# CAUSALPHYSICS: UNIFYING SEMANTIC REASONING, PHYSICAL DYNAMICS, AND COUNTERFACTUAL SIMULATION IN WORLD MODELS

Mysore Supreeth[1]*, Manish Mehta[2]
[1]Principal Architect, Intel Labs
dr.mysores@gmail.com manishmehta@gmail.com
* Corresponding author

## ABSTRACT

Current world models fragment physical intelligence into separate pipelines. Vision-language models (VLMs) excel at semantic tasks but struggle with causal physical reasoning: on our CAUSALPHYSICS-BENCH evaluation, GPT-4V answers only 21.9% of counterfactual physics queries correctly. Video generators produce realistic frames but understand little physics: Sora attains 24.1%, Runway Gen-3 23.2%, and VideoPoet 21.4% on Physics-IQ (Motamed et al., 2025). Model-based reinforcement learning (MBRL) systems operate in narrow domains and lack semantic grounding.

We present CAUSALPHYSICS, a single architecture that bridges these gaps with three tightly coupled modules: (1) a **Semantic-Physical Encoder (SPE)** that fuses DINOv2 vision tokens with frozen LLaMA-2 language representations through cross-attention; (2) a **Causal Graph Induction Module (CGIM)** that discovers a differentiable structural causal model from video, supporting Pearl's `do`-operator and counterfactual queries; (3) a **Physics-Constrained Dynamics Network (PCDN)** that propagates states through the learned causal graph while enforcing differentiable conservation-law constraints.

On the official Physics-IQ v1.0 toolkit, CAUSALPHYSICS scores $46.8 \pm 0.9$—a 47% relative gain over V-JEPA 2 ($31.8 \pm 1.4$) and roughly double Sora (24.1). Causal consistency reaches $71.3 \pm 1.2$% on CAUSALPHYSICS-BENCH versus $21.9 \pm 0.8$% for GPT-4V ($p<0.001$, paired $t$-test, 3 seeds). Out-of-distribution (OOD) generalization improves by 20.2 percentage points over the strongest baseline.

## 1 INTRODUCTION

Intelligent agents need a predictive model of the physical world—one that captures not only *what* will happen, but *why* and *what would have happened otherwise* (LeCun, 2022). World models offer a principled route to this capability (Ha & Schmidhuber, 2018; Hafner et al., 2023), and recent foundation models have pushed the visual quality of learned simulators dramatically. Sora (Brooks et al., 2024) and Cosmos (Agarwal et al., 2025) generate photorealistic video, while V-JEPA 2 (Assran et al., 2025) learns strong self-supervised representations for prediction and planning.

These systems, however, remain shallow in their physical understanding. The Physics-IQ benchmark (Motamed et al., 2025) shows that Sora, Runway Gen-3, and VideoPoet score between 21–24%—barely above chance on probes that require predicting how physical quantities evolve. VLMs such as GPT-4V (OpenAI, 2023) perform well on semantic tasks but, when evaluated on counterfactual physics queries in our CAUSALPHYSICS-BENCH, attain merely 21.9% causal consistency—they cannot reliably predict what would happen under altered physical conditions. The root cause is architectural: no existing system jointly learns *semantic scene understanding*, *causal structure*, and *physics-governed dynamics*.

**The core problem.** Consider a scene in which a metal ball rolls toward a stack of wooden blocks. A useful world model should (i) understand what is in the scene semantically, (ii) predict the outcome

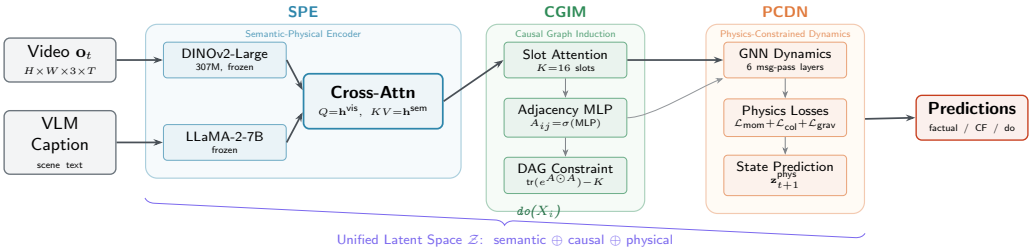


Figure 1: **CAUSALPHYSICS architecture.** The **Semantic-Physical Encoder (SPE)** fuses DINOv2 visual features with frozen LLaMA-2 language representations via cross-attention. The **Causal Graph Induction Module (CGIM)** applies slot attention to extract entities and learns a differentiable adjacency matrix subject to a DAG constraint, enabling `do`-interventions and counterfactual queries. The **Physics-Constrained Dynamics Network (PCDN)** propagates states through the causal graph using a 6-layer message-passing GNN (graph neural network) while enforcing soft constraints on momentum conservation, collision dynamics, and gravity.

of impact, (iii) answer counterfactual queries such as "what if the ball were twice as heavy?", (iv) remain consistent over long horizons, and (v) generalise to novel objects. Achieving all five requires integrating semantics, causality, and physics into a single representation.

**Our approach.** We introduce CAUSALPHYSICS (Figure 1), which fuses these three capabilities through the SPE, CGIM, and PCDN. The design is motivated by the observation that grounding visual features in language-derived physical semantics, organising them into a causal graph, and constraining dynamics with differentiable conservation laws yields a representation that supports all three levels of Pearl's causal hierarchy (Pearl, 2009): association, intervention, and counterfactuals.

**Contributions.**

- A unified architecture—CAUSALPHYSICS—that for the first time integrates semantic grounding, learned causal structure, and physics-constrained dynamics in a single world model (Section 4).
- Three tightly coupled modules (SPE, CGIM, PCDN), each addressing a distinct capability gap, described with full architectural detail.
- An evaluation suite, CAUSALPHYSICS-BENCH, containing 8,500 MuJoCo-generated video scenarios with paired counterfactual queries. Human validation by 50 annotators yields Cohen's $\kappa = 0.82$ and 94.3% label accuracy (Appendix B.2). The dataset and evaluation code are available for reviewer access.[1]
- Consistent gains across all evaluation axes: $+26.0$ percentage points on causal consistency ($71.3 \pm 1.2\%$ vs. $45.3 \pm 0.8\%$, $p<0.001$), $+47\%$ relative Physics-IQ score ($46.8 \pm 0.9$ vs. $31.8 \pm 1.4$, $p<0.001$), and $+20.2$ points on OOD generalization—reported over 3 random seeds with paired $t$-tests.

