half-normal priors $c_1, c_2 - c_1 \sim \text{HalfNormal}(1)$. It keeps the estimates stable when data are limited (e.g., when one model wins every comparison against another) and keeps w_r, λ_p, and β close to zero unless the data indicate otherwise. Since only score differences affect η, the scores have no absolute zero point; we fix it by constraining the scores to have mean zero, which is why some models receive negative scores. We solve this optimization with a gradient-based method (L-BFGS). Once the scores are estimated, $s_b - s_a$ is simply the difference between the estimated scores of two models.

Judgments of 0 are kept in the estimation: they indicate that the two models have similar scores, rather than being discarded or treated as half a win for each side. We rank the models by their estimated scores s_m to obtain the overall human rank. This ranking closely agrees with the YoCausal ranking (CCI rank + RSI rank, with ties broken by RSI rank), with a Kendall's τ of 0.667 ($p < 0.001$, exact permutation test). The complete ranking results are presented in Table 7.

I Detailed Numerical Results

In the main paper, we primarily present the RSI and CCI results in graphical form to facilitate intuitive comparison of performance differences and trends across models. To enable precise numerical lookup

Table 7: **Overall ranking of human causality preference study.** All 13 models are ranked by their human preference score, estimated with the Davidson model from 1,000 pairwise judgments collected from 40 participants (higher is better). For reference, we also list each model's YoCausal rank (CCI rank + RSI rank). The two rankings are strongly correlated (Kendall's $\tau = 0.667$, $p < 0.001$).

Rank	1	2	3	4	5	6	7
Model	Wan2.2-A14B	Wan2.1-14B	HunyuanVideo	Mochi	LTXV-2B	Wan2.2-5B	LTXV-13B
Human Preference Score	1.431	1.263	0.575	0.319	0.174	0.167	0.088
YoCausal Rank	1	2	6	7	3	10	5
Rank	8	9	10	11	12	13	
Model	CogVideoX-5B	Wan2.1-1.3B	CogVideoX1.5-5B	AnimateDiff-SD1.5	CogVideoX-2B	AnimateDiff-SDXL	
Human Preference Score	−0.055	−0.123	−0.459	−0.827	−0.973	−1.580	
YoCausal Rank	4	9	8	12	11	13	

and further analysis by the reader, this section provides the complete experimental data in tabular form. Specifically, Tab. 8 reports the RSI scores of all 13 evaluated models and the human baseline across the four subsets ($\mathcal{D}_{General}$, $\mathcal{D}_{Physics}$, $\mathcal{D}_{Human}$, $\mathcal{D}_{Animal}$) as well as the overall dataset. Tab. 9 provides the disaggregated RSI scores on the causal subset RSI($\mathcal{D}_c$) and the non-causal subset RSI($\mathcal{D}_{nc}$), along with the resulting CCI scores. We additionally include the human baseline as a reference for both metrics.

Table 8: **Numerical results of RSI.** This table shows RSI scores of all 13 evaluated models and the human baseline across four subsets ($\mathcal{D}_{General}$, $\mathcal{D}_{Physics}$, $\mathcal{D}_{Human}$, $\mathcal{D}_{Animal}$) and the overall average. Models are sorted by average RSI (RSI($\mathcal{D}$)) in ascending order. Bold values indicate the best-performing model in each column.

		level 1: Reverse Surprise Index (RSI)↑				
	Release Date	$\mathcal{D}_{General}$	$\mathcal{D}_{Physics}$	$\mathcal{D}_{Human}$	$\mathcal{D}_{Animal}$	$\mathcal{D}$
AnimateDiff-SDXL	04/2024	27.80%	41.67%	48.73%	46.50%	41.18%
CogVideoX-2b	08/2024	33.20%	40.15%	56.64%	36.00%	41.50%
Wan2.1-T2V-1.3B	03/2025	29.80%	59.09%	59.15%	34.00%	45.51%
AnimateDiff-SD-1.5	06/2023	33.00%	32.58%	61.68%	55.50%	45.69%
CogVideoX1.5-5b	11/2024	28.80%	62.12%	62.91%	33.50%	46.83%
Mochi-1-preview	10/2024	37.80%	43.18%	76.50%	39.00%	49.12%
CogVideoX-5b	08/2024	31.10%	67.42%	63.16%	38.00%	49.92%
Wan2.2-TI2V-5B	07/2025	34.40%	71.97%	63.75%	37.50%	51.91%
HunyuanVideo	11/2025	25.80%	64.39%	**86.50%**	31.50%	52.05%
Wan2.1-T2V-14B	03/2025	37.60%	70.45%	66.92%	38.00%	53.24%
Wan2.2-T2V-A14B	07/2025	36.80%	**77.27%**	66.17%	36.50%	54.19%
LTX-Video-13b-0.9.8	07/2025	**61.20%**	47.73%	47.50%	**69.50%**	56.48%
LTX-Video-2b-0.9.6	04/2025	58.60%	57.58%	54.25%	65.00%	**58.86%**
Human		76.60%	91.7%	76.0%	72.0%	79.08%