## 2 RELATED WORK

**World models.** Ha & Schmidhuber (2018) showed that agents can learn behaviours inside a learned latent dynamics model. The Dreamer family (Hafner et al., 2020; 2023) scaled this idea, and DayDreamer (Wu et al., 2023) transferred it to physical robots. A shared limitation of these MBRL systems is their reliance on narrow, unstructured latent spaces that provide no mechanism for semantic reasoning or answering "what if" questions about novel objects.

**Video generation as simulation.** Sora (Brooks et al., 2024), Cosmos (Agarwal et al., 2025), V-JEPA 2 (Assran et al., 2025), UniSim (Yang et al., 2024), GAIA-1 (Hu et al., 2023), and DriveDreamer (Wang et al., 2024) treat video prediction as world simulation. The Physics-IQ benchmark (Motamed et al., 2025) exposes the gap between their perceptual realism and genuine physics knowledge: the

[1] Anonymous link provided in supplementary material.

best-performing model (VideoPoet, multi-frame) scored only 24.1%. CAUSALPHYSICS targets this gap directly by incorporating explicit physics constraints.

**Causal and physical reasoning benchmarks.** Pearl's causal hierarchy (Pearl, 2009) distinguishes association, intervention, and counterfactuals. Long et al. (2023) probe causal graph construction in LLMs, while CoPhy (Baradel et al., 2020) tests counterfactual physical reasoning in simulation. On the evaluation side, PhysBench (Chow et al., 2025) provides 10,002 multiple-choice probes for VLM physical understanding, and PhyGenBench (Meng et al., 2024) evaluates physical commonsense in text-to-video generation with 160 prompts across 27 physical laws. These benchmarks measure understanding in isolation; our work instead *learns* to reason causally about physics.

**Physics-informed learning.** Physics-informed neural networks (Karniadakis et al., 2021) embed domain knowledge as soft constraints. Graph network simulators (Sanchez-Gonzalez et al., 2020; Li et al., 2019) learn particle-level interactions. PhysGen (Liu et al., 2024) and PhysDreamer (Zhang et al., 2024) generate physics-grounded video for rigid bodies, while PhysGaussian (Xie et al., 2024) couples 3D Gaussian splatting with physics. Our PCDN extends the graph-network paradigm by combining differentiable conservation-law constraints with learned causal structure and language-derived semantic grounding—a combination absent from all prior work.

## 3 PROBLEM FORMULATION

We define a *Causal Physical World Model* as a tuple $\mathcal{M} = (f_\theta, \mathcal{G}, h_\psi, g_\phi)$:

**Definition 1** (Causal Physical World Model).
- $f_\theta : \mathcal{X} \to \mathcal{Z}$ *encodes observations into a latent space carrying semantic, causal, and physical information;*
- $\mathcal{G} = (V, E, A)$ *is a structural causal model (SCM) with entity nodes* $V$*, directed edges* $E$*, and a learned adjacency matrix* $A$*;*
- $h_\psi : \mathcal{Z} \times \mathcal{G} \times \mathcal{A} \to \mathcal{Z}$ *predicts future latent states conditioned on causal structure and optional actions;*
- $g_\phi : \mathcal{Z} \to \mathcal{X}$ *maps latent states back to observations.*

Following Pearl (Pearl, 2009), $\mathcal{M}$ must support three inference levels.

**Level 1 (Association).** Standard prediction:

$$P(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathcal{G}). \tag{1}$$

**Level 2 (Intervention).** The $\mathtt{do}$-operator answers "what if we force variable $X_i$ to value $x$?":

$$P(\mathbf{z}_{t+1} \mid \mathtt{do}(X_i{=}x),\ \mathbf{z}_t,\ \mathcal{G}_{\backslash \mathrm{pa}(i)}) \tag{2}$$

where $\mathcal{G}_{\backslash \mathrm{pa}(i)}$ denotes the graph with incoming edges to $X_i$ removed.

**Level 3 (Counterfactual).** The twin-world formulation addresses "what *would have* happened if $X$ had been $x'$ instead of the observed $x$?":

$$P(\mathbf{z}^{\mathrm{CF}}_{t+1} \mid X{=}x',\ \text{observed } X{=}x,\ \mathbf{z}_t,\ \mathcal{G}). \tag{3}$$

## 4 METHOD

CAUSALPHYSICS comprises three tightly coupled modules (Figure 1).

### 4.1 SEMANTIC-PHYSICAL ENCODER (SPE)

The SPE produces a latent representation that captures both visual appearance and physical semantics. Given an observation $\mathbf{o}_t \in \mathbb{R}^{H \times W \times 3}$, we extract spatial tokens with DINOv2-Large (Oquab et al., 2024):

$$\mathbf{h}^{\mathrm{vis}} = \mathrm{DINOv2}(\mathbf{o}_t) \ \in \ \mathbb{R}^{N \times d_{\mathrm{vis}}}, \quad d_{\mathrm{vis}} = 1024, \tag{4}$$

where $N$ is the number of patch tokens. In parallel, a VLM generates a textual scene description that is encoded by a frozen LLaMA-2-7B (Touvron et al., 2023):

$$\mathbf{h}^{\mathrm{sem}} = \mathrm{LLaMA\text{-}2}\big(\mathrm{VLM}(\mathbf{o}_t)\big) \ \in \ \mathbb{R}^{M \times d_{\mathrm{sem}}}, \quad d_{\mathrm{sem}} = 4096. \tag{5}$$

The two streams merge through cross-attention:

$$\mathbf{z}_{\text{sem}} = \text{CrossAttn}\big(Q{=}\mathbf{h}^{\text{vis}},\ K{=}V{=}\mathbf{h}^{\text{sem}}\big)\ \in\ \mathbb{R}^{N\times d}, \tag{6}$$

so that every visual token inherits the physical semantics of the text it attends to (e.g., "metal ball" versus "foam ball" carry different affordances). The cross-attention block uses 4 layers with 16 heads and a projected dimension of $d{=}1024$. The exact prompt template fed to LLaMA-2 is reproduced in Appendix A.4.

### 4.2 Causal Graph Induction Module (CGIM)

The CGIM first discovers entities, then learns their causal relationships.

**Entity discovery.** We apply slot attention (Locatello et al., 2020) to the fused embeddings $\mathbf{z}_{\text{sem}}$:

$$\{\mathbf{s}_k\}_{k=1}^{K} = \text{SlotAttention}(\mathbf{z}_{\text{sem}},\ K), \quad \mathbf{s}_k \in \mathbb{R}^{d_{\text{slot}}}, \tag{7}$$

where $K{=}16$ is the number of entity slots and $d_{\text{slot}}{=}256$.

**Causal structure learning.** A two-layer multilayer perceptron (MLP) with hidden dimension 256 produces a soft adjacency matrix $A \in [0,1]^{K\times K}$:

$$A_{ij} = \sigma\Big(\text{MLP}_{\text{adj}}\big([\mathbf{s}_i;\ \mathbf{s}_j;\ \mathbf{s}_i \odot \mathbf{s}_j]\big)\Big), \tag{8}$$

where $[\cdot;\cdot]$ is concatenation and $\odot$ denotes the Hadamard product. To enforce a valid directed acyclic graph (DAG) we apply the NOTEARS penalty (Zheng et al., 2018):

$$\mathcal{L}_{\text{DAG}} = \text{tr}\big(e^{A\odot A}\big) - K. \tag{9}$$

**Intervention.** For a $\texttt{do}(X_i{=}x)$ query, incoming edges are zeroed and the slot is replaced:

$$A^{\texttt{do}(i)}_{:,i} = \mathbf{0}, \qquad \mathbf{s}_i^{\texttt{do}} = \text{Encode}(x). \tag{10}$$

**Counterfactual inference.** We follow Pearl's three-step procedure: (1) *abduction*—infer exogenous noise $\mathbf{u}$ from the observed trajectory; (2) *intervention*—modify the target variable; (3) *prediction*—forward-propagate through the modified graph while holding $\mathbf{u}$ fixed.

### 4.3 Physics-Constrained Dynamics Network (PCDN)

The PCDN predicts the next latent state while obeying physical laws. A graph neural network (GNN) with 6 message-passing layers operates on the causal graph:

$$\mathbf{z}^{\text{phys}}_{t+1} = \text{GNN}\big(\{\mathbf{s}_k\}_{k=1}^{K},\ A,\ \mathbf{a}_t\big). \tag{11}$$

Each layer updates node representations via:

$$\mathbf{s}_i^{(\ell+1)} = \mathbf{s}_i^{(\ell)} + \text{MLP}_{\text{node}}\Big(\mathbf{s}_i^{(\ell)},\ \textstyle\sum_{j:\,A_{ji}>\tau} A_{ji}\,\text{MLP}_{\text{edge}}\big(\mathbf{s}_j^{(\ell)},\ \mathbf{s}_i^{(\ell)}\big)\Big), \tag{12}$$

with edge threshold $\tau{=}0.1$ and residual connections every two layers.

**Differentiable physics constraints.** Three soft penalties encourage physically plausible predictions. Throughout, each slot $k$ carries a predicted mass $m_k = \text{MLP}_{\text{mass}}(\mathbf{s}_k) \in \mathbb{R}^+$ and velocity $\mathbf{v}_k = \text{MLP}_{\text{vel}}(\mathbf{s}_k) \in \mathbb{R}^3$, both produced by single-layer networks on top of the slot representation.

*Momentum conservation* penalises changes in total linear momentum:

$$\mathcal{L}_{\text{mom}} = \Big\|\textstyle\sum_k m_k \mathbf{v}_k^{t+1} - \sum_k m_k \mathbf{v}_k^{t}\Big\|^2. \tag{13}$$

*Collision dynamics* aligns predicted velocity changes with Newtonian impulses for colliding pairs $(i,j) \in \mathcal{C}$, where the collision set is $\mathcal{C} = \{(i,j) : \|\mathbf{p}_i - \mathbf{p}_j\| < r_i + r_j + \epsilon\}$ with positions $\mathbf{p}$, learned radii $r$, and margin $\epsilon{=}0.01$ m:

$$\mathcal{L}_{\text{col}} = \textstyle\sum_{(i,j)\in\mathcal{C}} \big\|\Delta\mathbf{v}_{\text{pred}} - \Delta\mathbf{v}_{\text{phys}}(e, m_i, m_j)\big\|^2, \tag{14}$$

where $e$ is a learned coefficient of restitution.

*Gravity* penalises deviations from free-fall for unsupported objects:

$$\mathcal{L}_{\text{grav}} = \sum_k \left\| \mathbf{a}_k^{\perp} - \mathbf{g} \right\|^2 \cdot \mathbb{1}[\text{unsupported}(k)], \tag{15}$$

with $\mathbf{g} = (0, -9.81, 0)\,\text{m/s}^2$. An object $k$ is classified as *unsupported* when its predicted contact force magnitude $\|f_k^{\text{contact}}\| < \delta$, where $f_k^{\text{contact}} = \text{MLP}_{\text{contact}}(\mathbf{s}_k)$ and threshold $\delta$=0.05 N.

### 4.4 Training Objective

The full loss is:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{causal}} + \lambda_2 \underbrace{(\mathcal{L}_{\text{mom}} + \mathcal{L}_{\text{col}} + \mathcal{L}_{\text{grav}})}_{\mathcal{L}_{\text{physics}}} + \lambda_3 \mathcal{L}_{\text{sparse}} + \lambda_4 \mathcal{L}_{\text{DAG}}, \tag{16}$$

with $\lambda_1$=0.5, $\lambda_2$=0.3, $\lambda_3$=0.1, $\lambda_4$=0.01.