Better perceives arrow of time ↓

J Scaling Law and Generational Evolution in Causal Cognition

Within the prevailing research trajectory of advancing video generation models toward world models, scaling laws have been regarded as a guiding principle (Yin et al., 2025; Liang et al., 2024; Kaplan et al., 2020). We therefore investigate whether model size and release date correlate with causal cognition by computing the correlation between these factors and our aggregate rank. As shown in Fig. 13, both release date and model parameters exhibit significant correlations with the aggregate ranking (Pearson's $r = 0.6226$ and Pearson's $r = 0.6972$, respectively). This demonstrates that increasing parameter count does enhance causal understanding to a meaningful degree, validating the effectiveness of scaling laws for this higher-level cognitive capability. Furthermore, models have progressively improved in causal cognition across generations. For instance, the transition from UNet-based architectures to DiT-based architectures has yielded substantial improvements in causal perception for most model families.

Table 9: **Numerical results of CCI.** This table shows CCI scores of all 13 evaluated models and the human baseline, along with the disaggregated RSI on the causal subset RSI($\mathcal{D}_c$) and non-causal subset RSI($\mathcal{D}_{nc}$). Normalized CCI rescales raw CCI values relative to the human baseline. Models are sorted by CCI in ascending order. Bold values indicate the best-performing model in each column.

		level 2: Causality Cognition Index (CCI) ↑			
	Release Date	CCI(D) ↑	Normalized CCI(D)	RSI(D_c)	RSI(D_{nc})
AnimateDiff-SD-1.5	06/2023	-5.21%	-60.09%	43.40%	48.61%
AnimateDiff-SDXL	04/2024	-5.07%	-58.48%	38.93%	44.00%
LTX-Video-13b-0.9.8	07/2025	-4.32%	-49.83%	54.65%	**58.97%**
Wan2.2-TI2V-5B	07/2025	-2.12%	-24.45%	50.90%	53.02%
HunyuanVideo	11/2025	-0.29%	-3.34%	51.15%	51.44%
LTX-Video-2b-0.9.6	04/2025	-0.20%	-2.31%	**57.95%**	58.15%
CogVideoX-2b	08/2024	0.93%	10.73%	41.11%	40.18%
Mochi-1-preview	10/2024	3.85%	44.41%	49.11%	45.26%
CogVideoX1.5-5b	11/2024	4.85%	55.94%	48.46%	43.61%
CogVideoX-5b	08/2024	5.09%	58.71%	51.36%	46.27%
Wan2.1-T2V-1.3B	03/2025	5.36%	61.82%	46.92%	41.56%
Wan2.2-T2V-A14B	07/2025	5.51%	63.55%	55.73%	50.22%
Wan2.1-T2V-14B	03/2025	**5.91%**	**68.17%**	54.80%	48.89%
Human		8.67%	100.00%	85.09%	76.42%

Better understand causality ↓

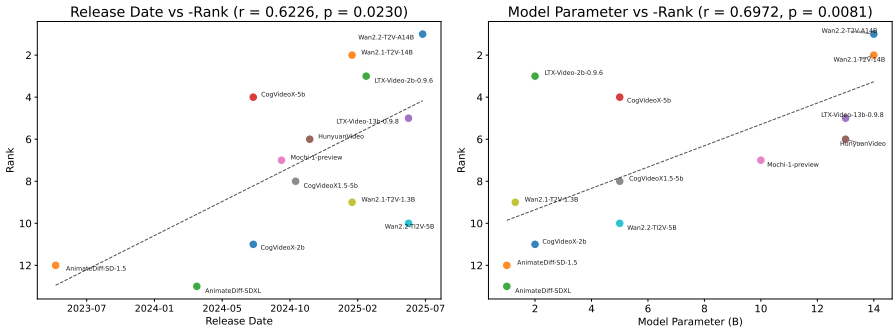


Figure 13: **Scaling laws and generational trends in causal cognition.** Aggregate causal cognition rank correlates positively with both release date (Pearson's r=0.6226) and parameter count (Pearson's r=0.6972), indicating that larger and newer models exhibit stronger causal understanding.