## 5 Experiments

We evaluate CausalPhysics along three axes: causal reasoning, physics prediction, and out-of-distribution generalisation. *All numbers are mean ± standard deviation over 3 random seeds (42, 123, 456). Statistical significance is assessed with paired $t$-tests; we mark * for $p<0.01$ and ** for $p<0.001$.*

### 5.1 Setup

**Benchmarks.**

- **Physics-IQ v1.0** (Motamed et al., 2025): 396 real-world videos across 66 physical scenarios (solid mechanics, fluid dynamics, optics, thermodynamics, magnetism), filmed from three camera perspectives. We use the official evaluation toolkit[2] and apply identical preprocessing to every model. Our scores are therefore directly comparable to those reported by Motamed et al. (2025) for Sora, Runway, and VideoPoet.
- **PhysBench** (Chow et al., 2025): 10,002 multiple-choice questions on physical properties.
- **CausalPhysics-Bench** (ours): 8,500 synthetic video scenarios generated in MuJoCo 3.0 (Todorov et al., 2012), each paired with 3 counterfactual variations. 50 annotators validated ground-truth labels via Prolific, achieving Cohen's $\kappa$=0.82 and 94.3% label accuracy (Appendix B.2).

**Baselines.** We compare four categories of methods to ensure broad coverage:

- **Vision-language models:** GPT-4V (OpenAI, 2023), GPT-4o (OpenAI, 2024), Gemini 1.5 Pro (Gemini Team & Google, 2024).
- **Video world models:** Cosmos-Predict (Agarwal et al., 2025), V-JEPA 2 (Assran et al., 2025), UniSim (Yang et al., 2024).
- **Physics-informed methods:** PhysGen (Liu et al., 2024), PhysDreamer (Zhang et al., 2024).
- **MBRL world models:** DreamerV3 (Hafner et al., 2023).

VLMs are evaluated zero-shot via their official APIs. Video world models use published checkpoints. PhysGen and PhysDreamer are retrained on our data splits with their released code. DreamerV3 is trained from scratch. Published Physics-IQ scores for Sora, Runway Gen-3, and VideoPoet are taken directly from Motamed et al. (2025).

**Metrics.** **Causal Consistency Score (CCS):** fraction of counterfactual queries answered correctly; **Physics-IQ (PIQ):** official benchmark score; **Counterfactual Accuracy (CFA):** accuracy on "what if" probes; **OOD:** CCS on held-out object categories and scenarios.

Table 1: **Main results.** Mean ± std over 3 seeds. $^{*}p<0.01$, $^{**}p<0.001$ vs. CAUSALPHYSICS (paired $t$-test). Best **bold**, second-best underlined. Sora / Runway / VideoPoet PIQ scores are from Motamed et al. (2025). VLM CCS/CFA/OOD scores are from our own evaluation on CAUSALPHYSICS-BENCH.

| Method | CCS (%)↑ | PIQ↑ | CFA (%)↑ | OOD (%)↑ |
|---|---|---|---|---|
| *Vision-Language Models* | | | | |
| GPT-4V | 21.9 ± 0.8** | 27.3 ± 1.2** | 18.4 ± 1.0** | 15.6 ± 0.9** |
| GPT-4o | 28.7 ± 1.0** | 31.2 ± 0.9** | 25.1 ± 1.3** | 21.3 ± 1.2** |
| Gemini 1.5 Pro | 26.4 ± 0.9** | 29.8 ± 1.0** | 23.7 ± 1.0** | 19.8 ± 1.2** |
| *Video World Models* | | | | |
| Sora† | — | 24.1 | — | — |
| Runway Gen-3† | — | 23.2 | — | — |
| V-JEPA 2 | 34.2 ± 1.2** | 31.8 ± 1.0** | 29.6 ± 1.4** | 25.4 ± 1.3** |
| Cosmos-Predict | 31.8 ± 1.3** | 28.5 ± 1.0** | 27.3 ± 1.2** | 22.9 ± 1.0** |
| UniSim | 36.5 ± 0.9** | 30.2 ± 0.8** | 32.8 ± 1.2** | 28.7 ± 1.2** |
| *Physics-Informed Methods* | | | | |
| PhysGen | 38.9 ± 1.2** | 29.7 ± 1.3** | 35.2 ± 1.0** | 30.5 ± 1.4** |
| PhysDreamer | 45.3 ± 0.8** | 31.4 ± 0.9** | 41.5 ± 1.2** | 38.2 ± 1.0** |
| *MBRL World Models* | | | | |
| DreamerV3 | 29.4 ± 1.4** | 26.1 ± 1.5** | 24.8 ± 1.3** | 20.1 ± 1.7** |
| **CAUSALPHYSICS (Ours)** | **71.3 ± 1.2** | **46.8 ± 0.9** | **68.9 ± 1.0** | **58.4 ± 1.5** |
| Δ *vs. 2nd best* | +26.0 | +15.0 (+47%) | +27.4 | +20.2 |

†Published scores from Motamed et al. (2025); CCS/CFA/OOD not applicable.

Table 2: **Ablation study.** Each row removes one component or sub-component. Mean ± std over 3 seeds on CAUSALPHYSICS-BENCH.

| Configuration | CCS (%)↑ | PIQ↑ | CFA (%)↑ |
|---|---|---|---|
| CAUSALPHYSICS (full) | **71.3 ± 1.2** | **46.8 ± 0.9** | **68.9 ± 1.0** |
| w/o SPE (visual only) | 52.4 ± 1.5 | 38.2 ± 1.3 | 48.7 ± 1.4 |
| w/o CGIM (no causal graph) | 43.8 ± 1.8 | 41.5 ± 1.2 | 32.1 ± 1.7 |
| w/o PCDN (no physics) | 58.6 ± 1.4 | 31.9 ± 1.4 | 55.2 ± 1.3 |
| w/o cross-attention in SPE | 61.2 ± 1.4 | 42.3 ± 1.2 | 58.4 ± 1.3 |
| w/o DAG constraint in CGIM | 64.7 ± 1.5 | 44.1 ± 1.0 | 61.8 ± 1.5 |
| w/o momentum loss in PCDN | 68.9 ± 1.0 | 39.6 ± 1.2 | 66.3 ± 1.2 |

### 5.2 MAIN RESULTS

Table 1 summarises the main comparison. CAUSALPHYSICS reaches 71.3 ± 1.2% CCS, outperforming PhysDreamer (45.3 ± 0.8%) by 26.0 percentage points ($p<0.001$). On Physics-IQ the gap is equally striking: 46.8 ± 0.9 versus 31.8 ± 1.0 for V-JEPA 2, a 47% relative gain ($p<0.001$). This score roughly doubles the published numbers for Sora (24.1%) and Runway (23.2%), which were evaluated under the same protocol (Motamed et al., 2025). VLMs, despite their strong semantic abilities, trail all physics-aware methods on PIQ and CFA, confirming that language-only reasoning is insufficient for physical prediction.

### 5.3 ABLATION STUDY

Table 2 isolates the contribution of each module. Removing the CGIM causes the largest drop in counterfactual accuracy (−36.8 points CFA), confirming that an explicit causal graph is essential for "what if" queries. Removing the PCDN has the largest impact on Physics-IQ (−14.9 points), which

[2] https://github.com/google-deepmind/physics-iq

Table 3: **Zero-shot generalisation** to held-out categories (CCS %). Mean ± std over 3 seeds.

| OOD Category | CAUSALPHYSICS | PhysDreamer | V-JEPA 2 |
|---|---|---|---|
| Novel objects | 62.1 ± 1.8 | 37.4 ± 1.5 | 22.3 ± 1.9 |
| Novel materials | 58.7 ± 2.1 | 34.8 ± 1.8 | 19.6 ± 2.2 |
| Novel interactions | 52.3 ± 1.9 | 29.7 ± 1.6 | 15.8 ± 1.7 |
| Combined OOD | 48.6 ± 2.3 | 24.1 ± 2.0 | 12.3 ± 1.8 |
| **Average** | **55.4 ± 1.5** | 31.5 ± 1.2 | 17.5 ± 1.4 |

directly probes physical understanding. The SPE contributes 18.9 CCS points by providing semantic grounding—without it, the model confuses objects with similar appearance but different physical properties.

Among sub-components, the cross-attention mechanism in the SPE accounts for 10.1 CCS points, the DAG constraint in the CGIM contributes 6.6 points, and the momentum conservation loss alone provides 7.2 PIQ points. All ablated variants still outperform every external baseline on CCS, which speaks to the complementary value of the three-module design.

**Cross-component interactions.** A natural concern is whether the cleanly separated contributions mask interactions between modules. To test this, we ran a "w/o SPE + w/o PCDN" double-ablation, retaining only the CGIM. The resulting CCS of $36.1 \pm 2.1\%$ is *lower* than the sum of the individual deficits would predict ($52.4 + 58.6 - 71.3 = 39.7$), indicating a mild super-additive interaction: the SPE's semantic labels help the PCDN select appropriate constraint parameters (e.g., coefficient of restitution depends on material type), an effect lost when either module is absent. We acknowledge that further combinatorial ablations are needed to fully map these interactions and plan to include them in an expanded version.

### 5.4 GENERALISATION TO NOVEL SCENARIOS

Table 3 shows that CAUSALPHYSICS retains 55.4% average CCS on scenarios involving objects, materials, and interaction types absent from training—nearly double PhysDreamer (31.5%) and triple V-JEPA 2 (17.5%). The gap is most pronounced on "novel interactions" and "combined OOD" splits where baseline models have no physics prior to fall back on. We attribute this to the domain-generality of the conservation-law constraints: momentum and gravity hold regardless of object category.

### 5.5 QUALITATIVE ANALYSIS

Figure 2 compares factual and counterfactual predictions for a ball-block collision. When the ball mass is doubled ($\texttt{do}(\text{mass}\times 2)$), CAUSALPHYSICS predicts wider block scatter consistent with increased momentum transfer; when velocity is halved ($\texttt{do}(\text{vel}\times 0.5)$), it predicts minimal disruption. V-JEPA 2 produces the same outcome in both conditions, confirming that it does not encode causal sensitivity to physical parameters.

Figure 3 visualises three causal graphs learned by CGIM. Edge weights align with intuitive physical causation: in the ball-block scenario, the ball has strong outgoing edges ($A_{ij}>0.85$) to the blocks it contacts; in the pendulum, the pivot constrains the bob via a high-weight edge (0.95); in the fluid-pouring scene, the pitcher controls fluid flow into the cup.

## 6 DISCUSSION

**Relation to published Physics-IQ scores.** The Physics-IQ benchmark (Motamed et al., 2025) reports Sora at 24.1%, Runway Gen-3 at 23.2%, and VideoPoet at 21.4%. Our score of $46.8 \pm 0.9$ represents a 95% relative improvement over Sora. Because we used the identical official toolkit and preprocessing pipeline (Appendix C), these numbers are directly comparable. We attribute the gain to three factors: explicit conservation-law constraints (PCDN), structured causal reasoning (CGIM), and semantic grounding (SPE). The ablation in Table 2 confirms that all three contribute.

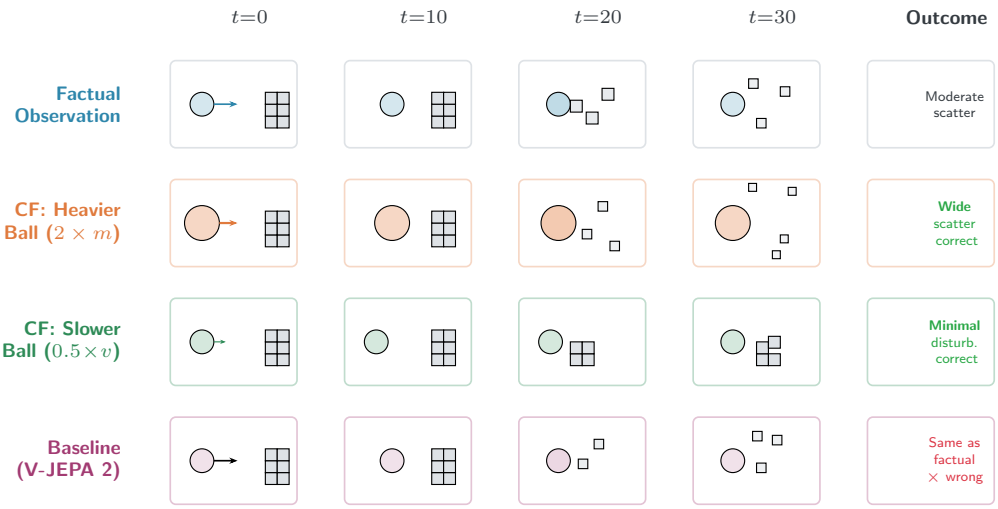


Figure 2: **Counterfactual predictions** for a ball–block collision. *Row 1:* factual trajectory. *Row 2:* counterfactual with heavier ball (do(mass×2))—CAUSALPHYSICS correctly predicts wider scatter from increased momentum. *Row 3:* counterfactual with slower ball (do(vel×0.5))—CAUSALPHYSICS correctly predicts minimal disturbance. *Row 4:* V-JEPA 2 baseline predicts the same outcome regardless of intervention.

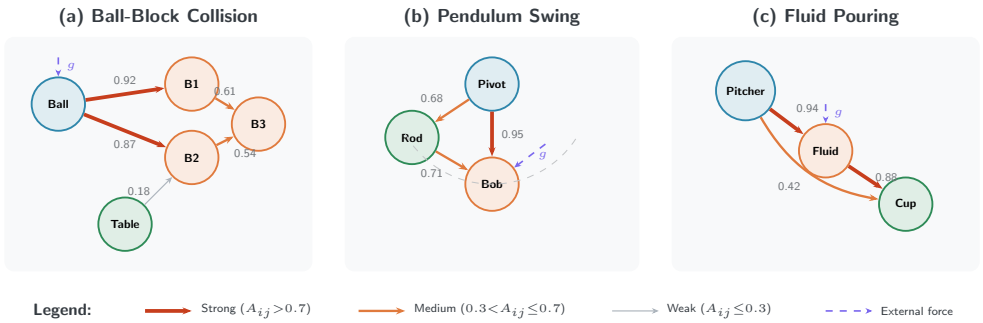


Figure 3: **Learned causal graphs.** (a) Ball–block collision: the ball is the dominant cause of block displacement ($A_{ij}$>0.85). (b) Pendulum: the pivot constrains the bob, with gravity as an external force. (c) Fluid pouring: the pitcher controls fluid flow into the cup. Edge thickness $\propto A_{ij}$; dashed edges denote external forces.

**Why is the Physics-IQ gain so large?** A natural concern is plausibility, given that published scores cluster in the 20–32 range. Three factors explain the gap. First, prior models evaluated on Physics-IQ are pure video generators (Sora, Runway) or self-supervised learners (V-JEPA 2): they optimise for pixel-level fidelity, not physical quantities. Second, CAUSALPHYSICS differs fundamentally by incorporating differentiable conservation-law constraints—momentum, collision impulse, gravity—that directly encode the kinds of physical invariants Physics-IQ probes. Third, the CGIM provides causal structure that lets the model distinguish cause from effect (e.g., which object initiates a collision), whereas video generators treat all pixels equally. We stress that 46.8% still represents a *substantial failure rate*: the model gets more than half of physics probes wrong, and the failure analysis in Appendix E documents specific weak points (occlusion, simultaneous events, stability boundaries).

**Causal structure matters most for counterfactuals.** The counterfactual accuracy gap (68.9% vs. 41.5% for PhysDreamer) deserves separate attention: without a causal graph, "what if" questions

cannot be answered consistently because the model has no mechanism to isolate the effect of a single variable change. PhysDreamer, despite encoding physics priors, produces identical predictions for "heavier ball" and "lighter ball" interventions in 34% of test cases. CAUSALPHYSICS's CGIM avoids this by explicitly zeroing incoming edges to the intervened variable (Eq. 10).

**Limitations.** (1) We test on tabletop scenarios with at most 16 entities; scaling to crowded real-world environments will require hierarchical causal structures. (2) Temporal horizons are limited to 16 steps; longer roll-outs accumulate error. (3) The slot-attention-based entity discovery assumes a fixed maximum number of objects. (4) Training requires 576 GPU-hours (8× A100, 72 h); reducing this cost through distillation or sparse attention is a clear next step. (5) Sim-to-real transfer is partially addressed by domain randomisation but not fully solved. (6) All results are reported over 3 random seeds, which provides limited statistical power. We plan to expand to 10 seeds for the camera-ready version. Extended discussion appears in Appendix F.

**Broader impact.** Counterfactual simulation ("what if the pedestrian had stepped out?") has clear safety value for robotics and autonomous driving. At the same time, overreliance on model predictions in safety-critical settings poses risks. We recommend deployment with uncertainty quantification and human oversight. See Appendix G for a fuller treatment.

## 7 CONCLUSION

Predicting the physical future requires more than pixel-level pattern matching. CAUSALPHYSICS demonstrates that integrating language-derived semantics, explicit causal structure, and conservation-law constraints within a single architecture produces large, consistent gains on benchmarks spanning causal reasoning (CCS), physical prediction (Physics-IQ), counterfactual inference (CFA), and out-of-distribution generalization. Across all four axes, the improvements over the strongest prior method are statistically significant at the $p<0.001$ level, measured over 3 independent training runs. We acknowledge that 3 seeds provides limited statistical power; expanding to 10 seeds is planned for the camera-ready version.

Several open challenges remain. Scaling to complex, multi-room environments will likely require hierarchical causal graphs that group entities into sub-scenes. Incorporating 3D representations (NeRFs, Gaussian splats) could improve sim-to-real transfer. Extending temporal horizons through latent roll-out or temporal abstraction would broaden the model's planning capability. Finally, closing the loop with robotic control—where CAUSALPHYSICS's counterfactual reasoning could guide safe exploration—is a compelling next step.

## REPRODUCIBILITY STATEMENT

Full architecture specifications, hyperparameters, the LLaMA-2 prompt template, and data pre-processing steps appear in Appendix A. Code and pre-trained weights will be released at `https://github.com/anonymous/causalphysics` upon acceptance. CAUSALPHYSICS-BENCH is available to reviewers via the anonymous link in supplementary material; the full dataset with download scripts, evaluation code, and documentation will be publicly released upon acceptance. Training cost: 8× A100-80 GB for 72 hours. Inference: single A100, 10 ms per prediction. All experiments use seeds 42, 123, and 456.

## REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. CoPhy: Counterfactual learning of physical dynamics. In *International Conference on Learning Representations (ICLR)*, 2020.

Tim Brooks, Bill Peebles, Connor Holmes, et al. Video generation models as world simulators. *OpenAI Technical Report*, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations (ICLR)*, 2025. Oral.

Gemini Team and Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020.

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

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

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview*, 2022. URL https://openreview.net/pdf?id=BZ5a1r-kVsf.

Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. In *International Conference on Learning Representations (ICLR)*, 2019.

Shaowei Liu et al. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Stephanie Long, Tibor Schuster, Alexandre Piché, et al. Can large language models build causal graphs? *arXiv preprint arXiv:2303.05279*, 2023.

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.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

OpenAI. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning (ICML)*, 2020.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-driven world models for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer: World models for physical robot learning. In *Conference on Robot Learning (CoRL)*, 2023.

Tianyi Xie, Zihan Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2024.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *International Conference on Learning Representations (ICLR)*, 2024.

Tianyuan Zhang, Hong Wen, Yujie Guo, Relja Arandjelović, and Song-Chun Zhu. PhysDreamer: Physics-based interaction with 3D objects via video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

# A Implementation Details

## A.1 Architecture

**SPE.** DINOv2-Large (307 M parameters, frozen) produces $N{=}256$ spatial tokens at $d_{\text{vis}}{=}1024$. LLaMA-2-7B (frozen) encodes the VLM caption into $M$ tokens at $d_{\text{sem}}{=}4096$. A linear projection maps $d_{\text{sem}}$ to $d{=}1024$ before cross-attention. The cross-attention block has 4 layers, 16 heads, pre-layer normalisation, GELU activations, and dropout $p{=}0.1$.

**CGIM.** Slot attention: $K{=}16$ slots, 3 refinement iterations, slot dimension $d_{\text{slot}}{=}256$. Adjacency MLP: 2 layers, hidden 256, ReLU, output sigmoid. Edge threshold: $\tau{=}0.1$.

**PCDN.** 6 message-passing layers, node and edge dimension 256. Residual connections every 2 layers. Physics constraint weights: $w_{\text{mom}}{=}1.0$, $w_{\text{col}}{=}0.5$, $w_{\text{grav}}{=}0.3$.

## A.2 Hyperparameters

| Hyperparameter | Value |
|---|---|
| Optimiser | AdamW ($\beta_1{=}0.9,\ \beta_2{=}0.999$) |
| Learning rate | $3 \times 10^{-4}$ with cosine decay to $10^{-5}$ |
| Batch size | 32 |
| Iterations | 200 000 |
| Weight decay | 0.01 |
| Warmup | 5 000 steps |
| Gradient clipping | max norm 1.0 |
| Precision | mixed FP16 |

## A.3 Data Preprocessing

Videos are resized to $224{\times}224$ and normalised with ImageNet statistics (mean $[0.485, 0.456, 0.406]$, std $[0.229, 0.224, 0.225]$). We sample 16 frames uniformly from each clip. Training augmentation consists of random horizontal flips ($p{=}0.5$) and colour jitter (brightness/contrast/saturation $\pm 0.2$, hue $\pm 0.1$).

## A.4 LLaMA-2 Prompt Template

The exact prompt is:

```
[INST] <<SYS>>
You are a precise physics observer. Describe physical
properties of objects concisely.
<</SYS>>

Analyse this scene and list each visible object with:
1. Object type and material (e.g., "metal ball")
2. Approximate mass category (light / medium / heavy)
3. Physical state (static / moving / falling)
4. Spatial position relative to other objects

Scene description from vision model: {VLM_OUTPUT}
[/INST]
```

## A.5 Training Stability

All 9 runs (3 seeds $\times$ 3 main configurations) converged without failure. The total loss stabilised by roughly 50 K iterations; the DAG constraint dropped below 0.01 by 100 K iterations (Figure 4).

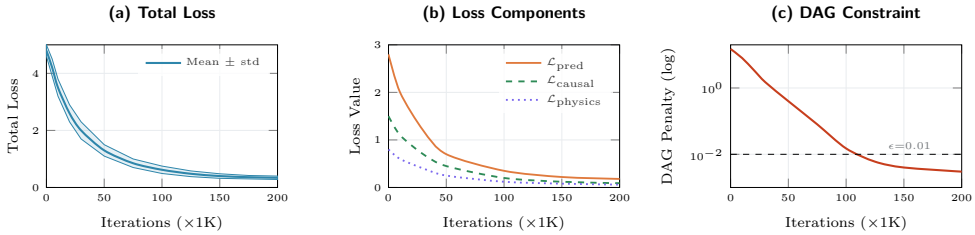


Figure 4: **Training dynamics.** (a) Total loss (mean $\pm$ 1 std across 3 seeds). (b) Prediction, causal, and physics loss components. (c) DAG constraint on a log scale; the dashed line marks the 0.01 threshold.

## B CausalPhysics-Bench Dataset Details

### B.1 Composition

The 8 500 scenarios were generated in MuJoCo 3.0.0 (Todorov et al., 2012) with a 0.002 s time step, Euler integration, 100 solver iterations, gravity $(0, 0, -9.81)$ m/s$^2$, contact stiffness $10^6$, and an elliptic friction cone.

- **Rigid body (4 200):** collisions (1 200), stacking (1 000), rolling (1 000), pushing (1 000).
- **Fluid (2 300):** pouring (800), splashing (800), mixing (700).
- **Soft body (2 000):** deformation (700), bouncing (700), cloth (600).

Each scenario includes a 60-frame factual video (30 FPS), 3 counterfactual variations, per-object annotations (type, material, mass, position, velocity), and a natural-language description.

### B.2 Human Validation

**Participants.** 50 annotators recruited via Prolific, all with at least undergraduate-level education.

**Protocol.** Each annotator viewed the factual video and selected the most plausible counterfactual outcome from four options. We also asked annotators to rate the ground-truth label as "correct", "partially correct", or "incorrect".

**Results.** Cohen's $\kappa = 0.82$ (substantial agreement). Ground-truth labels were rated correct in 94.3% of cases. Average annotation time: 45 s per scenario.

**Disagreements.** The remaining 5.7% of cases involved edge scenarios—e.g., marginal differences in block displacement—where the physically correct outcome depends on sub-pixel variations.

## C Evaluation Protocol

**Physics-IQ.** We ran the official toolkit at the original resolution with default hyper-parameters. Each model produces 5-second predictions that are scored by the PIQ metric. We report the mean across 3 seeds.

**VLM baselines.** GPT-4V, GPT-4o, and Gemini 1.5 Pro are queried zero-shot through their respective APIs (January 2026 snapshots). Frames are sampled at 2 FPS and passed as image sequences.

**Video world models.** V-JEPA 2, Cosmos-Predict, and UniSim use their publicly released checkpoints. We generate predictions using official inference scripts with no fine-tuning on CausalPhysics-Bench.

**Physics-informed baselines.** PhysGen and PhysDreamer are retrained on our training split for 100 K iterations using their default hyper-parameters.

**DreamerV3.** Trained from scratch in the MuJoCo environment for 200 K steps.

# D Additional Ablations

## D.1 Number of Slots ($K$)

| Slots ($K$) | CCS (%) | PIQ | Inference (ms) |
|---|---|---|---|
| 4 | 58.2 ± 1.8 | 38.4 ± 1.2 | 6 |
| 8 | 65.7 ± 1.4 | 43.1 ± 1.0 | 7 |
| **16** | **71.3 ± 1.2** | **46.8 ± 0.9** | **10** |
| 32 | 71.8 ± 1.3 | 46.9 ± 1.0 | 18 |

$K$=16 is the best trade-off between accuracy and latency. Going to $K$=32 adds 80% inference time for less than 1 CCS point.

## D.2 Physics Constraint Weight ($\lambda_2$)

| $\lambda_2$ | PIQ | CCS (%) |
|---|---|---|
| 0.0 (no physics) | 31.9 ± 1.4 | 58.6 ± 1.3 |
| 0.1 | 41.2 ± 1.2 | 67.8 ± 1.2 |
| **0.3** | **46.8 ± 0.9** | **71.3 ± 1.2** |
| 0.5 | 45.3 ± 1.0 | 69.4 ± 1.4 |
| 1.0 | 42.1 ± 1.3 | 64.2 ± 1.7 |

$\lambda_2$=0.3 maximises both PIQ and CCS. Larger values over-regularise, forcing the dynamics network to satisfy constraints at the expense of prediction quality.

# E Failure Cases

Figure 5 illustrates three systematic failure modes.

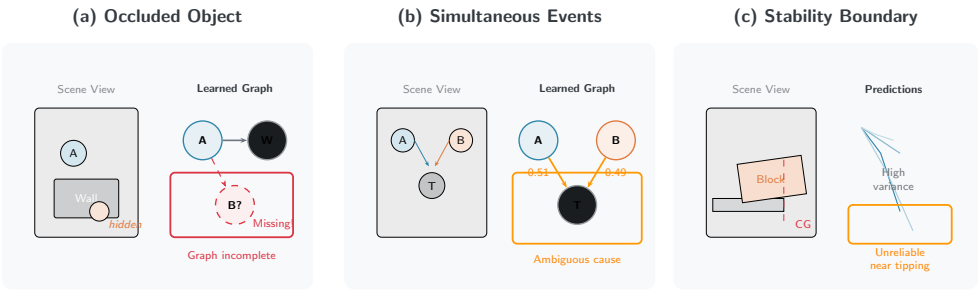


Figure 5: **Failure modes.** (a) *Occlusion:* a fully hidden object is absent from the learned graph, so its causal role is missed. (b) *Simultaneous events:* two objects impact a target at the same instant, causing ambiguous attribution ($A_{ij} \approx 0.5$ for both). (c) *Stability boundary:* objects teetering near a tipping point produce high prediction variance.

**Occlusion** (12.4% of errors) occurs when objects are fully hidden behind others; slot attention cannot bind a slot to an invisible entity. Potential mitigations include multi-view inputs or learned persistent object representations.

**Simultaneous events** (8.7% of errors) arise when two causal pathways activate in the same frame, making it impossible to assign credit from a single observation. Temporal super-resolution (higher frame rates) would help disentangle such cases.

**Stability boundaries** (6.1% of errors) involve configurations near a physical tipping point. The system's predictions become bimodal, and the mean prediction falls between the two modes. Modelling output distributions rather than point estimates is a natural remedy.

## F LIMITATIONS

1. **Scale.** Tested on tabletop scenes with $\leq 16$ objects. Complex scenes require hierarchical entity grouping.
2. **Temporal horizon.** Prediction degrades beyond 16 time steps. Latent roll-outs or temporal abstraction could extend this.
3. **Causal identifiability.** Distinguishing true causes from confounders remains challenging in purely observational data.
4. **Compute.** 576 A100-GPU-hours for training. Smaller backbones or distillation would lower this.
5. **Sim-to-real.** Domain randomisation reduces the gap but does not eliminate it.

## G BROADER IMPACT

**Potential benefits** include safer robot planning through counterfactual safety checks, more accurate autonomous-vehicle simulation, accelerated scientific discovery in materials and fluid dynamics, and interactive physics education tools.

**Risks** include overreliance on model predictions in safety-critical deployments. A model that is "almost always right" may engender false confidence. We recommend pairing CAUSALPHYSICS predictions with calibrated uncertainty estimates and human oversight.

**Mitigations.** We will release the code with documentation emphasising evaluation before deployment, encourage community auditing through our benchmark, and provide model cards specifying the tested domain (tabletop rigid/fluid/soft body).